UNIVERSITY OF SOUTHAMPTON

Faculty Of Medicine

Clinical and Experimental Science

Bioinformatics Approaches to Vaccine Design for Bacterial Pathogens

by

Ashley Ivan Heinson

Thesis for the degree of Doctor of Philosophy

31st March 2017

UNIVERSITY OF SOUTHAMPTON FACULTY OF MEDICINE

ABSTRACT

Thesis for the degree of Doctor of Philosophy

Bioinformatics Approaches to Vaccine Design for Bacterial Pathogens

Ashley Ivan Heinson

This thesis focused on bacterial vaccinology and employed a newly emergent branch of vaccinology; reverse vaccinology (RV). RV is an *in silico* process that predicts vaccine candidates from an entire bacterial proteome, thus enabling the realisation of a greater number of putative vaccine candidates when compared to conventional vaccinology approaches. A previous RV classifier that utilised the computational field of machine learning (ML) was used to predict bacterial protective antigens (BPAs) (i.e. vaccine candidates) *for Mycobacterium tuberculosis* (*Mtb*). *Mtb* was chosen as the initial focus for RV approaches in this thesis because one third of the world's population are infected with *Mtb* and in 2015 *Mtb* infection killed 1.8 million people. It is also being recognised that the only clinically licensed vaccine against *Mtb* infection, Bacille Calmette-Guérin (BCG), has varying rates of protection. Predicted BPAs by a published RV classifier were synthesised as DNA vaccines and tested in a mouse model of *Mtb* infection. However, the predicted BPAs were shown not to generate protection in repeat animal trials. To address the negative result obtained when testing BPAs predicted by a previous RV classifier, enhancements were made to the previously published RV classifier (i.e. nested leave-tenth-out cross-validation, subcellular localisation bias removal, increased size of training dataset and increased type of protein annotation tools used to generate features). Finally, the enhanced RV classifier, developed in this thesis, was assessed using a more biologically revealing metric termed recall in the proteomes of *Mtb* and *Neisseria meningitidis* serogroup B (*MenB*). *MenB* was chosen to assess the recall metric as it enabled comparisons to the BEXSERO vaccine, which was the first clinically licensed vaccine developed using RV. In summary, this thesis has developed a biologically relevant RV classifier that can now be used to predict BPAs for any bacterial pathogen with a sequenced genome. It is envisaged that these predicted BPAs could then be used to facilitate the rapid formulation of novel subunit vaccines.

Table of Contents

List of Tables

List of Figures

List of Appendices Appearance in Text

Declaration of Authorship

I, Ashley Ivan Heinson, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Part of this work has been published as:

Heinson AI, Woelk CH, Newell ML; The promise of reverse vaccinology. *International Health* 2015;7(2):85-9. – This was incorporated as part of the Introduction

Heinson AI, Gunawardana Y, Moesker B, Hume CC, Vataga E, Hall Y, Stylianou E, McShane H, Williams A, Niranjan M, Woelk CH; Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology. *International Journal of Molecular Sciences* 2017; **18**(2) – This was incorporated as Chapter 3.

Signed:

A. I. Heinson

Date: 31/03/2017

Acknowledgements

First and foremost I would like to thank Dr. Christopher H. Woelk for his help and guidance with this work and for shaping my scientific future. He has been a constant inspiration and encouragement to me and I look forward to working with him going forward. In particular I would like to thank him for taking a chance on recruiting a non-computational scientist for a PhD in Bioinformatics.

Many thanks also go to my co-supervisor Mahesan Niranjan, he has provided many hours of help and enthusiasm, in particular with respect to the machine learning approaches applied within this thesis.

A lot of gratitude goes to the members of Dr. Christopher H. Woelk's research group, they have been a constant source of encouragement and counsel with regards to my research but also have become close friends: Akul Singhania, Yawwani Gunawardana, Bastiaan Moesker, Jeongmin Woo and not least Michael Breen.

Yawwani Gunawardana also deserves an individual mention for her continued guidance with regards to the machine learning techniques employed in this thesis.

I would also like to thank other collaborators for continued input and scientific guidance, such as Dr. Ann Rawkins and Professor Helen McShane.

Abbreviations

aa	Amino acid
ACC	Auto Cross Covariance
APC	Antigen Presenting Cell
AUC	Area Under the Curve
BCG	Bacille Calmette-Guérin
BHK-21	Baby Hamster Kidney Cells
bp	Base Pair
BPA	Bacterial Protective Antigen
BPAD	Bacterial Protective Antigen Dataset
BPADb	Bacterial Protective Antigen Database
C	Cost Function
CFU	Colony Forming Unit
cm	Centimetre
DA-PLS	Discriminant Analysis by Partial Least Squares
DOMV	Detergent Extracted Outer Membrane Vesicle
E. Coli	*Escherichia coli*
ELISA	Enzyme-linked Imunosorbent Assay
eBPAD	Extracellular Subset of BPAD200
FACS	Fluorescence Activated Cell Sorting
fHbp	Factor H Binding Protein
FPR	False Positive Rate
iBPAD	Intracellular Subset of BPAD200
LOBOV	Leave-one-bacteria out-validation
LTOCV	Leave-tenth-out cross-validation
MDR	Multiple Drug Resistance
MenB	*Meningococcal meningitidis* sero group B
MHC	Major Histocompatibility Complex
µg	Microgram
µl	Microlitre
ML	Machine Learning
Mtb	*Mycobacterium tuberculosis*
NadA	Neisserial Adhesion A
NERVE	New Enhanced Reverse Vaccinology Environment
ng	nanogram
NHBA	Neisseria Heparin Binding Antigen

NHS	National Health Service
NN	Neural Network
PAMP	Pathogen-associated molecular pattern
PBS	Phosphate Buffer Solution
PCR	Polymerase Chain Reaction
Pfam	Protein Families Database
PHE	Public Health England
PorA	Porin A
PRR	Pattern Recognition Receptor
RBF	Radial Bias Function
ROC	Receiver Operating Characteristic
RV	Reverse Vaccinology
SBA	Serum Bactericidal Activity
SVM	Support Vector Machine
TAE	Tris-acetate with EDTA
TB	Tuberculosis
Th cells	T-helper cells
TPR	True Positive Rate
TLR	Toll Like Receptor
VC	Vaccine Candidate
XDR	Extensively Drug Resistant

Chapter 1: Introduction

1.1: Vaccinology

Vaccinology has long been an important arm of modern medicine and is the only form of research that has completely eradicated an infectious disease (i.e. Smallpox)[1]. Vaccination against smallpox was also the first documented use of vaccination and was undertaken in 1796 in Gloucestershire (England) by Edward Jenner[2]. Jenner noticed that a milkmaid infected with a bovine disease, cowpox, had gained protection against the more virulent smallpox virus. Pus was extracted from the milkmaid's arm and used to inoculate an 8 year old boy, James Phipps. After the inoculation Jenner tested the vaccination by exposing Phipps to the smallpox virus. Despite exposure to smallpox Phipps remained healthy, James Phipps was immune to the virus[1-3]. It should be noted that the eventual eradication of smallpox came as a result of using a vaccinia virus vaccine which replaced the use of cowpox as a vaccination agent against smallpox[4]. This early example of vaccinology still typifies our modern ideologies of a successful vaccine, which is that a lasting immune response will be generated. A lasting immune response is able to confer protection from future exposure to the disease. As well as eradicating smallpox, vaccine research has also rendered 26 infectious diseases preventable[2]. It is hoped that with greater levels of research and funding other diseases such as HIV, Malaria and TB will also one day be completely preventable by vaccination.

Successes achieved by vaccinology should not be under stated, it is attributed with saving the lives of three million children annually[1]. Yet despite this there is room for tangible improvements. An example of this is that by using the vaccines currently licensed for clinical use in a more effective manner, it is estimated an additional two million children could be saved annually[1]. Vaccinology is currently being thrust to the forefront of modern medicine due to the ever-increasing drug resistance of pathogens. In the United States alone, two million people are infected with antibiotic resistant bacteria per annum, which directly costs the US healthcare system in excess of $20 billion[5]. This growing resistance has been slow to be acted upon with funding bodies eager to fund therapeutics rather than preventative research[1]. However there is now a wider acceptance that vaccinology offers one of the most cost effective and realistic chances for controlling infectious diseases. Advances could be made through improvements in the use of current vaccines and also through the development of novel vaccines for diseases that currently have a partially preventative vaccine or for which there is currently no effective vaccine. This thesis contributes towards generating

novel vaccines, by exploring methods using *in silico* screening (i.e. Reverse vaccinology (RV)) to identify novel proteins in bacterial pathogens that act as protective antigens and may be incorporated into subunit vaccines.

1.2: Vaccine Immunology

Despite a lack of knowledge about the underlying mechanism the early attempts to confer protection through vaccination were similar to our practices today. In order to be able to design more effective vaccines with fewer side effects it is important to have an understanding of what occurs when a pathogen infects a human host and what types of immune responses results in protection. The immune system is split into two parts, the innate and the adaptive responses[3].

1.2.1: The Innate Immune Response

Innate immunity is activated first and is a more general response than the adaptive immune response, which is specific for different antigens. The innate immune system allows the host to rapidly respond to the site of an infection. It also plays an important role in activating the adaptive immune system[6]. When bacteria infect the body, the innate immune system will recognise the foreign pathogen and begin the process of disrupting the infection. This is accomplished in part by the innate immune system being responsible for killing pathogenic cells. One way that the innate immune system can cause the killing of a pathogen is by the complement system. The complement system is made up of proteins that are found the blood and act as a rapid response to a pathogen, as well as supporting antibodies in the adaptive branch of the immune system[7]. Complement activation triggers an enzymatic pathway that results in the recruitment of inflammatory cells, opsonisation of pathogens and the killing of pathogenic cells by membrane-attack complexes creating holes in the lipid bilayer of cell membranes[7]. Another way in which the innate immune system kills pathogens is through a process known as phagocytosis, which is carried out by Neutrophils and Macrophages. Cells infected with a pathogen are recognised by the innate immune system using molecules called pattern recognition receptors (PRRs)[6, 8]. The PRRs recognise well conserved bacterial pathogen-associated molecular patterns (PAMPs)[6]. When pathogen infected cells are recognised by the innate immune system inflammation engulfs the surrounding area. Inflammation is caused by the activation of macrophage cells due to the release of cytokines and chemokine's upon PRR activation. Released chemokine's and cytokines also play a role in the activation of the second branch of the immune system, the adaptive immune system[6].

There are many ways of activating the adaptive immune system and a common path is via the innate immune response's professional antigen presenting cells (APC) such as dendritic cells. Dendritic cells respond to PAMPs through PRRs called toll like receptors (TLR)[6]. They phagocytose the bacteria and then present it's antigens on the surface of the dendritic cell in a major histocompatibility complex (MHC). MHCs are made up of two types, MHC-I and MHC-II. Dendritic cells present the antigens on an MHC-II molecule and this is why they are known as professional APCs. Once a dendritic cell has been activated through pathogen interaction[9] it migrates to the lymph nodes, through the lymphatic system and performs its primary function which is to activate CD4 T-helper cells (Th cells)[10]. Dendritic cells with antigen presenting MHC-II molecules bind to T cell receptors on naive CD4 T cells to stimulate the production of activated CD4 Th cells[11].

1.2.2: The Adaptive Immune Response

The adaptive immune response is slower than the innate immune response but it is specific to each antigen and has a "memory" that means with subsequent exposures to a pathogen the adaptive immune response will be faster and stronger[8]. The adaptive immune response can be split into two parts, cellular and humoral.

Cellular immunity consists of two main types of T cell, CD4 and CD8. All T cells recognise antigens that have been processed and presented through an MHC molecule via T cell receptors. T cell receptors have randomly generated binding sites that allow specific recognition of antigens presented on MHC molecules, which allows a T cell response to individual pathogens as opposed to general bacterial PAMPS. CD8 T cells recognise antigens presented on MHC class I (MHC-I), this type of MHC is found on all nucleated cells[8]. Once CD8 T (T-cytotoxic) cells are activated they destroy any cell presenting the antigen that the CD8 T cell recognises (i.e. infected cells)[12]. CD8 T cells are only involved in the cellular branch of the adaptive immune system but CD4 T cells are involved in both the cellular and the humoral response and are commonly called T-helper cells (Th cells). Th cells recognise antigens presented in the MHC-II molecule and can be grouped as, Th1 or Th2 cells[13]. Th1 cells activate cellular immune responses by releasing chemicals such as interferon gamma. Th1 cells also become memory T cells that persist for long periods of time within the body thus providing the "memory" of the cellular immune response.

In the humoral arm of the adaptive immune response, activated Th2 cells cause the differentiation of B cells to form plasma cells (antibody releasing cells) and memory B cells. Plasma cells release antibodies that can bind to antigens from the pathogen[12]. B cell antibodies bind to the antigens in their native form and do not require the antigen

to have undergone processing and presentation on an MHC molecule[5]. Firstly the plasma cells produce the antibody IgM, however after a process of clonal expansion (to generate more clones of the effective antibody producing B-cell) this antibody is replaced by an IgG type, which is more effective at binding to antigens[8]. Antibodies do not break down the pathogen directly; they inhibit the pathogen by binding to a bacterial antigen and tag the pathogen for other parts of the immune system to degrade. Antibodies target the bound pathogen for destruction either by phagocytes (i.e. Neutrophils, Macrophages) or the complement system[3]. It is because of the memory B and T-cells that are generated by the Th cells that the adaptive immune response can react more quickly to successive exposures of the same pathogen. In summary, this section has described how the human immune system reacts to an infection with a bacterial pathogen and how long lasting immunity is generated. The goal of a vaccine is to generate such long lasting immunity through the formation of specific T- and B-memory cells. This is most successfully achieved by stimulating a robust immune response that involves both the innate and the adaptive immune system.

1.2.3: Types of Vaccine

A successful vaccine is able to deliver pieces of a pathogen to the host immune system so that immunological memory is triggered, and there are many methods of achieving this (i.e. live attenuated, inactivated and subunit vaccines). Live attenuated vaccines are an injection of the living pathogen into the host. The pathogen has been significantly weakened in the laboratory through mutations to vital areas of it's DNA or through culture techniques (serial passage) that force the pathogen to adapt to different conditions than those found within the host (i.e. growing in different species, growing pathogen at lower temperatures)[14]. When this "altered" pathogen is injected into the host as a vaccine the pathogen will not replicate efficiently to cause a dangerous infection but will still induce immunological memory[14]. Modern laboratory techniques are able to induce mutations in desired locations that limit pathogenicity. The major advantage of live attenuated vaccination is that a strong immune response is induced as the immune system encounters the whole pathogen and has to mount a defence for this "controlled" infection. Due to the robust immune response generated live attenuated vaccines need a low number of doses to generate a lasting protection. The major drawback however is that live attenuated vaccines have a higher level of risk due to the pathogen's potential ability to revert back to a virulent phenotype[8]. An example of this reversion was seen in a vaccine against the viral pathogen polio, the Sabin type 3 polio vaccine was shown to revert to a virulent polio infection[15]. Live

attenuated vaccines are more difficult to develop against bacterial pathogens due to large genomes and associated numbers of genes. Thus bacterial pathogens are more difficult to control reversion back to a pathogenic strain and therefore generate a safe vaccine. Another drawback is the need to refrigerate live attenuated vaccines. In summary, live attenuated vaccines are comprised of a whole pathogen and generate robust immune responses but do come with a risk of reversion to a pathogenic phenotype.

Inactivated vaccines contain killed copies of the disease causing pathogen. The pathogen will have been killed in a laboratory either using chemicals, heat or radiation[8]. Inactivated vaccines are safer than live attenuated, as the dead pathogen can never revert back to a pathogenic phenotype. However inactivated vaccines stimulate a weaker immune response and may require booster vaccinations to maintain immunity[8]. The main advantage of inactivated vaccines is that they do not require refrigeration and can therefore be more easily distributed. Inactivated vaccines primarily induce humoral immunity but a drawback is that repeated doses are required to generate protection. In summary, inactivated vaccines are killed pathogens and an example of an inactivated vaccine is the cholera vaccine, Shanchol (Shantha Biotechnics) which is made up of killed whole *Vibrio cholerae* cells[16].

Subunit vaccines are composed of one or more purified components, commonly epitopes, proteins or polysaccharides combined with an adjuvant to boost the immune response[17, 18]. Advantages of subunit vaccines are that they are well tolerated, as they do not contain elements of the whole pathogen that could cause severe immune reactions. Furthermore, using components from multiple strains of the pathogen, subunit vaccines can induce protection against multiple strains and subtypes[17, 19]. The ability to incorporate complete proteins into subunit vaccines, allows proteins to fold into natural 3D structures such that discontinuous epitopes may be reconstituted[19]. Discontinuous epitope representation is important as it is estimated that up to 90% of B-cell epitopes are conformational and that host antibodies bind to conformational epitopes with a stronger neutralising effect than linear epitopes[20]. The components making up a subunit vaccine alter the effect on the immune system that these vaccines elicit. Historically polysaccharide vaccines could not elicit a T cell immune response, but due to advances in vaccine design this is now possible by conjugation to protein molecules[8]. This increases the robustness of the immune responses and the immunological memory elicited by such polysaccharide vaccines[14]. The disadvantage of subunit vaccines is that multiple doses are required to generate lifelong immunity[14]. An example of a subunit vaccine is the BEXSERO vaccine (Novartis) against *Neisseria meningitidis* serogroup B, which is composed of five antigens present as three proteins

(i.e. two of the three proteins are fusion proteins that represent two antigens in one protein, for a full explanation please see **Seciton 1.8**) as well as a detergent extracted outer membrane vesicle[21].

This thesis employed the field of RV in an attempt to generate novel vaccine candidates for eventual inclusion into subunit vaccines. To facilitate an expedient, novel impact of RV subunit vaccines an organism for which traditional vaccine research has currently not resulted in an effective vaccine was targeted, *Mycobacterium tuberculosis* (*Mtb*)[22].

1.3: *Mycobacterium tuberculosis*

Tuberculosis (TB) is caused by infection of an intracellular bacterial pathogen, *Mtb*. This pathogen most commonly infects the lungs (pulmonary tuberculosis)[23, 24] and is the cause of the most deaths due to infectious disease worldwide[23]. TB causes the death of four people every minute[24] and these high rates of infection persists despite one of the most comprehensive vaccine strategies worldwide. Even with a massively distributed vaccine, the physical numbers of TB disease have not decreased in the past decade[24] (**Section 1.4**) and therefore it is of upmost importance that this trend is halted. It should be noted that the lack of success in this field is not due to a lack of funding; an estimated 6.3 billion US dollars went into TB research in 2014[25]. New and emerging avenues of research such as RV provide exciting methods for generating novel candidates that might provide efficacy against TB.

Almost one third of the world's population is living with TB in a latent phase of infection. Approximately 90% of *Mtb* infections will be contained by the hosts immune system in a latent state within a mass inside the lung called a granuloma[24]. When in this latent infectious stage *Mtb* cannot be transmitted to another host[23]. If the TB infection does progress to an active stage then 45% of otherwise healthy individuals will die without treatment[23]. TB can enter the active infection stage of its life cycle due to the immune system being depleted, such as if an individual contracts HIV, develops diabetes, or becomes malnourished[23]. When in the active stage of infection the bacteria can be passed to another host, this occurs by droplets containing *Mtb* being transferred to another host through mechanisms such as; sneezing, coughing or laughing when in close proximity with others. During active infection, symptoms can be mild for months and this enables the spread of TB by the infected host before they realise the severity of their infection and seek treatment[23].

Traditionally TB was treated with drugs such as isoniazid and rifampicin that have provided a strong protection against active infection[25]. Multiple drug resistant (MDR) strains of TB infection are becoming increasingly common. These strains are resistant

to at least isoniazid and rifampicin as well as other common anti-TB drugs[23]. In addition, new strains of extensively drug resistant TB (XDR TB) are emerging which are resistant to nearly all drug treatments[26]. XDR TB has already been confirmed in 58 countries[24] and in 2013 480,000 infections with MDR/XDR TB were reported[25]. With MDR and XDR TB on the rise a more efficacious vaccine is required to prevent the spread of *Mtb* infection.

1.4: Current Vaccines Against Tuberculosis Infection

Drug resistant strains of *Mtb* are becoming more prevalent and therefore more emphasis is being put upon the prevention of infection, commonly by vaccination. The most widely administered vaccine worldwide confers protection against TB and this is the Bacille Calmette-Guérin (BCG). The BCG was developed in 1921 and approximately four billion doses have been administered[27]. However, the BCG vaccine exhibits largely different performances in efficacies (0-80%) in the pulmonary form of the TB[28]. It is widely accepted that in infants the BCG does confer protection but in adults (the most at risk age group) the BCG efficacy levels vary. Acceptance of this has resulted in the BCG being administered to new-borns and not teenagers in secondary schools as of 2005 in the UK[29]. Not only does the BCG confer varying rates of protection in adults from pulmonary TB, it has even been suggested that vaccination with the BCG allows the continued dissemination of *Mtb* by preventing childhood mortality (i.e. disease that would not be transmitted) but not infective adult pulmonary TB[30].

Due to the front line vaccination (i.e. BCG) exhibiting variable protection it is desired that new preventative (prophylactic) therapies be developed for use against *Mtb*. Large amounts of research has been conducted using potential vaccine candidates (VCs) that provide immunity against TB with 16 currently in clinical trials[27]. However even promising VCs for *Mtb* infection have faltered in large scale clinical trials, a prime example of this is the antigen AG85A. Attempts were made to boost the efficacy of the BCG which resulted in protection within guinea pig models of TB and successful phase one clinical trials[31]. This led to a strain of BCG being engineered that over expressed the antigens AG85A, AG85B and TB10.4[32]. However this only exhibited very slight improvement over wild-type BCG that was not significant[32]. Animal models of *Mtb* infection represent an important step, through which to identify novel potential VCs. It was with the aim of identifying novel VCs against *Mtb* infection that putative VCs predicted by RV were assayed in a murine model of *Mtb* infection in this thesis (**Chapter 2**).

The rising levels of drug resistant infections are not exclusive to *Mtb.* Drug resistance is echoed throughout bacterial pathogens in general as reported in a European intergovernmental conference, which called for a specific focus on vaccine research due to the rise of antimicrobial resistance[33]. Currently this trend, of antibiotic resistance, is becoming more threatening and recently the emergence of *Escherichia coli (E. coli)*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* that are resistant to the antibiotic colistin[34] has been threatening enough to break into the mainstream news (BBC and the Guardian). Through the work in this thesis I have developed a broad RV approach to predict VCs for all bacterial pathogens (**Chapter 3**). The focus of RV was applied to *Mtb* to test predictions of previous RV approaches (**Chapter 2**) and when evaluating the performance of a newly developed RV classifier (**Chapter 4**).

1.5: Reverse Vaccinology

Conventional vaccinology remains an active field of research but is increasingly supplemented with new branches of vaccinology, such as RV. Conventional vaccinology cultures a pathogen in the laboratory and purifies proteins or materials from that pathogen to be used as potential VCs. This process is time consuming and costly and also involves an extensive amount of laboratory work (i.e. culture, isolating VCs and validation). A typical RV pipeline occurs *in silico* and involves evaluating the whole of the genome/proteome of a pathogen and selecting proteins that will then become VCs through computational methods.

In this thesis the bioinformatics immunotherapeutic approach taken was focussed on bioinformatic vaccine design (RV) with the intention of discovering novel vaccine candidates. To this end RV builds upon initial immunotherapeutic approaches by combining multiple sources of information to create predictions that model not only epitope prediction but a wide range of biological phenomena. Bioinformatic vaccine design can be achieved through a number of computational approaches and they are largely grouped under the umbrella of RV. For the purposes of this thesis RV will refer to the process of identifying whole proteins for incorporation into subunit vaccines, which has already led to the successful licensing of the subunit vaccine BEXSERO[35] and this is discussed in detail in **Chapter 1.5.1**. Other examples of RV include, epitope vaccine prediction and structural vaccinology[36]. Epitope vaccine prediction builds upon bioinformatics approaches to predict epitopes and aims to stimulate protection through forming a vaccine out of a series of epitopes. The epitopes to be included in potential vaccines are identified from the genome/proteome of a pathogenic species either by bioinformatics epitope prediction tools or by comparison to laboratory verified epitopes. Commonly the chosen epitopes are then compared across different strains of the

pathogenic organism by a Blast search[37]. Epitopes that show cross strain conservation are then incorporated into potential vaccines and their hypothetical population coverage is predicted utilising tools such as EPISOPT[38]. Some examples of epitope vaccine approaches are: "Towards the knowledge-based design of universal influenza epitope ensemble vaccines"[39], PEPVAC[40] and "Prediction of Epitope-Based Peptides for the Utility of Vaccine Development from Fusion and Glycoprotein of Nipah Virus Using *In Silico* Approach"[41].

The other main type of bioinformatics vaccine design, sometimes considered as RV, is structural vaccinology. Structural vaccinology is a process where the 3D structures of known epitopes are taken into account to generate novel antigens but more commonly to enhance vaccine candidates. It has been suggested that structural vaccinology should be combined with predictions made by other branches of RV to enhance the vaccine candidate's antigenicity and vaccine formulation approaches[19]. Resolving the crystal structures of proteins containing one or more protective epitopes led to the idea of structural vaccinology and this technique can be used to design minimised antigens which retain one or more key epitopes. The main advantage of structural vaccinology is that it can utilise 3D structure and can overcome some differences that are seen at the amino acid (aa) level but not on a 3D structure level. The other main advantage of structural vaccinology, which is also applicable to other branches of vaccinology, is that they can be used to target antigen variable pathogens by incorporating antigens/epitopes from multiple strains of the pathogen. In the particular case of structural vaccinology it has been demonstrated that multiple strain specific epitopes can be engineered onto a single immunogen[42]. Some examples of structural vaccinology are, "Structural vaccinology: structure-based design of influenza A virus hemagglutinin subtype-specific subunit vaccines"[43], and "Exploiting the *Burkholderia pseudomallei* Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology"[44].

As documented above, the RV approaches described within this thesis focus on identifying novel antigens that can confer protection against bacterial pathogens. From here onwards in this thesis RV refers to a bioinformatics vaccinology approach that identifies whole bacterial proteins as potential vaccine candidates. RV has been primarily focused on bacterial over viral pathogens due to the complexities of their protein coding genomes[45-51]. A major limitation of the RV approach is that only protein vaccine candidates can be identified[19]. In summary, RV is a branch of vaccinology that has led the successful development of a vaccine (BEXSERO) and has the potential to discover VCs for any pathogen with a sequenced genome.

1.5.1: Reverse Vaccinology Early Success

BEXSERO is the first vaccine developed from an RV approach that is in widespread clinical use[35]. BEXSERO is a subunit vaccine that protects against *Neisseria meningitidis* serogroup B (*MenB*) that was the most common form of bacterial meningitis in Europe[52]. Meningitis occurs when the bacteria infect the meninges of the body and can cause inflamation and swelling in this area, which results in increased pressure on the brain, and can lead to death. Patients who suffer from meningococcal meningitis suffer a mortality rate of between 10 to 15% and regular epidemics occur along the African menigitis belt[53]. Vaccines against meningococcal meningitis are able to have a positive impact in areas where they have been succesfully administered[54]. Vaccines for meningococcal serogroups (A, C, Y and W135) had already been created via conventional vaccinology approaches, using the polysaccharide capsule[55]. However, for *MenB* these capsular polysaccharides were poor immunogens and caused autoimmunity in humans[52, 55, 56]. Pizza and colleagues in the laboratory of Rino Rappuoli were the pioneers of an early RV approach, which applied a filtering method to the genome of *MenB* strain MC58[52]. In their process they identified the protein coding genes in the genome of MC58 and used bioinformatics programs to predict subcellular localisation. The tools used to predict subcellular localisation were PSORTb[57], ProDom[58] and the Blocks[59] database. After these initial steps they were left with 570 proteins, which were cloned and expressed in *E. coli*. Of the 570 proteins, 350 were expressed as recombinant proteins, purified, used for immunogenicity assays and had their predicted surface expression confirmed. These characteristics were interrogated by enzyme-linked imunosorbent assay (ELISA) and fluorescence activated cell sorting (FACS) techniques. Finally, three proteins (*i.e.*, Neisseria heparin binding antigen (NHBA, NMB2132), Factor H binding protein (fHbp, NMB1870), Neisserial adhesion A (NadA, NMB1994)) were chosen that showed conservation across multiple *MenB* strains[52, 60]. Novartis Vaccines incorporated these proteins to formulate a vaccine named BEXSERO. The final subunit vaccine BEXSERO was comprised of three recombinant proteins, NMB1994, NMB2132 as a fusion protein with NMB1030, and NMB1870 as a fusion protein with NMB2091[21, 60, 61]. These three recombinant proteins (five antigens) were combined with a detergent extracted outer membrane vesicle (DOMV) suspension. This DOMV was derived from the *Neisseria meningitidis* strain NZ98/254 and the primary antigenic component of the suspension was Porin A (PorA)[62, 63]. The BEXSERO vaccine is now being incorporated into the National Health Service (NHS) childhood vaccination program[64, 65] and has also been licensed for clinical use in the EU, Canada and Australia[35]. It

should be stated that the predicted advantage of increased speed in an RV approach has yet to be realised. BEXSERO took over a decade to develop from RV predicted VCs to a final vaccine. However, RV does enable the entire bacterial proteome to be evaluated for the discovery novel VCs.

1.5.2: Filtering Approaches to Reverse Vaccinology

Filtering approaches to RV started with the entire proteome of a bacterial pathogen and passed the constituent proteins through a series of filters until a small number of putative VCs remained[19]. A step incorporated into all filtering approaches was the removal of cytoplasmic proteins. The first automated RV filtering pipeline was The New Enhanced Reverse Vaccinology Environment (NERVE) (**Figure 1.1**)[47]. The NERVE pipeline was implemented in two stages and utilised eight separate perl scripts. The two stages of the NERVE pipeline were data production and data selection. Data production utilised protein annotation tools to annotate proteins within a pathogen's proteome and consisted of; subcellular localisation, predicted by PSORTb[57]. Topology predictions as predicted by HMMTOP[66]. Adhesin prediction as predicted by SPAAN[67]. Similarity to human proteins, which utilised the blast algorithm[37]. Following the data production stage the data selection stage was undertaken, this is where the proteome was filtered. First proteins that were predicted to be cytoplasmic and had more than two trans-membrane domains were removed. Secondly a filter to remove proteins with a low prediction of being an adhesin and with a risk of causing autoimmunity was applied. The cut offs for the similarity to human proteins and probability of being an adhesin filters were determined by tuning these values in 10 proteomes containing known antigens. Proteins that remained after all of the filtering stages were classed as VCs and were presented in an html table. The next RV filtering pipeline, Vaxign[45], built upon NERVE by creating a user friendly web interface for the pipeline as well as adding MHC class I and MHC class II epitope prediction to the evaluation of predicted VCs[45].

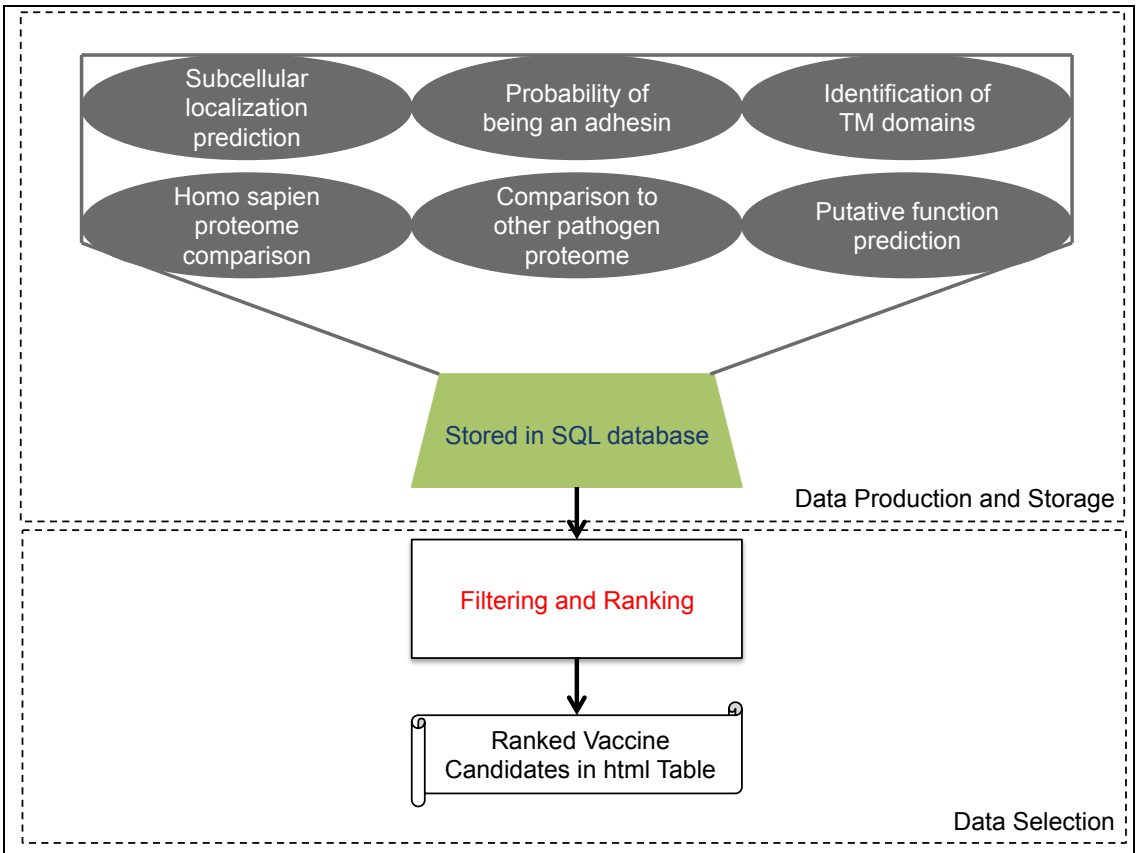


Figure 1.1: NERVE Reverse Vaccinology Pipeline. NERVE used protein annotation tools to generate information to describe proteins in a proteome for vaccine candidate (VC) prediction. This information consisted of subcellular localisation, predicted by PSORTb[57]. Topology predictions as predicted by HMMTOP[66]. Adhesin prediction as predicted by SPAAN[67]. Similarity to human proteomes, which utilised the blast algorithm[37]. Initially cytoplasmic proteins with more than two trans-membrane domains were removed. Next proteins with a low chance of being an adhesin and a high similarity to human proteins were excluded. Cut off values were tuned on ten proteomes that contain known antigens. The proteins that were predicted to be good VCs were then presented in an html table. Adapted from Vivona et al[68].

The most recent RV filtering pipeline is the Jenner-predict server (**Figure 1.2**), which can be found online http://14.139.240.55/vaccine/home.html[46]. The Jenner-predict pipeline achieved greater accuracies than all previous filtering based approaches and took into account host pathogen interactions. This was achieved by using known functional domains and the protein classes that they relate to (i.e., adhesin, virulence, invasion, porin, flagellin, colonisation, toxin, choline-binding, penicillin-binding, transferring-binding, fibronectin-binding and solute binding). The Jenner-predict pipeline first removed cytosolic proteins, predicting if a protein was cytosolic or not by using the subcellular localisation predictor PSORTb[57]. If the protein passed this filter HMMTOP[69] was used to predict the presence of trans-membrane helices. If a protein had less than 3 trans-membrane helices then it passed to the next filter. The Pfam[70] (protein families database) filter used hidden Markov models to compare the potential domains within the protein. The Jenner-predict pipeline characterised a "master list" of protein domains that are involved with host-pathogen interactions and pathogenesis. Proteins without domains present in the master list were considered poor VCs and were removed. After the initial filtering stages (i.e. subcellular localisation, trans-membrane domains and protein family domains) three measures were used to assess the remaining proteins vaccine potential, immunogenicity, autoimmunity and conservation. Immunogenicity was assessed by comparing the proteins to known B-cell and T-cell epitopes in the IEDB database using blast[37]. Proteins were considered a match for an epitope if a blast of 80% identity match with a minimum of nine amino acids (aa) length. Autoimmunity of potential VCs was evaluated using blastp[37] through two methods, implementing a cut off of 35% identity in at least 80 aa length or a continuous identical matching sequence of 9 or more aa in the alignment. Cross-strain conservation of VCs was evaluated using blastp[37] to compare proteins across strains of the same pathogen. If a protein was found with a blastp cut off of greater than 85% sequence identity and a minimum of 90% query coverage then cross stain conservation was deemed to be positive. Predicted VCs are then output in a ranked table that used the immunogenicity, autoimmunity and cross conservation scores for ranking. This ranking was achieved by proteins with more epitope matches being ranked more highly, and proteins with homology to humans being marked down.

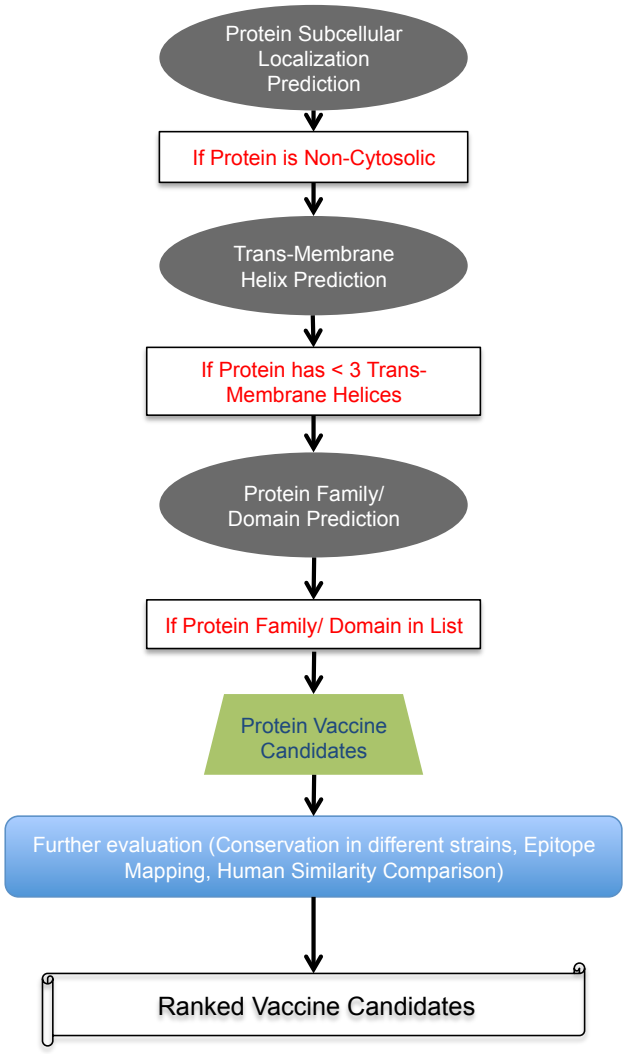


Figure 1.2: Jenner Predict Reverse Vaccinology Pipeline. This pipeline first removes proteins based on subcellular localisation (PSORTb[57]), number of trans-membrane helices (HMMTOP[66]) and known immunogenic protein families (Pfam[70]). Once putative vaccine candidates (VCs) have been selected for a ranking criteria was then generated, using epitope mapping (experimentally verified B and T cell epitopes were downloaded from IEDB, http://www.iedb.org), conservation across strains (blast[37]) and similarity to human proteins (blast[37]). Adapted from Jenner-predict server: prediction of protein VCs in bacteria based on host-pathogen interactions[46].

The main drawback of filtering approaches to date was that they followed the conventional school of vaccinology thought that all VCs would be located on the exterior of the bacterial pathogen. In **Chapter 3**, a literature curation of 200 protective antigens validated in animal models was shown to contain 14% that were predicted to be localised within the cytoplasm (PSORTb[57]). All filtering approaches would have removed such cytoplasmic proteins, and thus these proven protective antigens would have been excluded from the analysis. This was a major motivation for the development of Machine Learning (ML) approaches to RV that are able to identify putative VCs throughout the entire proteome of bacterial pathogens regardless of subcellular localisation.

1.6: Machine Learning

ML in RV is a small but promising field of research and to enable a thorough understanding of this area of research ML is described in this section before moving on to ML in RV. ML is a branch of computer science that uses algorithms to learn generalisable patterns in datasets. ML has a wide range of applications from spam filters to stopping credit card fraud and even predicting the winner of the football world cup[71]. The actual task of undertaking ML can be done in a number of different ways, using a number of different algorithms. Due to the wide breadth of ML applications and implementations it is hard to come up with a strict definition of ML. I would describe it as using a computer to learn rules from training data and then applying those rules to unseen data (i.e., a test dataset). Briefly, a ML classification pipeline for two classes (i.e. a positive class and a negative class) is conducted by first generating training data. The training data is comprised of the combined positive and negative training datasets. From this training data, information (annotation features) is passed to the ML algorithm to train the ML classifier. The aim when training a ML classifier is to learn generalisations from the training data, that distinguish between the positive and negative classes. This trained classifier is then validated to assess the performance of the trained classifier. For a working example of an ML classifier please see **Section 1.7**.

1.6.1: Algorithms for Classification

There are many types of ML algorithms that can be implemented when conducting ML classification, with new variations being developed continually. Some algorithms are more common and these include, Decision Trees, Neural Networks and Support Vector Machines[72].

Decision tree classification can suffer from lower accuracies than other algorithms but the main advantage of using decision trees is that due to their linearity,

the classification decision can be easily viewed and interpreted[73]. A method to increase the accuracies obtained by decision tree classifiers is known as random forests[74]. Random forests are a combination of decision trees, where multiple decision trees are used to predict a classification and the average prediction is used as the classifiers output[74]. Random forests enable a boosting of accuracies obtained by decision trees but also lose some interpretability when compared to individual decision tree classification. Decision trees are created from a root node and then from the root node, new nodes are created by the data undergoing iterative feature selection steps. These steps segregate samples through rules such as determined threshold cut-offs, which best divide the feature space[73, 75]. This process repeats until the terminal node is generated and a classification decision is made. Due to the number of layers in decision trees, they are not very computationally efficient classifiers[75]. The advantages of decision trees are that they are robust to noise and are simple classifiers to build. The disadvantages are, they can become very complex, needing many trees to classify the data and are therefore not efficient, also they are unable to consistently achieve as high accuracies as some of the other classification algorithms. Decision trees are interpretable classifiers, but suffer from lower accuracies than some of the other common classification algorithms.

The Neural Network (NN) algorithm is thought of as a sophisticated option for classification problems. NN classifiers are made up of many layers, an input layer, output layer, and any number of hidden layers. The more hidden layers that a NN has the more abstract features an NN can learn (i.e. deep NN), however, most commonly NNs are made up of 3 layers in total (1 hidden layer). The input layer takes a vector of the features to be learnt from and passes these on to the hidden layers, which apply a different weight for each hidden unit, a bias term and transforms the data using a non-linear function. This process repeats for every hidden layer within a network until the algorithm, reaches the output layer. During the training process of this algorithm the bias and weighting terms have to be learnt, this is conducted by minimising a loss function. The loss function is minimised through a process known as backpropagation[73, 76]. Advantages of NNs are that they can model complex relationships that are not obvious in the original dataset due to transformations at the nodes and if trained correctly they can be very accurate. This is especially true when using deep NNs. Disadvantages are that they are not understandable and with many layers will require a large amount of computational time. NNs and deep NNs represent a very accurate classification algorithm but require considerable expertise and time to train correctly.

One of the most widely utilised classification algorithms is the Support Vector Machine (SVM). Vapnik first introduced SVMs in 1963 and this approach to classification has since been shown to be very robust, even when dealing with noisy data[77]. SVM is one of the most commonly used classification algorithms, particularly in bioinformatics and often achieves superior accuracies to other classification techniques[75]. The drawbacks to SVM classification are that a kernel function needs to be selected and the model is hard to interpret. It uses a "black box" decision module whereby it is not clear how the SVM learns rules to classify data. SVMs were employed for the ML RV approach detailed in this thesis due to, high accuracies in classification, ease of implementation and the comparability of results to previous work[48].

An SVM classifies data by finding the maximum margin hyperplane between two classes[73, 75, 78]. A visualisation of this can be achieved by assuming that a classification problem has two features and classification classes are separable by a straight line (a two dimensional hyperplane). One can draw several lines (**Figure 1.3,** L1, L2 and L3), which separate the two groups, but the optimal SVM classifier will set the decision margin as the line which gives the maximal margin between the two classes (**Figure 1.3,** L1). This is known as the maximum margin hyperplane (**Figure 1.3**). When training an SVM there is one parameter that applies for all types of SVM and this is the cost function (*C*). *C* determines how much to penalise a classifier for misclassifications. Using a high *C* results in a complex model that may not generalise well to other data. Choosing a low value *C* will result in a model with high variance that has not modelled the training data correctly. *C* therefore directly affects how hyperplanes are drawn and is optimised when training a SVM[79]. For example, in certain data the separating margin may not be separable with a straight line due to contamination, sample miss labelling or anomalous results, an SVM deals with these samples by implementing something that is known as the soft margin. The soft margin allows the miss classification of some samples to maintain a greater overall classification with a better maximal margin hyperplane. The soft margin can be adjusted by the *C* value and is a balance between miss-classifying too many examples and maintaining a good, generalisable hyperplane.

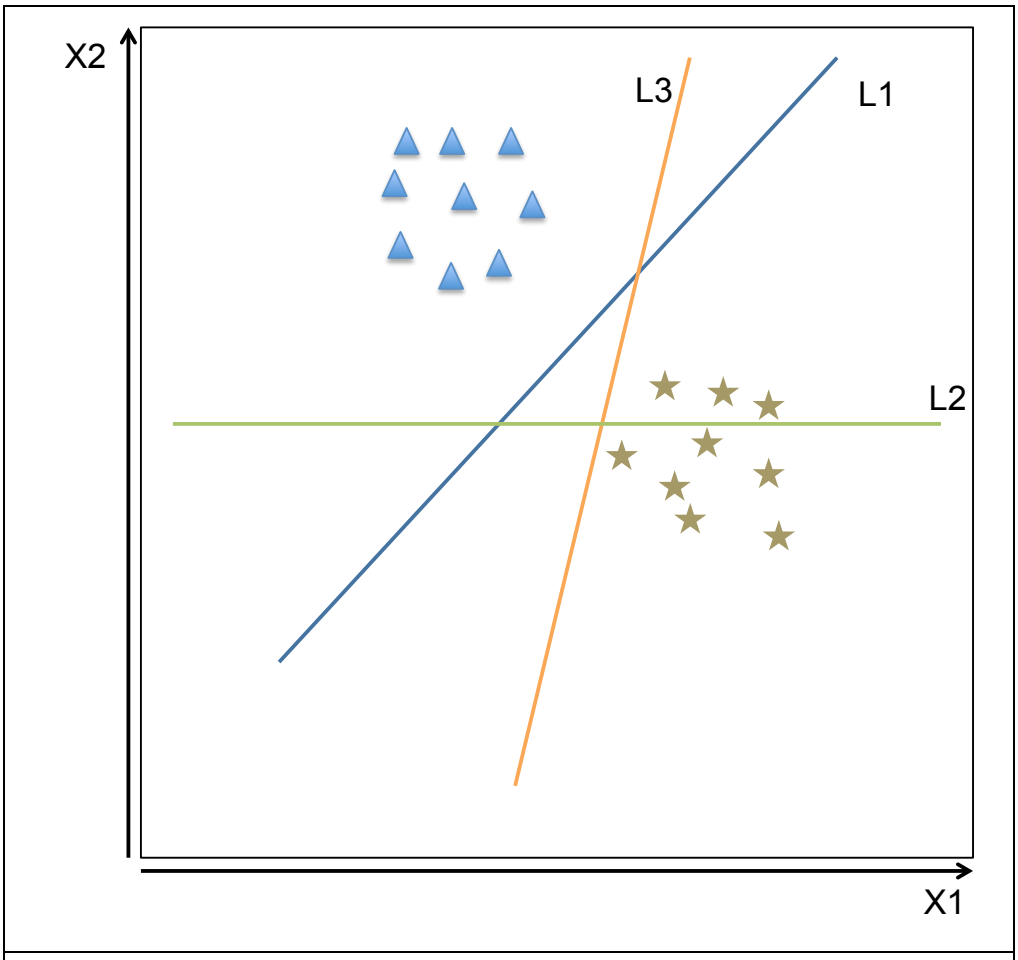


Figure 1.3: A Representation of Possible Hyperplanes generated by a Support Vector Machine (SVM) Trained on Two Features. In two-dimensional space there are many lines (hyper planes in multi dimensions) that would separate these two groups of data (blue triangles and dark yellow stars). An SVM selects the line with the largest distance between the two groups (i.e. the maximal margin hyperplane). In this example that would be the line L1. X1 and X2 represent two features used to train a SVM. Adapted from[78].

Another important concept of SVM's is the kernel function. Sticking with the 2 features example in 2D space some classes may not be separable by a straight line. The SVM kernel function projects this two dimensional data onto higher dimensions, this will result in it being separable with a straight line. It is possible to prove that for any data set with consistent labels a kernel function exists that will separate the data linearly. However, with an increasing number of features the possible solutions for this increases exponentially and this is known as the "curse of dimensionality"[78]. When raising the data to higher and higher dimensions it becomes harder for an algorithm to find the answer and this can lead to overfitting. The kernel is a vital part of the SVM and can be manipulated to give better accuracies; it is common practice to try simple kernels when first training a classifier to see which gives better accuracies. The kernel dramatically increases the power of an SVM by enabling its use on linearly un-separable data. To enable non-linear SVM classification the SVM classifiers implemented in this thesis utilised a radial bias function kernel (RBF). Utilising the RBF kernel enabled a direct comparison to the previous RV classifier[48], on which this thesis built. Despite the ability of ML to classify even noisy datasets (i.e. SVM) often a stage of feature selection is required to achieve maximal classification accuracies.

1.6.2: Feature Selection

Feature selection removes less informative or noisy features to leave the smallest number of features that can achieve maximal accuracies. It is desirable to remove features for several reasons. Firstly, that removing noisy features results in higher accuracies for ML classifiers. Secondly, by removing features you make the model less complex and this speeds up the process of ML (i.e. training the classifier and making predictions on unknown data). Thirdly, utilising a smaller number of features can make it easier to understand the ML classifier and how class predictions are being made. There are methods of feature selection that independently evaluate each feature's deterministic ability and there are also methods that take into account the cumulative effects of features. Since it has been shown that features that are not informative individually can be informative when evaluated together[80], this thesis employed a feature selection strategy that took into account the cumulative effect of features when performing feature selection. This strategy was greedy backward feature elimination. Greedy backward feature elimination is a computationally exhaustive method of feature selection. The greedy backward feature elimination algorithm starts off by including all features in the dataset, removing one feature at a time and building a classifier. The removed feature is then replaced and another feature is removed. This continues until every feature has been omitted once. The feature that has the least

effect (the least informative) on the accuracy of the classifier is then discarded (eliminated). If more than one feature is deemed as equally the least informative, then one of the joint least informative features is eliminated at random. The process then repeats until all features are removed or a desired optimal feature number is reached. Due to the random process incorporated in greedy backward feature elimination it is often recommended that the procedure be repeated iteratively to break ties generated through the random step in greedy backward feature elimination.

1.6.3: Validation

To assess the accuracies of ML classifiers a separate test dataset must be utilised. Commonly, in conventional ML studies, the training data is split into a training dataset and a test dataset, where the test dataset is left out when training the classifier and performing feature selection (i.e. Hold out validation). This test dataset can then be used to evaluate the classifiers ability to make predictions on unseen data for a low amount of computational time. Another example of a validation method that is used to estimate the performance of a ML classifier is k-fold cross-validation[81]. K-fold cross-validation has been shown to model the generalisation error better than the more traditional hold out validation process, but is more computationally expensive[81]. When applying ML to biological research there is often not a large enough dataset to perform a holdout validation and k-fold cross-validation is implemented. The RV classifier developed in this thesis (**Chapter 3**) used a k-fold cross-validation, specifically, k=10, leave-tenth-out cross-validation (LTOCV). LTOCV is performed by removing one tenth of the training dataset when training a classifier, this "left out" tenth then becomes a test dataset. The process repeats until the whole dataset has been used as a test dataset, one tenth at a time. LTOCV cross validation enables a realistic estimate of a classifiers performance on unseen data even when limited training data is available. A final example of how one could validate an ML classifier in a practical way is by testing the predictions of the ML classifier. An example of this occurs in this thesis (**Chapter 2**) where predictions from an RV classifier of proteins that would confer protection against *Mtb* (i.e. bacterial protective antigens) were used for testing in mouse models of infection.

It is becoming increasingly common to encounter ML in the field of Bioinformatics and there is a growing library of software packages that facilitate this. Some packages, languages and programs that enable ML without one having to code the algorithms themselves are: Matlab[82], Python packages (SciKit-Learn[83]), R[84] (i.e. libsvm[85]) packages and also individual stand alone programs such as WEKA[86]. The growth of big data analysis using ML can be shown in many fields, but the potential of RV with

the power of ML could result in putative VCs being taken into clinical trials for vaccines within the next few years.

1.7: Machine Learning & Reverse Vaccinology

ML approaches to RV are able to take into account all proteins within a pathogens proteome and make predictions for each protein as to whether it should be considered as a putative VC. Two published studies utilise ML in RV. The pioneering study of Doytchinova and Flower resulted in the VaxiJen classifier[87]. A positive training dataset of 100 known antigens was generated through a literature curation effort. This curation defined proteins as known antigens if the protein or part of the protein had been shown to induce a protective response in an appropriate animal model after immunisation. Construction of an ML classifier relies on an equal number of each prediction group being represented in the training data, in this case known antigens and non-antigens. To generate a matching negative training dataset (i.e. non-antigens) proteins were randomly sampled from the same bacterial species as each protein in the positive training dataset. By chance this random sampling of a bacterial proteome could include antigens that had yet to be described. In an attempt to limit the inclusion of undescribed antigens a blastp similarity comparison was implemented where any newly sampled non-antigen with an expectation value (E-value) > three to a protein in the positive or negative training datasets the protein was resampled. Next features were generated from the positive (i.e. known antigens) and the negative (i.e. non-antigens) training datasets, this was achieved by using auto cross covariance (ACC) transformations. These ACC transformations transformed the proteins to a uniform length and captured information from the proteins such as molecular size, hydrophobicity and weight. Then a discriminant analysis by partial least squares (DA-PLS) was performed and the VaxiJen classifier was able to achieve an accuracy of 82% when discriminating non-protective from protective proteins.

Bowman et al[48] built upon Doytchinova and Flower's (i.e. VaxiJen)[87] initial approach by creating a RV classifier that was used to distinguish bacterial protective antigens (BPAs) from non-BPAs. BPAs were different from the antigens curated in Doytchinova and Flower's approach to ML in RV[87] in that a BPA was only included if the whole protein was shown to be protective in an animal model and not part of the protein. The specific definition of a BPA was a bacterial protein that when injected into an animal model gave significant protection ($p < 0.05$) following immunisation and subsequent challenge with the bacterial pathogen (i.e. bacterial load reduction or survival assay). Through incorporating protective proteins from the previous ML RV approach that met the definition of a BPA and undertaking a literature curation 136

BPAs were identified to make up the positive training dataset. In accordance with Doytchinova & Flower's approach[87], to generate the negative training dataset this study used randomly selected proteins from the same bacterial species as each positive BPA. Similarly to the Doytchinova and Flower's approach[87] this study also implemented a blastp[37] screen whilst generating negative training data. Blastp was used to exclude randomly sampled non-BPAs that had > 98% identity with any antigen in the positive training dataset. To generate a diverse negative training dataset, if a selected negative training protein (non-BPA) had an E-value of < 10E-3 when compared to other non-BPAs, it was discarded and re-sampled. When the proteins to be included in the classifier (136 BPAs and 136 non-BPAs) had been selected annotation features that the ML classifier could learn were needed. Instead of implementing ACC transformations to generate features as in Doytchinova and Flower's RV classifier (i.e. VaxiJen[87]) Bowman et al[48] generated features from the training data by running the proteins through a series of protein annotation tools. The outputs from these tools were then parsed. Nineteen protein annotation tools were used to derive 122 annotation features (**Appendix A**). Bowman et al were able to classify BPAs from non-BPAs with a maximal accuracy of 92%[48].

As the work detailed in this thesis builds upon the RV classifiers developed by Doytchinova et al (Vaxijen)[87] and Bowman et al[48], a description of the work carried out in the most recent ML RV manuscript follows (i.e. Bowman et al[48]). Firstly Bowman et al sought to show improvements over the VaxiJen classifier[87]. The new method of generating features was compared to the previous RV approach[87] (i.e. protein annotation tools as opposed to ACC transformations) (**Figure 1.4**). It was shown that using the ACC descriptions gave lower accuracies than using features derived from protein annotation tools. It was also observed that combining the ACC descriptors and protein annotation tool derived features gave no improvement in accuracies. Bowman et al proceeded using only the features generated using protein annotation tools.

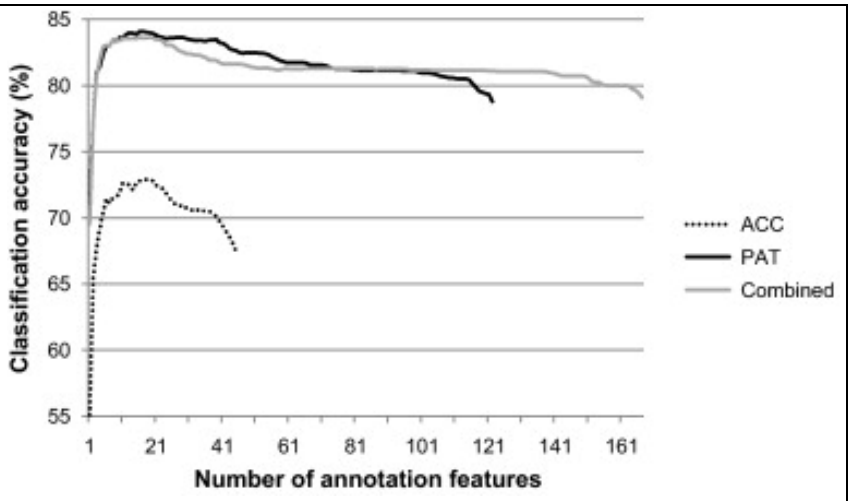


Figure 1.4: Different Types of Annotation Features have an Effect on the Classification Accuracies of BPAs from non-BPAs. Shows the ability of the DA-PLS classifier to discriminate the Doytchinova and Flower[87] training data using ACC annotations and Bowman et al[48] annotations derived from protein annotation tools, as well as the accuracies obtained when both annotation sets are combined.

After generating an annotated dataset several ML algorithms; SVM, DA-PLS (used in VaxiJen[87]) and Linear Regression were compared to see which obtained the greatest accuracy (**Figure 1.5**). Bowman et al also compared the training dataset in their study with the previous ML study in this field (VaxiJen classifier[87]). Bowman et al found that SVM classification consistently yielded the highest accuracies when classifying BPAs and non-BPAs. They also showed that there were negligible differences in accuracies between the VaxiJen[87] annotated dataset and the dataset generated in their research (Bowman et al[48]). It was decided to proceed by exploring classifiers built using the SVM classification method on the newly developed, larger dataset of BPAs and non-BPAs[48].

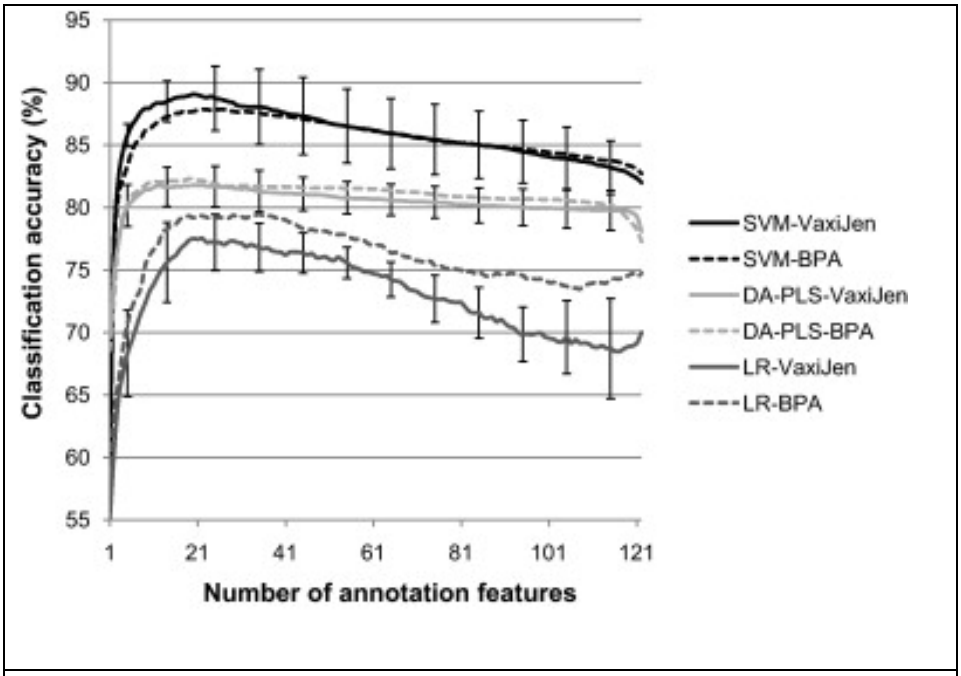


Figure 1.5: Evaluation of Different Machine Learning Approaches and Datasets on Classifying BPAs and non-BPAs. Comparing the effect of using Support Vector Machines, Linear Regression and DA-Partial Least Square Regression when making classifications of BPAs or non-BPAs. This figure also compares two different training sets, Bowman et al[48] and Doytchinova and Flower (i.e. VaxiJen)[87].

Next the study attempted to minimise the impact of randomly incorporating un-described BPAs into the negative training dataset of non-BPAs. To mitigate this, five negative training datasets were generated to evaluate which dataset was able to train the classifier to achieve the highest accuracy (i.e. has the greatest ability to separate BPAs from non-BPAs). Negative training dataset three trained the most accurate SVM classifier (**Figure 1.6A**). The 136 BPAs curated from the literature when combined with the 136 non-BPAs from the negative training dataset three were named BPAD136. After proceeding with BPAD136 Bowman et al showed that a maximal accuracy of 92% could be achieved when classifying BPAs and non-BPAs (**Figure 1.6B**).

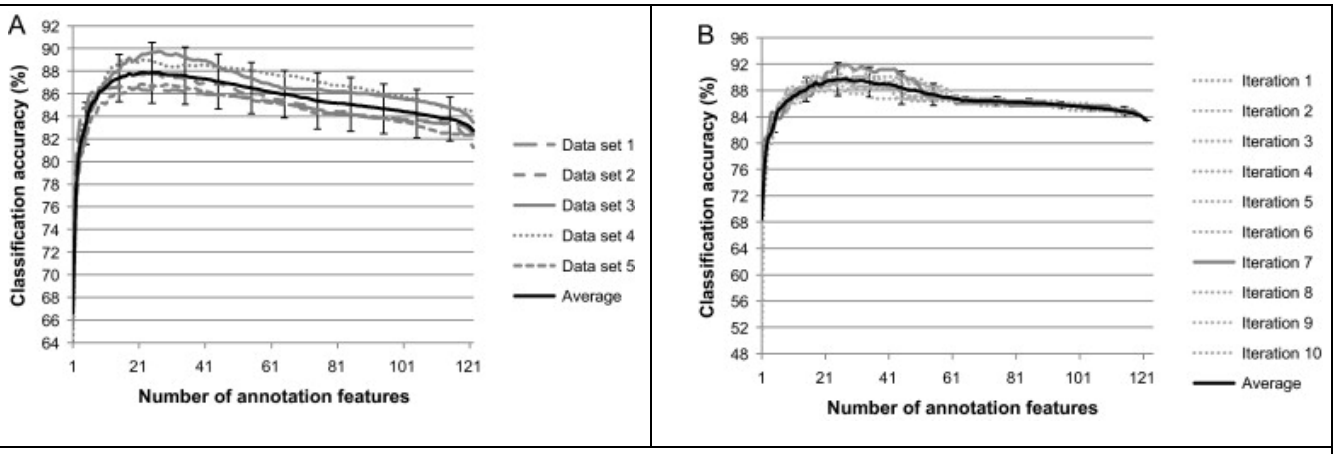


Figure 1.6: Support Vector Machine Classification Accuracies when Classifying BPAs and non-BPAs Utilising Different Negative Training Datasets and Across Multiple Iterations: A Shows the accuracies obtained when using different randomly generated negative training datasets, with a constant positive training dataset. **B** Shows the maximal achievable accuracies when employing the Bowman et al RV pipeline[48].

Next the annotation features utilised by the SVM trained on BPAD136, obtaining the highest accuracies were used to try and elucidate what determines the difference between BPAs and non-BPAs. To do this an F score was calculated for each annotation feature. The F score compares the ratio of variability between the two groups (BPA and non-BPA) to the variability within each group[88, 89]. Features with the largest F scores were determined to be more powerful in discriminating BPAs from non-BPAs. The main drawback of the F score metric was that it did not take into account cumulative effects of features. The top 10 annotation features were represented in a pie chart (**Figure 1.7**).

The most deterministic feature between BPAs and non-BPAs (**Figure 1.7**) was "PSORTb-ProbCytoplasm", which was a feature that predicted whether or not the protein is localised to the cytoplasm. The main perceived benefit of ML in RV as opposed to filtering approaches was that ML RV took into account the entire proteome (i.e. proteins from all subcellular localisations) when predicting BPAs (i.e. VCs). Bowman et al[48] revealed that the most deterministic feature of predicting a non-BPA was that the protein is predicted to have a subcellular localisation of cytoplasmic (**Figure 1.7**). This suggested that ML RV utilising entire proteomes for predictions of BPAs was not an advantage. Filtering approaches to RV all exclude proteins with a subcellular localisation predicted as cytoplasmic as an early filtering criterion. As Bowman et al[48] deemed that a protein predicted to be localised in the cytoplasm is the number one most deterministic feature between BPAs and non-BPAs this would

appear to substantiate previous filtering RV methodologies. **Chapter 3** of this thesis improved upon this method, constructing a more biologically reflective classifier, from which it was possible to infer differences between BPAs and non-BPAs that reflect protective immunity and not subcellular localisation.

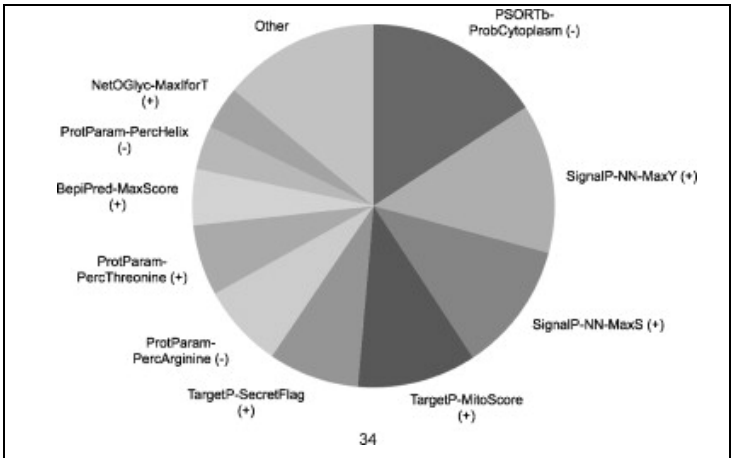


Figure 1.7: The Ten Most Discriminative Features for Classifying BPAs and non-BPAs. Pie chart showing the top 10 most discriminative annotation features by F score, the + or – sign represents the features relating to protection or not respectively[48].

Recall was first described by Bowman et al[48] and measures the ability of a classifier to recall known BPAs when in a background of an entire bacterial proteome. To undertake recall a classifier was used to rank every protein in a pathogens proteome for the predicted probability of being a BPA. The number of known BPAs recalled above a threshold (i.e. top 100 predicted BPAs) was assessed using a hypergeometric test (i.e. Fishers exact test). The hypergeometric test evaluated whether the classifier was significantly enriching the top 100 predicted BPA lists with known BPAs. Bowman et al[48] again compared their RV classifier to the previous RV classifier Vaxijen[87]. The Bowman et al[48] classifier was shown to recall all BPAs across six pathogenic proteomes with a higher rank than the Vaxijen classifier[87], accept for one BPA (**Figure 1.8**). Bowman et al then went on to show that the top 100 predicted BPA lists produced by the RV classifier that they developed was significantly enriched for known BPAs in all instances[48].

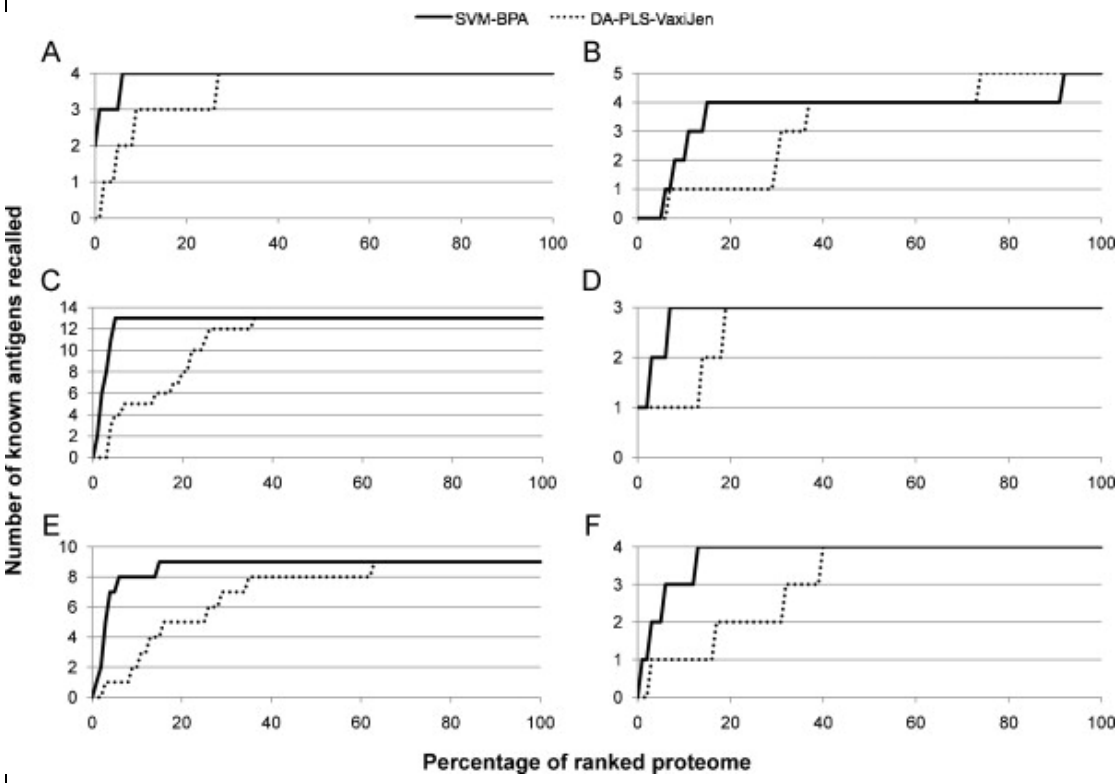


Figure 1.8: Recall Curves Generated when Recalling Known BPAs when in the Background of Bacterial Proteomes. Bowman et al RV classifier[48] and the VaxiJen RV classifier[87] were used to rank proteins in pathogenic proteomes for their predicted probability of being a bacterial protective antigen (BPA). The position of known BPAs in the ranked proteome was expressed as a percentage of the entire proteome. Pathogenic species (A) *Borrelia burgdorferi*, (B) *Helicobacter pylori*, (C) *Mycobacterium tuberculosis*, (D) *Staphylococcus aureus*, (E) *Streptococcus pneumoniae*, and (F) *Treponema pallidum*.

1.8: Chapter Overview

Chapter 2: Laboratory Validation of Vaccine Candidates Previously Predicted by Reverse Vaccinology

This chapter attempted to generate novel VCs for inclusion into a subunit vaccine for protection against *Mtb* infection. An RV classifier was used to make predictions of BPAs from the proteome of *Mtb.* Six predicted BPAs from the RV classifier were evaluated in a murine model of *Mtb* infection.

Hypothesis 2.1: The six VCs selected for animal challenge experiments would confer significant levels of protection against infection with the pathogen *Mtb.*

Chapter 3: Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology

This chapter addressed the lack of protection generated by the six BPAs tested in a murine model of infection (**Chapter 2**) by revising the RV approach.

Hypothesis 3.1: BPAs curated from the literature, contain a signal for protective antigens compared to randomly permuted data.

Hypothesis 3.2: Using a correctly nested leave-tenth-out cross-validation would reduce the accuracies achieved when classifying BPAs and non-BPAs.

Hypothesis 3.3: Increasing the size of the training data and the number and breadth of annotation tools would increase the accuracies obtained when classifying BPAs and non-BPAs.

Chapter 4: Evaluating the Ability of Enhanced Classifiers for Recalling Known Protective Proteins from Bacterial Proteomes

This chapter evaluated the enhanced RV classifier (BPAD200+N+B+AF) generated in **Chapter 3** using a biologically relevant metric, recall. BPA predictions from the enhanced classifier (developed in **Chapter 3**) were also compared to the six predicted BPAs from a previous RV classifier that were shown not to be protective through murine models of protection (**Chapter 2**).

Hypothesis 4.1: The BPAD200+N+B+AF classifier would be able to significantly enrich for known antigens in top 100 predicted BPA lists for bacterial pathogens.

Hypothesis 4.2: The six proteins assayed for protective efficacy in **Chapter 2** would not be significantly enriched in the top 100 predicted BPAs for *Mtb* using the BPAD200+N+B+AF classifier.

Chapter 2: Laboratory Validation of Vaccine Candidates Previously Predicted by Reverse Vaccinology

2.1: Introduction

Tuberculosis (TB), caused by infection with the pathogen *Mycobacterium tuberculosis* (*Mtb*), is still a serious problem for global health care. TB is estimated to have killed 1.5 million people in 2013[23] and the rise of multi-drug resistant (MDR) TB has also made the disease more difficult to treat with commonly used antibiotics. High rates of infection with TB persist despite one of the largest vaccination campaigns in history being deployed against TB with 100 million new born babies every year being vaccinated against the disease[90]. Recently there has been an increase in efforts to generate more effective vaccination regimes to prevent infection with *Mtb*. Currently the only vaccine available is the Bacillus Calmette-Guérin (BCG), which is a live attenuated vaccine derived from passaged *Mycobacterium bovis*. BCG confers good protection against miliary and meningeal *Mtb* however offers unsatisfactory protection against pulmonary *Mtb* which accounts for up to 70% of all infections[91]. Many vaccine trials using BCG have been carried out and varying levels of protection have been reported in different clinical trials (i.e. 80% protection in the UK but 0% protection in South India)[92]. There are many theories as to why this variation exists but the two leading causes appear to be due to prior exposure to environmental mycobacteria and the age at which the vaccine is administered[92]. Due to the varying degrees of protection attributed to the BCG vaccine in pulmonary *Mtb*, current research efforts focus on finding a more effective vaccine[93].

A field of research that could be used to predict novel vaccine candidates (VCs) for *Mtb* is reverse vaccinology (RV). RV is a branch of *in silico* vaccine research where the entire bacterial proteome is considered whilst predicting putative VCs (**Chapter 1, Section 1.5**). The field of RV has incorporated machine learning (ML) and a full description of this can be found in **Chapter 1, Section 1.7.** The ML RV approach developed by Bowman et al[48] was used in this chapter to predict putative VCs for testing in an animal challenge. Briefly Bowman et al[48] built upon previous ML RV approaches[87] by increasing the training dataset size through a literature curation for bacterial protective antigens (BPAs), generating biologically descriptive annotation features and implementing a support vector machine (SVM) to distinguish between BPAs and non-BPAs. A BPA was defined as a bacterial protein that when used to

immunise an animal model conferred significant levels of protection ($p < 0.05$) following subsequent challenge with the bacterial pathogen (i.e. bacterial load reduction or survival assay). The RV study of Bowman et al[48] obtained an accuracy of 92% when classifying BPAs and non-BPAs.

The work detailed in this chapter utilised the classifier developed by Bowman et al[48] to predict novel BPAs in *Mtb*. For the purpose of this chapter predicted BPAs were called putative VCs when selected for formulation into DNA vaccines (**Section 2.2.1**). As *Mtb* has no none correlates of protection[94], a putative VC would only be considered a VC if animal challenge models resulted in significant levels of protection against *Mtb* infection, following vaccination with a putative VC. Protection in animal challenge models was the outcome (BPA) that the Bowman et al[48] classifier was shown to be able to predict. It was envisaged that predictions of BPAs by the RV classifier developed by Bowman et al[48] (**Chapter 1 Section 1.7**) would be able to confer protection in animal challenge models of *Mtb* infection. The work detailed in this chapter generated six putative VCs that were then used as DNA vaccines to evaluate protection generated in a mouse model of *Mtb* infection.

Hypothesis 2.1: The six VCs selected for animal challenge experiments would confer significant levels of protection against infection with the pathogen *M. tuberculosis*.

2.2: Methods

2.2.1: *In Silico* Selection of Vaccine Candidates

Six predicted BPAs (i.e. putative VCs) were selected using three filtering criteria from the top 100 predicted BPAs for the *Mtb* proteome. The top 100 predicted BPAs were generated from the Bowman et al[48] RV classifier (detailed in **Chapter 1.7**). The first filtering criteria identified novel predicted BPAs by excluding proteins protected by vaccine related patents as well as proteins in known immunogenic families, such as the PPE and PE protein families. The PPE protein family contains an 180 amino acid N-terminal domain with PR, Pro-Glu, at positions nine and ten, and the PE protein family is characterised by the presence of 110 amino acid N-terminal domain with PE, Pro-Glu, at positions nine and ten[95]. The predicted BPAs were then subjected to a second filtering criteria for predicted trans-membrane domains using TMHMM[96] and those with three or more predicted trans-membrane domains were removed. Proteins with multiple trans-membrane domains were removed due to the fact that they are more difficult to clone and express in the laboratory[46, 52, 68]. A final filtering criterion used previously published *Mtb* expression data[97-100] to ensure predicted BPAs were expressed by *Mtb* at the transcript level, which indicated that the protein was being translated and thus exposed to the host immune system. The literature studies were used to denote an expression cut off to form the final filtering criteria by comparing the filtered predicted BPAs to known BPA expression scores in *Mtb* infection. To calculate known BPA expression scores a value was assigned (0,1,2,3) depending on whether the protein was ranked in the 4^{th} (lowest), 3^{rd}, 2^{nd} or 1^{st} quartile of expression levels for each study. These were then averaged across four studies[97-100] to give a single expression score for each protein. An average expression cut-off of 1.2 was implemented and thus predicted BPAs with an average expression score of less than 1.2 were excluded. From the final filtered list six predicted BPAs were selected by a panel of collaborators to be taken forward into mouse challenge experiments. The collaboration that selected the six selected BPAs for laboratory trials was comprised of prominent members of the TB vaccine community, laboratory animal trial researchers and ML in RV practitioners (Prof Helen McShane, Dr. Ann Rawkins, Ms. Yper Hall, Dr Elena Sylianou, Dr Christopher Woelk and myself).

2.2.2: DNA Amplification of Putative Vaccine Candidates

In order to amplify the DNA for incorporation of the six putative VCs into DNA vaccines, primers for polymerase chain reaction (PCR) were designed for each of the putative VCs. This was undertaken in collaboration with Public Health England (PHE) using

their protocol DGM015 (**Appendix B**). Briefly, gene sequences for the six putative VCs were obtained from tuberculist[101] (An *Mtb* database that integrates genome, protein, structural, transcriptome and drug information), the start and stop codons were removed and 18 bp from the start and the end of the gene were used as the primers. A pre-designed leader sequence (forward leader, GGGGACAAGTTTGTACAAAAAAGCAGGCT reverse leader, GGGGACCACTTTGTACAAGAAAGCTGGGT) was then added to these primers. These products were used as the primers for the PCR reaction. PCR was performed using KAPA HiFi as per the manufacturer's instructions; KAPA HiFi HotStart ReadyMix 75 µl, Forward Primer 4.5 µl, Reverse Primer 4.5 µl, genomic DNA (*Mtb* H37Rv) 12 µl, water (PCR-grade) 54 µl. This was split up into three reactions (three tubes) for each putative VC. The PCR reaction ran for the cycles listed in **Table 2.1**.

Step	Temperature	Duration	Cycles
Initial Denaturation	95 °C	3 min	1
Denaturation	98 °C	20 sec	35
Annealing	60 °C	15 sec	
Extension	72 °C	1 min	
Final Extension	72 °C	1 min	1

Table 2.1: Polymerase Chain Reaction Cycling parameters. Abbreviations: *min*; minute(s), *sec*; seconds.

DNA amplification of the putative VCs was confirmed by gel electrophoresis, which assessed the size of DNA fragments. Products from the PCR were run using a 1% agarose in Tris-acetate with EDTA (TAE) gel, using sybersafe gel stain (Cat No; 163795-75-3, Sigma, Missouri, USA). The wells were loaded with 5 µl DNA (i.e. putative VC), 3.3 µl of cyan yellow (Cat No: 10482035, Thermo Fisher Scientific, Waltham, USA) and 11.7 µl of water to give a total reaction volume of 20 µl. In an empty well 20 µl of ladder (1 µg/µl) was loaded (Invitrogen 1 kb plus DNA ladder). This

was run at 100 volts for 30 minutes. Putative VC1 was run on a second gel, using 1% agarose in TAE gel, using sybersafe gel stain (Cat No; 163795-75-3, Sigma, Missouri, USA). The wells were loaded with: 8.33 µl cyan yellow (Cat No: 10482035, Thermo Fisher Scientific, Waltham, USA), 50 µl of DNA (i.e. Putative VC1) and 40 µl of this mixture was loaded into the wells, in another lane 40 µl of ladder (1 µg/µl) was loaded (Invitrogen 1kb plus DNA ladder). This was run at 100 volts for 1 hour. All amplified DNA products (i.e. putative VCs) were purified using QIAquick PCR Purification kit (Cat No: 28104, Qiagen, Venlo, Netherlands) as per the manufacturer's instructions except for putative VC1. The DNA for putative VC1 was purified using a gel extraction kit 250 QiAquick, (Cat No: 28706, Qiagen, Venlo, Netherlands) this was used as per the manufacturer's instructions.

2.2.3: Cloning DNA into pVAX DNA Vaccines

The purified PCR products (i.e. Putative VCs) were transferred into a vector (pVax) to be used as a DNA vaccine. This reaction was carried out in two stages, first a BP reaction and secondly the LR reaction. This was achieved following the PHE's One tube Gateway Cloning protocol, DGM014 (**Appendix C**) using the Gateway Technology reagent kit (Cat No: 12535-019, Invitrogen, Carlsbad, USA). The complete BP reaction mix is listed in **Table 2.2** and was combined in a microcentrifuge tube and mixed by vortexing. The BP reaction mix was incubated at 25 °C overnight. The LR reaction was performed using 10 µl of the BP reaction mix and combining the LR reaction mix as detailed in **Table 2.3**, mixing again by vortexing. Next LR mix was incubated for 4 hours at 25 °C. After the incubation OneShot Top 10 *E. coli* cells (Cat No: C404010 Invitrogen, Carlsbad, USA) were subjected to heat shock to take up the plasmid that had been generated by the LR reaction. Heat shock to enable the uptake to the plasmid was achieved by adding 2 µl of proteinase K solution (Gateway Technology reagent kit, Invotrogen, Carlsbad, USA) to LR reaction mix and this was incubated at 37 °C for 10 minutes. The LR reaction (1.5 µl) was added to OneShot Top10 *E. coli* and incubated on ice for 5 minutes. The cells were then subjected to heat shock 42°C for 30 seconds and then returned to ice for 2 minutes. Next 250 µl of SOC medium (Cat No: 15544034, Thermo Fisher Scientific, Waltham, USA) was added and this mix was incubated for 1 hour at 37 °C at 220 rpm. Finally the *E. coli* was plated on L-agar plates containing 10mg/ml kanamycin at differing colony forming unit (CFU) concentrations by using differing amounts of the transformed *E. coli*. For each gene, 1 (µl), 10 (µl), 100 (µl) and the rest of the OneShot Top 10 *E. coli* tube was plated and left at 37 °C overnight. The gateway cloning procedure ensures selection of transformed colonies. If DNA was not inserted into the plasmid then there would have

been an active suicide gene within the plasmid and this would kill the bacteria. Another layer of protection built into the gateway system is that the bacteria only survived if the pVax plasmid was present, this contained a gene that conferred resistance to the kanamycin present in the growth media. Resulting colonies on this plate would therefore have a copy of the pVax plasmid present and also an insertion at the desired site within this plasmid. The plasmid map of pVax can be seen in **Figure 2.1**.

Protein	TE Buffer (µl)	pDONR Zeo (µl)	PCR Product (µl)	BP Clonase II (µl)	Total Volume (µl)
Putative VC1	5.7	1.3	5	3	15
Putative VC2	9.2	1.3	1.5	3	15
Putative VC3	9.2	1.3	1.5	3	15
Putative VC4	9.2	1.3	1.5	3	15
Putative VC5	9.2	1.3	1.5	3	15
Putative VC6	9.2	1.3	1.5	3	15

Table 2.2: Reaction Mixtures for BP reaction. Reaction mixtures used in the BP reaction of the gateway cloning protocol. Abbreviations: *TE Buffer*; Tris-HCL and EDTA, *pDONR Zeo*; forms the entry clone used in the gateway reaction finally used to insert the desired DNA into the pVax destination vector in the LR reaction.

Protein	BP Reaction Mix (µl)	pVax (µl)	LR Clonase II (µl)	Total (µl)
Putative VC1	10	2	3	15
Putative VC2	10	2	3	15
Putative VC3	10	2	3	15
Putative VC4	10	2	3	15
Putative VC5	10	2	3	15
Putative VC6	10	2	3	15

Table 2.3: Reaction Mixtures for LR reaction. Reaction mixtures used in the LR reaction of the gateway cloning protocol.

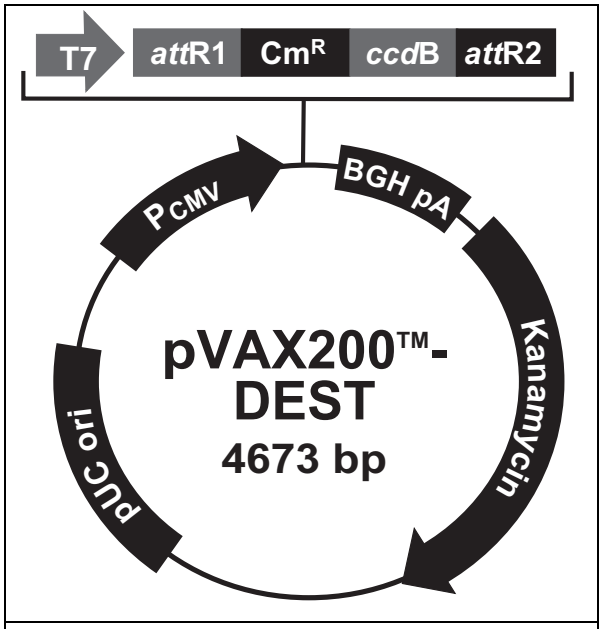


Figure 2.1: Pvax Plasmid Map: Plasmid map showing elements in the Pvax plasmid used as DNA vaccines. Abbreviations: *Pcmv*; Cytomegalovirus promoter, *BGH pA*; Polyadenylation signal, *pUC ori*; Origin of Replication, *ccdB*; Toxin ccdB gene, *attR*; Insertion sites for Gateway Cloning Protocol, *T7*; Prokaryotic promoter, Cm^R; chloramphenicol resistance gene.

The resulting DNA vaccines (pVax plasmid plus inserted putative VC) were purified using mini prep plasmid purification kits as per the manufacturers protocol (Cat No: 27106, Qiagen, Venlo, Netherlands). Following purification validation of the desired insert was confirmed by a restriction digest, which was again carried out as per the manufacturer's protocol provided by New England BioLabs (Ipswich, USA). The reaction was left to run for 15 minutes. Empty pVax (destination vector) and pDONR Zeo (entry clone) were used as controls. Products of the restriction digests were subjected to gel electrophoresis for size separation. A gel of 1% agarose in TAE, using sybersafe gel stain (Cat No: 163795-75-3, Sigma, Missouri, USA) was run. The restriction digest product for each putative VC (10 µl) was combined with 2 µl of gel loading buffer. This gel was run at 100V for 30 minutes**.** All putative VCs met the expected sizes of fragments, listed in **Table 2.4**. To confirm the presence of the desired inserts (i.e. putative VC) with no point mutations purified pVax vectors (i.e. DNA vaccines) were sent for sequencing at Beckman Coulter Genomics Sequencing (Essex, UK). Following confirmation that there were no point mutations and that the desired putative VCs had been successfully cloned into DNA vaccines, the amount the

DNA vaccines was amplified. Amplification of desired DNA vaccines was achieved by using confirmed, sequenced, colonies as a starter culture for 8 hours at 37 °C and the DNA vaccine plasmids being purified in larger quantities as per the manufacturers instructions in Qiagen-Endofree-Plasmid-Purification, Giga prep plasmid purification kits (Cat No: 12391, Qiagen, Venlo, Netherlands). Restriction digests were again used to confirm successful purification of the desired DNA vaccines, using the same methods and gel visualisation as above. Again, all putative VCs met the expected sizes of fragments, listed in **Table 2.5**.

Vector	Restriction Enzymes used for Restriction Digest	Size of Fragments	Total size of pVax plasmid in base pairs
pDONR Zeo	XmaI	2315, 1976	4291
Gateway pVAX Vector	HindIII, Xma I	3392, 1442	4834
pENTR-VC1	XmaI	2442	2442
pVAX-VC1	HindIII	3549	3549
pEntr-VC2	XmaI	2616	2616
pVAX-VC2	HindIII	3723	3723
pEntr-VC3	XmaI	2853	2853
pVAX-VC3	HindIII	3960	3960
pEntr-VC4	XmaI	2033, 859, 99	2991
pVAX-VC4	HindIII, XmaI	3381, 618, 99	4098
pEntr-VC5	HindIII, XmaI	1805, 973, 447	3225
pVAX-VC5	HindIII, XmaI	3495, 447, 390	4332
pEntr-VC6	XmaI	2628, 653, 439	3720
pVAX-VC6	HindIII, XmaI	3614, 1213	4827

Table 2.4: Expected Fragment Sizes Following Restriction Digest After Miniprep. Digest of the isolated predicted vaccine candidates cloned into a pVax vector. After the gateway cloning procedure and Miniprep plasmid purification.

Protein	Restriction Enzymes used for Restriction Digest	Size of Fragments
Putative VC1	PvuII	360, 3189
Putative VC2	PvuII	240, 360, 570, 2553
Putative VC3	PvuII	360, 954, 2646
Putative VC4	HindIII, XmaI	99, 618, 3381
Putative VC5	HindIII, XmaI	390, 447, 3495
	HindIII	3942, 390
Putative VC6	HindIII, XmaI	439, 1213, 3175

Table 2.5: Expected Fragment Sizes Following Restriction Digestion After Gigaprep: Digest of the putative vaccine candidates, following the second restriction digest after the Gigaprep procedure.

2.2.4: DNA Vaccine Expression in Mammalian Cells

Following the successful creation of pVax DNA vaccines the expression of the plasmids in mammalian cells was confirmed following transfection into Hamster Kidney Cells (BHK-21) using western blotting. The PHE Protocol "DGM18: BHK-21 Transfection Protocol for the Evaluation of pVax DNA Vaccines" (**Appendix D**) was used. Briefly, the DNA vaccine was diluted in OptiMEM (Cat No: 31985-062, Invitrogen, Carlsbad, USA), and then 5µl lipofectamine (Cat No: 18324-012, Invitrogen, Carlsbad, USA) was added to 20 µl OptiMEM/DNA mix. Cultured BHK-21 cells had the growth media removed and then 0.2 ml of MEM (Minimum Essential Media, Thermo Fisher Scientific, Waltham, USA) plus glutamine was added to the cells. The DNA OptiMEM complex (0.2 ml) was added to grown BHK-21 cells and left for 5 hours at 37°C. After this the media was replaced with fetal bovine serum (Cat No: F9665, Sigma, Missouri, USA) and kanamycin. The cells were grown for 2 days at 37°C. This protocol transfected the DNA vaccines into BHK-21. The expressed proteins were then visualised using western blotting. Another mycobacterial protein (i.e. Rv3537) that had previously been proven to successfully express in BHK-21 cells was used as a positive control. This was conducted as per the PHE DGM07 SDS PAGE and Western Blotting protocol (**Appendix E**). First separating the proteins on a gel, (Cat No: NP0321BOX Invitrogen, Carlsbad, USA), adding 10 µl MagicMark protein ladder for western blot

visualisation (Cat No: LC5952, Invitrogen, Carlsbad, USA) and 5 µl of SeeBlue pre-stained protein standard for gel electrophoresis visualisation (Cat No: LC5952, Invitrogen, Carlsbad, USA) in the same well and 20 µl of the transfection samples in other wells. Then the gel was run for 45 minutes at 200 Volts. Next the protein was transferred onto a nitrocellulose membrane (Cat No: RPN203D, GE healthcare, Little Chalfont, UK) inside a typical western blotting assembly (i.e. sponge, filter paper, gel, nitrocellulose membrane, filter paper, sponge, construction) this transfer occurred at 40 V for 60 minutes. The nitrocellulose membrane was then placed in blocking buffer (3% milk in PBS Tween) for 45 minutes. After removing the initial blocking buffer a solution of primary antibody (8.4 µl of primary antibody Cat No: MCA1360, AbD Serotec, Kidlington, UK) in 25 ml of blocking buffer was added and left for 45 minutes. Next the membrane was washed three times in PBS tween and exposed to the secondary antibody (8.4 µl, Cat No: A9044, Sigma, Missouri, USA) in 25 ml blocking buffer before a further three washes in PBS tween. The antibody tagged protein was then visualised using enhanced chemiluminescence kits as per the manufacturers instructions (Cat No: 32106 Thermo Fisher Scientific, Waltham, USA). Finally the DNA vaccine plasmids (pVax vectors) were diluted to 1 mg/ml using a NanoDrop 2000 (ThermoFisherScientific, Waltham, USA) in sterile phosphate buffer saline (PBS) solution. The DNA vaccines were aliquoted into six doses of 1.2 ml vaccines.

2.2.5: Mouse Challenge of *Mtb* Infection

To test the protective efficacy of the putative VCs, an infectious mouse model of *Mtb* was used[102]. Collaborators at The University of Oxford, Dr Helen McShane and Dr. Elena Stylianou conducted these experiments. This challenge experiment was comprised of eight groups of eight, six to eight week old female CB6F1 mice; two control groups, BCG (six weeks before challenge, $4*10^5$ CFU) and a naïve group, with six groups that were immunised with one of the six putative VCs. A total of 1 mg of pVax construct (i.e. DNA vaccine) in each vaccine was injected into a hind leg (intramuscular), three times with two week intervals. One month after the last immunisation, mice were subjected to aerosol challenge with 50-100 colony forming units of *Mtb* (**Figure 2.2**). The differences in CFUs in the lungs were then statistically determined using a Mann-Whitney test[103]. This challenge experiment was repeated in a second run after the first experiment generated a significant level of protection for one of the novel putative VCs. All procedures were performed in accordance with the Animals (Scientific Procedures) Act 1986 under project license number 30/2889 granted by the Home Office in the UK.

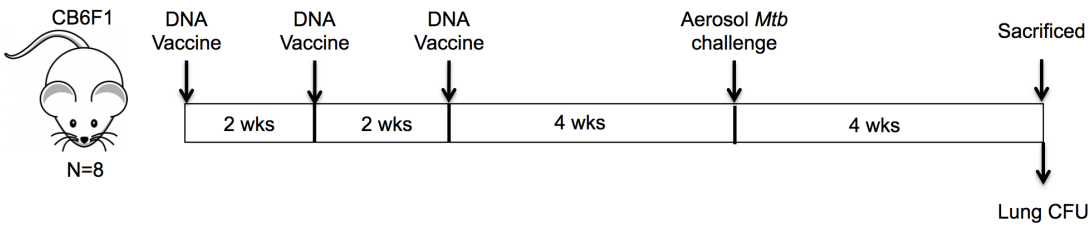


Figure 2.2: Mice Challenge Experimental Timeline. This figure shows the challenge experiment schedule testing the putative VCs. Mice (CB6F1 strain) were given 1 ml of 1 mg/ml of DNA vaccines 3 times with 2 week intervals. They were then challenged with *Mycobacterium tuberculosis (Mtb)*. BCG and naïve groups were used as controls. This experiment was repeated twice. Abbreviations: *CFU*; Colony forming units.

2.3: Results

2.3.1: Six Predicted BPAs were Selected for Validation in Mouse Models of *Mtb* Infection

First six putative VCs were selected from the top 100 predicted BPAs for *Mtb as* predicted by Bowman et al[48]. Three filters were applied to the top 100 predicted BPA for *Mtb* list, firstly proteins in known immunogenic families (PE and PPE) and described in vaccine related patents were removed, this left 62 BPAs from the top 100 predicted BPAs in *Mtb* list. Secondly predicted BPAs with more than two trans-membrane domains were excluded, 58 predicted BPAs remained after this filter. Finally a filter that measured the expression of predicted BPAs compared to known BPAs as an average across four microarray studies[97-100] (detailed in **2.2.1**) was implemented. Any remaining predicted BPA with an expression score of less than 1.2 was removed, this left 33 predicted BPAs that met the criteria deemed necessary to be considered for laboratory validation. From this list of 33 predicted BPAs six were selected by collaborators as a sample of predicted BPAs and were used in laboratory trials. The six selected putative VCs were: Rv3886c, Rv2190c, Rv2068, Rv1857, Rv1677 and Rv0608c (a table of the six annotated VCs can be found **Appendix F**). Selected candidates and their expected base pair lengths are shown in **Table 2.6.**

Gene	PCR product (bp)
Putative VC1 (Rv0680c)	427
Putative VC2 (Rv1677)	601
Putative VC3 (Rv1857)	838
Putative VC4 (Rv2068)	976
Putative VC5 (Rv2190c)	1210
Putative VC6 (Rv3886c)	1705

Table 2.6: Size of PCR product for each putative vaccine candidate: RV numbers were obtained from the H37RV genome from the Tuberculist[101] *Mtb* database. The purified PCR products from putative VCs were all of the expected size. Abbreviations: *VC;* vaccine candidate, *PCR*; Polymerase Chain Reaction, *bp*; Base Pair.

2.3.2: DNA of the Selected Vaccine Candidates was Successfully Amplified

The desired DNA sequences of the putative VCs were created by amplifying the genes in the laboratory from the genome of *Mtb H37Rv* using PCR. The primers used for this can be seen in **Appendix G**. The products of this PCR reaction were then visualised on a gel to confirm that the products were of the expected sizes (**Figure 2.3**). In all of the putative VCs except for putative VC1 the bands that correspond to the desired sizes are much more prominent than any others (i.e. the desired sequence of DNA has been amplified cleanly). PCR purification was then used to extract the DNA from the PCR products for all of the products except for putative VC1. Putative VC1 was not the only band visualised in the gel. A second gel (data not shown) was run for putative VC1, for a longer time to allow greater separation of the DNA fragments. A greater separation of DNA fragment sizes enabled a DNA gel extraction kit to purify the desired fragment of DNA for putative VC1. The DNA sequence for the six putative VCs were successfully amplified from the *Mtb* genome and the concentrations of DNA isolated are listed in **Table 2.7.**

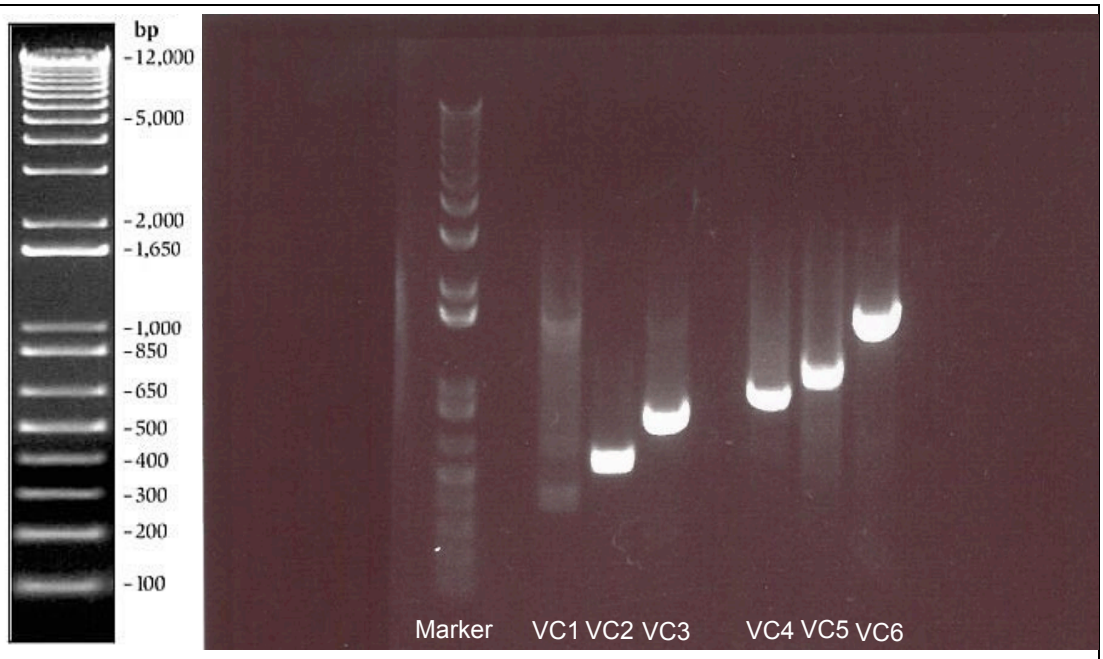


Figure 2.3: Electrophoresis Gel Depicting Polymerase Chain Reaction Products for each Vaccine Candidate (VC). DNA products from PCR reaction for each putative VC were run on an agarose gel to enable visualisation and determine fragment size. All DNA products were amplified successfully accept Putative VC1. Lanes left to right: Size key of Ladder, Ladder, Putative VC1, Putative VC2, Putative VC3, Putative VC4, Putative VC5 and Putative VC6

Protein	Concentration (ng/µl)	260/280	260/230
Putative VC1	25	1.96	0.003
Putative VC2	220.6	1.85	2.16
Putative VC3	320.4	1.85	2.13
Putative VC4	429.5	1.85	2.11
Putative VC5	379	1.84	2.11
Putative VC6	500.3	1.84	1.95

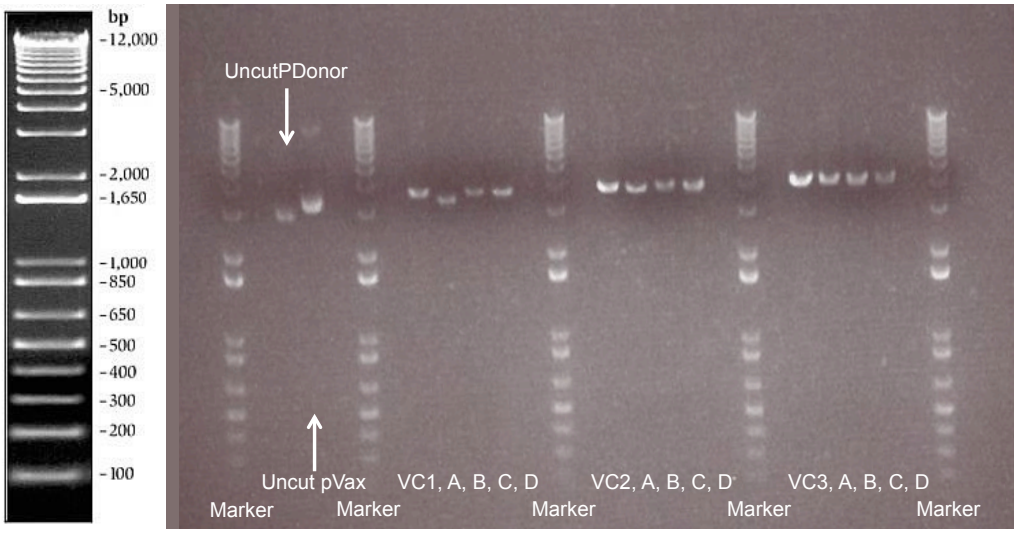
Table 2.7: Purified Vaccine Candidate DNA Concentrations. DNA concentrations Obtained after purification from polymerase chain reaction synthesis of vaccine candidates.

2.3.3: DNA was Successfully Cloned into pVAX DNA Vaccines

Next the amplified DNA of the putative VCs was inserted into pVax plasmids via the Gateway cloning procedure, to synthesise the final DNA vaccines. After the Gateway cloning procedure four colonies from each DNA construct (i.e. DNA vaccine) were selected and grown in culture. The plasmids (i.e. DNA vaccines) were then purified using Qiagen mini prep. As a quality control, a restriction digest was used to evaluate the inserts (i.e. putative VCs) within the pVax vector. Comparing the observed sizes in the gel (**Figure 2.4**) to the predicted fragment sizes confirmed that the correct gene (i.e. putative VC) was inserted into the plasmid and there were no cloning errors. To confirm point mutations had not randomly occurred the purified pVax vectors (i.e. DNA vaccines) were sent for sequencing from colonies 1D, 2D, 3A, 4A, 5A and 6D (**Figure 2.4**). All colonies contained the desired DNA insert with no point mutations. Once it was established that the correct DNA had been inserted into the pVax vectors, a gigaprep protocol was implemented to multiply and purify the DNA plasmids (i.e. DNA vaccine). A further restriction digest was used to confirm the presence of desired DNA inserts within the purified vaccines following gigaprep. The restriction digest

results following the gigaprep protocol were as expected and uncut plasmid DNA was predominantly supercoiled (**Figure 2.5**). Putative VC5 and putative VC3 were re-run due to low yields from the initial gigaprep protocols, the restriction digest after the second gigaprep can be seen in **Figure 2.5B**. All six putative VCs were successfully amplified and purified through gigaprep procedures and the following restriction digest again confirmed the presence of the desired DNA insert.

(A)



(B)

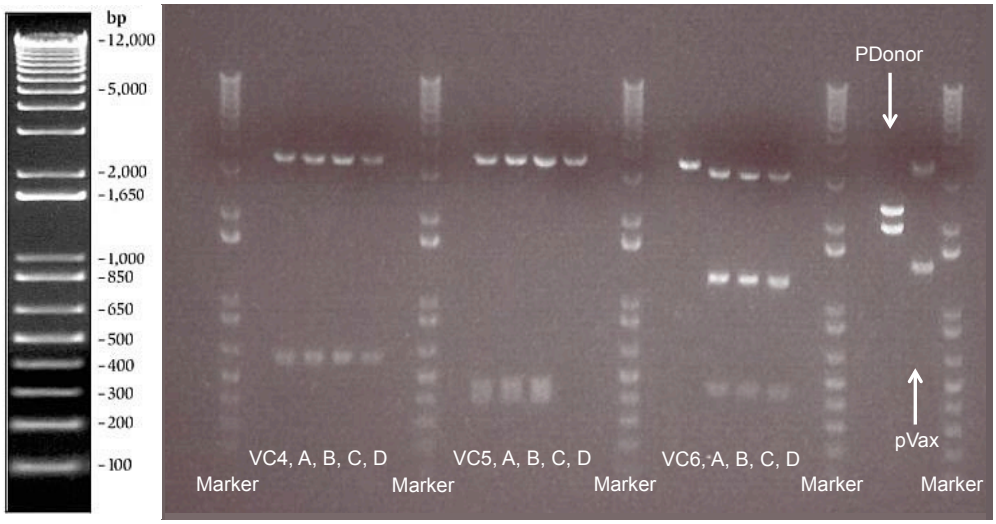
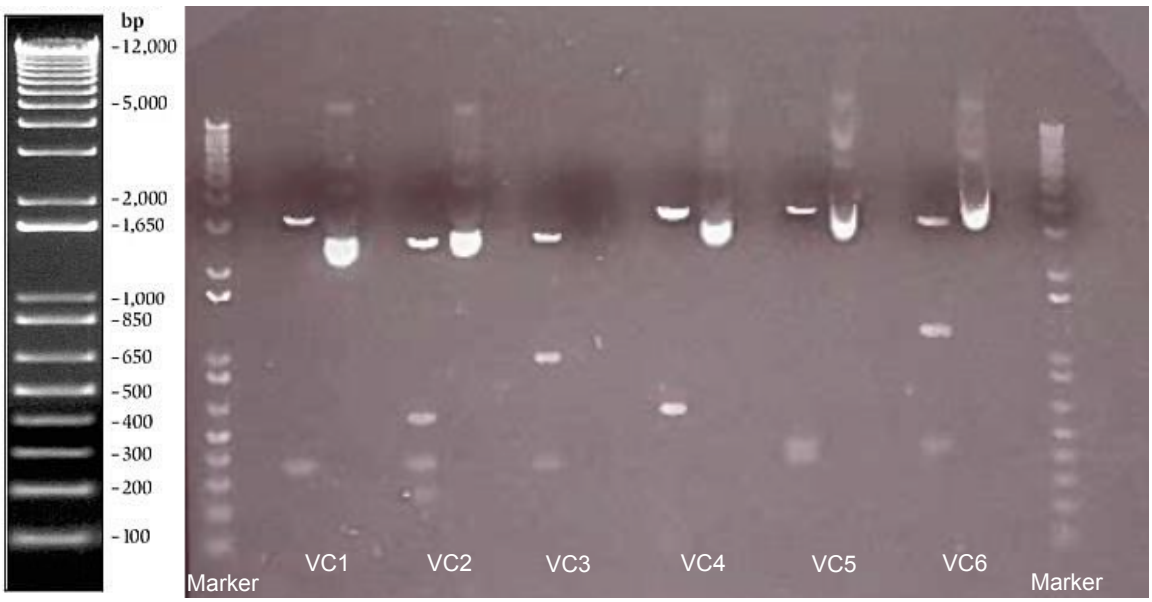


Figure 2.4: Restriction Digest Results for Isolated pVax DNA Vaccines. Figure showing the restriction digest on the miniprep for DNA vaccines. Four colonies were harvested for each vaccine candidate to maximise the chances of harvesting the correct insert. **(A).** Left to right: Ladder, Uncut pDONR Zeo, Uncut pVax, Ladder, (Putative VC1) 1A, 1B, 1C, 1D, Ladder, (Putative VC2) 2A. 2B, 2C, 2D, Ladder, (Putative VC3) 3A, 3B, 3C, 3D, Ladder. **(B)**. Left to right: Ladder, (Putative VC4) 4A, 4B, 4C, 4D, Ladder, (Putative VC5) 5A, 5B, 5C, 5D, Ladder, (Putative VC6) 6A, 6B, 6C, 6D, pDONR Zeo, pVax, Ladder.

(A)



(B)

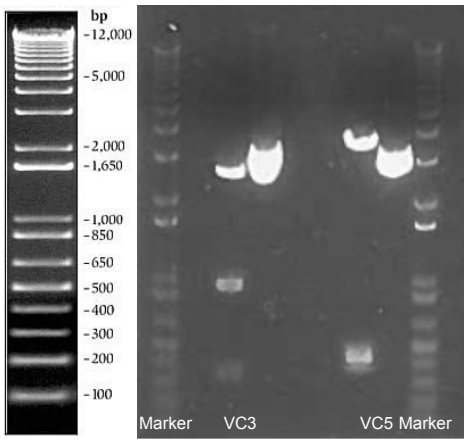


Figure 2.5: Restriction Digest of Gigaprep DNA Vaccines. Two separate gels were run for the vaccine candidates as too little DNA was generated from the gigaprep of putative VC3 and putative VC5. **(A).** Left to right reflects cut to uncut: putative VC1, putative VC2, putative VC3, putative VC4, putative VC5, and putative VC6. **(B)** left to right reflects cut to uncut, putative VC3 and putative VC5. Putative VC5 cut with different restriction enzymes in **(A)** (HindIII and XmaI) and **(B)** (HindIII) hence different bands.

2.3.4: DNA Vaccine Expression was shown in Mammalian Cells

A transfection protocol was used to confirm that all six DNA vaccines expressed protein in mammalian cells. For five of the six putative VCs there was a strong protein expression, however putative VC2 only exhibited a very faint signal (**Figure 2.6**). The faint signal in this lane could be due to low protein expression, however it is much more faint than for the other DNA vaccines. Our collaborators at PHE have previously observed efficacious vaccines that have failed this transfection assay, so it was still deemed worthy to carry this vaccine forward to animal trials. The DNA vaccines were then diluted and aliquoted into individual vaccine doses, they were to be administered at 1mg/ml (+/-5%) at an amount of 1 mg per mouse. Therefore the DNA vaccines were aliquoted into six 1.2 ml (1mg/ml) vaccine doses (**Table 2.8**).

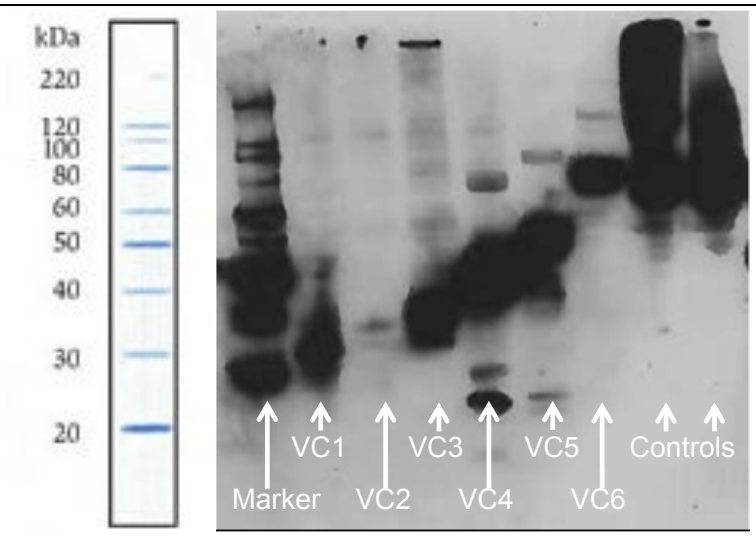


Figure 2.6: Chemiluminescence Detected on Film Following a Transfection Assay using Six DNA Vaccines. Western Blot analysis of vaccine candidates transfected into eukaryotic cells. Ladder for size comparison of protein bands. From left to right (expected size, kDa): Marker, putative VC1 (21.9), putative VC2 (27.9), putative VC3 (35.2), putative VC4 (41.2), putative VC5 (48.4), putative VC6 (64.3), and two duplicates of the positive control (60.6).

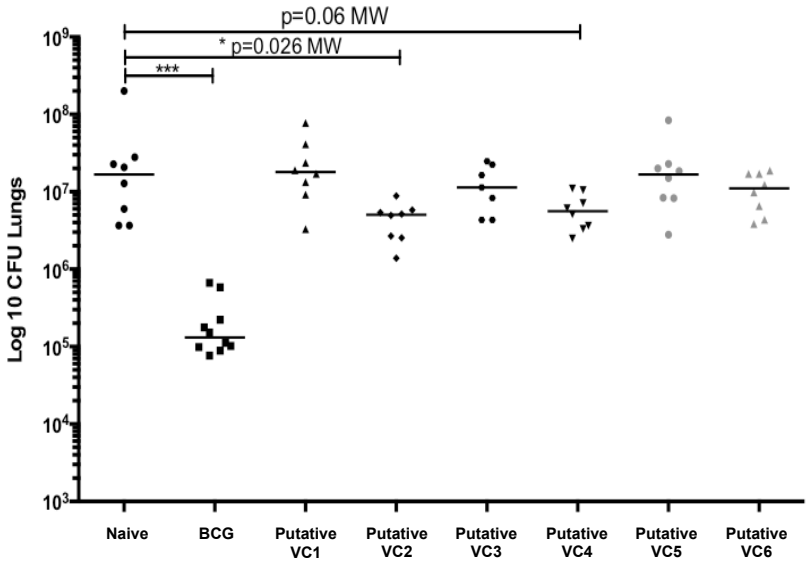
DNA Vaccine	Final DNA Vaccine Concentration (ng/µl)	A260/280	A260/230
Putative VC1	968.0	1.89	2.25
Putative VC2	989.8	1.87	2.27
Putative VC3	982.5	1.92	2.22
Putative VC4	958.0	1.88	2.26
Putative VC5	999.9	1.92	2.22
Putative VC6	988.5	1.89	2.20

Table 2.8: DNA Concentrations in Final DNA Vaccines. The concentration of DNA within the vaccines that were used in animal challenge experiments to assay protective efficacy of the putative VCs against *Mtb* infection.

2.3.5: Protective Capabilities of Vaccine Candidates in a Mouse Model of *Mtb* Infection

Testing of the protective efficacy of six DNA vaccines against TB was conducted in a murine model of *Mtb* infection[102]. The first animal challenge yielded positive results and the levels of protection conferred following *Mtb* challenge are shown in **Figure 2.7 A.** Two putative VCs stood out, VC2 and VC4, with putative VC2 achieving significant levels of protection against *Mtb* infection using a Mann Witney test (p = 0.026) and VC4 approaching significant levels of protection (p = 0.06, Mann Witney). The animal challenge assay was repeated for a second time (**Figure 2.7 B**) to confirm these results but none of the tested putative VCs conferred significant protection against *Mtb* challenge. Overall putative VC2 stimulated some level of protection, but this was not robust enough to be observed in repeat experiments.

(A)



(B)

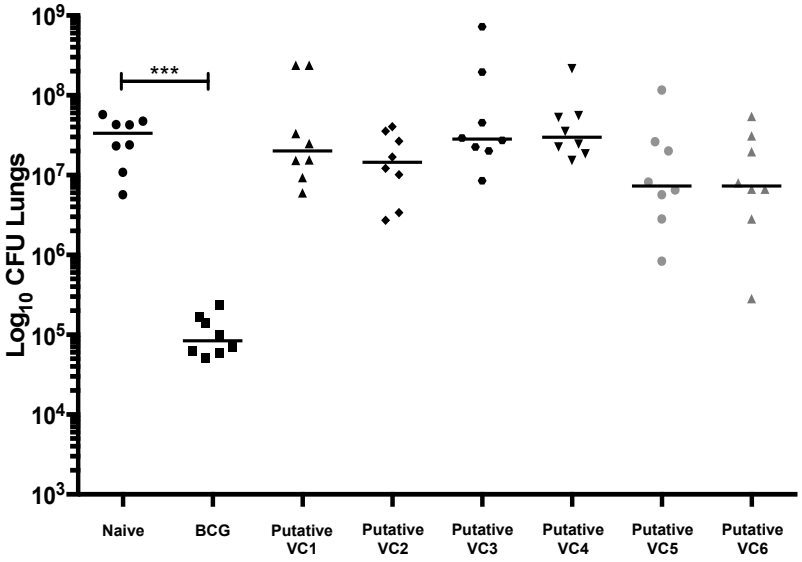


Figure 2.7: Colony Forming Units Measured in the Lungs of a Mouse Model of *Mycobacterium tuberculosis* (*Mtb*) Infection. Results for a repeated animal challenge of *Mtb* infection after DNA vaccination regimens. Each putative Vaccine Candidate was administered within a DNA vaccine three times at two week intervals. One month after the last immunisation, mice were exposed to aerosol challenged with *Mtb*. Mice were then sacrificed and the colony forming units (CFU) were measured from the lungs. MW = Mann-Whitney, ***p<0.001, *p<0.05. Abbreviations: *BCG*; Bacillus Calmette-Guérin, *VC*; Vaccine Candidate, *CFU*; Colony Forming Unit.

2.4: Discussion

In this chapter, six putative VCs were generated as DNA vaccines for protection against *Mtb* infection and evaluated in mouse challenge model experiments. In one of the mouse challenge experiments, putative VC2, Rv1677 did confer significant levels of protection against infection with *Mtb* but in a repeat challenge experiment this result was not replicated. Therefore, **hypothesis 2.1**, that the six *in silico* predicted putative VCs will confer significant protection against *Mtb* infection was not supported.

An initial animal challenge experiment revealed the ability of putative VC2 to confer significant levels of protection against *Mtb* infection and this suggested that RV could be used to find novel VCs in the proteomes of even extremely well characterised pathogens (*Mtb*). This work selected six completely novel putative VCs from the top 100 predicted BPA list for *Mtb* from Bowman et al[48]. Novel putative VC selection was achieved by the exclusion of known BPAs and predicted BPAs from known immunogenic families as it had already been shown that the RV classifier developed by Bowman et al[48] can recall known BPAs in the *Mtb* proteome. By selecting only the most novel predicted BPAs, the aim was to discover putative VCs for use in subunit vaccines against *Mtb* infection.

The obvious limitation of the approach detailed in this chapter is the lack of protection generated by the putative VCs in mouse models of *Mtb* infection. It might have been worth taking a step backwards and conducting immunogenicity testing on the six putative VCs as well as less "novel" predicted BPAs (i.e. predicted BPAs from known protective protein families in *Mtb*) by Bowman et al[48]. Immunogenicity testing would have enabled further exploration and understanding of RV's ability to predict immune stimulating antigens. Although this would have been an academic exploration as *Mtb* is considered not to have good correlates of protection[104].

The reasons behind the lack of protection generated by the DNA vaccine candidates are not understood, but some possible causes are discussed below. The most likely cause of the lack of protection described in this chapter is that the proteins selected as putative VCs were not BPAs. This would stem from the method implemented to select putative VCs. This chapter utilised an ML RV approach to predict putative VCs and this approach was published by Bowman et al[48]. Despite Bowman et al generating the largest and most advance approach to ML in RV the work conducted in this chapter led to an in depth re-evaluation of ML in RV which is described in full in **Chapter 3** of this thesis. The work conducted in **Chapter 3** of this thesis revealed previously unknown errors and also implemented improvements to the Bowman et al[48] approach that led to an enhanced ML RV approach[105]. Briefly, it was shown that Bowman et al[48] had over estimated the accuracies obtained by their approach of predicting BPAs as a nested leave tenth out cross validation (LTOCV) approach had not been implemented. This over estimation of accuracies would mean that more predicted BPAs from this pipeline would have to

be tested in the laboratory to discover a novel BPA than previously thought. Next an artificial bias was discovered in the training data of the Bowman et al[48] pipeline, which captured a signal for extracellular proteins as opposed to the desired signature of a BPA. This subcellular bias was removed in the enhanced approach detailed in **Chapter 3**. After the corrections of nesting the LTOCV and removing the artificial subcellular bias were implemented to the Bowman et al[48] ML RV pipeline improvements (i.e. an increase in the size of the training data and an increase in the annotation tools used to describe the training data) were implemented which boosted the accuracies obtained by ML RV approaches. It is hypothesized that predictions made, by the enhanced ML RV classifier described in **Chapter 3,** would result in more likely putative VC candidates than the predictions of BPAs from the Bowman et al[48] pipeline that were used in this chapter. **Chapter 4** of this thesis explored differences between predictions of BPAs from the newly described (**Chapter 3**) enhanced ML RV classifier and the Bowman et al[48] classifier via the metric Recall[48].

Another possible cause for the lack of protection conferred by the DNA vaccines are that there was no guarantee that the putative VCs were expressed during *Mtb* infection. In an attempt to factor this into the selection of the putative VCs this chapter undertook a comparison of four microarray expression studies of *Mtb* culture[97-100]. These studies would have described genes that are constitutively expressed in *Mtb* under normal growth conditions, however the DNA vaccines generated in this chapter were intended to prime an immune response to proteins that are present at an infectious stage of the *Mtb* lifecycle. Ideally the gene expression or protein measurement of *Mtb* proteins would have been measured during *Mtb* infection within a human or animal model host. Unfortunately, this data was not available at the time of this research. To select putative VCs it was decided that showing that the putative VC is expressed in the culture of *Mtb* gave an indication that the protein is physically synthesised by *Mtb.*

Despite the corrections described in **Chapter 3** that have been since implemented to the Bowman et al[48] approach to ML in RV, the significant protection obtained in the first mouse challenge experiment pointed to RV's ability to identify novel VCs against *Mtb* infection. If this challenge were to be repeated it would be suggested that a positive control (known BPA, i.e. AG85[106]) should be generated as a DNA vaccine in parallel with the putative VCs described in this chapter to enable further validation of the DNA vaccine generation and animal challenge techniques conducted for the work in this chapter. Generating a previously described BPA as a DNA vaccine at the same time as the putative VCs described in this chapter would have further confirmed that the vaccination formulation (i.e. DNA vaccine) and immunisation protocol were able to generate significant levels of protection in mouse models of *Mtb* infection. That said using the current vaccination protocol a significant level of protection was seen in the first animal challenge experiment (i.e. putative VC2 $p = 0.026$, Mann Whitney test) and therefore it would be surmised that the DNA vaccination protocol used in this pipeline was successfully implemented.

Additionally if this challenge were to be repeated it would have been suggested that at least one protein from a known immunogenic family would have been incorporated into the experimentally validated putative VC list.

Another limitation of the work carried out in this chapter was that the mouse challenge experiments were only repeated twice, and each of the repeated experiments generated different results. Ideally a third repeat of the mouse challenge experiment would have been run to obtain a reproducible result. However as animal challenge experiments are expensive it was decided to investigate the predictions of BPAs *in silico* as opposed to running further repeats of the mouse challenge experiments. When investigating the RV classifier developed by Bowman et al[48] it was deemed that improvements to this pipeline could be made. **Chapter 3** of this thesis details the improvements made to the Bowman et al classifier. Briefly, it was discovered that Bowman et al[48] had overestimated the performance of their classifier when classifying BPAs from non-BPAS. An artificial bias was also removed from the RV classifier, with regards to subcellular localisation in the positive (BPA) and negative (non-BPA) training data. Finally an increase was made to both the amount of the training data and the type of protein annotation tools used to train RV classifiers. A full investigation and revision of the ML approach to RV was conducted in the next chapter.

Despite a hint of protection generated by putative VC2 in one repeat of the mouse challenge model of *Mtb* infection, a replicate signal could not be found in repeated experiments (**Figure 2.7**). Following the rejection of **hypothesis 2.1**, it was acknowledged that refinements needed to be made to the RV classifier[48], before it could routinely be used to select novel VCs for *Mtb*. To address the need for refinements the work discussed in **Chapter 3** of this thesis implemented a number of enhancements (i.e. nested cross validation, balanced training data, increased size of training data and increased number of annotation features) to the RV classifier developed by Bowman et al[48]. It was hypothesised that the improvements to ML in RV documented in **Chapter 3** would reveal more immunologically focussed top 100 predicted BPA lists for the pathogen *Mtb*.

2.5: Statement of Contribution for Research in this Chapter.

In this chapter, I carried out the selection of the six VCs with input from my collaborators Dr. Ann Rawkins (PHE) and Prof. Helen McShane (University of Oxford) and Dr Christopher Woelk. I then went to PHE Porton Down, UK to fabricate the DNA vaccines under the supervision of Dr Ann Rawkins and Yper Hall. In addition, I received laboratory training from Sofiri Daminabo. The DNA vaccines were then sent to Prof. Helen McShane's Laboratory in Oxford and Dr. Elena Stylianou completed the assessment of the protective efficacy of the six VCs in a mouse model of *Mtb* infection.

Chapter 3: Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology

3.1: Introduction

Vaccine strategies and development continue to be at the forefront of current scientific research. An emergent new field within vaccinology is reverse vaccinology (RV). When RV was first introduced it promised many improvements over conventional vaccinology research such as increased speed, reduced cost, and the ability to screen entire proteomes to produce increased numbers of novel vaccine candidates. Whilst some of these advantages have been realised the much lauded speed increase of vaccine design has not yet materialised. However the advantages afforded by RV mean that this branch of vaccinology contains the potential to be used to generate novel subunit vaccines.

Methodologies in RV vary but the most advanced and sophisticated methods published use machine learning (ML)[48, 87]. Doytchinova and Flower[87] were the first to apply ML to the field of RV. In this study a training dataset of 100 known antigens was generated through a literature curation that defined a known antigen as a protein (or part of a protein) that, "has been shown to induce a protective response in an appropriate animal model after immunisation". A negative training dataset was constructed by randomly sampling 100 proteins or non-antigens from the same bacterial species that corresponded to each known antigen in the positive training dataset. The proteins in this training dataset were annotated with auto cross-covariance (ACC) transformations, which reflect hydrophobicity, molecular size, and polarity. The annotated proteins were used to train a classifier based on discriminant analysis by partial least squares (DA-PLS), which was able to achieve an accuracy of 82% when distinguishing non-antigens from known antigens.

In an extension to Doytchinova and Flower's[87] work, Bowman et al[48] focused exclusively on bacterial protective antigens (BPAs) defined as, "a whole protein that led to significant protection ($p < 0.05$) in an animal model (i.e. bacterial load reduction or survival assay) following immunisation and subsequent challenge with the bacterial pathogen". Bowman et al[48] expanded upon the previous ML RV approach[87] by focusing on antigens from bacterial species, utilising protein annotation tools for greater numbers of features relevant to biological annotation and implemented a support vector machine (SVM). The positive BPAs were obtained by a literature

curation for proteins that fitted the description of a BPA. This was combined with a randomly generated negative training dataset (non-BPAs) to create the BPAD136 training dataset. To extract information from these proteins, annotation tools that described biological features were used to annotate the training datasets. Next a SVM was applied for the classification of the two groups (BPA and non-BPA). This SVM was able to achieve a maximum accuracy of 92% (for more details see **Chapter 1, Section 1.7**).

Chapter 2 of this thesis tested six putative BPAs predicted by the Bowman et al[48] classifier in a murine model of *M. tuberculosis* infection. However these six predicted BPAs were shown not to confer protection. The aim of this chapter (i.e. **Chapter 3**) was to improve upon previous ML approaches to RV, to construct a more accurate classifier, from which biological inferences about differences between BPAs and non-BPAs could be made. It was observed that the previous publication (Bowman et al[48]) had not correctly implemented a nested cross-validation and therefore had overestimated the accuracies obtained by their application of ML in RV. When training an SVM classifier two main stages occur, feature selection and parameter optimisation (described in **Chapter 1, Section 1.6**). In the Bowman et al[48] study the whole dataset (136 BPAs and non-BPAs) was used when conducting feature selection and parameter optimisation. The whole dataset was then split into leave-tenth-out cross-validation (LTOCV) training and test datasets. The Bowman et al[48] RV classifier had been exposed to the test dataset at the feature selection and parameter optimisation stages; therefore the classifier was over estimating the accuracies achieved when classifying BPAs and non-BPAS. This was not a nested approach to cross-validation, which is currently state of the art for cross-validation in the ML field.

This chapter sought to establish that ML in RV is able to predict protection (i.e. BPAs) and to improve upon the classifier developed by the previous ML RV approach[48]. This chapter first proved the ability of a signal for protection to be captured by comparing classification on a dataset of curated BPAs and non-BPAs to a dataset that represented noise (permutation testing). The following improvements were implemented compared to the previous RV classifier[48]: a nested approach to cross-validation, removal of an artificial bias, increased size of the training data by 64 BPAs and 64 non-BPAs and the incorporation of new protein annotation tools that are able to model different aspects of biology (e.g. T-cell epitope prediction and Adhesin prediction).

Hypothesis 3.1: BPAs curated from the literature, contain a signal for protective antigens compared to randomly permuted data.

Hypothesis 3.2: Using a correctly nested leave-tenth-out cross-validation would reduce the accuracies achieved when classifying BPAs and non-BPAs.

Hypothesis 3.3: Increasing the size of the training data and the number and breadth of protein annotation tools would increase the accuracies obtained when classifying BPAs and non-BPAs.

3.2: Methods

3.2.1: Overview of Changes from Bowman et al[48] to Heinson et al[105]

The alterations that form part of the finalsed Heinson et al[105] RV ML approach were a small part of many research pathways. Before presenting the finalised research pipeline implemented by Heinson et al[105], it was deemed worthwhile to provide an overview of the main research directions that were taken when investigating RV ML approaches and these can be seen in **Figure 3.1.** Briefly the previous ML RV approach, published by Bowman et[48] al, in which an ML classifier was built to distinguish BPAs and non-BPAs was explored and expanded. Firstly the annotations derived from the bioinformatics protein annotation tools were parsed in a standardised manner across all tools. Secondly the amount of training data (BPAs, non-BPAs) was increased to 200 BPAs and 200 non-BPAs (detailed in **3.2.2**). Thirdly the number of protein annotation tools utilised to generate information about the BPAs and non-BPAS was F to 34 and these were parsed to generate 525 annotation features (detailed in **3.2.3**). Next, spurious differences in the randomly generated negative training data (non-BPAs) were explored by generating five non-BPA negative training datasets. It was shown that the variation caused by different random negative datasets was smaller than the variations across repeated iterations of feature selection on the same training data. When comparing the negative training data (non-BPAs) to the positive training data (BPAs) an artificial subcellular localization bias was apparent. This subcellular localization bias had been introduced by the random method of selecting negative training data (non-BPAs). To enable ML RV approaches to capture a signature for immunological protection this subcellular localization bias was removed by matching BPAs (positive training data) and non-BPAs (negative training data) for subcellular localization and thus a new negative training dataset was formed (detailed in **3.2.2**). Following the removal of the subcellular localisation bias from the training data, it was observed that alterations to ML RV classifiers utilised different feature numbers when obtaining maximum accuracies when classifying BPAs and non-BPAs. To enable a direct comparison of the alterations made to the ML RV approach the number of features utilized by the ML RV classifiers was fixed. Conducting a permutation analysis revealed that the number of features which best separated the protective signal from random noise was 10. Therefore 10 feature classifiers were utilised to asses changes to ML RV classifiers (detailed in **3.2.5**). Finally a comparison of un-optimised different ML algorithms was conducted (detailed in **3.2.4**) which

revealed that a Support Vector Machine (SVM) with a radial bias function achieved equally the highest accuracies when distinguishing between BPAs and non-BPAs. The accrual of the above alterations to the Bowman et al[48] approach to ML in RV is described in detail below and published in the Heinson et al[105] ML RV manuscript.

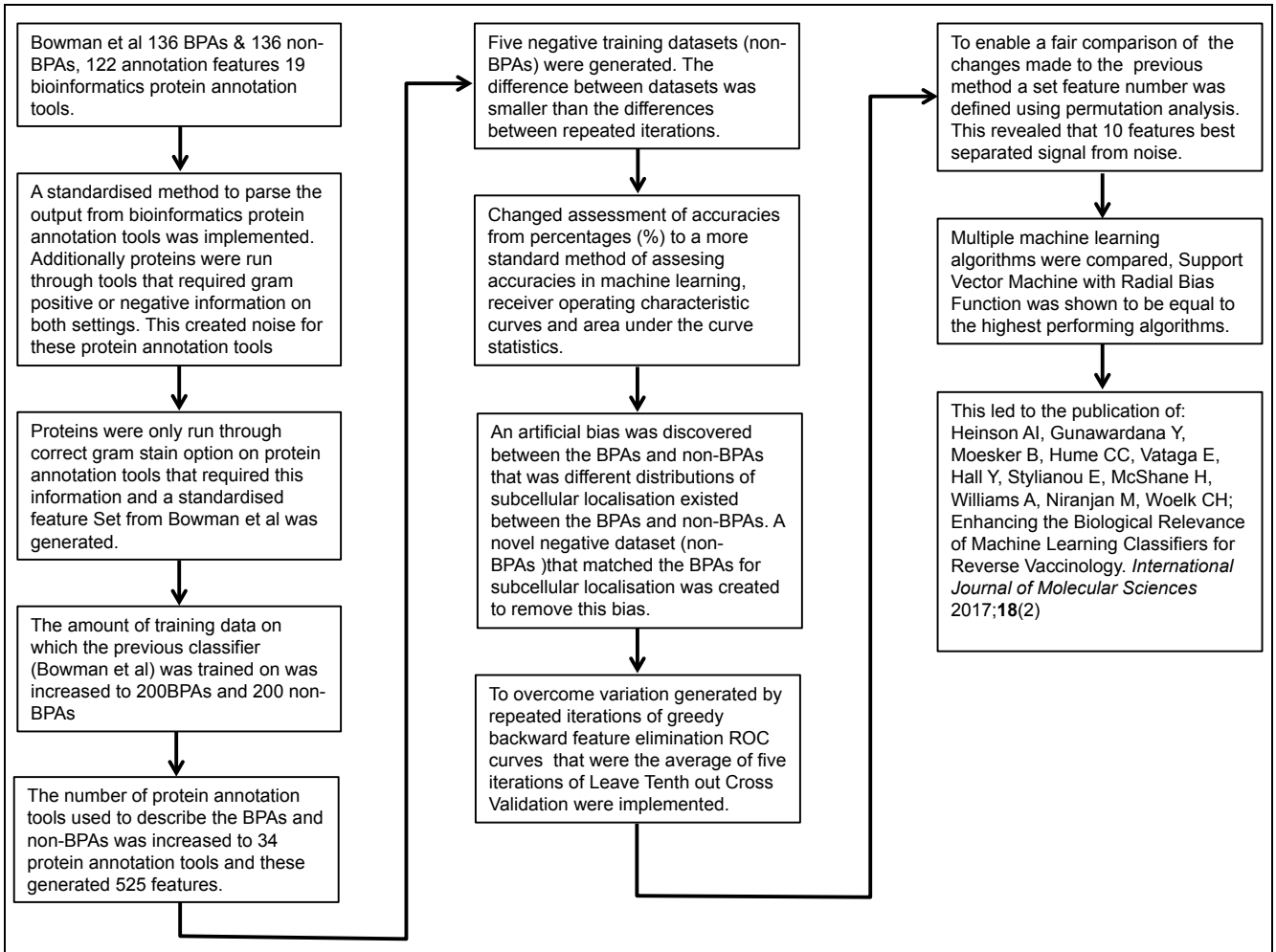


Figure 3.1: Overview of the Research Directions Taken Whilst Generating a New Machine Learning Approach to Reverse Vaccinology: Showing an overview of the research process when conducting revisions to the Bowman et al[48] Machine Learning (ML) Reverse Vaccinology (RV) pipeline in the process of generating the Heinson et al[105] ML RV approach.

3.2.2: Training Data

A literature curation identified 64 new BPAs which were combined with 136 previously characterised BPAs[48] for a positive training dataset totaling 200 BPAs. A BPA is a bacterial protein that has led to significant protection ($p < 0.05$) in an animal model (i.e., bacterial load reduction or survival assay) following immunisation and subsequent challenge with the bacterial pathogen. A bacterial protein was only retained as a BPA if two independent researchers agreed that it had met this definition. This extensive literature curation was undertaken by utilising Google Scholar. Search terms were: "Potential Vaccine Candidate", "Protective antigen", "Bacterial Vaccine Candidate", "Bacterial Vaccine Antigen", "Bacterial Protein Antigen", "Novel vaccine Candidate", "Bacterial Vaccine Candidate", and "Vaccine Candidate Mouse". The curation effort searched through the first ten result pages for each one of these search terms. When a potential BPA had been found the information was curated to fully characterise the proteins in catagories such as; "NCBI Protein ID", "Species", "Refference". For a full list of the curation effort please see a table of 136 BPAs curated by Bowman et al[48] and a table describing the 64 newly curated BPAs in **Appendix H**.

A negative dataset of non-BPAs was generated to match each BPA in the positive training dataset by randomly selecting a protein from the same bacterial species as each BPA. This study also matched the subcellular localisation, such that a non-BPA was selected from the same subcellular localisation as the corresponding BPA. The bioinformatics program PSORTb was used to predict subcellular localization as this was the most accurate subcellular localization prediction tool available, achieving stated accuracies of 97.9% for gram positive and 98.3% for gram negative bacterial organisms[57]. A blastp[50] was used to discard any non-BPAs that matched to known BPAs (i.e. > 98% similarity) or non-BPAs already selected (E-value < $10E^{-3}$). A BPA (ACF35754.1) had a subcellular localisation of extracellular. When selecting non-antigens from the proteome of *Salmonella enterica* a paired exracellular protein that matched the blastp inclusion criteria could not be found and therefore a protein with unknown subcellular localisation was sampled instead for inclusion in the negative training dataset. Datasets that have balanced positive (i.e. BPA) and negative (i.e. non-BPA) training datsets for subcellular localisation were given the "+B" tag. In summary, a dataset consisting of 200 BPAs and 200 non-BPAs was constructed and referred to throughout this thesis as BPAD200. BPAD200, as used to train ML classifiers can be accessed at http://www.mdpi.com/1422-0067/18/2/312#supplementary by clicking on the download link, Supplementary File 1 (The "Zip-Document" link) and then from this

folder, supplementary_Table_2. Additionaly a FASTA list of all proteins in the BPAD200 dataset can be found in the appendices of this thesis (**Appendix I**).

3.2.3: Increased Number of Protein Annotation Tools

A second literature curation identified new protein annotation tools to generate novel annotation features for training SVM classifiers. This study increased the number of protein annotation tools used to train previous RV classifiers[48] from 19 to 34, and the output from these tools was parsed to generate 525 annotation features. In the previous approach[48] there was a lack of consistency applied when parsing the information from protein annotation tools. Certain tools were normalised for protein length whilst others were not. This study applied a standardised parsing method to all tools and this can be seen in **Appendix J**. The criteria for inclusion of the additional protein annotation tools utilised in this chapter was that they had to be able to be downloaded, installed and run locally as well as having a high throughput method that could run through entire bacterial proteomes using the tools standard settings. Additional bioinformatics protein annotation tools were included to generate novel biological annotation when compared to the previous approach[48]. Some examples of additional bioinformatics protein annotation tools included in the novel approach detailed in this chapter describe biological phenomena such as: T-Cell epitope prediction, Dbox and Ken box prediction, Small ubiquitin like modifiers, Surface accessibility of amino acids and Adhesin prediction. An example of how a tools output was parsed is given using the bioinformatics protein annotation tool GPS-SNO[107], the features generated from this tool were: GPS-SNO_max_Score, GPS-SNO_Count, GPS-SNO_Max_CutOffDiff, GPS-SNO_Avg_Score, GPS-SNO_Avg_CutOffDiff, GPS-SNO_Length_Count, GPS-SNO_Length_Average, and GPS-SNO_Length_Avg_CutOffDiff. Classifiers trained on datasets including additional annotation features not previously utilised in ML approaches to RV have the "+AF" tag, a full list of protein annotation tools and annotations features can be found in **Appendix J.**

3.2.4: Machine Learning Classification

To enable comparison to previous RV classifiers (i.e. Bowman et al[48]) SVMs with a radial bias function (RBF) kernel were used to construct all classifiers in this chapter and implemented using the python package *libsvm*[85]. It should also be noted that when conducting a preliminary pass of the classification of BPAs from non-BPAs a range of non-optimised ML algorithms were implemented in the software package Matlab[82] such as; Nearest Neighbor, Decision Tree, Discriminant Analysis, SVM with

RBF Kernel, SVM with Quadratic Kernel and a Linear SVM. Of the non-optimised classifiers SVM with a RBF kernel yielded the joint highest accuracies. An optimised implementation of the SVM with a RBF kernel required the optimisation of two parameters, *γ* and *C*. Where, *C* is the cost function and *γ* is particular to an RBF kernel such that its value affects how much each training example can influence the decision margin. For more details see **Chapter 1 (Section 1.6.1)**. Optimisation for both of these parameters was conducted within the *libsvm* package using a grid search method. The datasets were comma separated variable formatted tables and were comprised of rows of proteins (BPA or non-BPA) and columns that represent annotation features. These annotation features were on differing scales and distributions and were scaled as individual features between -1 and 1 as is standard before training classifiers[85]. BPAD200, as used to train the ML classifier can be accessed at http://www.mdpi.com/1422-0067/18/2/312#supplementary by clicking on the download link Supplementary File 1 *(*The *"*Zip-Document*"* link*)* and then from this folder supplementary_Table_2. All implementations of the SVM in this chapter used a non-specific filtering step based on *F*-score (**Figure 3.2**)[88, 89] to reduce the number of annotation features used to train SVM classifiers, from 525 to 200.

$$F(i) \equiv \frac{\left(\bar{\boldsymbol{x}}_i^{(+)} - \bar{\boldsymbol{x}}_i\right)^2 + \left(\bar{\boldsymbol{x}}_i^{(-)} - \bar{\boldsymbol{x}}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(\boldsymbol{x}_{k,i}^{(+)} - \bar{\boldsymbol{x}}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(\boldsymbol{x}_{k,i}^{(-)} - \bar{\boldsymbol{x}}_i^{(-)}\right)^2}$$

Figure 3.2: F Score Used for Feature Selection: F Score was utilised to lower the number of features from 525 to 200 before submission to greedy backward feature elimination and support vector machine training and classification. The F score compares the variability between the two classes (Bacterial Protective Antigen (BPA) and Non-BPA) (Numerator) against the variability within the two classes (denominator)[88].

The remaining 200 features then underwent greedy backward feature elimination (as described in **Chapter 1, Section 1.6.2**) to determine the most informative feature set. When features had equal information content the algorithm randomly selected

which feature to remove, SVM classifiers were trained across five separate iterations to overcome the impact of randomly breaking such ties.

For all instances, except when implementing the previous RV classifier[48] ("BPAD136", detailed in **Section 3.1**), classifiers were evaluated using a nested LTOCV model[108]. Classifiers trained using a nested LTOCV approach were given the "+N" tag. This enabled an SVM score to be obtained for each protein within the training data of BPAs and non-BPAs. The first step in a fully nested approach was to split the data into 10 parts and isolate one of these tenths as the test dataset. Feature selection and parameter optimisation were then only applied to the remaining training dataset (i.e. the remaining 9/10ths of the data). An SVM classifier was then built on only the training dataset and used to predict the class (BPA or non-BPA) of the test data in the tenth left out. This process was repeated a further nine times leaving the remaining tenths of the training data out, one at a time.

3.2.5: Permutation Analysis

Permutation analysis was used to determine the optimal feature number for SVM classifiers. Labels (BPA and non-BPA) were randomly permuted five times, generating five datasets. The AUCs (Area Under the Curve) achieved when training SVM classifiers using greedy backward feature elimination on these datasets were averaged and compared to the average AUC obtained over five iterations of greedy backward feature elimination of BPAD200+N+B+AF (BPAD200, nested cross validation "+N", balanced for subcellular localisation "+B" and using additional annotation features "AF"). This process was repeated with SVM classifiers trained using different numbers of features.

3.2.6: Statistics

Receiver operating characteristic (ROC) curves were used to evaluate the performance of SVM classifiers in this study. ROC curves were generated by plotting the true positive rate (TPR, i.e. sensitivity) over the false positive rate (FPR, i.e. 1−specificity)[109]. Area under the curve (AUC) values were calculated from ROC curves and differences between curves was assessed with the DeLong[110] statistical test. *P*-values of <0.05 were considered significant. ROC curves for LTOCV were generated as an average SVM score across the five iterations for each protein in the training dataset.

3.2.7: Hierarchical Clustering

The subcellular localisations of BPAs from BPAD200 were predicted using the protein annotation tool PSORTb[57]. A BPA was labeled as extracellular if it was predicted to be localised close to the surface of the cell (i.e. cell wall, extracellular, outer membrane, or periplasmic). BPAs with a predicted subcellular localisation of periplasmic were included in the extracellular group as these proteins clustered predominantly with the extracellular as opposed to intracellular group. If a BPA was predicted to be localised to the cytoplasm or cytoplasmic membrane it was defined as intracellular. Intracellular and extracellular BPAs from BPAD200 were clustered using a Euclidean distance calculation and a Ward clustering metric using the *ClassDiscovery*[111], and *Dendextend*[112] packages in R[84].

3.3: Results

3.3.1: Permutation Analysis Revealed a Strong Protective Signal for BPAs Curated from the Literature

Initially, the BPAD200 dataset was used to determine that a signal for protective efficacy had been captured when curating BPAs from the literature record. The difference in AUC values generated from the classifier trained on BPAD200 (i.e. BPAD200+N+B+AF classifier) and the randomly permuted data was evaluated when training classifiers built on different numbers of features (**Figure 3.3A**). Classifiers consisting of 10 features demonstrate the greatest and most significant separation in AUC (average $p = 1.13 \times 10^{-12}$, DeLong test[110]) between BPAD200 and the randomly permuted data (**Figure 3.3B**). These results clearly demonstrate that the literature curation of BPAs captured a strong protective signal and that 10 features was the optimal number for discriminating BPAs from non-BPAs as opposed to noise. Therefore, SVM classifiers containing 10 features were used to evaluate the changes in classification accuracies of each of the modifications to the RV approach (i.e. BPAD200, +N, +B, and +AF).

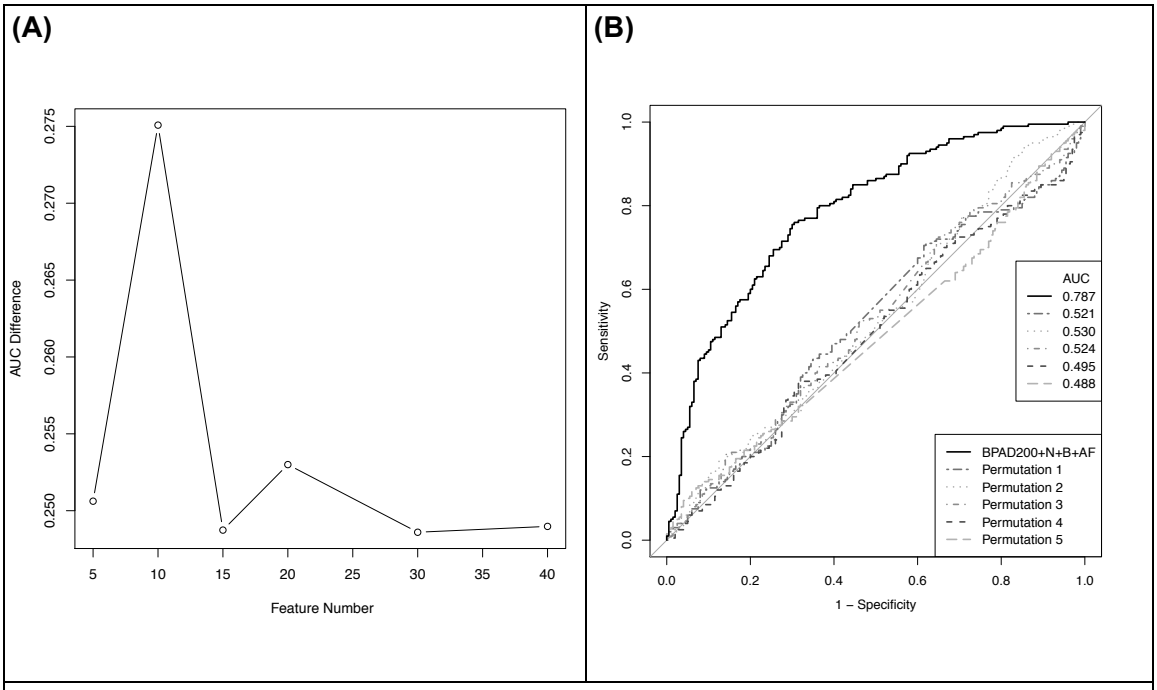


Figure 3.3 Comparison of the Difference in Area Under the Curve Between SVM Classifiers Trained on the Dataset BPAD200 and Datasets of Noise Generated using Randomly Permuted Data Labels: (**A**) Plot of the difference in area under the curve (AUC) between five iterations of SVM trained on the BPAD200 dataset (i.e. BPAD200+N+B+AF classifier) and five datasets with randomly permuted labels trained on increasing feature numbers. Support vector machine (SVM) classifiers were trained to discriminate bacterial protective antigens (BPAs) and non-BPAs. Receiver operator characteristic (ROC) curves generated from a nested leave-tenth-out cross-validation approach for different numbers of features selected by greedy backward feature elimination. Five iterations were performed to assess the random breakage of ties during greedy backward feature elimination and AUC was averaged across iterations for each feature set. This analysis was then repeated with five datasets where the BPA and non-BPA labels were randomly permuted. The average AUC was calculated across the five randomly permuted data sets for each number of features. (**B**) ROC curves for the average of the five iterations of the 10 feature SVM classifier trained on BPAD200 (i.e. BPAD200+N+B+AF) (black solid line) and from each of the five randomly-permutated datasets (dotted grey lines).

3.3.2: A Nested Approach had a Significant Impact on the Ability of SVMs to Classify BPAs

Having demonstrated that the BPAD200+N+B+AF classifier trained on datasets curated from the literature had captured a biological signal reflective of BPAs, the multiple modifications made to the RV classifier of Bowman et al[48] were assessed in a stepwise manner. The starting point for this assessment was the BPAD136 classifier from the previous study[48], consisting of 136 BPAs and 136 non-BPAs. Bowman et al[48] had not implemented a truly nested cross-validated approach. Implementing a nested approach caused a significant reduction in AUC (p = 9.69 × 10^{-5}, DeLong test[110]) when migrating from an overfit (BPAD136, AUC = 0.962) to a nested (BPAD136+N, AUC = 0.861) approach (**Figure 3.4**). Therefore, the implementation of a nested approach to cross-validation is recommended for RV studies as this enabled a better estimation of the performance of classifiers when distinguishing between BPAs and non-BPAs.

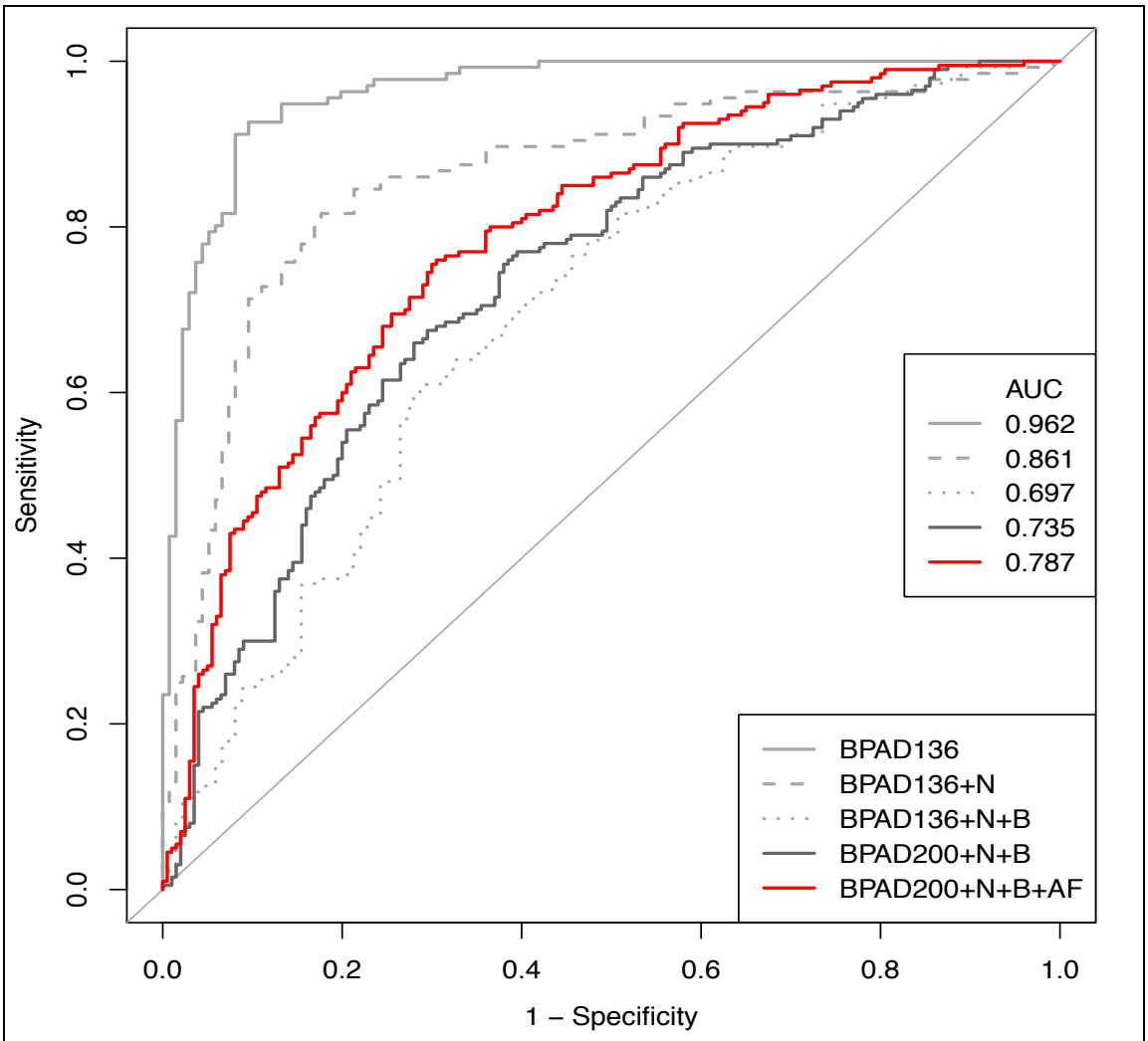


Figure 3.4 Receiver Operating Characteristic (ROC) Curves Assessing the Performances Obtained through Different Classifier Modifications: ROC curves were generated from support vector machine (SVM) classifiers utilising 10 features selected by greedy backward feature elimination in a leave-tenth-out cross-validation approach. Averages were plotted across five iterations of classifiers implemented to randomly break ties resulting from the greedy backward feature elimination procedure. The benchmark to assess these modifications was a non-nested, non-balanced training dataset of 136 BPAs and 136 non-BPAs annotated with 122 features from 19 protein annotation tools (BPAD136)[48]. Subsequent modifications were added in a stepwise fashion and included: a nested cross-validation approach (BPAD136+N), balanced selection of non-BPAs for predicted subcellular localisation (BPAD136+N+B), increased size of training data (BPAD200+N+B), and additional features (525 total) derived from an increased number of protein annotation tools (BPAD200+N+B+AF)

3.3.3: Correcting a Bias in the Selection of Negative Training Data Lowered the Accuracies of SVM Classifiers when Predicting BPAs

Comparing the subcellular localisations (as predicted by PSORTb[57]) of the BPAs and non-BPAs in BPAD136 revealed that the negative training data (non-BPAs) had a larger proportion of proteins predicted to be located in the cytoplasm (**Figure 3.5 A,B**). This resulted from the random selection of non-BPAs for the negative training data from a bacterial genomic context where the most common subcellular localisation is cytoplasmic. To correct this bias, a new negative training dataset was generated, where each non-BPA was selected to match not only the bacterial species of the corresponding BPA, but also its subcellular localisation (**Figure 3.5C**). Removing this bias in subcellular localisation decreased the ability of the SVM classifier to discriminate between BPAs and non-BPAs as reflected in a significant reduction in AUC (p = of 3.44×10^{-5}, DeLong test[110]) when comparing BPAD136+N to BPAD136+N+B (**Figure 3.4**). The performance of the SVM classifier was reduced because it could no longer utilise differences in subcellular localisation to discriminate BPAs from non-BPAs. This was reflected by the removal of features related to bacterial subcellular localisation (e.g. PSORTb and SignalP) from the top 10 features utilised by the classifier (data not shown).

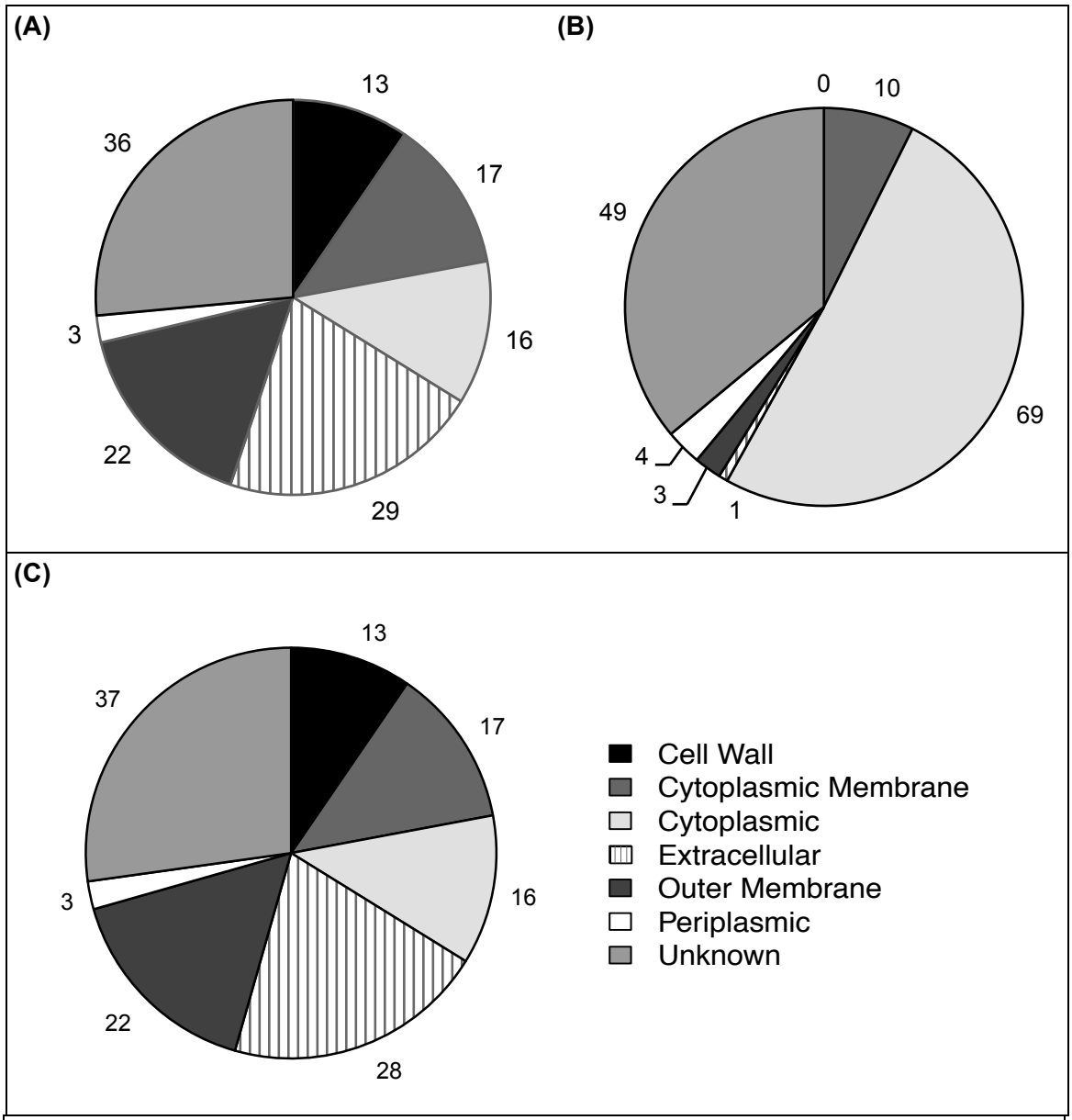


Figure 3.5 Pie Charts Comparing the Predicted Subcellular Localisation of the Positive and Negative Training Data for BPAD136: Pie charts showing subcellular localisation as predicted by PSORTb[57] for the numbers of BPAs and non-BPAs in the following subsets of the BPAD136 dataset: (**A**) Positive training data (i.e. 136 BPAs), (**B**) Negative training data (i.e. 136 non-BPAs), and (**C**) negative training data balanced for subcellular localisation (i.e. 136 non-BPAs).

3.3.4: Increasing the Size of the Training Data had a Positive Impact on the Ability of SVMs to Classify BPAs

A literature curation obtained 64 new BPAs. These BPAs expanded the positive training data set from 136 to 200 BPAs. The 64 additional BPAs were paired with 64 non-BPAs balanced for subcellular localisation to form the training data set of the classifier BPAD200+N+B. Increasing the size of the training data in this manner led to improvements in AUC (**Figure 3.4**, 0.697 to 0.735), although this change was not statistically significant.

3.3.5: Increasing the Number of Protein Annotation Tools Enhanced the Ability of SVMs to Classify BPAs

The effect of including new annotation features derived from protein annotation tools (e.g., Spaan[67], MHCpan[113], GPS-MBA[114]) on the ability of SVM classifiers to distinguish BPAs from non-BPAs was examined. An additional 15 protein annotation tools were included in comparison to the previous RV classifier (Bowman et al[48]). The total number of 34 protein annotation tools was used to generate the training dataset for the BPAD200+N+B+AF classifier, which was comprised of a total of 525 features. The incorporation of these additional features resulted in an SVM classifier (BPAD200+N+B+AF) with an increased AUC (0.787) when compared to BPAD200+N+B (**Figure 3.4**). Although this increase did not obtain significance, it should be noted that when the incorporation of additional BPAs and protein annotation tools were considered together (i.e., BPAD136+N+B to BPAD200+N+B+AF) there is a significant increase in AUC ($p = 2.11 \times 10^{-2}$, DeLong test[110]). Finally, the top 10 features selected by greedy backward feature elimination for the discrimination of BPAs and non-BPAs in the training data of the BPAD200+N+B+AF classifier, contained novel features derived from the additional protein annotation tools. These novel features were related to T-cell epitope prediction that had not previously been included in ML RV approaches[48, 87] (**Table 3.1**)

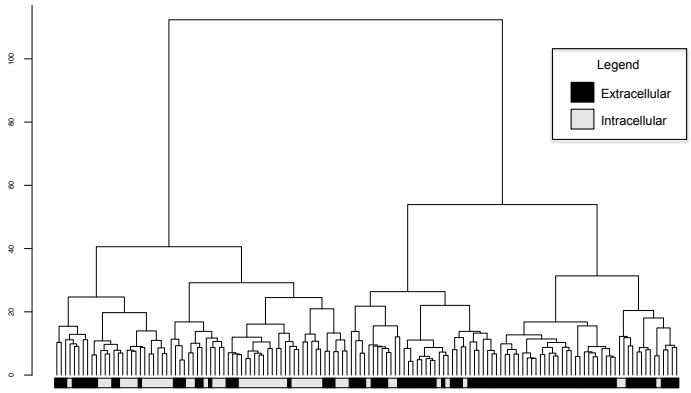
Rank	Feature	Name of Bioinformatics Tool	Protein Annotation Tool Type	Correlated with BPA or non-BPA
1	LipoP_Signal_Avg_Length	LipoP	Lipoprotein	BPA
2	YinOYang-T-Count	YinOYang	Glycosylation	BPA
3	NetPhosK-S-Count	NetPhosK	Phosphorylation	BPA
4	LipoP_SPI_Avr_Length	LipoP	Lipoprotein	BPA
5	**NetMhcPan-B-AvgRank**	**NetMhc**	T-Cell Epitope predictor (MhcClass II)	BPA
6	TargetP-SecretFlag	TargetP	Subcellular Compartmentalisation – in Eukaryotic Cells	BPA
7	YinOYang-Average-Difference1_Length	YinOYang	Glycosylation	Non-BPA
8	**MBAAgl7_CorCount**	**GPS-MBA**	T-Cell Epitope predictor	BPA
9	**PickPocket-Average_score**	**PickPocket**	MhcPeptide Binding	Non-BPA
10	PropFurin-Count_Score	ProP	Cleavage Sites in Eukaryotic Cells	BPA

Table 3.1 The Top 10 Annotation Features Selected by Greedy Backward Feature Elimination for Discrimination of BPAs from non-BPAs: Features selected by the SVM classifier trained on the training data for the BPAD200+N+B+AF classifier. Features in bold represent those derived from protein annotation tools that were added in this study compared to our previous approach[48]. For a full list of bioinformatics tools utilised in this study and the annotation features derived from them please see **Appendix J**.

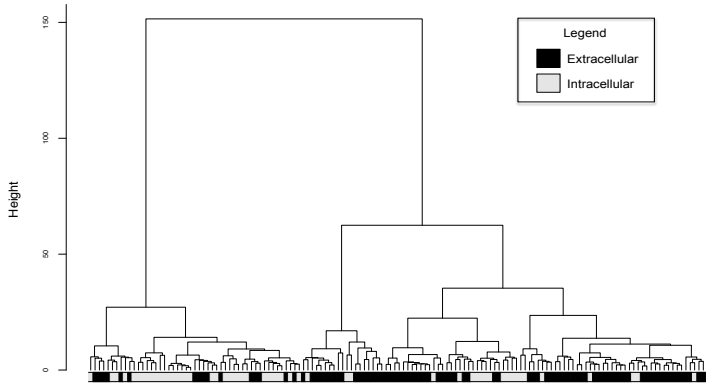
3.3.6: Intracellular and Extracellular BPAs Utilised Different Features for Classification

An unsupervised approach was used to further explore the biological signals in the features derived from protein annotation tools used to annotate the 200 BPAs curated from the literature record. Utilising a hierarchical clustering approach of 142 BPAs predicted to have subcellular localisation by PSORTb[57] other than "unknown" only revealed clear differences between intracellular and extracellular BPAs when clustering the antigens from BPAD200 using all 525 features (**Figure 3.6 A**). Clustering of the BPAs ustilising 200 F score selected (**Figure 3.6 B**) or 10 greedy backward feature selected (**Figure 3.6 C**) features did not reveal a difference between intracellular and extracellular BPAs. It was shown that as the features that best characterised the differences between BPAs and non-BPAs (defined on a combined subcellular dataset BPAD200) become more focused on the differences between BPAs and non-BPAs (**Figure 3.6 A to B to C**) the signal between intracellular and extracellular BPAs was masked. This was expected as both the F score and greedy backward feature elimination procedure select features that best distinguish between BPAs of all subcellular localisations and non-BPAs of all subcellular localisations. Therefore the techniques used to select the features utilised in (**Figure 3.6 B and C**) both select features that generalise across all subcellular localisations but are different between BPAs and non-BPAs. The features selected by the feature selection are masking the true signal that a difference exists between intracellular and extracellular BPAs. Hierarchical clustering using all 525 annotation features, revealed two groups that corresponded to predicted subcellular localisation, i.e. intracellular or extracellular (**Figure 3.6 A**). This suggested that intracellular and extracellular BPAs may have fundamentally different biological properties that are being described through the data curation and annotation approach detailed in this chapter. Therefore, it was hypothesized that greater accuracies would be achieved when classifiers were trained separately on intracellular and extracellular BPAs. To test this hypothesis, two training datasets were constructed from the training data of BPAD200+N+B+AF: intracellular BPAD51 (iBPAD51) consisting of 51 BPAs and 51 non-BPAs with subcellular localisation predicted as intracellular, and extracellular BPAD91 (eBPAD91) with 91 BPAs and 90 non-BPAs with subcellular localisation predicted as extracellular. A matching non-BPA balanced for subcellular localisation could not be found that met the inclusion criteria for ACF35754.1, detailed in methods (**Section 3.2.2**), thus, the negative training data was reduced by one.

(A)



(B)



(C)

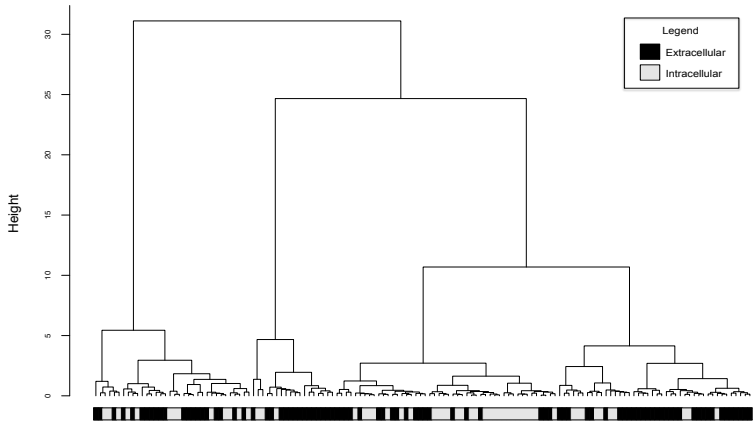


Figure 3.6 Hierarchical Clustering of BPAs from the Training Data of the BPAD200+N+B+AF Classifier: Hierarchical clustering of 142 BPAs from BPAD200 using all **(A)** 525 annotation features, **(B)** 200 most deterministic features based on F score of BPAs and non-BPAs in BPAD200+N+B+AF and **(C)** 10 most deterministic features based on BPAs and non-BPAs based on greedy backward feature elimination (**Table 3.1**). Distances between BPAs were calculated using Euclidean metrics and then clustered using the Ward algorithm. Grey labels at the branch tips refer to BPAs with subcellular localisation predicted by PSORTb [57] as intracellular (i.e., cytoplasm or cytoplasmic membrane) and black labels as extracellular BPAs (i.e., extracellular, periplasmic, outer membrane, cell wall).

Classifiers trained on iBPAD51 and eBPAD91 had lower AUC values when compared to the BPAD200+N+B+AF classifier (**Figure 3.7A**). The classifiers had been trained on datasets of different sizes and this might explain the superior performance of BPAD200+N+B+AF, which was the largest dataset. Therefore, classifiers were trained on randomly selected subsets of the BPAD200+N+B+AF training data of decreasing size (**Figure 3.7B**). This facilitated a better comparison of intracellular (trained on iBPAD51) or extracellular (trained on eBPAD91) classifiers, versus those trained using BPAs and non-BPAs from all subcellular localisations (BPAD200+N+B+AF). The accuracies derived from iBPAD51 and eBPAD91 were similar to data sets of similar size consisting of BPAs from all subcellular localisations (**Figure 3.7B**). This suggests there is no immediate benefit to training separate classifiers to recognise intracellular or extracellular BPAs.

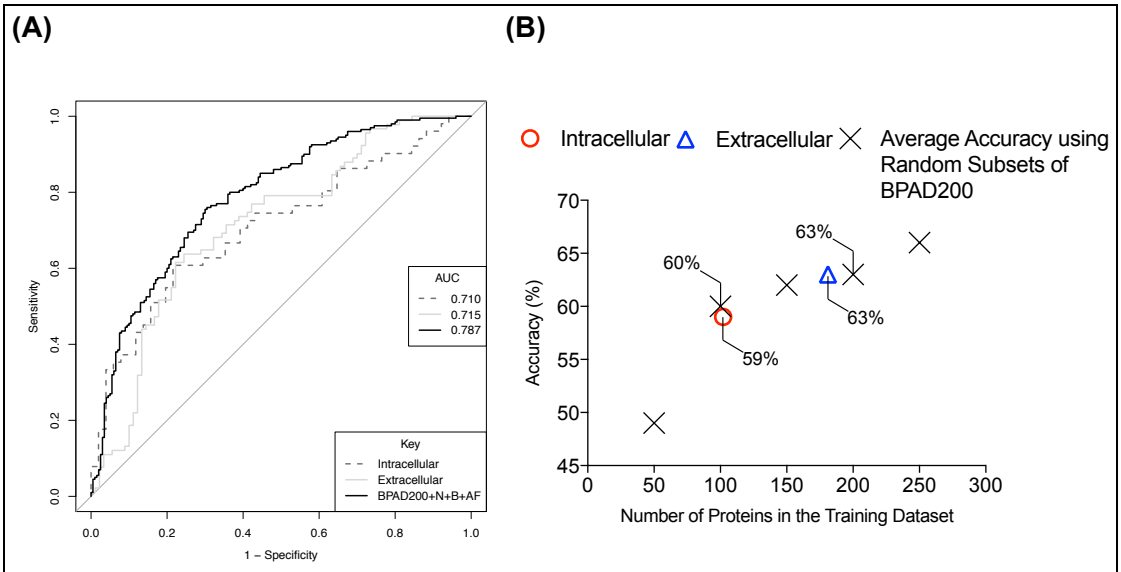


Figure 3.7 Plots Comparing the Performance of Intracellular, Extracellular and Combined Subcellular Localisation Classifiers: (**A**) ROC curves obtained from SVM classifiers trained to distinguish BPAs from non-BPAs in the following data sets: iBPAD51 (dotted line), eBPAD91 (solid grey line) and BPAD200+N+B+AF (black line). Curves were drawn by averaging results from five iterations of SVM classifiers consisting of 10 features selected by greedy backward feature elimination assessed in a LTOCV approach. (**B**) Plot showing the average percentage accuracy (five iterations) of SVM classifiers of 10 features trained on different sized subsets of BPAD200+N+B+AF for comparison to SVM classifiers derived from iBPAD51 and eBPAD91.

To fully explore differences between the classifiers trained on intracellular (iBPAD51) and extracellular (eBPAD91) BPAs, the top 10 features selected by greedy backward feature elimination for each dataset were interrogated (**Tables 3.2 and 3.3**). The classifier trained on eBPAD91 (extracellular classifier) utilised features derived from types of protein annotation tools that were not utilised by the classifier trained on iBPAD51 (intracellular classifier). The unique features utilised by the classifier eBPAD91 were related to the following diverse array of biological phenomena: an adhesin predictor SPAAN[67] (which describes if the protein adheres to the surface of cells), surface accessibility, and secondary structure predictor NetSurfP[115] (which predicts the relative likelihood of sections of each protein being exposed on the surface of a protein structure), lipoprotein prediction LipoP[116] (which predicts if a protein will

interact with lipids), as well as a cleavage site predictor NetChop[117] (which predicts if a protein will be cleaved by the human proteasome) (**Table 3.2**). Features derived from protein annotation tools that were unique to the classifier trained on iBPAD51 were largely related to immunogenicity and included: a B-cell epitope predictor[118] (predicts possible binding sites for B-cells), a T-cell epitope predictor[114] (predicts possible binding sites for T-cells) and a calpain cleavage predictor[119] (predicts a specific type of protein cleavage dependent on the presence of Ca^{2+}) (**Table 3.3**). In summary, SVM classifiers appear to utilise features from different protein annotation tools to discriminate intracellular and extracellular BPAs from their respective non-BPAs.

Rank	Feature	Protein Annotation Tool Type	Rank in Intracellular Classifier
1	Pad-value	**Adhesin**	42
2	DictOGlyc_Ser_Average_Threshold_Length	Glycosylation	189
3	LipoP_SPI_AvgScore	**Lipoprotein**	NF
4	Netsurfp_RSA_Exposed_AverageDiff	**Surface accesibility and secondary structure**	NF
5	PoloPhosphorylation_CorAvg	Phosphorylation	NF
6	Net_Chop_CorCount	**Predicts Cleavage sites**	NF
7	DictOGlyc-No_Score_Sites_Length	Glycosylation	NF
8	GPS_SUMO_Suolyation_Average_Score	Small Ubiquitin like modifiers (SUMOs) binding site prediction	9
9	ProtParam-PercIsoleucine	General Annotation	144
10	ProtParam-PercGlutamicAcid	General Annotation	NF

Table 3.2 The top 10 Annotation Features Selected by Greedy Backward Feature Elimination for Discrimination of Extracellular BPAs and Non-BPAs: The top 10 annotation features selected by greedy backward feature elimination utilised by SVM classifier trained on eBPAD91. Protein annotation tools listed in bold represent types of protein annotation not present in the intracellular classifier's top 10 annotation features, as selected by greedy backward feature elimination. NF: not found in the top 200 features of the intracellular classifier type that were submitted to the greedy backward feature elimination algorithm following non-specific F-score filtering.

Rank	Feature	Protein Annotation Tool Type	Rank in Extracellular Classifier
1	Bepipred-Count_Length	**B-Cell Epitope**	149
2	CCD_av_diff	**Calpain Cleavage**	NF
3	YinOYang-TAverage-Difference1_Length	Glycosylation	NF
4	ProtParam-GRAVY	General Annotation	35
5	NetOGlyc-T-Max-I	Glycosylation	196
6	YinOYang-T-Average_Length	Glycosylation	NF
7	ProtParam-PercAlanine	General Annotation	97
8	NetPhosL-Y-MaxScore	Phosphorylation	NF
9	GPS_SUMO_Sumoylaion_Average_Score	Small Ubiquitin like modifiers (SUMOs) binding site predictor	8
10	MBAAgl7_CorAvg	**T-Cell Epitope predictor**	NF

Table 3.3 The top 10 Annotation Features Selected by Greedy Backward Feature Elimination for Discrimination of Intracellular BPAs and Non-BPAs: The top 10 annotation features selected by greedy backward feature elimination utilised by SVM classifier trained on iBPAD51. Protein annotation tools listed in bold represent types of protein annotation not present in the extracellular classifier's top 10 annotation features, as selected by greedy backward feature elimination. NF: not found in the top 200 features of the extracellular classifier type that were submitted to the greedy backward feature elimination algorithm following non-specific F-score filtering.

3.4: Discussion

The work detailed in this chapter was consistent with **hypothesis 3.1**, that BPAs curated from the literature, annotated with protein annotation tools contain a signal for protective antigens compared to randomly permuted data (**Figure 3.3**). This chapter then made alterations to a previous ML in RV approach (Bowman et al[48], **Figure 3.8**) and proved **hypothesis 3.2** to be correct, i.e. that implementing a correctly nested LTOCV did lower the accuracies achieved when classifying BPAs and non-BPAs. Furthermore an artificial bias within the training data was removed, which further lowered the accuracies obtained when classifying BPAs and non-BPAs. However, removing the subcellular localisation bias enabled a protective signature to be captured (**Table 3.1**), which was confirmed through permutation analysis (**Figure 3.3**). Finally this chapter proved **hypothesis 3.3** to be correct, that increasing the size of the training data and the breadth of annotation tools used to describe the training data significantly improved the classification accuracies when distinguishing BPAs from non-BPAs. In summary this chapter has addressed errors and biases in the most recent ML RV approach[48] and implemented improvements to this corrected approach resulting in a new RV classifier, i.e. BPAD200+N+B+AF (**Figure 3.8**).

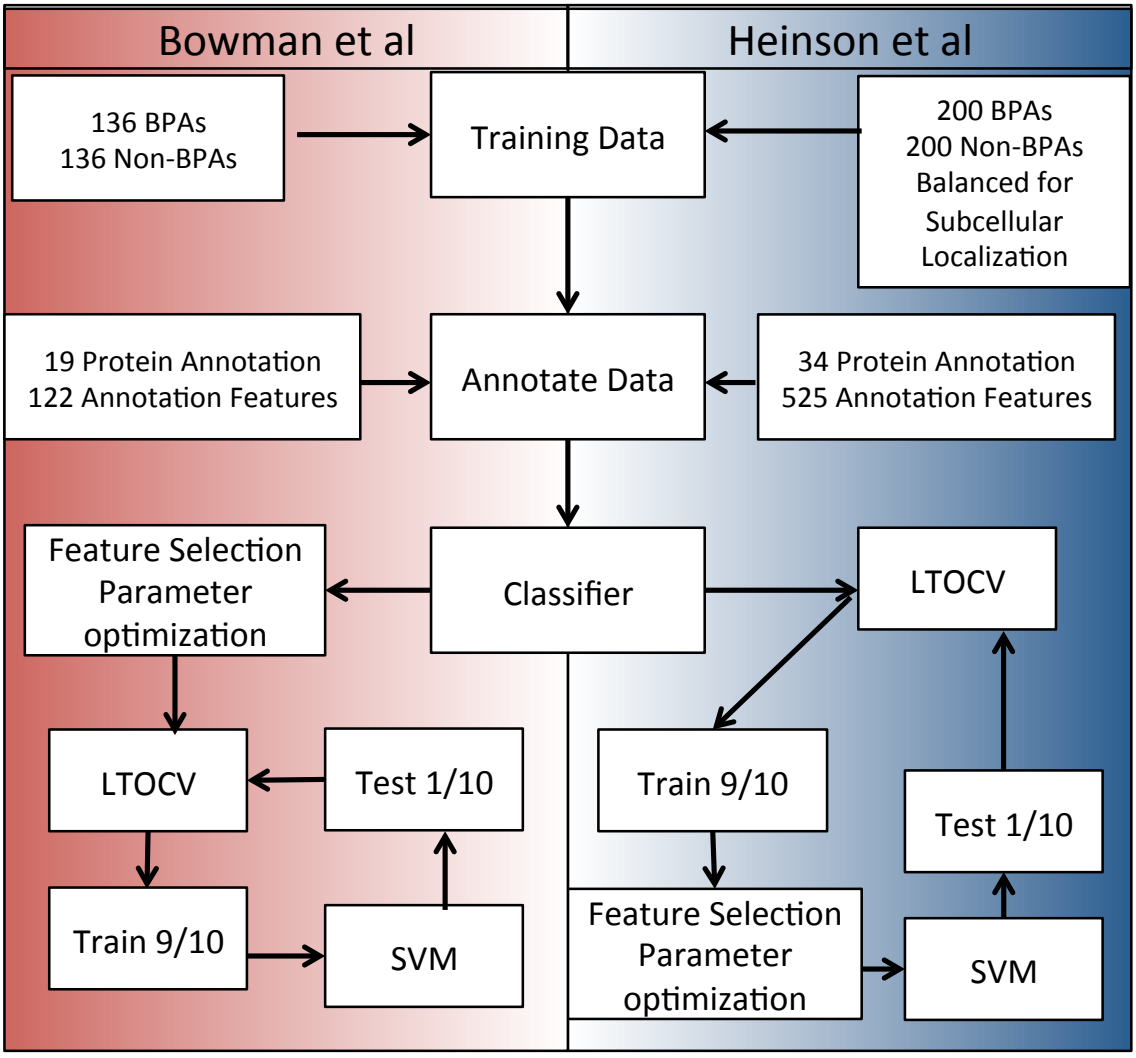


Figure 3.8: Schematic Depicting Improvements made to the Previous Reverse Vaccinology (RV) Classifier: Side by side comparison of the Bowman et al[48] RV classifier and enhancements that were made to this approach[105]. Abbreviations: *BPA*; Bacterial Protective Antigen defined as a bacterial protein that when injected into an animal model gave significant protection ($p<0.05$) following immunisation and subsequent challenge with the bacterial pathogen (i.e. bacterial load reduction or survival assay), *LTOCV*; Leave-tenth-out cross-validation, *SVM*; Support vector machine.

The proposed advantage gained by utilising ML in RV, specifically, that an organism's entire proteome could be interrogated when predicting BPAs, was achieved through this work. Filtering approaches to RV had heavily relied upon removing any proteins that are unlikely to be expressed on the surface of a cell or externally. The previous ML RV approach (Bowman et al[48]) appeared to corroborate this view in that the most informative feature to discriminate between BPAs and non-BPAs was whether

a protein is located in the cytoplasm ("PSORTb-ProbCytoplasm (-)"[48]). If the protein was cytoplasmic that corresponded heavily with the protein being predicted as a non-BPA. A classifier that deemed the most deterministic feature between the negative training data (non-BPAs) and positive training data (BPAs) as whether the protein is predicted to be localised in the cytoplasm would miss already validated BPAs that had a predicted subcellular localisation as cytoplasmic (**Figure 3.5**).

Through data visualisation it was shown that the negative training data (non-BPAs) had a much larger proportion of cytoplasmic proteins than the positive training data (BPAs) (**Figure 3.5 A and B**). This was due to the fact that the negative training data (non-BPAs) was selected randomly from the proteome of a bacterial species. Bacteria have more proteins predicted to have a cytoplasmic subcellular localisation than other subcellular compartments. By selecting a protein at random from a bacterial proteome there was a higher chance that the protein would have a cytoplasmic subcellular localisation than other localisations. The distribution of bacterial protein subcellular localisation generated an artificial subcellular localisation bias between the positive (BPA) and negative (non-BPA) training data for the previous RV classifier[48]. In this approach the subcellular localisation bias was removed and despite significantly lowering the accuracies obtained when classifying BPAs and non-BPAs (**Figure 3.4**), it was a crucial step because this enabled the classifier to more cleanly characterise a protective signal (**Table 3.1**).

The features used by the BPAD200+N+B+AF classifier could be examined to determine which features are reflective of a BPA (**Table 3.1**). The main signal related to protective efficacy in the 10 most deterministic features between BPAs and non-BPAs was the prediction of T-cell epitopes (bioinformatics tools NetMHC[113] and MBAAgl7[114]), a greater number of epitopes correlated with protection. In addition, annotation related to general biological processes, such as threonine glycosylation, phosphorylation, and lipoproteins, were positively linked with being a BPA. These processes have been implicated in controlling both the humoral and cellular immune responses[120-124]. For example, lipoproteins have been shown to increase the influence of major histocompatibility complex-II (MHC-II) activation on T-helper cells (Th cells)[121]. This is achieved by lipid rich microdomains co-localising and increasing the MHC-II molecules concentration on the cell surface, resulting in more efficient Th cell activation while requiring less antigen[121]. Lipoproteins have also been found to be a common component of bacterial membranes and often perform adhesin functions to host cells and thus are easily exposed to the immune system[125]. This is supported by the use of lipoproteins in licensed vaccines and this has been attributed to lipoproteins ability to stimulate both innate and adaptive immune responses[125]. Additionally it has also been

shown that the glycosylation of proteins can be represented on both the Mhc I and Mhc II complexes and thus cause the stimulation of an adaptive immune response[122]. Specifically with a focus towards developing subunit vaccines changes to glycosylation patterns have also been shown to impact the immunogenicity of subunit vaccines[126]. In summary, the classifier BPAD200+N+B+AF is using features that make biological sense with regards to predicting the protective qualities of BPAs. Furthermore the top ten most deterministic features between BPAs and non-BPAs (**Table 3.1**) demonstrated the value of epitope prediction tools but that other tools predicting general protein annotation features should also be considered and continue to be incorporated in future enhancements.

Clustering BPAs in the training data of BPAD200+N+B+AF based on all annotation features derived from protein annotation tools revealed that BPAs grouped based on extracellular and intracellular predicted subcellular localisation (**Figure 3.6 A**). Unexpectedly, separate intracellular (trained on iBPAD51), extracellular (trained on eBPAD91), and combined (BPAD200+N+B+AF) classifiers achieved similar accuracies (**Figure 3.7 B**). However, it was theorised that intracellular and extracellular classifiers would be selecting different features in order to capture biological differences whilst making predictions of BPA or non-BPA. A logical hypothesis was that extracellular classifiers utilise features related to B-cell epitopes since this antigen type was surface exposed and that intracellular classifiers require features related to T-cell epitopes since this antigen type was internalised. However, this was not the case, the intracellular classifier utilised features from both B-cell and T-cell epitope predictors but the extracellular classifier did not utilise features from any epitope prediction tools (**Tables 3.2 and 3.3**). This could have been due to the difficulty in predicting conformational epitopes from amino acid sequences[127]. It is estimated that 90% of B-cell epitope binding is conformational[20]. To model this information CBTOPE[128] (a conformational B-cell epitope predictor) was included, however annotation features derived from CBTOPE were not present in the top ten annotation features used in classification for any of the classifiers constructed in this chapter. It is possible that with more advanced techniques these conformational epitopes will be more accurately predicted from amino acid sequences and may become an important part of the classification for extracellular BPAs. Instead the extracellular classifier used annotation features derived from more general protein annotation tools (e.g., adhesin prediction, surface accessibility, and general cleavage site prediction). Although the utility of separate intracellular and extracellular classifiers has not been demonstrated in this chapter it is clear that these classifier types were modeling different aspects of

protective biology and future studies should further explore this as more data becomes available.

There are a number of limitations that may be affecting the ability of SVM classifiers to discriminate BPAs from non-BPAs in this chapter. Firstly, it is possible that the random selection of non-BPAs for the negative training data may have resulted in the mistaken addition of BPAs that simply have not yet been tested and documented in the literature record. To help to negate this limitation, non-BPAs with homology to BPAs were discarded. Furthermore, permutation analysis demonstrated that the non-BPAs represent a useful negative training dataset since there was clearly a discriminatory signal between BPAs and non-BPAs compared to random noise (**Figure 3.3B**).

Another limitation was that most antigens previously tested and confirmed in the literature are predicted to be of unknown (29%) or extracellular (45.5%, eBPAD of BPAD200) subcellular localisation. This led to a small training dataset (iBPAD51) when building the intracellular classifier (51 BPAs and 51 non-BPAs). Incorporating more proteins in the training datasets may have enabled a better description of differences in protection derived through intracellular or extracellular proteins.

It is envisaged that the application of ML approaches to RV will build upon the success of filtering approaches that led to the BEXSERO® vaccine. The SVM classifier constructed in this study (BPA200+N+B+AF) discriminates BPAs from non-BPAs and through comparisons to randomly permuted data clearly demonstrated that a signal for protective efficacy has been curated from the literature record. Future studies should concentrate on the expansion of the training data through the addition of more BPAs and the incorporation of new protein annotation tools since these enhancements were shown collectively to significantly increase the accuracy of the classifier (p-value = 2.11×10^{-2}, Delong test). This will increase the accuracy with which BPAs are predicted and reduce the number of laboratory assays that need to be performed in order to identify novel vaccine candidates. The BPAD200+N+B+AF classifier developed in this chapter has enhanced ML approaches to RV and could now be used to predict BPAs for all pathogens with a sequenced genome, which could ultimately lead to the development of novel subunit vaccines. **Chapter 4** aimed to test the enhanced classifier's (BPAD200+N+B+AF) ability to predicted novel BPAs using a metric termed recall that determines the biological relevance of RV classifiers.

3.5: Statement of Contribution for Research in this Chapter

I personally carried out all of the work detailed in this chapter. Dr. Christopher Woelk, Dr. Carmen Denman-Hume and Dr. Bastiaan Moesker acted as the additional curators when confirming BPAs for inclusion in this study.

Chapter 4: Evaluating the Ability of Enhanced Classifiers for Recalling Known Protective Proteins from Bacterial Proteomes

4.1: Introduction

Following improvements to machine learning (ML) classifiers utilised in reverse vaccinology (RV)[48, 87] (**Chapter 3**), resulting in the development of an enhanced classifier (BPAD200+N+B+AF), this chapter evaluated the biological relevance of the newly developed classifier. **Chapter 3** of this thesis has enabled more realistic predictions of the accuracies obtained by ML RV approaches when separating bacterial protective antigens (BPAs) from non-BPAs. A BPA was defined as a bacterial protein that when used to immunise an animal model confers significant protection ($p < 0.05$) following subsequent challenge with the bacterial pathogen (i.e. bacterial load reduction or survival assay)[48]. This approach improved upon an older RV classifier (Bowman et al[48]) by implementing a nested cross-validation approach, correcting an artificial bias for subcellular localisation, increasing the size of the training data and the breadth of features on which the classifier was trained. Having shown that increasing the number of features and the breadth of annotation features used by RV classifiers significantly increased the area under the curve when classifying BPAs and non-BPAS (BPAD136+N+B to BPAD200+N+B+AF, $p = 2.11 \times 10^{-2}$, Delong test, **Chapter 3**), the BPA predictions of the enhanced classifier (BPAD200+N+B+AF) was assessed in a biologically relevant metric termed recall[48].

Recall, as defined by Bowman et al[48], was used to evaluate the biological relevance of a classifier. This was achieved by measuring the ability of the classifier to recall known BPAs when in a background of the entire bacterial proteome. Recall was assessed by using the classifier to rank all proteins in a bacterial proteome for the predicted probability of being a BPA. The number of known BPAs above a particular cut-off (e.g. top 100 predicted BPAs) for each bacterium was then evaluated using a hypergeometric test (i.e. Fishers exact test). This recall statistic established a practical confidence in the real world use of ML classifiers for RV. If a classifier was able to significantly recall known BPAs, then by inference, other highly ranked proteins may be considered as BPAs and thus potential vaccine candidates.

In this chapter, the recall technique[48] was applied to the bacterial pathogens *Neisseria meningitidis* serogroup B MC58 (*MenB*) and *Mycobacterium tuberculosis* H37RV (*Mtb*). *MenB* was selected because a previous RV approach (Pizza et al[52])

used the genome sequence of this pathogen to identify antigens for the BEXSERO vaccine. Proteins incorporated in the BEXSERO vaccine represented antigens that can be used as a test dataset to evaluate classifier accuracy. The immunology of *MenB* has also been extensively studied and *MenB* has robust immune correlates of protection[129]. Immune correlates of protection refer to a laboratory assay that is more cheaply and rapidly implemented compared to protection assays in animal models, but a positive result will likely predict protection in an animal model. For *MenB* a positive result in a serum bactericidal activity (SBA) assay strongly correlates with protection[129, 130]. The SBA assay was shown to be a robust enough correlate of protection that vaccines for *N. meningitidis* serogroups A,C,W,Y and B have been licensed on data accumulated from this assay as opposed to traditional efficacy trials[131]. This chapter utilised *MenB* proteins that had been shown to have positive results in SBA assays to generate larger test samples on which the recall metric could be evaluated.

Evaluating the recall metric in *MenB* enabled comparisons to RV's previous success story and also to test the classifier's (BPAD200+N+B+AF) ability to generate novel vaccine candidates the recall metric was applied to *Mtb*. This important pathogen is currently seeing a revival in vaccinology research due to the emergence of multiple and extensively drug resistant strains[132]. Recall in *Mtb* was evaluated using BPAs curated from a literature search that were present in the training data of the BPAD200+N+B+AF classifier. Furthermore, as a negative control, the recall metric was also used to evaluate the six predicted BPAs that failed to confer protection in a murine model of *Mtb* challenge in **Chapter 2**. In summary, this chapter assessed the biological relevance of the BPAD200+N+B+AF classifier derived in **Chapter 3** of this thesis and this was achieved using the recall metric[48] by testing the following hypotheses:

Hypothesis 4.1: The BPAD200+N+B+AF classifier would be able to significantly enrich for known antigens in top 100 predicted BPA lists for bacterial pathogens.

Hypothesis 4.2: The six proteins assayed for protective efficacy in **Chapter 2** would not be significantly enriched in the top 100 predicted BPAs for *Mtb* using the BPAD200+N+B+AF classifier.

4.2: Methods

4.2.1: Recall

A recall metric, as described in Bowman et al[48] was implemented. Recall was evaluated by using a SVM classifier to output the predicted probability of each protein in a pathogenic proteome being a BPA. The proteins in the pathogenic proteome were then ranked by their predicted probability of being a BPA. The ability of a classifier to recall known antigens within the top 100 predicted BPAs was then assessed using a hypergeometric test (**Section 4.2.8**).

4.2.2: Proteomes

Proteomes for *Mtb* (ID: NC_000962) and *MenB* (ID: NC_003112) were downloaded from the NCBI genome database. The ID (NC_000962 or NC_003112) was entered at "https://www.ncbi.nlm.nih.gov/" and then the genome entry was selected. From this page the proteome was downloaded using the link "Download sequences in FASTA format for genome, proteome". Files were downloaded as text files containing FASTA sequences, *Mtb* N = 3906 proteins and *MenB* N = 1943 proteins.

4.2.3: Proteome Annotation

FASTA sequences of each protein in the *Mtb* and *MenB* proteomes were annotated with protein annotation tools as described in **Chapter 3**. The proteome of *MenB* was annotated using the top 10 classification features as determined for the BPAD200+N+B+AF classifier in **Chapter 3, Table 3.1**. *Mtb* was used to explore the influence of feature numbers on recall and utilised different feature combinations depending on the iteration. In total *Mtb* was annotated with 24 protein annotation tools and the output parsed to provide several annotation features per tool. For a full list of annotation tools and the parsed features for the BPAD200+N+B+AF classifier please see **Appendix J.**

4.2.4: Classification

The BPAD200+N+B+AF classifier, developed in **Chapter 3,** was used to make a classification prediction for each protein in the *Mtb* and *MenB* proteomes. This classification resulted in a SVM-score, which was used to rank the proteins in each proteome for their probability of being a BPA. The recall carried out on the *MenB* genome used the top ten classification features determined for the BPAD200+N+B+AF classifier as described in **Chapter 3, Table 3.1**. To compare recall statistics using

different numbers of features in *Mtb* a feature selection step was implemented. This was conducted utilising a non-specific filter based on F score to reduce the feature numbers to 200. Then greedy backward feature elimination (detailed in **Chapter 1, Section 1.6.1**) was used on the entire training dataset of BPAD200+N+B+AF, before training of the classifier. Greedy backward feature elimination had a random step that was related to removing features that were deemed to have the same information content. When eliminating features one at a time sometimes multiple features were deemed to be the least informative to the classification of BPAs and non-BPAs, when this was the case one of these features was eliminated at random. To determine the effect of breaking these random ties, several iterations of the BPAD200+N+B+AF classifier was evaluated using the recall metric. The *Mtb* proteome was used to explore the effect that different numbers of features had on the recall metric due to the large number of known BPAs for *Mtb* (N = 16) within the training dataset of the BPAD200+N+B+AF classifier.

4.2.5: Known Protective Proteins used for Recall

The five antigens incorporated in the BEXSERO vaccine (Gene ID: NMB2132, NMB1030, NMB2091, NMB1870 and NMB1994[21]) were used to assess recall in *MenB*. Antigens for *Mtb* recall were found through a literature curation and met the definition of a BPA (i.e. proteins that led to significant protection (p-value < 0.05) when used to immunise an animal model and subsequently challenged with *Mtb*). A full explanation of the literature curation effort is detailed in **Chapter 3, Section 3.2.1**. These 16 known BPAs for *Mtb* were previously included in the training dataset of the BPAD200+N+B+AF classifier and were used to assess recall in this chapter. Finally, as a negative control, the ability of the BPAD200+N+B+AF classifier to recall the six predicted BPAs (Rv3886c, Rv2190c, Rv2068, Rv1857, Rv1677 and Rv0608c) that failed to elicit a protective immune response in **Chapter 2** was also evaluated.

4.2.6: Serum Bactericidal Activity Positive Proteins used for Recall

The SBA assay is a measure of complement-mediated killing via antibody[130]. For immune responses to bacterial pathogens like *MenB* that rely on antibody killing for long-term protection this has been shown to be a very robust correlate of protection[130]. This chapter utilised a literature curation of SBA assays in the bacterial species of *N. meningitidis*. This curation was conducted by collaborator, Prof. Myron Christodoulides for a review article (in preparation[133]). The literature curated SBA assays were carried out under slightly differing experimental conditions but an overview of the SBA assays

follows. SBA assays were conducted using serum collected from a pre-exposed (vaccine) patient or animal. The innate complement activity of the serum sample was removed; commonly this was achieved by heat inactivation. The removed complement was then replaced by a naïve complement, commonly from rabbit or human. The survival of *MenB* was then assessed after incubation with different dilutions of the processed serum by plating the SBA assay reaction mix and counting the colony forming units (CFUs) of *N. meningitidis.* Typically the bactericidal titre of the antigen was measured. The bactericidal titre of a SBA assay was measured as the dilution of the test serum that resulted in at least a 50% decrease in CFUs per ml of *Neisseria meningitidis*[134]. This chapter utilised proteins from a literature curation undertaken by Prof. Christodoulides to form a test dataset (i.e. SBA positive proteins). A protein was incorporated into the SBA positive protein dataset if they were deemed to generate significant bactericidal activity against *Neisseria meningitidis* in a SBA assay, and had an orthologue in the *MenB* proteome (**Section 4.2.7**). A protein was deemed to generate significant bactericidal activity in a SBA assay if the CFUs of bacteria were significantly reduced (p = < 0.05) when compared to unvaccinated controls. SBA positive proteins were used to test the ability of the BPAD200+N+B+AF classifier to recall antigenic proteins when in the background of the entire *MenB* proteome. To avoid re-testing antigens from **Section 4.2.5,** the five antigens included in the BEXSERO vaccine were not incorporated into the SBA positive protein dataset. The SBA positive protein dataset was comprised of 18 proteins and a list of the SBA positive proteins can be seen in **Appendix K**. A subset of these 18 proteins (N=3) were also known BPAs included in the training data of the BPAD200+N+B+AF classifier, the 15 proteins within the SBA positive proteins dataset, not included in the training data of BPAD200+N+B+AF, were also used as an individual test set to evaluate the recall metric.

4.2.7: Orthologue Identification

Known BPAs and SBA positive proteins were sometimes described in different strains of *M. tuberculosis* and *N. meningitidis* to those used for the recall metric in this chapter (*Mtb: H37RV* and *MenB: MC58*). An example of this being the known BPA for *Mycobacterium tuberculosis*: CAA88138.1, which was described as a BPA in the Erdman strain[106]. This known BPA was mapped onto the *Mtb* genome (i.e. *Mycobacterium tuberculosis H37RV*) using blastp[50] and the orthologue YP_177768.1 with an E value of < 0.001 was used in the recall metric for *Mtb* in the place of CAA88138.1. As described above, literature curated proteins from the same bacterial species, but different pathogenic strains, were incorporated when conducting recall.

Orthologues were found to match these proteins, in the bacterial strain used for recall (i.e. *Mtb: H37RV* or *MenB: MC58*), and were identified using blastp. If an orthologue was discovered in the recall strain of the bacterial pathogen with a blastp E value < 0.001 it was included in the recall statistic test datasets (i.e. known BPA or SBA positive proteins).

4.2.8: Recall Statistics

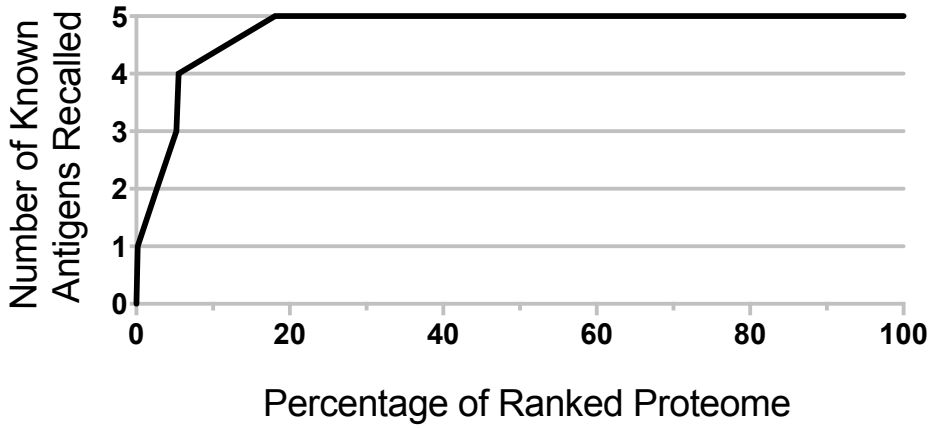
A Hypergeometric test was implemented in the statistical package R[135] using *dhyper* which tested the enrichment of known BPAs or SBA positive proteins in the top 100 predicted BPA recall lists. The input parameters to *dhyper* were *x*, *m*, *n* and *k*. Where *x* = number of known BPAs recalled in the top 100, *m* = the number of known BPAs in the entire proteome, *n* = the number of non-BPAs in the entire proteome, and *k* = subsets of the proteome under assessment (i.e. top 100). *P*-values < 0.05 were considered statistically significant.

4.3: Results

4.3.1: The Enhanced Classifier (BPAD200+N+B+AF) was able to Significantly Recall Proteins in the BEXSERO Vaccine

The RV classifier, BPAD200+N+B+AF, was able to generate significantly enriched recall lists for the antigens in the subunit vaccine BEXSERO from a background of the entire proteome of *MenB.* Out of the five antigens incorporated in the BEXSERO vaccine, the BPAD200+N+B+AF classifier predicted four as BPAs. The fifth protein (NP_274866.1, NMB1870) was not predicted as a BPA and was recalled at a rank equivalent to 18.12% of the *MenB* proteome. Overall the classifier BPAD200+N+B+AF predicted 18.10% of the *MenB* proteome as BPAs. It should be noted that 18.10% of an organism's genome being a biologically functioning BPA is unlikely but the selection of highly predicted BPAs is predicted to be useful for discovering novel BPAs. A hypergeometric test revealed that two out of five antigens from the BEXSERO vaccine were significantly recalled in the top 100 predicted BPAs (p = 0.022) (**Figure 4.1**). Therefore, this enhanced RV classifier (BPAD200+N+B+AF) was able to significantly recall antigens that had previously been incorporated into the BEXSERO vaccine.

A.



B.

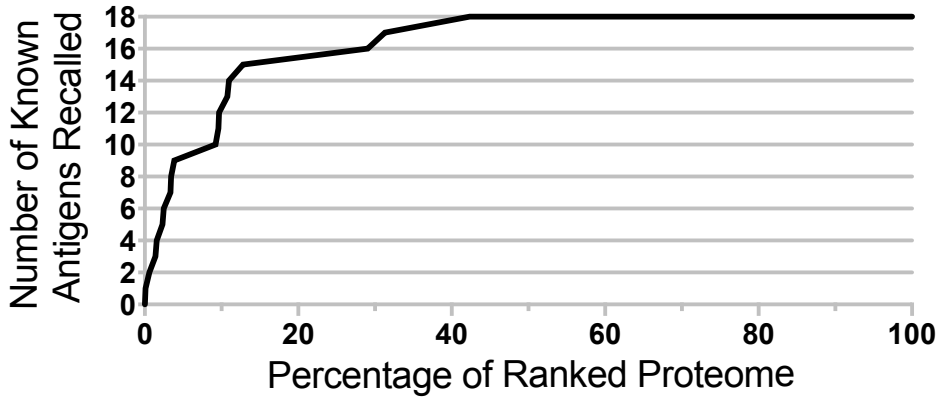


Figure 4.1 Plot Depicting the Ability of the BPAD200+N+B+AF Classifier to Recall (A) Antigens Incorporated in the Vaccine BEXSERO and (B) Proteins that were Positive in a Serum Bactericidal Activity (SBA) Assay: The BPAD200+N+B+AF classifier was used to classify each protein within the proteome of *MenB* as a BPA or non-BPA. Proteins were also ranked on the predicted probability of being a BPA. The position of each protein in the ranked proteome of *MenB* was expressed as a percentage (N = 1943 proteins).

4.3.2: The Enhanced Classifier (BPAD200+N+B+AF) was able to Significantly Recall Proteins that were Positive in a Serum Bactericidal Assay for *MenB*

The ability of the BPAD200+N+B+AF classifier to recall antigenic proteins, not incorporated in the BEXSERO vaccine, in the background of the *MenB* proteome was assessed (**Figure 4.1 B**). Specifically, the ability of the BPAD200+N+B+AF classifier to enrich for proteins that have been shown to be positive in a SBA assay against *MenB,* in the top 100 predicted BPAs was assessed (N=18, a list of the SBA positive proteins can be seen in **Appendix K**). The BPAD200+N+B+AF classifier predicted 15 out of the 18 (83.33%) SBA positive proteins as BPAs. Nine of the 18 curated SBA positive proteins were recalled in the top 100 predicted BPAs by the classifier BPAD200+N+B+AF, reflecting a significant enrichment (p = 5.61E-08, Hypergeometric test). A subset of the SBA positive proteins test dataset (N=3), were known BPAs used to train the BPAD200+N+B+AF classifier. To evaluate the enhanced classifier's (BPAD200+N+B+AF) ability to predict novel SBA positive proteins, recall was calculated whilst leaving the three SBA positive proteins, known BPAs out of the metric (N=15). When evaluating the ability of the BPAD200+N+B+AF classifier to recall the 15 SBA positive proteins not used to train the BPAD200+N+B+AF classifier, a significant enrichment of SBA positive proteins was still achieved. In the top 100 predicted BPAs, eight of the 15 were recalled (p = 1.71E-07, Hypergeometric test). The recall results for the SBA positive proteins test dataset can be seen in **Appendix K**. In summary, the BPAD200+N+B+AF classifier was able to significantly recall SBA positive proteins in the top 100 predicted BPAs regardless of whether the SBA positive proteins tested were included in the training data of the classifier or not.

4.3.3: The Enhanced Classifier (BPAD200+N+B+AF) was able to Significantly Recall Known BPAs in *Mtb*

Recall statistics were evaluated in *Mtb* and it was observed that classifiers utilising 10 features exhibited a large amount of variation in the number of known BPAs recalled in the top 100 predicted BPA lists across repeated iterations (**Table 4.1**). When evaluating the ability of the BPAD200+N+B+AF classifiers, constructed using 10 features, to enrich top 100 predicted BPA lists for known BPAs, three out of five iterations did obtain significance (p = < 0.05, Hypergeometric test). To test whether increasing the number of features generated a more stable recall metric, BPAD200+N+B+AF classifiers were trained using 10, 20, 30, and 40 features (**Table 4.1**). From five iterations it was not entirely clear whether using a larger number of features generated a more stable recall metric. However, classifiers built with 40 features were able to significantly recall known BPAs across all five iterations. The higher performance of 40 feature classifiers was also visible in the majority of plots when comparing recall curves to those generated with 10 features (**Figure 4.2**). This suggests that classifiers with 40 features should be used for *Mtb* recall.

Feature Number	Recall Metric	Iteration 1	2	3	4	5	Standard Deviation Across Five iterations of Recall
10	Known BPAs in Top 100.	6	6	5	1	0	2.88
	P-value	**1.520E-06**	**1.520E-06**	**3.320E-05**	2.780E-01	6.600E-01	
20	Known BPAs in Top 100.	8	8	2	2	0	3.74
	P-value	**1.480E-09**	**1.480E-09**	5.450E-02	6.600E-01	6.600E-01	
30	Known BPAs in Top 100.	6	6	3	3	1	2.17
	P-value	**1.520E-06**	**1.520E-06**	**6.570E-03**	**6.570E-03**	2.780E-01	
40	Known BPAs in Top 100.	10	9	7	4	9	2.39
	P-value	**5.350E-13**	**3.190E-11**	**5.380E-08**	**5.460E-04**	**3.190E-11**	

Table 4.1. Assessing the Recall Metric in *Mycobacterium tuberculosis* Across Five Iterations and Different Numbers of Features Utilised by the BPAD200+N+B+AF Classifier: Showing the number of known bacterial protective antigens (BPA) recalled in the top 100 predicted BPAs of the ranked *Mtb* proteome. *P*-values were obtained from hypergeometric tests evaluating the enrichment of known BPAs in the top 100 predicted BPA recall lists. Values in **bold** represent a *p*-value < 0.05.

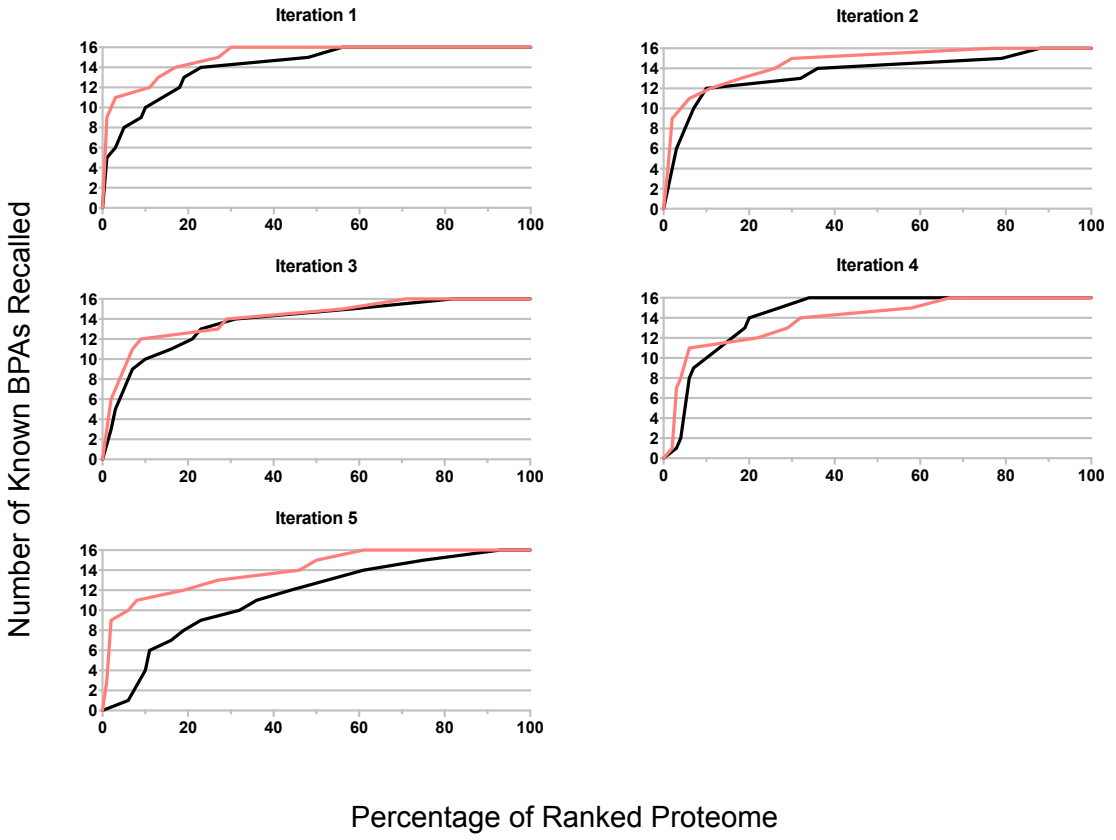


Figure 4.2 Plots Depicting the Ability of Classifiers to Recall Known Bacterial Protective Antigens (BPAs) from the *Mycobacterium tuberculosis* (*Mtb*) proteome: BPAD200+N+B+AF classifiers of different numbers of features were constructed with greedy backward feature elimination in order to rank proteins in the *Mtb* proteome for their probability of being a BPA. This was repeated over five iterations. The position at which known BPAs are recalled in the ranked proteome was expressed as a percentage of the entire proteome. Red = classifier utilising 40 features. Black = classifier utilising 10 features. There are 16 known BPAs of *Mtb* in the training data of BPAD200+N+B+AF.

4.3.4: Recall of Previously Tested Laboratory Proteins

In **Chapter 2** of this thesis, six proteins that ranked in the top 100 predicted BPAs for *Mtb* using the RV classifier of Bowman et al[48] were tested in the laboratory and shown not to confer protection in a murine challenge model of *Mtb* infection (**Figure 2.6**). The ability of BPAD200+N+B+AF classifiers trained on 40 features to recall these six proteins in the top 100 predicted BPAs was evaluated. It was shown that in three of the five iterations small subsets of the six proteins were included in the top 100 predicted BPA lists (iteration 1 N=1, Iteration 2 N=1, Iteration 3 N=2, iteration 4 N=0, iteration 5 N=0). Only one of the iterations showed a significant enrichment for the six proteins previously tested in the laboratory (Iteration 3, p = 0.0088, Hypergeometric test). The majority of the recall iterations show that the enhanced classifier, BPAD200+N+B+AF, was not actively selecting the six proteins that failed to induce a protective response, this suggested that different BPAs were elevated within the top 100 predicted BPAs of the BPAD200+N+B+AF classifier when compared to the RV classifier developed by Bowman et al[48].

4.4: Discussion

The work in this chapter applied the recall metric[48] to demonstrate the biological relevance of the newly developed RV classifier, BPAD200+N+B+AF (**Chapter 3**). As hypothesised (**Hypothesis 4.1**), the BPAD200+N+B+AF classifier was able to successfully recall known antigens in a background of an entire proteome for both *MenB* and *Mtb.* A significant enrichment was also seen (p = 1.71E-07, Hypergeometric test) in the top 100 predicted BPAs for the proteome of *MenB* when evaluating 15 SBA positive proteins that were not present in the training data of the BPAD200+N+B+AF classifier. This strongly suggested that the BPAD200+N+B+AF classifier was able to identify novel SBA positive proteins in *MenB*. Since there is such a strong correlation between SBA positive proteins and protection (i.e. BPA) in *MenB*[129] it would be hypothesised that the BPAD200+N+B+AF classifier could be used to predict novel BPAs for *MenB*. It is envisaged that novel BPA predictions could ultimately be incorporated into a subunit vaccine to boost the efficacy of current *MenB* vaccines (i.e. BEXSERO). It was also shown that the BPAD200+N+B+AF classifier was able to significantly (p-value < 0.05) recall known BPAs in the *Mtb* proteome across five iterations when utilising 40 features. These 40 feature classifiers were employed to evaluate **Hypothesis 4.2**. **Hypothesis 4.2** of this chapter predicted that the six proteins that failed to elicit protection in a mouse model of *Mtb* in **Chapter 2** would not be significantly enriched for in the top 100 predicted BPAs for *Mtb* using the BPAD200+N+B+AF classifier. This hypothesis (i.e. **Hypothesis 4.2)** was not proven, when utilising the highest performing (40 feature) classifiers, a significant enrichment was obtained in one iteration of recall for the six proteins that failed in protection assays in **Chapter 2** (Iteration 3, p = 0.0088, Hypergeometric test). However, the iteration (i.e. iteration 3) that significantly enriched the top 100 predicted BPAs, for the six previously tested BPAs only recalled two out of the six. Furthermore, four out of five iterations of BPAD200+N+B+AF classifiers trained on 40 features did not significantly enrich for these six proteins within the top 100 predicted BPAs, suggesting that the BPAD200+N+B+AF classifier was making different predictions of BPAs when compared to the previous RV classifier[48]. This implied that the top 100 predicted BPA lists from the BPAD200+N+B+AF classifier represented a novel resource of predicted BPAs for *Mtb*.

Whilst further exploring the recall metric it was necessary to utilise a greater number of features to consistently, significantly enrich known BPAs in the top 100 predicted BPAs for *Mtb* (**Table 4.1**). Across five iterations, when assessed using the recall metric, BPAD200+N+B+AF classifiers trained on 40 features appear to out

perform BPAD200+N+B+AF classifiers trained on smaller numbers of features (**Table 4.1, Figure 4.2**). Despite BPAD200+N+B+AF classifiers trained on 40 features significantly recalling known BPAs in *Mtb* across all five iterations the number of recalled known BPAs in the top 100 predicted BPA lists varied. Variation was also seen when assesing the recall of BPAD200+N+B+A+F classifiers on six proteins that failed to elicit a protective immune response (**Chapter 2**), where one of five iterations did significantly enrich for the six proteins but the other four iterations did not. This variation was undesired, as different iterations result in different predicted BPAs and different abilities to recall known BPAs. This variation was a major limitation of this study. The most probable cause of the variable performance, of the BPAD200+N+B+AF classifiers to recall known BPAs displayed in this chapter, is the greedy backward feature elimination utilised to select features for each iteration of the classifiers. Greedy backward feature elimination included a random step in which, if the least informative features had equal discriminatory abilities between the two classes (BPA and non-BPA), then one was removed at random. It was suspected that it is a difficult and noisy problem to predict BPAs and one feature was not strongly discriminatory between the two groups (BPA and non-BPA). If this random step incorporated different features then the make up of the final classifiers would have been altered. That said the selected features sets often reflect the same biological phenomena (i.e. a different feature but generated from the same protein annotation tool). This was seen in the fact that across all five iterations of the 40 feature classifiers only 24 different protein annotation tools were used to generate the features.

The simplest way of overcoming the variability associated with known BPA recall across iterations would be to fix the feature set. It was important to run different iterations and leave-tenth-out cross-validation to gain an understanding of how the classifier performed generally on unseen data. However, once the overall ability of the BPAD200+N+B+AF classifier to predict known BPAs was ascertained (**Chapter 3**), one could take the feature set derived from the iteration that gave the highest level of recall for future BPA predictions. A potentially more sensitive method, to overcome the variability associated with BPA recall by fixing a feature set, could be isolating a specific feature set for each pathogen. Pathogens have different methods of invading a host and evading their immune system. Currently there is not enough data to train classifiers for each pathogen. Due to difficulties in training dataset generation an option that would potentially need a smaller training dataset size increase would be to utilise a leave-one-bacteria-out validation (LOBOV). This would entail the known BPAs and non-BPAs of one bacterium being left out of the classifier's training data and used as a test set to see how the classifier performs predicting BPAs in each specific bacterial

species. By performing LOBOV the ability of the BPAD200+N+B+AF classifier would be assessed on a wide range of bacterial species independently. One could envisage that predicting BPAs using ML RV approaches may be more successful in certain species over others. An example of a pathogen that would be of interest for a LOBOV approach would be *Mtb.* The pathogen *Mtb* is an unusual organism, as it is not classed as Gram positive or negative. Largely the training data for the BPAD200+N+B+AF classifier comprised of Gram positive (32.5%) or negative (57.5%) BPAs as opposed to mycobacterial (10%) BPAs. Due to fundamental differences in cell structure it could be hypothesised that separate classifiers for each of the types, Gram positive, Gram negative or mycobacterial would achieve better accuracies when classifying BPAs and non-BPAs. However a signal for Gram stain (i.e. Gram positive, Gram negative or mycobacterial) was not found when clustering the BPAs in the dataset BPAD200, differences were only revealed between BPAs when labelled by subcellular localisation (**Figure 3.4**). Furthermore, the construction of generalisable classifiers for each Gram stain is currently still hampered by the size of the training data. However, a generalisable signal of protection is being learnt by the combined Gram stain classifier (BPAD200+N+B+AF), which can be shown in *Mtb*. Despite mycobacterial being the smallest Gram stain class present in the training data for the BPAD200+N+B+AF classifier, a significantly enriched predicted BPA list was still generated when assessing the recall metric in the pathogen *Mtb* (**Table 4.1**). Due to Gram stain differences not being seen whilst interrogating the classifier's (BPAD200+N+B+AF) training data, it is suggested that in the future as new protein annotation tools are incorporated into ML RV approaches, LOBOV should be considered to evaluate how generalisable the signal for BPAs are across individual bacterial pathogens. It is hypothesised that as new protein annotation tools become available their incorporation into ML RV approaches may reveal differences between bacterial species, which have not been realised in the current dataset.

Once species differences can be detected in the ML RV training approaches then LOBOV should be conducted to gain an insight into the ability of classifiers to classify BPAs from bacterial species not present within the training dataset. The ability of RV classifiers to predict BPAs for bacterial species not present in the training data would enable the rapid prediction of BPAs for newly emergent pathogens. A scenario could be envisaged where a newly emergent pathogen[136], for which a genome sequence has been established but no protective antigens have been identified and RV would represent a rapid approach to identify BPAs. An example of this occurred in 1976, when a bacterial outbreak of unknown origin infected 221 individuals and killed 34 at the American Legion convention in Philadelphia[137]. It took six months to identify the

pathogenic cause as the bacterium *Legionella pneumophila*. If a similar situation arose today, the pathogen could be sequenced from infected individuals and could be exposed to an RV approach to facilitate the prediction of BPAs and subunit vaccine formulation.

Future work to this approach of ML in RV is first suggested to focus on the bacterial pathogens (*MenB* and *Mtb*) interrogated in this chapter. The subunit vaccine BEXSERO has recently been incorporated into the NHS childhood vaccination program and is a successful vaccine[138]. However, early predictions of protection are at 66.1% of vaccinated individuals, which leaves room for improvement[138]. It was shown that the BPAD200+N+B+AF classifier would have predicted four out of five proteins in the BEXSERO vaccine as protective and therefore is describing a protective signature for *MenB*. With that in mind, it is hypothesised that including additional predicted BPAs from the RV classifier tested in this chapter (BPAD200+N+B+AF) could potentially boost the protective efficacy of *MenB* vaccines. The field of *Mtb* vaccine research is again returning to prominence as a method to combat the spread of drug resistant strains of Tuberculosis[132], due to the variable protective efficacies witnessed in adults by the BCG[27]. It is envisaged that future collaboration with vaccine experts in the field of *Mtb* would lead to potential novel vaccine candidates being discovered. Selected predicted BPAs from BPAD200+N+B+AF should undergo animal challenge assays. If the predicted BPAs conferred protection then they could be incorporated into a new subunit vaccine against *Mtb*. In summary the work in this chapter has shown that the BPAD200+N+B+AF classifier is a biologically focussed classifier, which can be used to predict novel BPAs for any bacterial pathogen with a sequenced genome.

4.5: Statement of Contribution for Research in this Chapter

I personally carried out all of the work detailed in this chapter. A list of literature curated SBA assays in *N. meningitidis* was provided by Prof. Myron Christodoulides; proteins from this list were used to form the SBA positive protein training dataset for recall.

Chapter 5: General Discussion, Limitations and Future Work

5.1: General Discussion

The research in this thesis followed a logical experimental workflow. First, predicted bacterial protective antigens (BPAs) from the reverse vaccinology (RV) pipeline of Bowman et al[48] were tested in a mouse model of tuberculosis (TB) infection (**Chapter 2**). When the desired protective efficacy was not achieved, the *in silico* RV approach was revised and errors related to implementing a nested cross validation approach and an artificial subcellular localisation bias were removed. After correcting for these, improvements in accuracy to this previous RV classifier were achieved through increasing the amount of training data on which the classifier was trained and the number of protein annotation tools used to annotate the training data (BPAD200+N+B+AF, **Chapter 3**). Finally, the ability of the BPAD200+N+B+AF classifier to recall known antigens in the background of entire proteomes was assessed for *Neisseria meningitidis* serogroup B (*MenB*) and *Mycobacterium tuberculosis* H37RV *(Mtb)* (**Chapter 4)**. Collectively this thesis represents an advance in the field of RV and demonstrated that machine learning (ML) in RV research shows great promise.

Laboratory testing of six predicted BPAs by the RV classifier of Bowman et al[48] was conducted in **Chapter 2** of this thesis. It was hypothesised (hypothesis 2.1) that the six putative vaccine candidates (VC) selected for animal challenge experiments would confer significant levels of protection against infection with *Mtb*. This was shown not to be the case, prompting enhancements to the RV ML classifier as described in **Chapter 3** of this thesis. It should be noted that despite an artificial bias and over estimations of obtained accuracies one of the predictions made by the Bowman et al[48] ML RV approach did show some level of protection in an mouse model of *Mtb* infection (**Figure 2.7 A**). However, this protection was not consistent across repeated animal challenge assays (**Figure 2.7 B**); improvements to the RV classifier were explored to address the lack of protection generated in the mouse models of *Mtb* infection utilised in **Chapter 2**.

Chapter 3 extended the field of RV and confirmed that a signal for antigenic protection can be attained from a literature curation of BPAs (**Figure 3.3,** hypothesis 3.1). In this chapter (**Chapter 3**) it was shown that implementing corrections (i.e. nested leave-tenth-out cross-validation (LTOCV) and removing an artificial subcellular localisation bias incorporated in the training data) to the previous RV classifier (Bowman et al[48]) lowered the estimated accuracies when classifying BPAs and non-

BPAs. This was shown when testing hypothesis 3.2, that implementing a nested LTOCV would reduce the accuracies achieved when classifying BPAs and non-BPAs. By implementing a nested LTOCV the overfitting (i.e. higher accuracies) of the previous ML RV approach was removed. Furthermore the accuracies obtained when predicting BPAs and non-BPAs were lowered through the removal of the subcellular localisation bias. This is a vital correction for RV classifiers, by removing this subcellular localisation bias, the major advantage of ML approaches to RV (i.e. all subcellular localisations are included and therefore BPAs are predicted for all subcellular localisations) when compared to traditional filtering approaches can be realised.

Previously, Bowman et al[48] had included an artificial bias that meant the negative training dataset had a much greater proportion of cytoplasmic proteins than the positive training dataset. This was reflected in the resulting classifier with the most deterministic feature between the positive and negative training datasets being whether a protein is localised to the cytoplasm ("PSORTb-ProbCytoplasm")[48]. If the protein was predicted to have a cytoplasmic subcellular localisation this strongly correlated with the protein being predicted as a non-BPA. In the newly developed approach (BPAD200+N+B+AF) this artificial bias was removed by balancing the positive and negative training data for subcellular localisation. For every protein in the positive training data (BPA) a non-BPA from the same bacterial species, with the same predicted subcellular localisation was randomly selected to form the negative training data. Classifiers balanced for subcellular localisation were able to better capture a protective signature as opposed to a signal for subcellular localisation (**Table 3.1**). By removing this bias it enabled the classifier (BPAD200+N+B+AF) developed in this thesis to be utilised to explore the immunological signal for protection. Using greedy backward feature elimination on the BPAD200+N+B+AF training dataset revealed the top ten features (most deterministic between the positive training data i.e. BPA and the negative training data i.e. non-BPA) as deemed by the RV classifier developed in this thesis. Expected annotations such as epitope predictors were included in the top 10 most deterministic features but also features describing more general biological phenomena were strongly represented (**Table 3.1**). The biological annotation in the top 10 most deterministic features, between the positive and negative training data, consisted of glycosylation, phosphorylation and lipoproteins, all of which have been implicated in a protective immune response[120, 122, 123]. The mix of biological annotation included in the top 10 most deterministic features of BPAD200+N+B+AF highlights the complex mechanisms through which protection is conferred and underlines the fact that the specific attributes needed to confer protection require further research.

Ultimately these features (**Table 3.1**) suggest that epitope prediction on it's own is not as informative when classifying BPAs and non-BPAs, as combining both general biological annotation and epitope predicting annotation. The final hypothesis (hypothesis 3.3) tested in **Chapter 3** was that increasing the size of the training data and increasing the type of annotation tools incorporated when training classifiers to predict BPAs from non-BPAs would both individually increase the performance of RV classifiers and this was shown to be correct (**Figure 3.4**). This increase in the ability of RV classifiers was attributed to the enhanced BPAD200+N+B+AF classifier learning better generalisations, of the differences between BPAs and non-BPAs, from the expanded training datasets when compared to the previous RV classifier[48].

Chapter 4 tested the biological relevance of the RV classifier, BPAD200+N+B+AF, using a metric termed recall[48]. Firstly, hypothesis 4.1, that BPAD200+N+B+AF was able to significantly enrich for known antigens in top 100 predicted BPA lists for bacterial pathogens, was proven correct (**Figures 4.1 and 4.2**). Secondly, hypothesis 4.2 that the six proteins assayed for protective efficacy in **Chapter 2** would not be significantly enriched in the top 100 predicted BPAs for *Mtb* using the BPAD200+N+B+AF classifier was found not to be correct. A significant enrichment of the six proteins assayed for protective efficacy in **Chapter 2** was obtained in one out of the five iterations of the BPAD200+N+B+AF classifier trained on 40 features (Iteration 3 p = 0.0088, Hypergeometric test). However, four out of five iterations of the BPAD200+N+B+AF classifier trained on 40 features were shown not to significantly enrich for the six proteins previously tested in **Chapter 2.** Furthermore in the iteration of recall that significantly enriched for the six proteins previously tested **Chapter 2**, only a small subset (N=2) of the six previously predicted BPAs were present. Due to the fact that four out of five iterations of recall did not significantly enrich for the six previously predicted BPAs and in the iteration that they were enriched for they were only present as a small subset, it was suggested that the BPAD200+N+B+AF classifier was making novel BPA predictions. The work carried out in **Chapter 4** has shown that the BPAD200+N+B+AF classifier developed in this thesis (**Chapter 3**) was able to generate biologically relevant predicted BPA lists that were different from the previous RV classifier[48].

The organisms selected for testing in **Chapter 4** were *Mtb* and *MenB*. *Mtb* is the causative agent of TB and was selected for targeting with an RV approach as a response to a renewed focus on finding novel prophylactic vaccines for TB. This renewed focus on vaccine research in *Mtb* has primarily been a reaction to the rising levels of drug resistant *Mtb* infection, which is rapidly becoming resistant to common antibiotics[24, 132]. Currently the leading vaccine for preventing TB is the BCG[27] and

despite approximately four billion doses having been administered, rates of drug resistant TB infections are on the rise[132]. The BCG vaccine is good at protecting infants from all forms and adults from extrapulmonary forms of TB[27]. However the primary infective site of TB is the lungs (pulmonary) and therefore to prevent the continued spread of drug resistant strains a new prophylactic vaccine for *Mtb* is required. **Chapter 4** confirmed the ability of the BPAD200+N+B+AF classifier to recall known BPAs in *Mtb* and by extension ML in RV's ability to predict putative BPAs for protection studies against *Mtb* infection. The BPAs predicted in the proteome of *Mtb* by the BPAD200+N+B+AF classifier should be explored with experts in the field of TB vaccine research to facilitate the discovery of novel *Mtb* VCs. *MenB* was selected to enable a comparison to the success story attributed to RV (i.e. BEXSERO). Vaccines for the serogroups of A, C, W, and Y had been formulated using capsular polysaccharides but in serogroup B this was not possible. It was the application of early RV filtering approaches[52] that led to the identification of novel antigens that were incorporated into the subunit vaccine BEXSERO. Despite only recently being incorporated into the NHS childhood vaccination regime (2015) coverage of the BEXSERO vaccine is predicted at 66.1% protection[138]. This is a large improvement compared to no vaccine but also could still be improved upon. In **Chapter 4** of this thesis it was shown that the BPAD200+N+B+AF classifier was able to characterise a signal for protection in *MenB* (**Figure 4.1**). Specifically the BPAD200+N+B+AF classifier predicted four out of five of the antigens incorporated in the vaccine BEXSERO as BPAs. It is envisaged that utilising the predictions of BPAs made by the BPAD200+N+B+AF classifier, novel BPAs for *MenB* could be discovered. Novel BPAs that are confirmed through laboratory trials could then be incorporated into subunit vaccines of the future to boost the efficacy of vaccines against *MenB*. In summary, future collaborations with experts in the field of vaccinology should be pursued to interrogate BPAs predicted by the BPAD200+N+B+AF classifier in animal protection assays in order to develop novel subunit vaccines in the future.

5.2: Limitations And Future Work

Despite improving the application of ML to RV there were a number of limitations associated with the work in this thesis. General limitations that were present throughout this thesis and the approach to using ML in RV relate to the collection of the training data. Despite an intensive literature curation effort the training data for the BPAD200+N+B+AF classifier was composed of 200 positive training examples (BPAs) and 200 negative training examples (non-BPAs). This was not a large dataset in terms of traditional ML studies. It should be highlighted however, that this was the largest dataset of BPAs ever assembled, improving upon the previous datasets[48, 87]. This increase in training data was shown to increase accuracies obtained when classifying BPAs and non-BPAs (**Figure 3.4**).

Due to the limited ability of curating BPAs, the training data of BPAD200+N+B+AF did not contain examples of BPAs from all bacterial species. The ability of the classifier to make predictions of BPAs for bacterial species not present in the training data was a potential limitation. Future work to evaluate the ability of the RV classifier to predict BPAs for bacterial species not included in the training data of BPAD200+N+B+AF could be focussed around a new metric termed leave-one-bacteria out-validation (LOBOV). LOBOV would be conducted by removing a bacterial species incorporated in the training data of a classifier. Separate classifiers with each bacterial species in turn left out and used as a test dataset (detailed in **Chapter 4**) would then be used to evaluate the classifier's ability to predict BPAs in bacterial species not present in the training data of the classifier. It would be hypothesised that the BPAD200+N+B+AF classifier would perform consistently across all bacterial species, as the training data represents multiple classes of bacterial pathogens and is comprised of Gram negative, Gram positive and mycobacterial species. The smallest class of the training data is mycobacterial and it has been shown that the BPAD200+N+B+AF classifier can recall known BPAs in *Mtb* (**Table 4.1 and Figure 4.2**) despite being trained on a dataset mostly comprised of Gram positive and negative proteins. This suggests that the BPAD200+N+B+AF classifier has learnt a generalisable signature for bacterial protection (i.e. BPA) and this can be extrapolated across different classes (i.e. Gram negative, Gram positive or mycobacterial) of bacterial pathogens.

Another limitation is that the training dataset may contain noise. This noise would have been incorporated during the selection of negative training data and also by the bioinformatics annotation tools that were used to generate features from the training data. First, non-BPAs were selected randomly from the same proteome as each known

BPA and this may have resulted in the mistaken addition of BPAs into the negative training data (non-BPAs) that have not yet been tested and confirmed in the literature record. To mitigate this, non-BPAs with homology to BPAs were discarded. A second source of noise could be the way in which the training features were generated. Proteins (BPAs and non-BPAs) in the training dataset were annotated with protein annotation tools. The annotation tools utilised to generate features did not achieve 100% accuracies for the biological characteristic that they were describing. For example, when predicting subcellular localisation using PSORTb, 26.5% of the BPAD136 positive training data was identified as having an unknown subcellular localisation (**Figure 3.5**). Therefore, some noise was introduced into the features that were used to train the RV classifiers in this thesis. Despite both known potential sources of noise (i.e. negative training data generation and the bioinformatics annotation tools used to generate features) included when training RV classifiers, it has been shown that SVM classifiers are able to overcome this noise (**Figure 3.3)**. **Figure 3.1** demonstrated that the methods used to generate the training data and the annotation features, were able to capture a biological signal for protection in contrast to the result of randomly permuting the labels BPA and non-BPA (i.e., classifiers were not capturing randomly generated noise). In summary, the ML RV techniques detailed in this thesis may have included potential noise within the training datasets; despite this classifiers trained using these datasets (BPAD200+N+B+AF) have been shown to capture a biological signal for protection and not noise (**Figure 3.3**). Future studies should focus on maximising this signal to noise ratio.

Future work to improve the ability of the BPAD200+N+B+AF classifier to capture a protective signature should be related to increasing the size of the training dataset (number of BPAs and non-BPAs) as well as increasing the range of protein annotation tools (used for feature generation). It was shown that by increasing the size of the training data and the breadth of protein annotation tools used to annotate the training data an increase in the accuracies obtained when classifying BPAs and Non-BPAs was achieved (**Figure 3.4**).

A final limitation in this thesis was the attempted laboratory validation of predicted BPAs from the RV classifier of Bowman et al[48] (**Chapter 2**), which were shown not to be protective in mouse models of *Mtb* infection. Revisions to the RV classifier developed by Bowman et al[48] detailed in **Chapter 3** were made to address the failure of the attempted laboratory validation for predicted BPAs. It would be informative to utilise predictions from the newly developed BPAD200+N+B+AF classifier (**Chapter 3**) for laboratory testing and this should be a goal for future work. Despite not being able to test novel BPAs predicted by the BPAD200+N+B+AF classifier in animal challenge

assays, when using the recall metric this newly developed classifier was shown to have biological relevance (**Chapter 4**).

Additional future work could explore the impact of utilising different ML algorithms in RV approaches. The SVM algorithm was chosen in this application of ML to RV, as it was a widely applicable algorithm that achieved high accuracies even on noisy data and is still widely used in bioinformatics research[48, 72, 78]. An example of an ML algorithm that could match or surpass the accuracies achieved by SVMs is neural networks (NN) and more specifically deep NNs[72]. Deep NNs are an algorithm that is loosely based on a model of the human brain that is made up of layers known as input layer, a hidden layer and an output layer. The hidden layer applies functions and weights to each input (i.e. features from the input layer) this hidden layer is then utilised to generate an output. Deep NNs have several hidden layers that enable the algorithm to learn abstract and complex interactions by applying transformations to already transformed features, thus achieving high classification accuracies[73]. This additional work is beyond the scope of this thesis due to the biological focus and the desire to create a practical RV classifier that offers the ability to be utilised to generate novel VCs for laboratory validation.

To expand the impact of this work and disseminate the predictions from the newly developed RV classifier, BPAD200+N+B+AF, future work would be to use the classifier to predict BPAs (i.e. putative VCs) for every bacterial pathogen with a sequenced genome. BPA predictions could then be dispersed throughout the vaccine community via a database named the bacterial protective antigen database (BPADb). By making BPA predictions available it will enable the rapid laboratory validation of predicted BPAs from this RV classifier. If successful, this platform will have the potential to significantly speed up vaccine discovery that would lead to the creation of many more subunit vaccines against bacterial infectious diseases. It is envisaged that laboratory testing of predicted BPAs would be reported to this database. This would enable novel laboratory confirmed BPAs to be incorporated into the training data of BPAD200+N+B+AF, further improving the classifier's ability to distinguish between BPAs and non-BPAs. In summary, despite the limitations above, I have shown that a signal describing biological protection can be captured through a literature curation for BPAs. I have improved upon previous ML RV approaches and the ability of BPAD200+N+B+AF to identify novel potential BPAs has been shown through the recall metric.

5.3: Concluding Remarks

The major finding of this thesis was that the newly developed BPAD200+N+B+AF classifier was able to capture a signal for biological protection through a literature curation of BPAs (**Figure 3.3**). After proving that a signal for protection could be curated from a literature search, **Chapter 3** implemented enhancements to the previous ML RV approach[48]. First, by correcting the assessment of classifier performance by implementing a nested LTOCV and then removing an artificial subcellular localisation bias from the training data. After these initial corrections, an increase in the performance of the classifier when distinguishing BPAs and non-BPAs was achieved by increasing the amount of training data and increasing the type of protein annotation tools used to generate features from the training data (**Figure 3.4**).

The final chapter of this thesis (**Chapter 4**) established the biological relevance of the BPAD200+N+B+AF classifier and future work should first focus on testing novel predicted BPAs in laboratory assays. Upon successful confirmation of protection for predicted BPAs they can be incorporated in the formulation of new subunit vaccines. In conclusion, with the success of the BEXSERO (*MenB*) vaccine RV has emerged as a wing of vaccinology in its own right. I believe, that as our understanding grows of what makes a protein a BPA, based on the features selected by ML classifiers, that RV will routinely lead to the development of subunit vaccines that reduce mortality and morbidity in human populations.

Bibliography

1. André FE; Vaccinology: past achievements, present roadblocks and future promises. *Vaccine* 2003;**21**(7):593-595

2. Stern AM, Markel H; The history of vaccines and immunization: familiar patterns, new challenges. *Health Affairs* 2005;**24**(3):611-621

3. Murphy KM. *Janeway's immunobiology*: Garland Science, 2011.'I.S.B.N' 1136665218

4. Nikesh A Patel DD. *Vaccinia, Infectious Diseases*. http://emedicine.medscape.com/article/231773-overview (21/03/2017 'Date Accessed')

5. U.S. Department of Health and Human Services Centers for Disease Control CfDC, and Prevention. *Antibiotic Resistance Threats in the United States, 2013*. http://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf (12/03/2017 'Date Accessed')

6. Medzhitov R; Recognition of microorganisms and activation of the immune response. *Nature* 2007;**449**(7164):819-826

7. Janeway Jr CA, Travers P, Walport M, Shlomchik MJ; Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001. The complement system and innate immunity. Available from: https://http://www.ncbi.nlm.nih.gov/books/NBK27100/. 2001

8. Clem AS; Fundamentals of vaccine immunology. *Journal of Global Infectious Diseases* 2011;**3**(1):73

9. Sousa CRe; Activation of dendritic cells: translating innate into adaptive immunity. *Current Opinion in Immunology* 2004;**16**(1):21-5

10. Wieder E; Dendritic cells: a basic review. *International Society for Cellular Therapy* 2003

11. Luckheeram RV, Zhou R, Verma AD, Xia B; CD4(+)T cells: differentiation and functions. *Clinical and Developmental Immunology* 2012;**2012**:925135

12. Huether SE, McCance KL. *Understanding pathophysiology*: Elsevier Health Sciences, 2013.'I.S.B.N' 0323293433

13. Guermonprez P, Valladeau J, Zitvogel L, Théry C, Amigorena S; Antigen presentation and T cell stimulation by dendritic cells. *Annual review of immunology* 2002;**20**(1):621-667

14. Baxter D; Active and passive immunity, vaccine types, excipients and licensing. *Occupational Medicine* 2007;**57**(8):552-556

15. Cann A, Stanway G, Hughes P, Minor P, Evans D, Schild Gt, Almond J; Reversion to neurovirulence of the live-attenuated Sabin type 3 oral pollovirus vaccine. *Nucleic Acids Research* 1984;**12**(20):7787-7792

16. Sur D, Kanungo S, Sah B, Manna B, Ali M, Paisley AM, Niyogi SK, Park JK, Sarkar B, Puri MK; Efficacy of a low-cost, inactivated whole-cell oral cholera vaccine: results from 3 years of follow-up of a randomized, controlled trial. *PLoS Neglected Tropical Diseases* 2011;**5**(10):e1289

17. Hansson M, Nygren PA, Stahl S; Design and production of recombinant subunit vaccines. *Biotechnology and Applied Biochemistry* 2000;**32 (Pt 2)**:95-107

18. Coffman RL, Sher A, Seder RA; Vaccine adjuvants: putting innate immunity to work. *Immunity* 2010;**33**(4):492-503

19. Heinson AI, Woelk CH, Newell ML; The promise of reverse vaccinology. *International health* 2015;**7**(2):85-9

20. Huang Jian H, Wataru; CED: a conformational epitope database. *BMC Immunology* 2006;**7**(1)

21. Christodoulides M; Neisseria proteomics for antigen discovery and vaccine development. *Expert Review of Proteomics* 2014;**11**(5):573-591

22. Russell DG, Barry CE, Flynn JL; Tuberculosis: what we don't know can, and does, hurt us. *Science* 2010;**328**(5980):852-856

23. WHO. *Tuberculosis Fact sheet N°104*. http://www.who.int/mediacentre/factsheets/fs104/en/ (12/03/2017 'Date Accessed')

24. Kaufmann SH; Fact and fiction in tuberculosis vaccine research: 10 years later. *The Lancet Infectious Diseases* 2011;**11**(8):633-40

25. Zumla A, George A, Sharma V, Herbert RH, Oxley A, Oliver M; The WHO 2014 global tuberculosis report, further to go. *The Lancet Global Health* 2015;**3**(1)

26. Jassal M, Bishai WR; Extensively drug-resistant tuberculosis. *The Lancet Infectious Diseases* 2009;**9**(1):19-30

27. WHO. *Tuberculosis vaccine developments, Immunization, Vaccines and Biologicals*. http://www.who.int/immunization/research/development/tuberculosis/en/ (19/11/2015 'Date Accessed')

28. Monterrubio-Lopez GP, Gonzalez YMJA, Ribas-Aparicio RM; Identification of Novel Potential Vaccine Candidates against Tuberculosis Based on Reverse Vaccinology. *BioMed research international* 2015;**2015**

29. Crocheton N; Stopping routine vaccination for tuberculosis in schools. *BMJ* 2005;**331**:647-8

30. Elkington PT; Tuberculosis: time for a new perspective? *Journal of Infection* 2013;**66**(4):299-302

31. Ottenhoff THM, Kaufmann SHE; Vaccines against Tuberculosis: Where Are We and Where Do We Need to Go? *PLoS Pathogens* 2012;**8**(5):e1002607

32. Sun R, Skeiky YA, Izzo A, Dheenadhayalan V, Imam Z, Penn E, Stagliano K, Haddock S, Mueller S, Fulkerson J; Novel recombinant BCG expressing perfringolysin O and the over-expression of key immunodominant antigens; pre-clinical characterization, safety and protection against challenge with Mycobacterium tuberculosis. *Vaccine* 2009;**27**(33):4412-4423

33. Finch R, Hunter PA; Antibiotic resistance--action to promote new technologies: report of an EU Intergovernmental Conference held in Birmingham, UK, 12-13 December 2005. *The Journal of Antimicrobial Chemotherapy* 2006

34. Liu YY, Wang Y, Walsh TR, Yi LX, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X, Yu LF, Gu D, Ren H, Chen X, Lv L, He D, Zhou H, Liang Z, Liu JH, Shen J; Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *The Lancet Infectious Diseases* 2016;**16**(2):161-8

35. Friend R, Staton T; Novartis Bexsero® meningitis B vaccine receives clinical recommendation for use in infants and adolescents in Australia. 2014

36. Rappuoli R; Bridging the knowledge gaps in vaccine design. *Nature biotechnology* 2007;**25**(12):1361

37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ; Basic local alignment search tool. *Journal of Molecular Biology* 1990;**215**(3):403-10

38. Molero-Abraham M, Lafuente EM, Flower DR, Reche PA; Selection of conserved epitopes from hepatitis C virus for pan-populational stimulation of T-cell responses. *Clinical and Developmental Immunology* 2013;**2013**

39. Sheikh QM, Gatherer D, Reche PA, Flower DR; Towards the knowledge-based design of universal influenza epitope ensemble vaccines. *Bioinformatics* 2016:btw399

40. Reche PA, Reinherz EL; PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands. *Nucleic Acids Research* 2005;**33**(suppl 2):W138-W142

41. Sakib MS, Islam MR, Hasan A, Nabi A; Prediction of epitope-based peptides for the utility of vaccine development from fusion and glycoprotein of nipah virus using in silico approach. *Advances in bioinformatics* 2014;**2014**

42. Kulp DW, Schief WR; Advances in structure-based vaccine design. *Current opinion in virology* 2013;**3**(3):322-331

43. Xuan C, Shi Y, Qi J, Zhang W, Xiao H, Gao GF; Structural vaccinology: structure-based design of influenza A virus hemagglutinin subtype-specific subunit vaccines. *Protein & cell* 2011;**2**(12):997-1005

44. Gourlay LJ, Peri C, Ferrer-Navarro M, Conchillo-Solé O, Gori A, Rinchai D, Thomas RJ, Champion OL, Michell SL, Kewcharoenwong C; Exploiting the Burkholderia pseudomallei acute phase antigen BPSL2765 for structure-based epitope discovery/design in structural vaccinology. *Chemistry & biology* 2013;**20**(9):1147-1156

45. He Y, Xiang Z, Mobley HL; Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *Journal of Biomedicine & Biotechnology* 2010

46. Jaiswal V, Chanumolu SK, Gupta A, Chauhan RS, Rout C; Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinformatics* 2013;**14**:211

47. Vivona S, Bernante F, Filippini F; NERVE: New Enhanced Reverse Vaccinology Environment. *BMC Biotechnology* 2006;**6**(1):35

48. Bowman BN, McAdam PR, Vivona S, Zhang JX, Luong T, Belew RK, Sahota H, Guiney D, Valafar F, Fierer J, Woelk CH; Improving reverse vaccinology with a machine learning approach. *Vaccine* 2011;**29**(45):8156-64

49. Delany I, Rappuoli R, Seib KL; Vaccines, reverse vaccinology, and bacterial pathogenesis. *Cold Spring Harbour Perspectives in Medicine* 2013;**3**(5)

50. Talukdar S, Zutshi S, Prashanth KS, Saikia KK, Kumar P; Identification of potential vaccine candidates against Streptococcus pneumoniae by reverse vaccinology approach. *Applied Biochemistry and Biotechnology* 2014;**172**(6):3026-41

51. Esposito S, Castellazzi L, Bosco A, Musio A, Stoddard J; Use of a multicomponent, recombinant, meningococcal serogroup B vaccine (4CMenB) for bacterial meningitis prevention. *Immunotherapy* 2014;**6**(4):395-408

52. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broeker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R; Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 2000;**287**(5459):1816-20

53. Agier L, Martiny N, Thiongane O, Mueller JE, Paireau J, Watkins ER, Irving TJ, Koutangni T, Broutin H; Towards understanding the epidemiology of Neisseria meningitidis in the African meningitis belt: a multi-disciplinary overview. *International Journal of Infectious Diseases* 2017;**54**:103-112

54. Moraes Cd, Moraes JCd, Silva GDMd, Duarte EC; Evaluation of the impact of serogroup C meningococcal disease vaccination program in Brazil and its regions: a population-based study, 2001-2013. *Memórias do Instituto Oswaldo Cruz* 2017;**112**(4):237-246

55. Shea MW; The long road to an effective vaccine for meningococcus group B (MenB). *Annals of Medicine and Surgery* 2013;**2**(2):53-56

56. Giuliani MM, Adu-Bobie J, Comanducci M, Arico B, Savino S, Santini L, Brunelli B, Bambini S, Biolchi A, Capecchi B, Cartocci E, Ciucchi L, Di Marcello F, Ferlicca F,

Galli B, Luzzi E, Masignani V, Serruto D, Veggi D, Contorni M, Morandi M, Bartalesi A, Cinotti V, Mannucci D, Titta F, Ovidi E, Welsch JA, Granoff D, Rappuoli R, Pizza M; A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences of the United States of America* 2006;**103**(29):10834-9

57. Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ; PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010;**26**(13):1608-1615

58. Corpet F, Servant F, Gouzy J, Kahn D; ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research* 2000;**28**(1):267-269

59. Henikoff S, Henikoff JG, Pietrokovski S; Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 1999;**15**(6):471-479

60. Gorringe AR, Pajon R; Bexsero: a multicomponent vaccine for prevention of meningococcal disease. *Human Vaccines & Immunotherapeutics* 2012;**8**(2):174-183

61. Novartis Vaccines and Diagnostics I; Product Monograph Bexsero Multicomponent Menigococcal B Vaccine (recombinant, absorbed) Bexsero Suspension for injection Active Immunizing Agent for the PRevention of Meningococcal Disease. ATC Code: J07AH09 - https://ca.gsk.com/media/1212390/bexsero.pdf. In: Novartis Vaccines and Diagnostics Is (ed), 2013.

62. Tani C, Stella M, Donnarumma D, Biagini M, Parente P, Vadi A, Magagnoli C, Costantino P, Rigat F, Norais N; Quantification by LC-MS(E) of outer membrane vesicle proteins of the Bexsero(R) vaccine. *Vaccine* 2014;**32**(11):1273-9

63. Bai X, Findlow J, Borrow R; Recombinant protein meningococcal serogroup B vaccine combined with outer membrane vesicles. *Expert Opinion on Biological Therapy* 2011;**11**(7):969-85

64. (NHS) NHS. *Meningitis B Vaccine*. http://www.nhs.uk/Conditions/vaccinations/Pages/meningitis-B-vaccine.aspx (04/10/2014 'Date Accessed')

65. Wise J; Meningitis B vaccine to be introduced in UK after U turn on its cost effectiveness. *BMJ* 2014;**348**

66. Tusnady GE, Simon I; The HMMTOP transmembrane topology prediction server. *Bioinformatics* 2001;**17**(9):849-50

67. Sachdeva G, Kumar K, Jain P, Ramachandran S; SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* 2005;**21**(4):483-91

68. Vivona S, Bernante F, Filippini F; NERVE: new enhanced reverse vaccinology environment. *BMC Biotechnology* 2006;**6**:35

69. Tusnady GE, Simon I; The HMMTOP transmembrane topology prediction server. *Bioinformatics* 2001;**17**(9):849-850

70. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL; The Pfam protein families database. *Nucleic Acids Research* 2004;**32**(suppl 1):D138-D141

71. Zeileis A, Leitner C, Hornik K; Home victory for Brazil in the 2014 FIFA World Cup -https://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2014-17.pdf. Working Papers in Economics and Statistics, 2014.

72. Fatima M, Pasha M; Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications* 2017;**9**(01):1

73. Tarca AL, Carey VJ, Chen X-w, Romero R, Draghici S; Machine learning and its applications to biology. *PLoS Computational Biology* 2007;**3**(6):e116

74. Liaw A, Wiener M; Classification and regression by randomForest. *R news* 2002;**2**(3):18-22

75. Danso S, Atwell E, Johnson O; A comparative study of machine learning methods for verbal autopsy text classification. *International Journal of Computer Science Issues* 2013;**10**(6)

76. Witten IH, Frank E. *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.'I.S.B.N' 008047702X

77. Smola AJ, Schölkopf B; A tutorial on support vector regression. *Statistics and Computing* 2004;**14**(3):199-222

78. Noble WS; What is a support vector machine? *Nature Biotechnology* 2006;**24**(12):1565-1567

79. Hsu C-W, Chang C-C, Lin C-J; A practical guide to support vector classification - http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. 2003

80. Guyon I, Elisseeff A; An introduction to variable and feature selection. *The Journal of Machine Learning Research* 2003;**3**:1157-1182

81. Anguita D, Ghio A, Ridella S, Sterpi D; K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines. *DMIN*, 2009, 291-297.

82. Mathworks T; Guide, MATLAB User's. http://www.mathworks.com 2017;**March 2017 Twenty eighth printing Revised for Version 9.2 (R2017a)**

83. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V; Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 2011;**12**:2825-2830

84. Ihaka R, Gentleman R; R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996;**5**(3):299-314

85. Chang C-C, Lin C-J; LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2011;**2**(3):27

86. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH; The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 2009;**11**(1):10-18

87. Doytchinova IA, Flower DR; VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007;**8:4**

88. Chen Y-W, Lin C-J; Combining SVMs with various feature selection strategies. *Feature extraction*: Springer, 2006, 315-324.

89. Polat K, Güneş S; A new feature selection method on classification of medical datasets: Kernel F-score feature selection. *Expert Systems with Applications* 2009;**36**(7):10367-10373

90. Dye C; Making wider use of the world's most widely used vaccine: Bacille Calmette-Guerin revaccination reconsidered. *Journal of the Royal Society Interface* 2013;**10**(87):20130365

91. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MP; The immune response in tuberculosis. *Annual Review of Immunology* 2013;**31**:475-527

92. Mangtani P, Abubakar I, Ariti C, Beynon R, Pimpin L, Fine PE, Rodrigues LC, Smith PG, Lipman M, Whiting PF; Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. *Clinical Infectious Diseases* 2014;**58**(4):470-480

93. Betts G, Poyntz H, Stylianou E, Reyes-Sandoval A, Cottingham M, Hill A, McShane H; Optimising immunogenicity with viral vectors: mixing MVA and HAdV-5 expressing the mycobacterial antigen Ag85A in a single injection. *PLoS One* 2012;**7**(12):e50447

94. Ellner JJ, Hirsch CS, Whalen CC; Correlates of protective immunity to Mycobacterium tuberculosis in humans. *Clinical Infectious Diseases* 2000;**30**:279-282

95. Mukhopadhyay S, Balaji KN; The PE and PPE proteins of Mycobacterium tuberculosis. *Tuberculosis* 2011;**91**(5):441-7

96. Kall L, Krogh A, Sonnhammer EL; A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* 2004;**338**(5):1027-36

97. Koul A, Vranckx L, Dhar N, Gohlmann HW, Ozdemir E, Neefs JM, Schulz M, Lu P, Mortz E, McKinney JD, Andries K, Bald D; Delayed bactericidal response of Mycobacterium tuberculosis to bedaquiline involves remodelling of bacterial metabolism. *Nature Communications* 2014;**5**:3369

98. Fu LM, Shinnick TM; Genome-wide exploration of the drug action of capreomycin on Mycobacterium tuberculosis using Affymetrix oligonucleotide GeneChips. *The Journal of Infection* 2007;**54**(3):277-84

99. Venkataraman B, Vasudevan M, Gupta A; A new microarray platform for whole-genome expression profiling of Mycobacterium tuberculosis. *Journal of Microbiological Methods* 2014;**97**:34-43

100. Arnvig KB, Comas I, Thomson NR, Houghton J, Boshoff HI, Croucher NJ, Rose G, Perkins TT, Parkhill J, Dougan G, Young DB; Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathogens* 2011;**7**(11):e1002342

101. Swiss Institute of Bioinformatics ÉPFDL. *Tuberculist*. http://tuberculist.epfl.ch/ (26/08/2015 'Date Accessed')

102. Hoang T, Aagaard C, Dietrich J, Cassidy JP, Dolganov G, Schoolnik GK, Lundberg CV, Agger EM, Andersen P; ESAT-6 (EsxA) and TB10. 4 (EsxH) based vaccines for pre-and post-exposure tuberculosis vaccination. *PLoS One* 2013;**8**(12):e80579

103. Ruxton GD; The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology* 2006;**17**(4):688-690

104. Bhatt K, Verma S, Ellner JJ, Salgame P; Quest for correlates of protection against tuberculosis. *Clinical and Vaccine Immunology* 2015;**22**(3):258-66

105. Heinson AI, Gunawardana Y, Moesker B, Hume CC, Vataga E, Hall Y, Stylianou E, McShane H, Williams A, Niranjan M, Woelk CH; Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology. *International Journal of Molecular Sciences* 2017;**18**(2)

106. Tanghe A, Lefèvre P, Denis O, D'Souza S, Braibant M, Lozes E, Singh M, Montgomery D, Huygen K; Immunogenicity and protective efficacy of tuberculosis DNA vaccines encoding putative phosphate transport receptors. *The Journal of Immunology* 1999;**162**(2):1113-1119

107. Xue Y, Liu Z, Gao X, Jin C, Wen L, Yao X, Ren J; GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. *PLoS One* 2010;**5**(6):e11290

108. Simon R, Radmacher MD, Dobbin K, McShane LM; Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003;**95**(1):14-18

109. Fawcett T; An introduction to ROC analysis. *Pattern Recognition Letters* 2006;**27**(8):861-874

110. DeLong ER, DeLong DM, Clarke-Pearson DL; Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**(3):837-45

111. Coombes K; ClassDiscovery: Classes and methods for "class discovery" with microarrays or proteomics. *R package*; **version 2.1 (2009)**

112. Galili T; dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* 2015;**doi: 10.1093/bioinformatics/btv428**

113. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S; NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* 2007;**2, e796**

114. Cai R, Liu Z, Ren J, Ma C, Gao T, Zhou Y, Yang Q, Xue Y; GPS-MBA: computational analysis of MHC class II epitopes in type 1 diabetes. *PloS one* 2012;**7**(3):e33884

115. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C; A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC structural biology* 2009;**9**(1)

116. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A; Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science : a Publication of the Protein Society* 2003;**12**(8):1652-62

117. Nielsen M, Lundegaard C, Lund O, Kesmir C; The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 2005;**57**(1-2):33-41

118. Larsen JE, Lund O, Nielsen M; Improved method for predicting linear B-cell epitopes. *Immunome Research* 2006;**2:2**

119. Liu Z, Cao J, Gao X, Ma Q, Ren J, Xue Y; GPS-CCD: a novel computational program for the prediction of calpain cleavage sites. *PLoS One* 2011;**6**(4):e19001

120. Norata GD, Pirillo A, Ammirati E, Catapano AL; Emerging role of high density lipoproteins as a player in the immune system. *Atherosclerosis* 2012;**220**(1):11-21

121. Norata GD, Catapano AL; HDL and adaptive immunity: a tale of lipid rafts. *Atherosclerosis* 2012;**225**(1):34-35

122. Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA; Glycosylation and the immune system. *Science* 2001;**291**(5512):2370-2376

123. Liu S, Cai X, Wu J, Cong Q, Chen X, Li T, Du F, Ren J, Wu YT, Grishin NV, Chen ZJ; Phosphorylation of innate immune adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science* 2015;**347**(6227)

124. Snapper CM, Rosas FR, Jin L, Wortham C, Kehry MR, Mond JJ; Bacterial lipoproteins may substitute for cytokines in the humoral immune response to T cell-independent type II antigens. *The Journal of Immunology* 1995;**155**(12):5582-5589

125. Kovacs-Simon A, Titball R, Michell SL; Lipoproteins of bacterial pathogens. *Infection and immunity* 2011;**79**(2):548-561

126. Li D, von Schaewen M, Wang X, Tao W, Zhang Y, Li L, Heller B, Hrebikova G, Deng Q, Ploss A; Altered Glycosylation Patterns Increase Immunogenicity of a Subunit Hepatitis C Virus Vaccine, Inducing Neutralizing Antibodies Which Confer Protection in Mice. *Journal of Virology* 2016;**90**(23):10486-10498

127. Haste Andersen P, Nielsen M, Lund O; Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein science : a Publication of the Protein Society* 2006;**15**(11):2558-67

128. Ansari HR, Raghava GP; Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Research* 2010;**Volume 6**

129. Borrow R, Balmer P, Miller E; Meningococcal surrogates of protection—serum bactericidal antibody activity. *Vaccine* 2005;**23**(17):2222-2227

130. McIntosh ED, Broker M, Wassil J, Welsch JA, Borrow R; Serum bactericidal antibody assays - The role of complement in infection and immunity. *Vaccine* 2015;**33**(36):4414-21

131. Pollard A; Correlates of protection against Neisseria Meningitidis. *Nature Reviews Immunology* 2016

132. Manjelievskaia J, Erck D, Piracha S, Schrager L; Drug-resistant TB: deadly, costly and in need of a vaccine. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2016;**110**(3):186-91

133. Myron Christodoulides JH; Novel Approaches to *Neisseria meningitidis* vaccine design. *Personal Communication (Pre Submission)* 2017

134. Gill CJ, Ram S, Welsch JA, Detora L, Anemona A; Correlation between serum bactericidal activity against Neisseria meningitidis serogroups A, C, W-135 and Y measured using human versus rabbit serum as the complement source. *Vaccine* 2011;**30**(1):29-34

135. Team RC; R: A language and environment for statistical computing. 2013

136. Vouga M, Greub G; Emerging bacterial pathogens: the past and beyond. *Clinical Microbiology and Infection* 2016;**22**(1):12-21

137. Altman LK. *In Philadelphia 30 Years Ago, an Eruption of Illness and Fear.* http://www.nytimes.com/2006/08/01/health/01docs.html?pagewanted=all&_r=0 (20/03/2017 'Date Accessed')

138. Brehony C, Rodrigues CM, Borrow R, Smith A, Cunney R, Moxon ER, Maiden MC; Distribution of Bexsero® Antigen Sequence Types (BASTs) in invasive meningococcal disease isolates: Implications for immunisation. *Vaccine* 2016;**34**(39):4690-4697

Appendix A: A list of the 19 protein annotation tools used to generate 122 annotation features for bacterial protective antigens (BPAs) and non-BPAs in the training dataset of Bowman et al "Improving Reverse Vaccinology with a Machine Learning Approach". This table was included in the supplemental information of the reverse vaccinology paper published Bowman et al.

Bowman BN, McAdam PR, Vivona S, Zhang JX, Luong T, Belew RK, Sahota H, Guiney D, Valafar F, Fierer J, Woelk CH; Improving reverse vaccinology with a machine learning approach. *Vaccine* 2011;**29**(45):8156-64

Program	Version	Location	Annotation	AF No.	AF Name	Description	URL	Ref.
NA	NA	NA	NA	1	Length	Protein length	NA	NA
DictyOGlyc	1.1	Web	Glycosylation	2	MaxScore	Maximum score of the predicted glycosylation sites	http://www.cbs.dtu.dk/services/DictyOGlyc/	[1]
				3	AvgScore	Average score of all predicted glycosylation sites		
				4	MaxDiff	Maximum difference between the site-score and threshold		
				5	AvgDiff	Average difference between the site-score and threshold		
				6	Count	Number of predicted glycosylation sites		
				7	CorrCount	Number of predicted glycosylation sites, normalized for protein length		
NetAcet	1.0	Web	Acetylation	8	A-Flag	1 if the acetylated residue is Alanine, otherwise 0	http://www.cbs.dtu.dk/services/NetAcet/	[2]
				9	G-Flag	1 if the acetylated residue is Glycine, otherwise 0		
				10	S-Flag	1 if the acetylated residue is Serine, otherwise 0		
				11	T-Flag	1 if the acetylated residue is Threonine, otherwise 0		
				12	Score	Score for the potential acetylation site		
				13	SiteFlag	1 if there is a potentially acetylated residue, otherwise 0		
				14	AcetFlag	1 if the residue is predicted to be acetylated, otherwise 0		
NetGlycate	1.0	Web	Glycation	15	MaxScore	Maximum score of predicted glycation sites	http://www.cbs.dtu.dk/services/NetGlycate/	[3]
				16	AvgScore	Average score of all predicted glycation sites		

				17	Count	Number of predicted glycation sites		
				18	CorrCount	Number of predicted glycation sites, normalized for protein length		
NetPhosBac	1.0	Web	Phosphorylation	19	MaxScore	Maximum score of predicted phosphorylation sites	http://www.cbs.dtu.dk/services/NetPhosBac-1.0/	[4]
				20	AvgScore	Average score of all predicted phosphorylation sites		
				21	Count	Number of predicted phosphorylation sites		
				22	CorrCount	Number of predicted phosphorylation sites, normalized for protein length		
NetPhosYeast	1.0	Web	Phosphorylation	23	MaxScore	Maximum score of predicted phosphorylation sites	http://www.cbs.dtu.dk/services/NetPhosYeast/	[5]
				24	AvgScore	Average score of all predicted phosphorylation sites		
				25	Count	Number of predicted phosphorylation sites		
				26	CorrCount	Number of predicted phosphorylation sites, normalized for protein length		
ProtParam	Biopython 1.53	Local	Basic Protein Stats	27	Isoelectric	Predicted isoelectric point	http://ca.expasy.org/tools/protparam.html	[6]
				28	Instability	Predicted instability index		
				29	MolecWeight	Predicted Molecular Weight		
				30	Aromaticity	Predicted protein aromaticity		
				31	GRAVY	Predicted grand average of hydropathy		
				32	PercTurn	Percent of protein predicted to be a part of a turn		
				33	PercHelix	Percent of protein predicted to be a part of an alpha-helix		
				34	PercSheet	Percent of protein predicted to be a part of a beta-sheet		
				35	PercAlanine	Percentage of residues that are alanine		
				36	PercCysteine	Percentage of residues that are cysteine		
				37	PercAsparticAcid	Percentage of residues that are aspartic acid		
				38	PercGlutamicAcid	Percentage of residues that are glutamic acid		

				39	PercPhenylalanine	Percentage of residues that are phenylalanine		
				40	PercGlycine	Percentage of residues that are glycine		
				41	PercHistidine	Percentage of residues that are histidine		
				42	PercIsoleucine	Percentage of residues that are isoleucine		
				43	PercLysine	Percentage of residues that are lysine		
				44	PercLeucine	Percentage of residues that are leucine		
				45	PercMethionine	Percentage of residues that are methionine		
				46	PercAsparagine	Percentage of residues that are asparagine		
				47	PercProline	Percentage of residues that are proline		
				48	PercGlutamine	Percentage of residues that are glutamine		
				49	PercArginine	Percentage of residues that are arginine		
				50	PercSerine	Percentage of residues that are serine		
				51	PercThreonine	Percentage of residues that are threonine		
				52	PercValine	Percentage of residues that are valine		
				53	PercTryptophan	Percentage of residues that are tryptophan		
				54	PercTyrosine	Percentage of residues that are tyrosine		
NetPhosK	3.1	Local	Phosphorylation	55	MaxScore	Maximum score of predicted phosphorylation sites	http://www.cbs.dtu.dk/services/NetPhosK/	[7]
				56	AvgScore	Average score of all predicted phosphorylation sites		
				57	Count	Number of predicted phosphorylation sites		
				58	CorrCount	Number of predicted phosphorylation sites, normalized for protein length		

YinOYang	1.2	Local	O-Linked Beta-N-acetylglucosamine	59	MaxRank	Maximum rank of predicted glc-n-ac sites	http://www.cbs.dtu.dk/services/YinOYang/	[8]
				60	AvgRank	Average rank of predicted glc-n-ac sites		
				61	MaxScore	Maximum score of predicted glc-n-ac sites		
				62	AvgScore	Average score of predicted glc-n-ac sites		
				63	MaxDiff1	Maximum difference between the prediction scores and the lower threshold		
				64	AvgDiff1	Average difference between the prediction scores and the lower threshold		
				65	MaxDiff2	Maximum difference between the prediction scores and the higher threshold		
				66	AvgDiff2	Average difference between the prediction scores and the higher threshold		
				67	Count	Number of predicted glc-n-ac sites		
				68	CorrCount	Number of predicted glc-n-ac sites, normalized for protein length		
LipoP	1.0a	Local	Lipoproteins & Signal Peptides	69	PeptidaseI	Predicted non-lipoprotein signal peptide cleavage site	http://www.cbs.dtu.dk/services/LipoP/	[9]
				70	PeptidaseII	Predicted lipoprotein signal peptide cleavage site		
				71	Transmembrane	Predicted n-terminal transmembrane helix		
				72	Cytoplasmic	No predicted signal protein - predicted localization to the cytoplasm		
				73	Score	Score associated with the above prediction		
TargetP	1.1	Local	Subcellular Localization	74	SecretScore	Score for secretory pathway signal peptide	http://www.cbs.dtu.dk/services/TargetP/	[10]
				75	MitoScore	Score for the mitochondrial targeting peptide		
				76	OtherScore	Score for non-secretory, non-mitochondrial localization		

				77	SecretFlag	1 if the program predicts that the protein is secreted, otherwise 0		
				78	MitoFlag	1 if the program predicts that the protein is localized to the mitochondria, otherwise 0		
				79	OtherFlag	1 if the program predicts that the protein is neither secreted nor mitochondrial, otherwise 0		
NetNGlyc	1.0a	Local	N-Glycosylation	80	MaxScore	Maximum score of predicted N-glycosylation sites	http://www.cbs.dtu.dk/services/NetNGlyc/	[11]
				81	AvgScore	Average score of predicted N-glycosylation sites		
				82	MaxRank	Maximum rank of predicted N-glycosylation sites		
				83	AvgRank	Average rank of predicted N-glycosylation sites		
				84	Count	Number of predicted N-glycosylation sites		
NetOGlyc	3.1d	Local	O-Glycosylation	85	MaxGforT	Maximum generalized score for all predicted threonine O-glycosylation sites	http://www.cbs.dtu.dk/services/NetOGlyc/	[12]
				86	MaxIforT	Maximum isolated score for all predicted threonine O-glycosylation sites		
				87	AvgGforT	Average generalized score for all predicted threonine O-glycosylation sites		
				88	AvgIforT	Averages isolated score for all predicted threonine O-glycosylation sites		
				89	HitsForT	Number of predicted threonine O-glycosylation sites		
				90	CountForT	Number of possible threonine O-glycosylation sites		
				91	MaxGforS	Maximum generalized score for all predicted serine O-glycosylation sites		
				92	MaxIforS	Maximum isolated score for all predicted serine O-glycosylation sites		
				93	AvgGforS	Average generalized score for all predicted serine O-glycosylation sites		
				94	AvgIforS	Averages isolated score for all		

						predicted serine O-glycosylation sites		
				95	HitsForS	Number of predicted serine O-glycosylation sites		
				96	CountForS	Number of possible serine O-glycosylation sites		
ProP	1.0c	Local	Cleavage Sites	97	Furin-MaxScore	Maximum score of predicted cleavage sites (furin mode)	http://www.cbs.dtu.dk/services/ProP/	[13]
				98	Furin-AvgScore	Average score of predicted cleavage sites (furin mode)		
				99	Furin-Count	Number of predicted cleavage sites (furin mode)		
				100	General-MaxScore	Maximum score of predicted cleavage sites (general mode)		
				101	General-AvgScore	Average score of predicted cleavage sites (general mode)		
				102	General-Count	Number of predicted cleavage sites (general mode)		
BepiPred	1.0b	Local	Linear B-cell epitopes	103	MaxScore	Maximum score of predicted linear epitopes	http://www.cbs.dtu.dk/services/BepiPred/	[14]
				104	AvgScore	Average score of predicted linear epitopes		
				105	PercEpitope	Percent of the protein predicted to be a part of a linear epitope		
TMHMM	2.0c	Local	Transmembrane Helices	106	TransAAcount	Number of amino acids predicted to be a part of trans-membrane helices	http://www.cbs.dtu.dk/services/TMHMM/	[15]
				107	StartAAcount	Number of amino acids, from the first 60, predicted to be a part of trans-membrane helices		
				108	Count	Number of predicted trans-membrane helices		
HMMTOP	2.1	Local	Transmembrane Helices	109	Count	Number of predicted trans-membrane helices	http://www.enzim.hu/hmmtop/	[16]
PSORTb	3.0	Local	Subcellular Localization	110	ProbCytoMem	Predicted probability that the protein is localized to the cytoplasmic membrane	http://www.psort.org/psortb/	[17]
				111	ProbCytoplasm	Predicted probability that the protein is localized to the cytoplasm		
				112	ProbPeriplasm	Predicted probability that the protein is localized to the periplasm		
				113	ProbExtracell	Predicted probability that the protein is exported from the cell		

				114	ProbOuterMem	Predicted probability that the protein is localized to the outer membrane		
				115	ProbCellWall	Predicted probability that the protein is localized to the cell wall		
SignalP	3.0	Local	Signal Peptides	116	NN-MaxC	Neural network predicted maximum cleavage site score	http://www.cbs.dtu.dk/services/SignalP/	[18]
				117	NN-MaxY	Combined neural network predicted scores for cleavage and signal peptide probability		
				118	NN-MaxS	Neural network predicted maximum signal peptide site score		
				119	NN-AvgS	Average of all signal peptide scores n-terminal to the predicted cleavage site		
				120	D-score	Average value of AvgS and MaxY, known to provide better discrimination than MaxY alone		
				121	HMM-MaxC	Alternate HMM-based prediction of the likelihood of protein cleavage		
				122	HMM-ProbS	HMM-based predicted probability of whether the protein has a signal peptide or not		

[1]In many cases several different annotation features could be derived from a single protein annotation tool and these are indicated in the table. Abbreviations: *AF No.*, refers to a number assigned to each of the 122 annotation features derived from the 19 protein annotation tools; *AF name*, refers to the label given to the annotation feature; *Ref.*, refers to the literature reference for the protein annotation tool listed below; *NA*, not applicable. Shading is for display purposes only.

References for Table:

[1] Gupta R, Jung E, Gooley AA, Williams KL, Brunak S, Hansen J. Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks. Glycobiology 1999;9:1009-22.

[2] Kiemer L, Bendtsen JD, Blom N. NetAcet: prediction of N-terminal acetylation sites. Bioinformatics 2005;21:1269 -70.

[3] Johansen MB, Kiemer L, Brunak S. Analysis and prediction of mammalian protein glycation. Glycobiology 2006;16:844 -53.

[4] Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. NetPhosBac – A predictor for Ser/Thr phosphorylation sites in bacterial proteins. PROTEOMICS 2009;9:116-25.

[5] Ingrell CR, Miller ML, Jensen ON, Blom N. NetPhosYeast: prediction of protein phosphorylation sites in yeast. Bioinformatics 2007;23:895 -7.

[6] Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 1999;112:531-52.

[7] Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. PROTEOMICS 2004;4:1633-49.

[8] Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput, 2002.
[9] Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 2003;12:1652-62.
[10] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005-16.
[11] Gupta R, Jung E, Brunak S. Prediction of N-glycosylation sites in human proteins. 2004.
[12] Julenius K, Molgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 2005;15:153-64.
[13] Duckert P, Brunak S, Blom N. Prediction of proprotein convertase cleavage sites. Protein Engineering Design and Selection 2004;17:107 -12.
[14] Larsen J, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Research 2006;2:2.
[15] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567-80.
[16] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849-50.
[17] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 2010;26:1608 -15.
[18] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved Prediction of Signal Peptides: SignalP 3.0. Journal of Molecular Biology 2004;340:783-95.

Title	Generation of TB latency vaccine candidates				
Protocol Ref.	DGM015	**Version**	01	**Date**	06/04/11
Associated protocols, Risk and GM Assessments					
Author(s)	Robert Watson				

Primer Design

View the gene sequence from Tuberculist - http://tuberculist.epfl.ch/

Copy the gene sequence into EditSeq. Copy the gene name into the information box below the sequence and save the file as the gene name

Remove the first 3 base pairs and the last 3 base pairs from the DNA sequence (start and stop codons)

The forward attB sequence = 5'GGGGACAAGTTTGTACAAAAAAGCAGGCT 3'. Add the attB forward sequence AS IT APPEARS IN THE FILEto the start of the gene ..\..\EXP RESULTS\Latency vaccines\attB Latency gene sequences\attB forward sequence to add to gene.seq

The reverse attB sequence = 5' GGGGACCACTTTGTACAAGAAAGCTGGGT 3' Add the attB reverse sequence AS IT APPEARS IN THE FILE to the end of the gene ..\..\EXP RESULTS\Latency vaccines\attB Latency gene sequences\attB reverse sequence to add to gene.seq

Save this file as gene name with attB. Also add this information to the information panel below the sequence.

Highlight the attB sequence at the start of the gene plus approx 18bp = 47 bp total. Copy this sequence into the database below.

For the reverse sequence the process is slightly more complicated. Highlight the attB sequence at the end of the gene plus approx 18bp = 47 bp total. Select Goodies from the EditSeq menu. Then choose ' Reverse Complement'. This will output the sequence in the correct orientation for ordering as a primer. Copy this sequence into the database below.

All designed primers are to be entered into the TB Latency vaccines database.xls document (\\Camr_dc2\new shares\research\Public Sector\Project 196\Experiment results\Planners and Databases\TB Latency vaccines database.xls)

The attB sequence will not anneal to the TB genome, but will be incorporated into the sequence during extension. These attB sites allow simple insertion of the insert into a

pDONR Zeo plasmid, using the gateway transfer system. For more information see www.invitrogen.com

Primers resuspended in Sigma water (W-4502) to achieve a 100µM stock solution. The volume to add will be provided with the supporting information for your primers. From this prepare a 5 µM working stock set for use in the PCR reactions as follows; add 10 µl of primer 1 to a tube. Add 10 µl of primer 2 to this tube. Add 180 µl of Sigma water (W-4502) to give a 5 µM primer set.

PCR conditions

Reaction:
1µl of 100 ng/ul H37Rv genomic DNA
0.4µl of AccuPrime Pfx (1U)
5µl of AccuPrim Pfx Buffer
6µl of attB 5 µM Forward+Reverse primer set
36.6 µl DNA/RNase free water

Thermocycler settings:

1. 95°C 5 minutes
2. 95°C 30 seconds
3. 62°C 30 seconds
4. 68°C 1 or 2 or 3+ minutes (1 minute per KB)
5. Repeat steps 2-4 30 times
6. 68°C 7 minutes
7. 4°C hold

Run 8µl of PCR product on 1% TAE agarose gel to check size and purity.

Purify successful reactions before use in BP Gateway reactions, by either gel extraction if products appear dirty (QIAgen cat# 28706) or PCR clean up if products appear clean (QIAgen cat# 28106).

Quantify the product after purification using the Nanodrop.

Title	One tube Gateway cloning				
Protocol Ref.	DGM014	**Version**	02	**Date**	28/03/11
Associated protocols, Risk and GM Assessments					
Author(s)	Robert Watson				

One Tube Format

To transfer an attB-flanked PCR product directly into an expression clone, you can easily combine the BP and LR reactions using the following protocol. This will potentially eliminate the transformation and DNA isolation of the Gateway® entry clone.

1. In a 1.5 ml microcentrifuge tube, prepare the following 15 µl BP reaction:

Purified attB PCR product (100 ng) 1.0-5.0 µl
pDONR Zeo™ vector (150 ng/µl) 1.3 µl
BP Clonase™ II enzyme mix 3.0 µl
TE Buffer, pH 8.0 add to a final volume of 15 µl

2. Mix well by vortexing briefly and incubate at 25°C overnight. Remove 10 µl of the reaction to a separate tube and use this aliquot to set up the LR reaction. Freeze the remaining BP reaction at -20°C in the box labelled Latency vaccines BP reactions
3. To the new tube containing the 10 µl BP reaction, add:

pVAX1 Gateway (150 ng/µl) 2.0 µl (in red top tube -20°C)
LR Clonase™ II enzyme mix 3.0 µl
Final volume 15 µl

4. Mix well by vortexing briefly and incubate at 25°C for 4 hours.
5. Add 2 µl of proteinase K solution. Incubate at 37°C for 10 minutes.
6. Add 1.5 µl of the LR reaction into OneShot Top10 E. coli. Incubate on ice for 5 minutes
7. Transform the cells by heat shock 42°C for 30 seconds.
8. Return the cells immediately to ice for 2 minutes
9. Add 250 µl SOC medium
10. Incubate for 1 hour at 37°C 220rpm
11. Plate 10 µl, 50 µl and 100 µl on LB plates containing Kanamycin
12. Incubate overnight at 37°C
13. Freeze remaining LR reaction at -20°C in the box labelled Latency vaccines LR reactions

Title	BHK-21 Transfection Protocol for the Evaluation of pVAX DNA Vaccines				
Protocol Ref.	DGM18	**Version**	01	**Date**	21 Jul 11
Associated protocols, Risk and GM Assessments			RA4999, RA5014, RD/0610-416 & amendments.		
Author(s)	S Daminabo				

- Seed a 24 well plate BHK-21, a Syrian Hamster kidney cell line (ATCC; cat# CCL-10). Incubate until near confluence, roughly 85-90%.

Reagent Preparation

- ***Ensure test DNA has been diluted to 0.08 mg/ml in Optimem.***

  32ul (vaccine dna stock) + 168ul (optimem) will give concentration of 0.08mg/ml

- Dilute DNA in OptiMEM (Invitrogen; 31985-062) and mix gently as follow in eppendorf:

  Per well:
  - 22.5 ul OptiMEM + 2.5 ul DNA = 0.2 µg in 25 µl.
  - 21.25 ul OptiMEM + 3.75 ul DNA = 0.3 µg in 25 µl.
  - 20 ul OptiMEM + 5 ul DNA = 0.4 µg in 25 µl.

For multiple wells, multiply the above.

- Dilute lipofectamine (Invitrogen; 18324-012) in OptiMEM as follows in eppendorf:

  Per well
  - 20 ul OptiMEM + 5 ul lipofectamine

- Mix the lipofectamine/Optimem and DNA/Optimem to form **complex**, pipette up and down slowly to mix, incubate for 15-45 minutes at room temperature (NB **complexes** are stable at RT for 6 hrs).
- Add 0.15 ml Optimem to the lipofectamine/DNA **complex**. Mix gently.

Transfection Stage

NB: Do not add antibiotics to the media during transfection. The cell wall is deliberately made permeable to allow uptake of plasmids into the cell. Adding antibiotics causes cellular toxicity as they accumulate in the cell in a higher than tolerable concentration, causing cell death.

- Remove growth media from cells
- Wash cells with PBS 1ml per well (2x)
- Add 0.2 ml of MEM + l-glut per well
- Add 0.2ml of complex and mix gently (by swirling or rocking plate)
- Incubate for 5 hours at 37°C in 5% CO_2 incubator.

- After 5 hours, remove media and replace with MEM, 2% FCS (Sigma; F9665) L-glut and 1x Penicillin/Streptomycin.
- Incubate plate at 37°C in 5% CO_2 for 2 days.

Harvesting and plating

- Harvest the cells after 2 days and then perform western blot

Title	SDS PAGE and Western Blotting				
Protocol Ref.	DGM07	**Version**	04	**Date**	14-Jun 2012
Associated protocols, Risk and GM Assessments			RA4999 RA5014		
Author(s)	Sofiri Daminabo				

Reagents and Materials

Item	Supplier	Code	Notes
NuPAGE MES SDS Running Buffer (20x)	Invitrogen	NP0002	500 ml bottles (add 1 to 10 L carbuoy and fill with dH20.)
NuPAGE 4-12% Bis-Tris gel 10 well	Invitrogen	NP0321BOX	
NuPAGE 4-12% Bis-Tris gel 15 well	Invitrogen	NP0323BOX	
SeeBlue Plus 2	Invitrogen	LC5952	
Magic Mark XP	Invitrogen	LC5602	
NuPAGE Transfer Buffer (20x)	Invitrogen	NP0006-1	
Methanol	Stores		
Filter paper	Whatman	FIL2081	
Hybond ECL nitrocellulose	GE healthcare	RPN203D	
PBS/Tween 20 0.01M (10x)	Cellgro	99-847-CM	
Skimmed milk powder	Becton Dickinson	DIFCO232100	
Mouse anti-V5	AbD Serotec	MCA1360	
Anti-mouse IgG HRP conjugated	Sigma	A9044	
ECL Plus / Pierce ECL	GE healthcare/ Thermo scientific	RPN2132/ 32109	
Developer	Invitrogrn	P7042-1GA	
Fixer	Invitrogen	P7167-1GA	

- **Transfer Buffer**
  - 200ml Methanol
  - 50ml NUpage Transfer buffer
  - 750ml distilled water – to make 1 litre, enough for one blot unit / tank

- **PBST Wash buffer**
  - PBS/Tween 10x dilution

- **Blocking buffer (5% Marvel solution)**
  - 5g non-fat milk powder
  - 95ml PBST wash buffer

SDS-PAGE

Denature sample at 70 degrees for 10 mins

Remove pre-cast gel from packaging. Remove comb and white sticker from the gel.
Set up in gel tank. Pour in running buffer ensuring that the central reservoir has formed a tight seal.
Prepare and run protein gel with protein ladder and Magicmark standards **(10ul**), 200 V, > 45 mins.
Select NuPage Gel
Add 20ul of sample to well of choice. 5ul of Magic mark and See blue to the same well..

Western Blot
Create Sponge – filter paper – gel – nitrocellulose membrane – filter paper – sponge sandwich, keeping everything moist with transfer buffer.
Place into transfer cassette so that current runs from gel to membrane (black to red).
Select NuPage Blot
Run at 40 V for > 60 min. (If Voltage does not reach 40V when transfer begins, go back and increase the current measured in Ampere)

Immuno-detection
Block in 3-5% milk in PBS/Tween (0.05%) for > 45 mins.
Incubate with primary antibody diluted in block for > 45 mins.>1:3000 – 8.4ul (Antibody) 24.9916ml (blocking buffer).

Wash 3 times, > 5 mins per wash, in PBS/Tween (0.05%).
Incubate with secondary antibody diluted in block for > 45 mins.> 1:3000 – 8.4ul (Antibody) 24.9916ml (blocking buffer).
Wash as above.

ECL detection
Mix ECL reagents A and B in a 40:1 ratio, e.g. 2ml solution A and 50µl solution B per membrane
Apply ECL as per manufacturer instructions.
Expose to film in dark room, place in developer, fixer and then water or use Bio-Rad scanner

Lysis Buffer Preparation (Based on volume required per plate)

NuPage sample reducing agent (10 x) (Invitrogen Cat: NP0009)

NuPage LDS sample buffer (4 x) (Invitrogen Cat: NP0007

Distilled water

Making Developer and Fixer

Fixer = 800 ml (Water) + 200 ml (Fixer replenisher)

Developer = 780 ml (Water) + 220 ml (Developer)

Appendix F: Annotated selected predicted bacterial protective antigens (i.e. Putative Vaccine candidates).

RV Number	Gene Symbol	Description	Predicted Antigen Score	Average Expression Score	Localization [1]	Localization Score [2]	Protein Length (aa)	Transmembrane domains	Google scholar	Erdman	BCG	Concern
Rv2068c	blaC	Class a beta-lactamase BlaC [Mycobacterium tuberculosis H37Rv]	0.996695	1.93	Extracellular	9.60	307	0	Approach for potential drug target identification in Mycobacterium tuberculosis (PMID: 19833135)	Yes	Yes	
Rv1677	dsbF	Probable conserved lipoprotein DsbF [Mycobacterium tuberculosis H37Rv]	0.975867	2.75	Extracellular	9.72	182	0	Structural and biochemical characterization of the essential DsbA-like disulfide bond forming protein from Mycobacterium tuberculosis (PMID: 24134223)	Yes	Yes	Patent titled; Compositions and methods for immunodominant antigens of Mycobacterium tuberculosis, I had previously discounted this one out but after the teleconference it may be fine. In the patent it is refered to in 3 sections - - The following sequences producing the antigens were determined to be sixth most preferential, characterized by ranks <25 by either CERNO or RF - The inventors have discovered various immunodominant antigens from M. tuberculosis that are suitable for diagnostic and therapeutic purposes. Particularly preferred immunodominant antigens and are those encoded by nucleic acids having a sequence according to SEQ ID NO:1 to SEQ ID NO:586, and it is generally contemplated that such antigens can be used as single antigens, or in combination (optionally also in combination with antigens from another pathogen) in the manufacture of various diagnostic devices, therapeutic compositions, and vaccines. Preferably, the immunodominant antigens suitable for diagnostic and therapeutic purposes are encoded by the sequences designated - For example, suitable diagnostic devices especially include those comprising one or more of the immunodominant antigens, fragments, or analogs thereof that are encoded by nucleic acids according to SEQ ID NO:1 to SEQ ID NO:586http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=Rv1677&OS=Rv1677&RS=Rv1677
Rv0680c	Rv0680c	Probable conserved transmembrane protein [Mycobacterium tuberculosis H37Rv]	0.932385	2.68	Unknown		124	2	Identification of proteins from tuberculin purified protein derivative (PPD) by LC-MS/MS (PMID: 19683472)	Yes	Yes	Transmembrane domains.
Rv1857	modA	Probable molybdate-binding lipoprotein ModA [Mycobacterium tuberculosis H37Rv]	0.939283	1.32	Unknown		261	0	Lipoproteins of Mycobacterium tuberculosis: an abundant and functionally diverse class of cell envelope components (PMID:	Yes	Yes	
Rv2190c	Rv2190c	Conserved hypothetical protein [Mycobacterium tuberculosis H37Rv]	0.980188	2.86	Extracellular	9.73	385	1	Rv2190c, an NlpC/P60 Family Protein, Is Required for Full Virulence of Mycobacterium tuberculosis (PMID: 22952680)	Yes	Yes	The Patent is titled; Immunogenic compositions against tuberculosis. This protein Rv number appears in table 11, which is described as; Replicate microarray hybridizations were performed for at least two biological samples of both the wild type and mutant strains and showed a high correlation level (r=0.9). Using a standard protocol for Bayesian statistics, significantly regulated genes with a probability of differential expression (PDE)>0.5 and >.+-.2 fold change between H37Rv and .DELTA.rv0348 mutant were identified. Using these criteria, a significant change in a set of 163 genes (Table 11) was identified between the transcriptomes for H37Rv and the .DELTA.rv0348 (rv0348-regulon). Listed in table 11 - http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=Rv2190c&OS=Rv2190c&RS=Rv2190c
Rv3886c	mycP2	Probable alanine and proline rich membrane-anchored mycosin MycP2 (serine protease) (subtilisin-like protease) (subtilase-like) (mycosin-2) [Mycobacterium tuberculosis H37Rv]	0.965291	1.96	Unknown		550	1	Mycobacterium Tuberculosis Proteases and Implications for New Antibiotics Against Tuberculosis (PMID: 22181704)	Yes	Yes	

Footer note - 1: Localization, the proteins were ran through a protein annotation tool called PSORTb and this was where it predicted the proteins to be localized within the cell. 2: score from the PSORTb protein annotation tool for the stated localizatio area. This score is a probability out of 10. 3: When searching for the Rv number on Google scholar, I have copied the title of the most relevant result on google scholar first page relating to vaccine/ immunology, if not relevant the top result on google scholar. 4: Why I have put this protein at this rank in the list.

Appendix G: Primers used to amplify sections of DNA to create putative DNA vaccines

Protein	Forward Primer	Reverse Primer
VC1 (Rv0680c)	GGGGACAAGTTTGTACAAAAAAGCAGGCTAAGTGGAACACCGTCGCC	GGGGACCACTTTGTACAAGAAAGCTGGGTGCCGACCCGCTGCAGTC
VC2 (Rv1677)	GGGGACAAGTTTGTACAAAAAAGCAGGCTACTCATTCCCGTCTGATT	GGGGACCACTTTGTACAAGAAAGCTGGGTACGGCTGGTTAACGCCG
VC3 (Rv1857)	GGGGACAAGTTTGTACAAAAAAGCAGGCTCGTTGGATCGGACTGTCA	GGGGACCACTTTGTACAAGAAAGCTGGGTGGGCTTGGCGAAACCCG
VC4 (Rv2190c)	GGGGACAAGTTTGTACAAAAAAGCAGGCTCGCAACAGAGGATTCGGT	GGGGACCACTTTGTACAAGAAAGCTGGGTTGCAAGCACACCGGCAAC
VC5 (Rv2190c)	GGGGACAAGTTTGTACAAAAAAGCAGGCTAGGCTCGACCAGAGGTGG	GGGGACCACTTTGTACAAGAAAGCTGGGTGTAACGGCGGGCGTCGT
VC6 (Rv3886c)	GGGGACAAGTTTGTACAAAAAAGCAGGCTGCTTCGCCACTAAACCGA	GGGGACCACTTTGTACAAGAAAGCTGGGTACGCGCCCTCCGTAGCCG

Appendix H (A): 136 BPAs curated by Bowman et al, Improving reverse vaccinology with a machine learning approach,
Bowman BN, McAdam PR, Vivona S, Zhang JX, Luong T, Belew RK, Sahota H, Guiney D, Valafar F, Fierer J, Woelk CH; Improving reverse vaccinology with a machine learning approach. *Vaccine* 2011;**29**(45):8156-64

Species	Gram	No.	Gene	Description	Acc. No.	Ref.
Bacillus anthracis	Positive	1	LF[I-III]	Lethal factor (LF) domains I-III	AAR88322	[1]
		2	PA83[delta]furin	Protective antigen 83 delta furin	AAR88321	[1]
Bordetella pertussis	Negative	3	fhaB	Filamentious haemagglutinin	NP_880571	[2]
		4	prnA	Pertactin	AAK92094	[3]
		5	ptxA	Pertussis toxin S1 subunit	NP_882282	[4]
Borrelia burgdorferi	Negative	6	dbpA	Decorin-binding protein A	YP_003110629	[5]
		7	ospA	Outer surface protein A (ospA)	NP_045688	[6]
		8	ospB	Outer surface protein B (ospB)	NP_045689	[7]
		9	ospC	Outer surface protein C (ospC)	YP_002364847	[8]
		10	p66	Outer membrane protein (Oms66)	NP_212737	[9]
		11	VraA	Virulent strain-associated repetitive antigen A (VraA)	NP_045547	[10]
Brucella abortus	Negative	12	dnaK	Molecular chaperone DnaK	ZP_05821969	[11]
		13	omp16	Outer membrane protein 16 (Omp16)	P0A3S9	[12]
		14	omp19	Outer membrane protein 19 (Omp19)	P0C109	[12]
		15	rplL	50S ribosomal protein L7/L12	AAD51621	[13]
		16	sodC	Superoxide dismutase [Cu-Zn]	YP_418725	[14]
		17	surA	Periplasmic peptidyl prolyl *cis-trans* isomerase (SurA)	ZP_05821343	[11]
Brucella melitensis	Negative	18	bp26	Periplasmic immunogenic protein (bp26)	NP_539453	[15]
		19	ialB	Invasion associated locus B (IalB) protein	NP_540501	[16]
		20	omp25	Outer membran protein 25 kDa (Omp25)	NP_540166	[17]
		21	omp31	Outer membrane protein 31 (Omp31)	NP_539319	[18]
		22	P39	Periplasmic binding protein (P39)	NP_541568	[19]
		23	tig	Trigger factor (TF)	NP_539986	[15]
Campylobacter coli	Negative	24	flaA	Flagellin FlaA	P27053	[20]
Campylobacter jejuni	Negative	25	cjaA	Surface antigen (CjaA)	CAA71822	[21]
Chlamydia muridarum	Negative	26	momp	Major outer membran protein (MOMP)	AAB07068	[22]
Chlamydophila pneumoniae	Negative	27	Eno (CPn0800)	Phosphopyruvate hydratase, Enolase (Eno)	NP_224995	[23]
		28	OmpH-like (CPn0301)	Outer membrane protein H-like (OmpH-like)	NP_224506	[23]
		29	Pmp10 (CPn0449)	Probable outer membran protein 10 (Pmp10)	Q9RB65	[23]
		30	Pmp2 (CPn0013)	Polymorphic outer membrane protein G family (Pmp2)	NP_224227	[23]
Clostridium difficile	Positive	31	tcdA	Toxin A	CAA63564	[24]
Clostridium	Positive	32	etx	Epsilon toxin	AAA23236	[25]

perfringens		33	plc	Phospholipase C (Alpha toxin)	YP_694509	[26]
Clostridium tetani	Positive	34	tetx	Tetanus toxin	NP_783831	[27]
Corynebacterium pseudotuberculosis	Positive	35	pld	Phospholipase D	P20626	[28]
Escherichia coli	Negative	36	c1275	Hypothetical protein c1275	AAN79749	[29]
		37	c5321	Hypothetical protein c5321	NP_757168	[29]
		38	ECOK1_0290	Bacterial Ig-like domain (group 1) protein	ADE88959	[29]
		39	ECOK1_3374	General secretion pathway protein K	ADE91828	[29]
		40	ECOK1_3385	Putative lipoprotein	ADE89421	[29]
		41	ECOK1_3457	TonB-dependent siderophore receptor	ADE91247	[29]
		42	ecp_3827	Hemolysin A	YP_671699	[29]
		43	fepA	Outer membrane receptor (FepA)	YP_003076587	[30]
		44	stxB2	Shiga toxin 2 subunit B	NP_049501	[31]
		45	eltB	Heat-labile enterotoxin B subunit	YP_003293996	[32]
		46	fimH	FimH adhesin of type 1 fimbriae	ACZ17766	[33]
Haemophilus influenzae	Negative	47	omp1 (fadL)	Outer membrane protein P1 (Omp1)	NP_438563	[34]
		48	Hap	Hap adhesin, autotransporter protein	AAN37923	[35]
		49	omp26	Outer membrane protein 26 (Omp26)	AAD23967	[36]
		50	TbpB	transferrin binding protein B	YP_248692	[37]
		51	ompP5	Outer membrane protein P5 (OmpP5)	YP_248824	[38]
		52	ompP6 (pal)	Outer membrane protein P6 (OmpP6)	NP_438542	[39]
Helicobacter pylori	Negative	53	cagA	Cytotoxin-associated antigen (CagA)	AAF17598	[40]
		54	gltA	Citrate synthase	NP_222744	[41]
		55	katA	Catalase (KatA)	AAC16068	[42]
		56	napA	Neutrophil activatin protein (NapA)	AAA67928	[43]
		57	ureB	Urease B	P69996	[44]
		58	VacA	Vacuolating cytotoxin (VacA)	AAD04290	[40]
		59	groES	Heat shock protein 10 kDa (Hsp10, GroES)	NP_222731	[45]
Legionella pneumophila	Negative	60	groEL	Heat shock protein 60 kDa (Hsp60, GroEL)	YP_094724	[46]
		61	OmpS	Outer membrane protein S (OmpS)	YP_096954	[46]
		62	MSP	Major secretory protein (MSP), zinc metalloprotease	YP_094511	[47]
Listeria monocytogenes	Positive	63	hly	Listeriolysin O (Hly)	ZP_05300691	[48]
Mycobacterium avium	Mycobacterium	64	groEL	Heat shock protein 65 kDa (Hsp65, GroEL)	AAA99670	[49]
Mycobacterium bovis	Mycobacterium	65	fbpA	Fibronectin binding protein antigen (FbpA, Ag85A)	CAA37206	[49]
Mycobacterium tuberculosis	Mycobacterium	66	esxA	Early secretory antigenic target, 6 kDa (ESAT-6)	YP_178023	[50]
		67	fbpB	Fibronectin binding protein antigen B (FbpB, Ag85B)	NP_216402	[51]
		68	hbha	Heparin-binding hemagglutinin (Hbha)	NP_214989	[52]

		69	katG	Catalase-peroxidase-peroxynitritase T (KatG)	NP_216424	[50]
		70	mpt63	Immunogenic protein (MPT63)	NP_216442	[53]
		71	mpt64	Immunogenic protein (MPT-64)	NP_216496	[54]
		72	mpt83	Immunogenic protein (MPT-83)	NP_217389	[53]
		73	PstS-2	Phosphate-binding protein 2 (PstS-2)	CAA88137	[55]
		74	PstS-3	Phosphate-binding protein 3 (PstS-3)	CAA88138	[55]
		75	esxB	Culture filtrate antigen 10 kDa (CFP10, EsxB)	NP_218391	[56]
		76	fbpD	Fibronectin binding protein, MPT51/MPB51 antigen 85 complex C	YP_002646905	[57]
		77	hspX	Heat shock protein (hspX), 14 kDa antigen	NP_216547	[58]
		78	Mtb8.4	low molecular weight T-cell antigen (Mtb8.4)	NP_215690	[59]
		79	PPE14	PPE family protein (PPE14, Mtb41)	NP_854596	[60]
		80	PPE44	PPE family protein (PPE44)	CAE55522	[61]
		81	PstS-1	Phosphate-binding protein 1 (PstS-1)	YP_177770	[62]
Neisseria meningitidis	Negative	82	LolB (NMB0873)	Outer membrane lipoprotein (LolB)	NP_273914	[63]
		83	NMB1163	Putative periplasmic lipoprotein	NP_274190	[63]
		84	nspA	Outer-membrane associated protein (NspA)	AAC36000	[64]
		85	tbpA	transferrin binding protein (TbpA)	AAF81744	[65]
		86	tbpB	transferrin binding protein (TbpB)	AAF81745	[65]
Orientia tsutsugamushi	Negative	87	Bor56	56-kDa type-specific antigen (Bor56)	YP_001248444	[66]
Pasteurella multocida	Negative	88	ToxA	Pasteurella multocida toxin (PMT, ToxA)	P17452	[67]
Porphyromonas gingivalis	Negative	89	PG32	Immunoreactive 43kD antigen (PG32)	AAD51067	[68]
		90	PG33	Immunoreactive 42kD antigen (PG33)	AAD51068	[68]
Pseudomonas aeruginosa	Negative	91	pcrV	Type III secretion protein (PcrV)	NP_250397	[69]
		92	$toxA_{mut}$[2]	Exotoxin A mutant	AAB59097	[70]
		93	oprF	Outer membrane porin F (OprF)	NP_250468	[71]
Rickettsia rickettsii	Negative	94	ompA	Outer membrane protein A (OmpA)	AAA26380	[72]
		95	ompB	Outer membrane protein B (OmpB)	CAA34403	[73]
Salmonella paratyphi A	Negative	96	h1a	Flagellum antigen phase 1a (H1a)	ACF35754	[74]
		97	spaO	Surface presentation of antigens protein (SpaO)	ACF35726	[74]
Salmonella typhimurium	Negative	98	iicA	putative cytoplasmic protein (IicA)	AAD45811	[75]
		99	mig-14	putative transcriptional activator (Mig-14)	AAG31201	[75]
		100	sseB	Secretion system effector B (SseB)	AAC28879	[75]
Staphylococcus aureus	Positive	101	cna	Collagen adhesin (Cna)	AAA20874	[76]
		102	mecA	Penicillin-binding protein 2 prime (MecA)	NP_373278	[77]
		103	sea_{mut}	Enterotoxin type A	1DYQ_A	[78]
		104	clfa	Clumping factor A (Clfa), Fibrinogen binding protein	YP_001331790	[79]
Streptococcus agalactiae	Positive	105	ap1-2a	Ancillary protein 1 of pilus 2a (AP1-2a)	NP_688406	[80]
		106	ap1-2b	Ancillary protein 1 of pilus 2b, subtilase (AP1-2b)	ZP_00785378	[80]

		107	bp-2a	Backbone protein of pilus 2a (BP-2a)	NP_688405	[80]
		108	bp-2b	Backbone protein of pilus 2b (Bp-2b)	ZP_00785385	[80]
		109	GBS104 (SAG0649)	Cell wall surface anchor family protein, putative (SAG0649)	AAM99541	[81]
		110	GBS80 (SAG0645)	Cell wall surface anchor family protein (SAG0645)	AAM99537	[81]
		111	sip	Surface immunogenic protein (Sip)	AAG18474	[82]
Streptococcus pneumoniae	Positive	112	cbpA	choline binding protein A (CbpA)	YP_817402	[83]
		113	pdb	pneumolysin (Pdb)	YP_817150	[83]
		114	prtA (Sp130)	Serine protease (PrtA)	NP_345151	[84]
		115	psaA	Pneumococcal surface adhesin A (PsaA)	AAB09440	[85]
		116	pspA	Pneumococcal surface protein A (PspA)	NP_357715	[85]
		117	lytB (Sp46)	Putative endo-beta-N-acetylglucosaminidase (LytB)	AAK19156	[84]
		118	PhpA	histidine triad protein (PhpA)	AAK26629	[86]
		119	phtA (Sp36)	histidine triad protein (PhtA)	AAK19155	[84]
		120	lytC (Sp91)	1,4-beta-N-acetylmuramidase (LytC)	YP_873931	[84]
		121	pvaA (Sp101)	Pneumococcal vaccine antigen A (pvaA), PXO2-08	AAK19158	[84]
Streptococcus pyogenes	Positive	122	fbpA	Fibronectin-binding protein (FbpA)	BAB62098	[87]
		123	sfb	Fibronectin-binding protein 1 (Sfb)	CAA48133	[88]
Streptococcus suis serotype 2	Positive	124	esa	Epidermal surface antigen (Esa)	ABP90422	[89]
		125	ibp	IgG-binding protein (Ibp)	ABP89196	[89]
		126	rfeA	RTX family exoprotein A (RfeA)	ABP89149	[89]
		127	sly	Suilysin (Sly)	ABP90369	[89]
Treponema pallidum	Negative	128	glpQ	Glycerophosphodiester phosphodiesterase (GlpQ)	NP_218698	[90]
		129	tmpB	treponemal membrane protein (TmpB)	NP_219206	[91]
		130	tprK	tpr protein K (tprK)	AAF45140	[92]
		131	tpF1	Bacterioferrin antigen (TpF1)	NP_219475	[93]
Yersinia pestis	Negative	132	caf1	Capsular antigen F1 (Caf1)	NP_395430	[94]
		133	lcrV	low calcium response protein V (LcrV), V antigen	NP_395165	[95]
		134	yapF (YPO0606)	Autotransporter protein (YapF)	YP_002345675	[96]
		135	ybtQ (YPO1914)	Inner membrane ABC-transporter (YbtQ)	YP_002346905	[96]
		136	yscF	Needle complex major subunit (yscF)	NP_857727	[97]

Descriptive information for the 136 bacterial protective antigens (BPAs) in the BPA training data set. *Gene*, refers to the gene symbol for the antigen; *Gram*, refers to Gram stain and is not applicable (NA) for antigens from *M. tuberculosis*; *Acc. No.*, refers to the accession number at the National Center of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) for the protein sequence of the antigen; *Ref.*, refers to the literature reference from which descriptive information was curated for each bacterial protective antigen (BPA). Shading is for display purposes only.

[2]Sequence data for $toxA_{mut}$ was obtained by manually editing the sequence data for the *toxA* protein of *P. aeruginosa* (strain PA103) available from GenBank with accession number AAB59097. Residue 553 was deleted and the histidine at residue 426 was replaced with lysine to convert *toxA* to $toxA_{mut}$ as described by Denis-Mize *et al.* [70].

References for Appendix H (A)

[1] Hermanson G, Whitlow V, Parker S, Tonsky K, Rusalov D, Ferrari M, et al. A cationic lipid-formulated plasmid DNA vaccine confers sustained antibody-mediated protection against aerosolized anthrax spores. Proc Natl Acad Sci U S A 2004;101:13601-6.

[2] Kimura A, Mountzouros KT, Relman DA, Falkow S, Cowell JL. Bordetella pertussis filamentous hemagglutinin: evaluation as a protective antigen and colonization factor in a mouse respiratory infection model. Infect Immun 1990;58:7-16.

[3] Roberts M, Tite JP, Fairweather NF, Dougan G, Charles IG. Recombinant P.69/pertactin: immunogenicity and protection of mice against Bordetella pertussis infection. Vaccine 1992;10:43-8.

[4] Kamachi K, Konda T, Arakawa Y. DNA vaccine encoding pertussis toxin S1 subunit induces protection against Bordetella pertussis in mice. Vaccine 2003;21:4609-15.

[5] Hanson MS, Cassatt DR, Guo BP, Patel NK, McCarthy MP, Dorward DW, et al. Active and passive immunity against Borrelia burgdorferi decorin binding protein A (DbpA) protects against infection. Infect Immun 1998;66:2143-53.

[6] Luke CJ, Carner K, Liang X, Barbour AG. An OspA-based DNA vaccine protects mice against infection with Borrelia burgdorferi. J Infect Dis 1997;175:91-7.

[7] Fikrig E, Barthold SW, Marcantonio N, Deponte K, Kantor FS, Flavell RA. Roles of OspA, OspB, and flagellin in protective immunity to Lyme borreliosis in laboratory mice. Infect Immun 1992;60:657-61.

[8] Scheiblhofer S, Weiss R, Durnberger H, Mostbock S, Breitenbach M, Livey I, et al. A DNA vaccine encoding the outer surface protein C from Borrelia burgdorferi is able to induce protective immune responses. Microbes Infect 2003;5:939-46.

[9] Exner MM, Wu X, Blanco DR, Miller JN, Lovett MA. Protection elicited by native outer membrane protein Oms66 (p66) against host-adapted Borrelia burgdorferi: conformational nature of bactericidal epitopes. Infect Immun 2000;68:2647-54.

[10] Labandeira-Rey M, Baker EA, Skare JT. VraA (BBI16) protein of Borrelia burgdorferi is a surface-exposed antigen with a repetitive motif that confers partial protection against experimental Lyme borreliosis. Infect Immun 2001;69:1409-19.

[11] Delpino MV, Estein SM, Fossati CA, Baldi PC, Cassataro J. Vaccination with Brucella recombinant DnaK and SurA proteins induces protection against Brucella abortus infection in BALB/c mice. Vaccine 2007;25:6721-9.

[12] Pasquevich KA, Estein SM, Garcia Samartino C, Zwerdling A, Coria LM, Barrionuevo P, et al. Immunization with recombinant Brucella species outer membrane protein Omp16 or Omp19 in adjuvant induces specific CD4+ and CD8+ T cells as well as systemic and oral protection against Brucella abortus infection. Infect Immun 2009;77:436-45.

[13] Oliveira SC, Splitter GA. Immunization of mice with recombinant L7/L12 ribosomal protein confers protection against Brucella abortus infection. Vaccine 1996;14:959-62.

[14] Onate AA, Vemulapalli R, Andrews E, Schurig GG, Boyle S, Folch H. Vaccination with live Escherichia coli expressing Brucella abortus Cu/Zn superoxide dismutase protects mice against virulent B. abortus. Infect Immun 1999;67:986-8.

[15] Yang X, Hudson M, Walters N, Bargatze RF, Pascual DW. Selection of protective epitopes for Brucella melitensis by DNA vaccination. Infect Immun 2005;73:7297-303.

[16] Commander NJ, Spencer SA, Wren BW, MacMillan AP. The identification of two protective DNA vaccines from a panel of five plasmid constructs encoding Brucella melitensis 16M genes. Vaccine 2007;25:43-54.

[17] Bowden RA, Cloeckaert A, Zygmunt MS, Dubray G. Evaluation of immunogenicity and protective activity in BALB/c mice of the 25-kDa major outer-membrane protein of Brucella melitensis (Omp25) expressed in Escherichia coli. J Med Microbiol 1998;47:39-48.

[18] Cassataro J, Velikovsky CA, Bruno L, Estein SM, de la Barrera S, Bowden R, et al. Improved immunogenicity of a vaccination regimen combining a DNA vaccine encoding Brucella melitensis outer membrane protein 31 (Omp31) and recombinant Omp31 boosting. Clin Vaccine Immunol 2007;14:869-74.

[19] Al-Mariri A. Protection of BALB/c mice against Brucella melitensis 16 M infection induced by vaccination with live Escherchia coli expression Brucella P39 protein. Vaccine 2009;28:1766-70.

[20] Lee LH, Burg E, 3rd, Baqar S, Bourgeois AL, Burr DH, Ewing CP, et al. Evaluation of a truncated recombinant flagellin subunit vaccine against Campylobacter jejuni. Infect Immun 1999;67:5799-805.
[21] Wyszynska A, Raczko A, Lis M, Jagusztyn-Krynicka EK. Oral immunization of chickens with avirulent Salmonella vaccine strain carrying C. jejuni 72Dz/92 cjaA gene elicits specific humoral immune response associated with protection against challenge with wild-type Campylobacter. Vaccine 2004;22:1379-89.
[22] Igietseme JU, Murdin A. Induction of protective immunity against Chlamydia trachomatis genital infection by a vaccine based on major outer membrane protein-lipophilic immune response-stimulating complexes. Infect Immun 2000;68:6798-806.
[23] Finco O, Bonci A, Agnusdei M, Scarselli M, Petracca R, Norais N, et al. Identification of new potential vaccine candidates against Chlamydia pneumoniae by multiple screenings. Vaccine 2005;23:1178-88.
[24] Sauerborn M, Leukel P, von Eichel-Streiber C. The C-terminal ligand-binding domain of Clostridium difficile toxin A (TcdA) abrogates TcdA-specific binding to cells and prevents mouse lethality. FEMS Microbiol Lett 1997;155:45-54.
[25] Oyston PC, Payne DW, Havard HL, Williamson ED, Titball RW. Production of a non-toxic site-directed mutant of Clostridium perfringens epsilon-toxin which induces protective immunity in mice. Microbiology 1998;144 (Pt 2):333-41.
[26] Williamson ED, Titball RW. A genetically engineered vaccine against the alpha-toxin of Clostridium perfringens protects mice against experimental gas gangrene. Vaccine 1993;11:1253-8.
[27] Norton PM, Wells JM, Brown HW, Macpherson AM, Le Page RW. Protection against tetanus toxin in mice nasally immunized with recombinant Lactococcus lactis expressing tetanus toxin fragment C. Vaccine 1997;15:616-9.
[28] Fontaine MC, Baird G, Connor KM, Rudge K, Sales J, Donachie W. Vaccination confers significant protection of sheep against infection with a virulent United Kingdom strain of Corynebacterium pseudotuberculosis. Vaccine 2006;24:5986-96.
[29] Moriel DG, Bertoldi I, Spagnuolo A, Marchi S, Rosini R, Nesta B, et al. Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic Escherichia coli. Proc Natl Acad Sci U S A 2010;107:9072-7.
[30] Baghal SM, Gargari SL, Rasooli I. Production and immunogenicity of recombinant ferric enterobactin protein (FepA). Int J Infect Dis 2010;14 Suppl 3:e166-70.
[31] Tsuji T, Shimizu T, Sasaki K, Tsukamoto K, Arimitsu H, Ochi S, et al. A nasal vaccine comprising B-subunit derivative of Shiga toxin 2 for cross-protection against Shiga toxin types 1 and 2. Vaccine 2008;26:2092-9.
[32] Mason HS, Haq TA, Clements JD, Arntzen CJ. Edible vaccine protects mice against Escherichia coli heat-labile enterotoxin (LT): potatoes expressing a synthetic LT-B gene. Vaccine 1998;16:1336-43.
[33] Langermann S, Palaszynski S, Barnhart M, Auguste G, Pinkner JS, Burlein J, et al. Prevention of mucosal Escherichia coli infection by FimH-adhesin-based systemic vaccination. Science 1997;276:607-11.
[34] Bolduc GR, Bouchet V, Jiang RZ, Geisselsoder J, Truong-Bolduc QC, Rice PA, et al. Variability of outer membrane protein P1 and its evaluation as a vaccine candidate against experimental otitis media due to nontypeable Haemophilus influenzae: an unambiguous, multifaceted approach. Infect Immun 2000;68:4505-17.
[35] Cutter D, Mason KW, Howell AP, Fink DL, Green BA, St Geme JW, 3rd. Immunization with Haemophilus influenzae Hap adhesin protects against nasopharyngeal colonization in experimental mice. J Infect Dis 2002;186:1115-21.
[36] Kyd JM, Cripps AW. Potential of a novel protein, OMP26, from nontypeable Haemophilus influenzae to enhance pulmonary clearance in a rat model. Infect Immun 1998;66:2272-8.
[37] Webb DC, Cripps AW. Immunization with recombinant transferrin binding protein B enhances clearance of nontypeable Haemophilus influenzae from the rat lung. Infect Immun 1999;67:2138-44.
[38] Bakaletz LO, Leake ER, Billy JM, Kaumaya PT. Relative immunogenicity and efficacy of two synthetic chimeric peptides of fimbrin as vaccinogens against nasopharyngeal colonization by nontypeable Haemophilus influenzae in the chinchilla. Vaccine 1997;15:955-61.
[39] Hotomi M, Saito T, Yamanaka N. Specific mucosal immunity and enhanced nasopharyngeal clearance of nontypeable Haemophilus influenzae after intranasal immunization with outer membrane protein P6 and cholera toxin. Vaccine 1998;16:1950-6.

[40] Marchetti M, Rossi M, Giannelli V, Giuliani MM, Pizza M, Censini S, et al. Protection against Helicobacter pylori infection in mice by intragastric vaccination with H. pylori antigens is achieved using a non-toxic mutant of E. coli heat-labile enterotoxin (LT) as adjuvant. Vaccine 1998;16:33-7.
[41] Dunkley ML, Harris SJ, McCoy RJ, Musicka MJ, Eyers FM, Beagley LG, et al. Protection against Helicobacter pylori infection by intestinal immunisation with a 50/52-kDa subunit protein. FEMS Immunol Med Microbiol 1999;24:221-5.
[42] Radcliff FJ, Hazell SL, Kolesnikow T, Doidge C, Lee A. Catalase, a novel antigen for Helicobacter pylori vaccination. Infect Immun 1997;65:4668-74.
[43] Satin B, Del Giudice G, Della Bianca V, Dusi S, Laudanna C, Tonello F, et al. The neutrophil-activating protein (HP-NAP) of Helicobacter pylori is a protective antigen and a major virulence factor. J Exp Med 2000;191:1467-76.
[44] Ermak TH, Giannasca PJ, Nichols R, Myers GA, Nedrud J, Weltzin R, et al. Immunization of mice with urease vaccine affords protection against Helicobacter pylori infection in the absence of antibodies and is mediated by MHC class II-restricted responses. J Exp Med 1998;188:2277-88.
[45] Ferrero RL, Thiberge JM, Kansau I, Wuscher N, Huerre M, Labigne A. The GroES homolog of Helicobacter pylori confers protective immunity against mucosal infection in mice. Proc Natl Acad Sci U S A 1995;92:6499-503.
[46] Weeratna R, Stamler DA, Edelstein PH, Ripley M, Marrie T, Hoskin D, et al. Human and guinea pig immune responses to Legionella pneumophila protein antigens OmpS and Hsp60. Infect Immun 1994;62:3454-62.
[47] Blander SJ, Horwitz MA. Vaccination with the major secretory protein of Legionella pneumophila induces cell-mediated and protective immunity in a guinea pig model of Legionnaires' disease. J Exp Med 1989;169:691-705.
[48] Sirard JC, Fayolle C, de Chastellier C, Mock M, Leclerc C, Berche P. Intracytoplasmic delivery of listeriolysin O by a vaccinal strain of Bacillus anthracis induces CD8-mediated protection against Listeria monocytogenes. J Immunol 1997;159:4435-43.
[49] Velaz-Faircloth M, Cobb AJ, Horstman AL, Henry SC, Frothingham R. Protection against Mycobacterium avium by DNA vaccines expressing mycobacterial antigens as fusion proteins with green fluorescent protein. Infect Immun 1999;67:4243-50.
[50] Li Z, Howard A, Kelley C, Delogu G, Collins F, Morris S. Immunogenicity of DNA vaccines expressing tuberculosis proteins fused to tissue plasminogen activator signal sequences. Infect Immun 1999;67:4780-6.
[51] Kamath AT, Groat NL, Bean AG, Britton WJ. Protective effect of DNA immunization against mycobacterial infection is associated with the early emergence of interferon-gamma (IFN-gamma)-secreting lymphocytes. Clin Exp Immunol 2000;120:476-82.
[52] Parra M, Pickett T, Delogu G, Dheenadhayalan V, Debrie AS, Locht C, et al. The mycobacterial heparin-binding hemagglutinin is a protective antigen in the mouse aerosol challenge model of tuberculosis. Infect Immun 2004;72:6799-805.
[53] Morris S, Kelley C, Howard A, Li Z, Collins F. The immunogenicity of single and combination DNA vaccines against tuberculosis. Vaccine 2000;18:2155-63.
[54] Delogu G, Howard A, Collins FM, Morris SL. DNA vaccination against tuberculosis: expression of a ubiquitin-conjugated tuberculosis protein enhances antimycobacterial immunity. Infect Immun 2000;68:3097-102.
[55] Tanghe A, Lefevre P, Denis O, D'Souza S, Braibant M, Lozes E, et al. Immunogenicity and protective efficacy of tuberculosis DNA vaccines encoding putative phosphate transport receptors. J Immunol 1999;162:1113-9.
[56] Wu Y, Woodworth JS, Shin DS, Morris S, Behar SM. Vaccine-elicited 10-kilodalton culture filtrate protein-specific CD8+ T cells are sufficient to mediate protection against Mycobacterium tuberculosis infection. Infect Immun 2008;76:2249-55.
[57] Miki K, Nagata T, Tanaka T, Kim YH, Uchijima M, Ohara N, et al. Induction of protective cellular immunity against Mycobacterium tuberculosis by recombinant attenuated self-destructing Listeria monocytogenes strains harboring eukaryotic expression plasmids for antigen 85 complex and MPB/MPT51. Infect Immun 2004;72:2014-21.
[58] Woolhiser L, Tamayo MH, Wang B, Gruppo V, Belisle JT, Lenaerts AJ, et al. In vivo adaptation of the Wayne model of latent tuberculosis. Infect Immun 2007;75:2621-5.

[59] Coler RN, Campos-Neto A, Ovendale P, Day FH, Fling SP, Zhu L, et al. Vaccination with the T cell antigen Mtb 8.4 protects against challenge with Mycobacterium tuberculosis. J Immunol 2001;166:6227-35.
[60] Skeiky YA, Alderson MR, Ovendale PJ, Lobet Y, Dalemans W, Orme IM, et al. Protection of mice and guinea pigs against tuberculosis induced by immunization with a single Mycobacterium tuberculosis recombinant antigen, MTB41. Vaccine 2005;23:3937-45.
[61] Romano M, Rindi L, Korf H, Bonanni D, Adnet PY, Jurion F, et al. Immunogenicity and protective efficacy of tuberculosis subunit vaccines expressing PPE44 (Rv2770c). Vaccine 2008;26:6053-63.
[62] Falero-Diaz G, Challacombe S, Banerjee D, Douce G, Boyd A, Ivanyi J. Intranasal vaccination of mice against infection with Mycobacterium tuberculosis. Vaccine 2000;18:3223-9.
[63] Pajon R, Yero D, Niebla O, Climent Y, Sardinas G, Garcia D, et al. Identification of new meningococcal serogroup B surface antigens through a systematic analysis of neisserial genomes. Vaccine 2009;28:532-41.
[64] Martin D, Cadieux N, Hamel J, Brodeur BR. Highly conserved Neisseria meningitidis surface protein confers protection against experimental infection. J Exp Med 1997;185:1173-83.
[65] West D, Reddin K, Matheson M, Heath R, Funnell S, Hudson M, et al. Recombinant Neisseria meningitidis transferrin binding protein A protects against experimental meningococcal infection. Infect Immun 2001;69:1561-7.
[66] Seong SY, Huh MS, Jang WJ, Park SG, Kim JG, Woo SG, et al. Induction of homologous immune response to Rickettsia tsutsugamushi Boryong with a partial 56-kilodalton recombinant antigen fused with the maltose-binding protein MBP-Bor56. Infect Immun 1997;65:1541-5.
[67] Seo J, Pyo H, Lee S, Lee J, Kim T. Expression of 4 truncated fragments of Pasteurella multocida toxin and their immunogenicity. Can J Vet Res 2009;73:184-9.
[68] Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, Rothel L, et al. Identification of vaccine candidate antigens from a genomic analysis of Porphyromonas gingivalis. Vaccine 2001;19:4135-42.
[69] Holder IA, Neely AN, Frank DW. PcrV immunization enhances survival of burned Pseudomonas aeruginosa-infected mice. Infect Immun 2001;69:5908-10.
[70] Denis-Mize KS, Price BM, Baker NR, Galloway DR. Analysis of immunization with DNA encoding Pseudomonas aeruginosa exotoxin A. FEMS Immunol Med Microbiol 2000;27:147-54.
[71] Gilleland HE, Jr., Gilleland LB, Matthews-Greer JM. Outer membrane protein F preparation of Pseudomonas aeruginosa as a vaccine against chronic pulmonary infection with heterologous immunotype strains in a rat model. Infect Immun 1988;56:1017-22.
[72] Crocquet-Valdes PA, Diaz-Montero CM, Feng HM, Li H, Barrett AD, Walker DH. Immunization with a portion of rickettsial outer membrane protein A stimulates protective immunity against spotted fever rickettsiosis. Vaccine 2001;20:979-88.
[73] Diaz-Montero CM, Feng HM, Crocquet-Valdes PA, Walker DH. Identification of protective components of two major outer membrane proteins of spotted fever group Rickettsiae. Am J Trop Med Hyg 2001;65:371-8.
[74] Ruan P, Xia XP, Sun D, Ojcius DM, Mao YF, Yue WY, et al. Recombinant SpaO and H1a as immunogens for protection of mice from lethal infection with Salmonella paratyphi A: implications for rational design of typhoid fever vaccines. Vaccine 2008;26:6639-44.
[75] Rollenhagen C, Sorensen M, Rizos K, Hurvitz R, Bumann D. Antigen selection based on expression levels during infection facilitates vaccine development for an intracellular pathogen. Proc Natl Acad Sci U S A 2004;101:8739-44.
[76] Nilsson IM, Patti JM, Bremell T, Hook M, Tarkowski A. Vaccination with a recombinant fragment of collagen adhesin provides protection against Staphylococcus aureus-mediated septic death. J Clin Invest 1998;101:2640-9.
[77] Ohwada A, Sekiya M, Hanaki H, Arai KK, Nagaoka I, Hori S, et al. DNA vaccination by mecA sequence evokes an antibacterial immune response against methicillin-resistant Staphylococcus aureus. J Antimicrob Chemother 1999;44:767-74.
[78] Nilsson IM, Verdrengh M, Ulrich RG, Bavari S, Tarkowski A. Protection against Staphylococcus aureus sepsis by vaccination with recombinant staphylococcal enterotoxin A devoid of superantigenicity. J Infect Dis 1999;180:1370-3.
[79] Mamo W, Boden M, Flock JI. Vaccination with Staphylococcus aureus fibrinogen binding proteins (FgBPs) reduces colonisation of S. aureus in a mouse mastitis model. FEMS Immunol Med Microbiol 1994;10:47-53.

[80] Margarit I, Rinaudo CD, Galeotti CL, Maione D, Ghezzo C, Buttazzoni E, et al. Preventing bacterial infections with pilus-based vaccines: the group B streptococcus paradigm. J Infect Dis 2009;199:108-15.
[81] Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, et al. Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science 2005;309:148-50.
[82] Brodeur BR, Boyer M, Charlebois I, Hamel J, Couture F, Rioux CR, et al. Identification of group B streptococcal Sip protein, which elicits cross-protective immunity. Infect Immun 2000;68:5610-8.
[83] Ogunniyi AD, Woodrow MC, Poolman JT, Paton JC. Protection against Streptococcus pneumoniae elicited by immunization with pneumolysin and CbpA. Infect Immun 2001;69:5997-6003.
[84] Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, et al. Use of a whole genome approach to identify vaccine molecules affording protection against Streptococcus pneumoniae infection. Infect Immun 2001;69:1593-8.
[85] Briles DE, Ades E, Paton JC, Sampson JS, Carlone GM, Huebner RC, et al. Intranasal immunization of mice with a mixture of the pneumococcal proteins PsaA and PspA is highly protective against nasopharyngeal carriage of Streptococcus pneumoniae. Infect Immun 2000;68:796-800.
[86] Zhang Y, Masi AW, Barniak V, Mountzouros K, Hostetter MK, Green BA. Recombinant PhpA protein, a unique histidine motif-containing protein from Streptococcus pneumoniae, protects mice against intranasal pneumococcal challenge. Infect Immun 2001;69:3827-36.
[87] Terao Y, Okamoto S, Kataoka K, Hamada S, Kawabata S. Protective immunity against Streptococcus pyogenes challenge in mice after immunization with fibronectin-binding protein. J Infect Dis 2005;192:2081-91.
[88] Schulze K, Medina E, Talay SR, Towers RJ, Chhatwal GS, Guzman CA. Characterization of the domain of fibronectin-binding protein I of Streptococcus pyogenes responsible for elicitation of a protective immune response. Infect Immun 2001;69:622-5.
[89] Liu L, Cheng G, Wang C, Pan X, Cong Y, Pan Q, et al. Identification and experimental verification of protective antigens against Streptococcus suis serotype 2 based on genome sequence analysis. Curr Microbiol 2009;58:11-7.
[90] Cameron CE, Castro C, Lukehart SA, Van Voorhis WC. Function and protective capacity of Treponema pallidum subsp. pallidum glycerophosphodiester phosphodiesterase. Infect Immun 1998;66:5763-70.
[91] Wicher K, Schouls LM, Wicher V, Van Embden JD, Nakeeb SS. Immunization of guinea pigs with recombinant TmpB antigen induces protection against challenge infection with Treponema pallidum Nichols. Infect Immun 1991;59:4343-8.
[92] Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Van Voorhis WC, et al. Treponema pallidum major sheath protein homologue Tpr K is a target of opsonic antibody and the protective immune response. J Exp Med 1999;189:647-56.
[93] Borenstein LA, Radolf JD, Fehniger TE, Blanco DR, Miller JN, Lovett MA. Immunization of rabbits with recombinant Treponema pallidum surface antigen 4D alters the course of experimental syphilis. J Immunol 1988;140:2415-21.
[94] Grosfeld H, Cohen S, Bino T, Flashner Y, Ber R, Mamroud E, et al. Effective protective immunity to Yersinia pestis infection conferred by DNA vaccine coding for derivatives of the F1 capsular antigen. Infect Immun 2003;71:374-83.
[95] Anderson GW, Jr., Leary SE, Williamson ED, Titball RW, Welkos SL, Worsham PL, et al. Recombinant V antigen protects mice against pneumonic and bubonic plague caused by F1-capsule-positive and -negative strains of Yersinia pestis. Infect Immun 1996;64:4580-5.
[96] Li B, Zhou L, Guo J, Wang X, Ni B, Ke Y, et al. High-throughput identification of new protective antigens from a Yersinia pestis live vaccine by enzyme-linked immunospot assay. Infect Immun 2009;77:4356-61.
[97] Wang S, Joshi S, Mboudjeka I, Liu F, Ling T, Goguen JD, et al. Relative immunogenicity and protection potential of candidate Yersinia Pestis antigens against lethal mucosal plague challenge in Balb/C mice. Vaccine 2008;26:1664-74.

Appendix H (B): 64 curated BPAs that were combined with 136 previously curated BPAs (Appendix H (A)) to form the positive training data of BPAD200 (200 BPAs), Heinson et al, Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology, International Journal of Molecular Sciences, 2017.

Species	Gram	No.	Gene	Description	ACC. No.	Ref.
Bacillus anthracis	Positive	137	GroEL	Molecular chaperone GroEL	YP_026537.1	[1]
		138	DNAK	Molecular chaperone DnaK	YP_030461	[1]
Brucella abortus	Negative	139	cobB	Cobyrinic acid A,C-diamide synthase	ENR47700.1	[2]
		140	L9	Ribosomal Protein L9	ENR50124.1	[3]
		141	BLS	luminaze synthase	1DI0_A	[4]
Burkholderia pseudomallei	Negative	142	fliC	Flagellin structural protein	EET06073.1	[5]
		143	Omp85	Outer Membrane Protein	ABN54438.1	[6]
Campylobacter jejuni	Negative	144	fspA2	Flagella Secreted Protein	EDK22348.1	[7]
		145	fspA1	Flagella Secreted Protein	YP_005657787.1	[7]
Chlamydia trachomatis	Negative	146	CT144	CT144	NP_219647.1	[8]
		147	CT823 (htrA)	DO serine protease (htrA)	NP_220344.1	[8]
		148	CPAF	Hypothetical Protein CPAF	WP_009872247.1	[9]
Chlamydophila psittaci	Negative	149	ompA	Major outer membrane protein	YP_328507.1	[10]
Escherichia coli	Negative	150	c2482	Copper exporting ATPase	NP_754374.1	[11]
		151	iutA	IutA protein	NP_755498.1	[11]
		152	c5174	Iron-regulated outer membrane virulence protein	NP_757022	[11]
Francisella tularensis	Negative	153	fopA	OmpA family protein	YP_007012122.1	[12]
Helicobacter pylori	Negative	154	HpaA	Flagella sheath adhesin protein HpaA	AEB91475.1	[13]
Leptospira interrogans	Negative	155	LA_0607	MCE-related protein	AAN47806.1	[14]
		156	LA_1118	Hypothetical protein LA_1118	AAN48316.1	[14]
		157	LA_1454	Hypothetical protein LA_1454	AAN48653.1	[14]
		158	LIF_A0192	OmpA family lipoprotein	AER01005.1	[15]
		159	LIF_A2958	OmpA family protein	AER03729.1	[15]
		160	ompA	OmpA family protein	AER04228	[15]
		161	Hap1	hemolysin-associated protein 1	AAL18599.1	[16]
Listeria monocytogenes	Positive	162	iap	P60 extracellular protein, invasion associated protein Iap	CAC98661	[17]
Mycobacterium tuberculosis	Mycobacterium	163	esxH	Low molecular weight protein antigen 7 EsxH	NP_214802.1	[18]
		164	PE3	PE family protein PE3	YP_177697.1	[19]
Neisseria meningitidis	Negative	165	PorA	Outer membrane protein PorA (NMB1429)	AAF41790	[20]

		166	exbB	Biopolymer transport protein	AAF42074	[20]
		167	LctP	L-lactate permease	NP_273588.1	[20]
Pasteurella multocida	Negative	168	ompH	Outer membrane protein	AAC02243.1	[21]
Pseudomonas aeruginosa	Negative	169	popB	Translocator Protein PopB	NP_250399.1	[22]
Salmonella typhimurium	Negative	170	ompL	Outer membrane porin L (OmpL)	NP_462896.1	[23]
Shigella Sonnei	Negative	171	IpaB	IpaB	AAA26522	[23]
		172	IpaD	IpaD	AAA26524	[23]
Staphylococcus aureus	Positive	173	hla	Hla alpha hemolysin	YP_001332107	[24]
Streptococcus agalactiae	Positive	174	bca	C protein alpha antigen	A46405	[25]
		175	Rib	Surface protein Rib	AAC44468.1	[25]
		176	BibA	Surface Protein, BibA	ACU44469	[26]
		177	BPS	Group B protective surface protein	CAB46338.1	[27]
Streptococcus pneumoniae	Positive	178	gltX	Glutamyl-tRNA Synthetase (gts).	NP_346492.1	[28]
		179	SrtA	Sortase A	YP_816547.1	[29]
		180	SP_0463	Rrgb Pilus Protein Clade 1	2Y1V_C	[30]
		181	phtD	Pht D histidine triad protein D	YP_002037641.1	[31]
		182	phtE	Pht E histidine triad protein E	YP_002037642.1	[31]
Streptococcus pyogenes	Positive	183	prtS	Putative cell envelope proteinase	AAK33444	[32]
		184	Spy2018	M protein Type 1	AAK34694	[32]
		185	SPy_0019	SPy0019 Putative secreted protein	AAK33158.1	[33]
		186	slo	SPy0167 Streptolysin O precursor	AAK33267.1	[33]
		187	SPy_0488	Hypothetical protein SPy_0488	AAK33494.1	[34]
		188	SPy_0872	Putative secreted 5'-nucleotidase	AAK33792.1	[34]
		189	SPy_0895	Histidine protein kinase	AAK33814.1	[34]
		190	inlA	SPy1361 Putative internalin A precursor	AAK34188.1	[33]
		191	mraW	Conserved hypothetical protein	AAK34428.1	[34]
		192	SPy_1727	Conserved hypothetical protein	AAK34472.1	[34]
		193	scpA	SPy2010 C5A peptidase precursor	AAK34691.1	[33]
		194	sagP	Streptococcal acid glycoprotein	AAM22954.1	[35]
		195	tig	Transcription regulator - Trigger factor	NP_665438.1	[35]
		196	SPy_0146	Putative regulatory protein	NP_268530.1	[34]
		197	prgA	Surface exclusion protein	NP_268623	[34]
		198	dacA	D-alanyl-D-alanine carboxypeptidase	NP_268639.1	[34]
Streptococcus suis	Positive	199	SSU05_0332	Hypothetical protein SSU05_0332	YP_001197700.1	[36]

Yersinia pestis	Negative	200	YopD	Secreted effector protein	NP_395162.1	[37]

Descriptive information for the 64 bacterial protective antigens (BPAs) in the BPAD 200 training dataset. *Gene*, refers to the gene symbol for the antigen *Gram*, refers to Gram stain and is not applicable (NA) for antigens from *M. tuberculosis*; *Acc. No.*, refers to the accession number at the National Center o Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) for the protein sequence of the antigen; *Ref.*, refers to the literature reference from which descriptive information was curated for each bacterial protective antigen (BPA). Shading is for display purposes only.

References for Appendix H (B)

1. Sinha K, Bhatnagar R; GroEL provides protection against Bacillus anthracis infection in BALB/c mice. *Mol Immunol* 2010;**48**(1-3):264-71
2. Fu S, Xu J, Li X, Xie Y, Qiu Y, Du X, Yu S, Bai Y, Chen Y, Wang T, Wang Z, Yu Y, Peng G, Huang K, Huang L, Wang Y, Chen Z; Immunization of mice with recombinant protein CobB or AsnC confers protection against Brucella abortus infection. *PLoS One* 2012;**7**(2):e29552
3. Jain S, Kumar S, Dohre S, Afley P, Sengupta N, Alam SI; Identification of a protective protein from stationary-phase exoproteome of Brucella abortus. *Pathog Dis* 2014;**70**(1):75-83
4. Velikovsky CA, Cassataro J, Giambartolomei GH, Goldbaum FA, Estein S, Bowden RA, Bruno L, Fossati CA, Spitz M; A DNA vaccine encoding lumazine synthase from Brucella abortus induces protective immunity in BALB/c mice. *Infect Immun* 2002;**70**(5):2507-11
5. Chen YS, Hsiao YS, Lin HH, Yen CM, Chen SC, Chen YL; Immunogenicity and anti-Burkholderia pseudomallei activity in Balb/c mice immunized with plasmid DNA encoding flagellin. *Vaccine* 2006;**24**(6):750-8
6. Su YC, Wan KL, Mohamed R, Nathan S; Immunization with the recombinant Burkholderia pseudomallei outer membrane protein Omp85 induces protective immunity in mice. *Vaccine* 2010;**28**(31):5005-11
7. Baqar S, Applebee LA, Gilliland TC, Jr., Lee LH, Porter CK, Guerry P; Immunogenicity and protective efficacy of recombinant Campylobacter jejuni flagellum-secreted proteins in mice. *Infect Immun* 2008;**76**(7):3170-5
8. Picard MD, Cohane KP, Gierahn TM, Higgins DE, Flechtner JB; High-throughput proteomic screening identifies Chlamydia trachomatis antigens that are capable of eliciting T cell and antibody responses that provide protection against vaginal challenge. *Vaccine* 2012;**30**(29):4387-93
9. Murthy AK, Cong Y, Murphey C, Guentzel MN, Forsthuber TG, Zhong G, Arulanandam BP; Chlamydial protease-like activity factor induces protective immunity against genital chlamydial infection in transgenic mice that express the human HLA-DR4 allele. *Infect Immun* 2006;**74**(12):6722-9
10. Harkinezhad T, Schautteet K, Vanrompay D; Protection of budgerigars (Melopsittacus undulatus) against Chlamydophila psittaci challenge by DNA vaccination. *Vet Res* 2009;**40**(6):61
11. Alteri CJ, Hagan EC, Sivick KE, Smith SN, Mobley HL; Mucosal immunization with iron receptor antigens protects against urinary tract infection. *PLoS Pathog* 2009;**5**(9):e1000586
12. Hickey AJ, Hazlett KR, Kirimanjeswara GS, Metzger DW; Identification of Francisella tularensis outer membrane protein A (FopA) as a protective antigen for tularemia. *Vaccine* 2011;**29**(40):6941-7
13. Sutton P, Doidge C, Pinczower G, Wilson J, Harbour S, Swierczak A, Lee A; Effectiveness of vaccination with recombinant HpaA from Helicobacter pylori is influenced by host genetic background. *FEMS Immunol Med Microbiol* 2007;**50**(2):213-9
14. Chang YF, Chen CS, Palaniappan RU, He H, McDonough SP, Barr SC, Yan W, Faisal SM, Pan MJ, Chang CF; Immunogenicity of the recombinant leptospiral putative outer membrane proteins as vaccine candidates. *Vaccine* 2007;**25**(48):8190-7
15. Yan W, Faisal SM, McDonough SP, Chang CF, Pan MJ, Akey B, Chang YF; Identification and characterization of OmpA-like proteins as novel vaccine candidates for Leptospirosis. *Vaccine* 2010;**28**(11):2277-83
16. Branger C, Chatrenet B, Gauvrit A, Aviat F, Aubert A, Bach JM, Andre-Fontaine G; Protection against Leptospira interrogans sensu lato challenge by DNA immunization with the gene encoding hemolysin-associated protein 1. *Infect Immun* 2005;**73**(7):4062-9

17. Grenningloh R, Darj A, Bauer H, zur Lage S, Chakraborty T, Jacobs T, Weiss S; Liposome-encapsulated antigens induce a protective CTL response against Listeria monocytogenes independent of CD4+ T cell help. *Scand J Immunol* 2008;**67**(6):594-602
18. Dietrich J, Aagaard C, Leah R, Olsen AW, Stryhn A, Doherty TM, Andersen P; Exchanging ESAT6 with TB10.4 in an Ag85B fusion molecule-based tuberculosis subunit vaccine: efficient protection and ESAT6-based sensitive monitoring of vaccine efficacy. *J Immunol* 2005;**174**(10):6332-9
19. Singh SK, Kumari R, Singh DK, Tiwari S, Singh PK, Sharma S, Srivastava KK; Putative roles of a proline-glutamic acid-rich protein (PE3) in intracellular survival and as a candidate for subunit vaccine against Mycobacterium tuberculosis. *Med Microbiol Immunol* 2013;**202**(5):365-77
20. Sun Y, Li Y, Exley RM, Winterbotham M, Ison C, Smith H, Tang CM; Identification of novel antigens that protect against systemic meningococcal infection. *Vaccine* 2005;**23**(32):4136-41
21. Tan HY, Nagoor NH, Sekaran SD; Cloning, expression and protective capacity of 37 kDa outer membrane protein gene (ompH) of Pasteurella multocida serotype B:2. *Trop Biomed* 2010;**27**(3):430-41
22. Wu W, Huang J, Duan B, Traficante DC, Hong H, Risech M, Lory S, Priebe GP; Th17-stimulating protein vaccines confer protection against Pseudomonas aeruginosa pneumonia. *Am J Respir Crit Care Med* 2012;**186**(5):420-7
23. Yang Y, Wan C, Xu H, Wei H; Identification and characterization of OmpL as a potential vaccine candidate for immune-protection against salmonellosis in mice. *Vaccine* 2013;**31**(28):2930-6
24. Bubeck Wardenburg J, Schneewind O; Vaccine protection against Staphylococcus aureus pneumonia. *J Exp Med* 2008;**205**(2):287-94
25. Larsson C, Stalhammar-Carlemalm M, Lindahl G; Experimental vaccination against group B streptococcus, an encapsulated bacterium, with highly purified preparations of cell surface proteins Rib and alpha. *Infect Immun* 1996;**64**(9):3518-23
26. Santi I, Maione D, Galeotti CL, Grandi G, Telford JL, Soriani M; BibA induces opsonizing antibodies conferring in vivo protection against group B Streptococcus. *J Infect Dis* 2009;**200**(4):564-70
27. Erdogan S, Fagan PK, Talay SR, Rohde M, Ferrieri P, Flores AE, Guzman CA, Walker MJ, Chhatwal GS; Molecular analysis of group B protective surface protein, a new cell surface protective antigen of group B streptococci. *Infect Immun* 2002;**70**(2):803-11
28. Mizrachi Nebenzahl Y, Bernstein A, Portnoi M, Shagan M, Rom S, Porgador A, Dagan R; Streptococcus pneumoniae surface-exposed glutamyl tRNA synthetase, a putative adhesin, is able to induce a partially protective immune response in mice. *J Infect Dis* 2007;**196**(6):945-53
29. Gianfaldoni C, Maccari S, Pancotto L, Rossi G, Hilleringmann M, Pansegrau W, Sinisi A, Moschioni M, Masignani V, Rappuoli R, Del Giudice G, Ruggiero P; Sortase A confers protection against Streptococcus pneumoniae in mice. *Infect Immun* 2009;**77**(7):2957-61
30. Gianfaldoni C, Censini S, Hilleringmann M, Moschioni M, Facciotti C, Pansegrau W, Masignani V, Covacci A, Rappuoli R, Barocchi MA, Ruggiero P; Streptococcus pneumoniae pilus subunits protect mice against lethal challenge. *Infect Immun* 2007;**75**(2):1059-62
31. Godfroid F, Hermand P, Verlant V, Denoel P, Poolman JT; Preclinical evaluation of the Pht proteins as potential cross-protective pneumococcal vaccine antigens. *Infect Immun* 2011;**79**(1):238-45
32. Rodriguez-Ortega MJ, Norais N, Bensi G, Liberatori S, Capo S, Mora M, Scarselli M, Doro F, Ferrari G, Garaguso I, Maggi T, Neumann A, Covre A, Telford JL, Grandi G; Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. *Nat Biotechnol* 2006;**24**(2):191-7
33. Bensi G, Mora M, Tuscano G, Biagini M, Chiarot E, Bombaci M, Capo S, Falugi F, Manetti AG, Donato P, Swennen E, Gallotta M, Garibaldi M, Pinto V, Chiappini N, Musser JM, Janulczyk R, Mariani M, Scarselli M, Telford JL, Grifantini R, Norais N, Margarit I, Grandi G; Multi high-throughput approach for highly selective identification of vaccine candidates: the Group A Streptococcus case. *Mol Cell Proteomics* 2012;**11**(6):M111.015693
34. Fritzer A, Senn BM, Minh DB, Hanner M, Gelbmann D, Noiges B, Henics T, Schulze K, Guzman CA, Goodacre J, von Gabain A, Nagy E, Meinke AL; Novel conserved group A streptococcal proteins identified by the antigenome technology as vaccine candidates for a non-M protein-based vaccine. *Infect Immun* 2010;**78**(9):4051-67
35. Henningham A, Chiarot E, Gillen CM, Cole JN, Rohde M, Fulde M, Ramachandran V, Cork AJ, Hartas J, Magor G, Djordjevic SP, Cordwell SJ, Kobe B, Sriprakash KS, Nizet V, Chhatwal GS, Margarit IY, Batzloff MR, Walker MJ; Conserved anchorless surface proteins as group A streptococcal vaccine candidates. *J Mol Med (Berl)* 2012;**90**(10):1197-207

36. Shao Z, Pan X, Li X, Liu W, Han M, Wang C, Wang J, Zheng F, Cao M, Tang J; HtpS, a novel immunogenic cell surface-exposed protein of Streptococcus suis, confers protection in mice. *FEMS Microbiol Lett* 2011;**314**(2):174-82
37. Andrews GP, Strachan ST, Benner GE, Sample AK, Anderson GW, Jr., Adamovicz JJ, Welkos SL, Pullen JK, Friedlander AM; Protective efficacy of recombinant Yersinia outer proteins against bubonic plague caused by encapsulated and nonencapsulated Yersinia pestis. *Infect Immun* 1999;**67**(3):1533-7

Appendix I: FASTA sequences for proteins in the BPAD200 dataset (200BPAs and 200 non-BPAs).

>CAA37206.1
MQLVDRVRGAVTGMSRRLVVGAVGAALVSGLVGAVGGTATAGAFSRPGLPVEYLQVPSPSMGRDIKVQFQ
SGGANSPALYLLDGLRAQDDFSGWDINTPAFEWYDQSGLSVVMPVGGQSSFYSDWYQPACGKAGCQTYKW
ETFLTSELPGWLQANRHVKPTGSAVVGLSMAASSALTLAIYHPQQFVYAGAMSGLLDPSQAMGPTLIGLA
MGDAGGYKASDMWGPKEDPAWQRNDPLLNVGKLIANNTRVWVYCGNGKPSDLGGNNLPAKFLEGFVRTSN
IKFQDAYNAGGGHNGVFDFPDSGTHSWEYWGAQLNAMKPDLQRALGATPNTGPAPQGA

>CAA34403.1
MAQKPNFLKKLISAGLVTASTATIVASFAGSAMGAAIQQNRTTNGAATTVDGAGFDQTAAPANVGVALNA
VITANANNGINFNTPAGSFNGLLLNTANNLAVTVSEDTTLGFITNVVHNAHSFNLTLNAGKTLTITGQGV
TNAQAAATKNAQNVVVQFNNGAAIDNNDLKGVGRIDFGAPASTLVFNLANPTTQKAPLILGDNAVIANGV
NGTLNVTNGFIQVSNKSFATVKAINIADGQGIIFNTDANNANTLNLQAGGTTINFTGTDGTGRLVLLSKH
AAATNFNITGSLGGNLKGVIEFNTVAVDGQLTANAGAANAVIGTNNGAGRAAGFVVSVDNGKVATIDGQV
YAKDMVIQSANATGQVNFRHIVDVGADGTTAFKTAASKVTITQDSNFGNTDFGNLAAQIKVPNAITLTGN
FTGDASNPGNTAGVITFDANGTLESASADANVAVTNNITAIEASGAGVVQLSGTHAAELRLGNAGSIFKL
ADGTVINGKVNQTALVGGALAAGTITLDGSATITGDIGNAGGAAALQRITLANDAKKTLTLGGANIIGAG
GGTIDLQANGGTIKLTSTQNNIVVDFDLAIATDQTGVVDASSLTNAQTLTINGKIGTIGANNKTLGQFNI
GSSKTVLSNGNVAINELVIGNDGAVQFAHDTYLITRTTNAAGQGKIIFNPVVNNGTTLAAGTNLGSATNP
LAEINFGSKGVNVDTVLNVGEGVNLYATNITTTDANVGSFVFNAGGTNIVSGTVGGQQGNKFNTVALENG
TTVKFLGNATFNGNTTIAANSTLQIGGNYTADCVASADGTGIVEFVNTGPITVTLNKQAAPVNALKQITV
SGPGNVVINEIGNAGNHHGAVTDTIAFENSSLGAVVFLPRGIPFNDAGNTMPLTIKSTVGNKTAKGFDVP
SVVVLGVDSVIADGQVIGDQNNIVGLGLGSDNGIIVNATTLYAGISTLNNNQGTVTLSGGVPNTPGTVYG
LGTGIGASKFKQVTFTTDYNNLGNIIATNATINDGVTVTTGGIAGIGFDGKITLGSVNGNGNVRFADGIL
SNSTSMIGTTKANNGTVTYLGNAFVGNIGDSDTPVASVRFTGSDSGAGLQGNIYSQVIDFGTYNLGIVNS
NIILGGGTTAINGKIDLVTNTLTFASGTSTWGNNTSIETTLTLANGNIGHIVILEGAQVNTTTTGTTTIK
VQDNANANFSGTQTYTLIQGGARFNGTLGSPNFAVTGSNRFVNYSLIRAANQDYVITRTNNAENVVTNDI
ANSPFGGAPGVDQNVTTFVNATNTAAYNNLLLAKNSANSANFVGAIVTDTSAAITNVQLDLAKDIQAQLG
NRLGALRYLGTPETAEMAGPEAGAISAAVAAGDEAIDNVAYGIWAKPFYTDAHQSKKGGLAGYKAKTTGV
VIGLDTLANDNLMIGAAIGITKTDIKHQDYKKGDKTDVNGFSFSLYGAQQLVKNFFAQGSAIFSLNQVKN
KSQRYFFDANGNMSKQIAAGHYDNMTFGGNLTVGYDYNAMQGVLVTPMAGLSYLKSSDENYKETGTTVAN
KQVNSKFSDRTDLIVGAKVAGSTMNITDLAVYPEVHAFVVHKVTGRLSKTQSVLDGQVTPCINQPDRTTK
TSYNLGLSASIRSDAKMEYGIGYDAQISSKYTAHQGTLKVRVNF

>P27053.3
MGFRINTNVAALNAKANSDLNSRALDQSLSRLSSGLRINSAADDASGMAIADSLRSQANTLGQAISNGND
ALGILQTADKAMDEQLKILDTIKTKATQAAQDGQSLKTRTMLQADINRLMEELDNIANTTSFNGKQLLSG
GFTNQEFQIGSSSNQTIKASIGATQSSKIGVTRFETGSQSFSSGTVGLTIKNYNGIEDFKFDSVVISTSV
GTGLGALAEEINRNADKTGIRATFDVKSVGAYAIKAGNTSQDFAINGVVIGKVDYSDGDENGSLISAINA
VKDTTGVQASKDENGKLVLTSADGRGIKITGSIGVGAGILHTENYGRLSLVKNDGRDINISGTGLSAIGM
GATDMISQSSVSLRESKGQISAANADAMGFNAYNGGGAKQIIFASSIAGFMSQAGSGFSAGSGFSVGSGK
NYSAILSASIQIVSSARSISSTYVVSTGSGFSAGSGNSQFAALRISTVSAHDETAGVTTLKGAMAVMDIA
ETAITNLDQIRADIGSVQNQITSTINNITVTQVNVKSAESQIRDVDFASESANYSKANILAQSGSYAMAQ
ANSSQQNVLRLLQ

>P20626.1
MREKVVLFLSIIMAIMLPVGNAAAAPVVHNPASTANRPVYAIAHRVLTTQGVDDAVAIGANALEIDFTAW
GRGWWADHDGIPTSAGATAEEIFKHIADKRKQGANITFTWLDIKNPDYCRDARSVCSINALRDLARKYLE
PAGVRVLYGFYKTVGGPAWKTITADLRDGEAVALSGPAQDVLNDFARSENKILTKQKIADYGYYNINQGF
GNCYGTWNRTCDQLRKSSEARDQGKLGKTFGWTIATGQDARVNDLLGKANVDGLIFGFKITHFYRHADTE
NSFKAIKRWVDKHSATHHLATVADNPW

>AAA23236.1
MKKNLVKSLAIASAVISIYSIVNIVSPTNVIAKEISNTVSNEMSKKASYDNVDTLIEKGRYNTKYNYLKR
MEKYYPNAMAYFDKVTINPQGNDFYINNPKVELDGEPSMNYLEDVYVGKALLTNDTQQEQKLKSQSFTCK
NTDTVTATTTHTVGTSIQATAKFTVPFNETGVSLTTSYSFANTNTNTNSKEITHNVPSQDILVPANTTVE
VIAYLKKVNVKGNVKLVGQVSGSEWGEIPSYLAFPRDGYKFSLSDTVNKSDLNEDGTININGKGNYSAVM
GDELIVKVRNLNTNNVQEYVIPVDKKEKSNDSNIVKYRSLYIKAPGIK

>AAB59097.1
MHLIPHWIPLVASLGLLAGGSSASAAEEAFDLWNECAKACVLDLKDGVRSSRMSVDPAIADTNGQGVLHY
SMVLEGGNDALKLAIDNALSITSDGLTIRLEGGVEPNKPVRYSYTRQARGSWSLNWLVPIGHEKPSNIKV
FIHELNAGNQLSHMSPIYTIEMGDELLAKLARDATFFVRAHESNEMQPTLAISHAGVSVVMAQTQPRREK
RWSEWASGKVLCLLDPLDGVYNYLAQQRCNLDDTWEGKIYRVLAGNPAKHDLDIKPTVISHRLHFPEGGS
LAALTAHQACHLPLETFTRHRQPRGWEQLEQCGYPVQRLVALYLAARLSWNQVDQVIRNALASPGSGGDL
GEAIREQPEQARLALTLAAAESERFVRQGTGNDEAGAANADVVSLTCPVAAGECAGPADSGDALLERNYP
TGAEFLGDGGDVSFSTRGTQNWTVERLLQAHRQLEERGYVFVGYHGTFLEAAQSIVFGGVRARSQDLDAI
WRGFYIAGDPALAYGYAQDQEPDARGRIRNGALLRVYVPRSSLPGFYRTSLTLAAPEAAGEVERLIGHPL
PLRLDAITGPEEEGGRLETILGWPLAERTVVIPSAIPTDPRNVGGDLDPSSIPDKEQAISALPDYASQPG
KPPREDLK

>AAA26380.1
MANISPKLFKKAIQQGLKAALFTTSTAAIMLSSSGALGVATGVIATNNNAAFSNNVGNNNWNEITAAGVA
NGTPAGGPQNNWAFTYGGDYTVTADAADRIIKAINVAGTTPVGLNITQNTVVGSIITKGNLLPVTLNAGK
SLTLNGNNAVAANHGFDAPADNYTGLGNIALGGANAALIIQSAAPSKITLAGNIDGGGIITVKTDAAING
TIGNTNALATVNVGAGTATLGGAVIKATTTKLTNAASVLTLTNANAVLTGAIDNTTGGDNVGVLNLNGAL
SQVTGDIGNTNSLATISVGAGTATLGGAVIKATTTKLTDAASAVKFTNPVVVTGAIDNTGNANNGIVTFT
GNSTVTGNVGNTNALATVNVGAGLLQVQGGVVKANTINLTDNASAVTFTNPVVVTGAIDNTGNANNGIVT
FTGNSTVTGDIGNTNALATVNVGAGTATLGGAVIKATTTKLTNAASVLTLTNANAVLTGAIDNTTGGDNV
GVLNLNGALSQVTGNIGNTNSLATISVGAGTATLGGAVIKATTTKLTDAASAVKFTNPVVVTGAIDNTGN
ANNGIVTFTGNSTVTGDIGNTNSLATISVGAGTATLGGAVIKATTTKLTNAASVLTLTNANAVLTGAIDN
TTGGDNVGVLNLNGALSQVTGDIGNTNSLATISVGAGTATLGGAVIKATTTKITNAVSAVKFTNPVVVTG
AIDSTGNANNGIVTFTGNSTVTGDIGNTNALATVNVGAGTATLGGAVIKATTTKLTNAASVLTLTNANAV
LTGAIDNTTGGDNVGVLNLNGALSQVTGDIGNTNSLATISVGAGTATLGGAVIKATTTKLTNAASVLTLT
NANAVLTGAVDNTTGGDNVGVLNLNGALSQVTGDIGNTNSLATISVGAGTATLGGAVIKATTTKLTNAAS
VLTLTNANAVLTGAIDNTTGGDNVGVLNLNGALSQVTGDIGNTNSLATISVGAGTATLGGAVIKATTTKL
TDAASAVKFTNPVVVTGAIDNTGNANNGIVTFTGNSTVTGNVGNTNALATVNVGAGLLQVQGGVVKANTI
NLTDNASAVTFTNPVVVTGAIDNTGNANNGIVTFTGNSTVTGNVGNTNALATVNVGAGLLQVQGGVVKAN
TINLTDNASAVTFTNPVVVTGAIDNTGNANNGIVTFTGNSTVTGDIGNTNALATVNVGAGITLQAGGSLA
ANNIDFGARSTLEFNGPLDGGGKAIPYYFKGAIANGNNAILNVNTKLLTASHLTIGTVAEINIGAGNLFT
IDASVGDVTILNAQNINFRARDSVLVLSNLTGVGVNNILLAADLVAPGADEGTVVFNGGVNGLNVGSNVA
GTARNIGDGGGNKFNTLLIYNAVTITDDVNLEGIQNVLINKNADFTSSTAFNAGAIQINDATYTIDANNG
NLNIPAGNIQFAHADAQLVLQNSSGNDRTITLGANIDPDNDDEGIVILNSVTAGKKLTIAGGKTFGGAHK
LQTILFKGAGDCSTAGTTFNTTNIVLDITGQLELGATTANVVLFNDAVQLTQTGNIGGFLDFNAKNGMVT
LNNNVNVAGAVQNTGGTNNGTLIVLGASNLNRVNGIAMLKVGAGNVTIAKGGKVKIGEIQGTGTNTLTLP
AHFNLTGSINKTGGQALKLNFMNGGSVSGVVGTAANSVGDITTAGATSFASSVNAKGTATLGGTTSFANT
FTNTGAVTLAKGSITSFAKNVTATSFVANSATINFSNSLAFNSNITGGGTTLTLGANQVTYTGTGSFTDT
LTLNTTFDGAAKSGGNILIKSGSTLDLSGVSTLALVVTATNFDMNNISPDTKYTVISAETAGGLKPTSKE
NVKITINNDNRFVDFTFDASTLTLFAEDIAADVIDGDFAPGGPLANIPNAANIKKSLELMEDAPNGSDAR
QAFNNFGLMTPLQEADATTHLIQDVVKPSDTIAAVNNQVVASNISSNITALNARMDKVQSGNKGPVSSGD
EDMDAKFGAWISPFVGNATQKMCNSISGYKSDTTGGTIGFDGFVSDDLALGLAYTRADTDIKLKNNKTGD
KNKVESNIYSLYGLYNVPYENLFVEAIASYSDNKIRSKSRRVIATTLETVGYQTANGKYKSESYTGQLMA
GYTYMMPENINLTPLAGLRYSTIKDKGYKETGTTYQNLTVKGKNYNTFDGLLGAKVSSNINVNEIVLTPE
LYAMVDYAFKNKVSAIDARLQGMTAPLPTNSFKQSKTSFDVGVGVTAKHKMMEYRINYDTNIGSKYFAQQ
GSVKVRVNF

>AAA20874.1
MNKNVLKFMVFIMLLNIITPLFNKNEAFAARDISSTNVTDLTVSPSKIEDGGKTTVKMTFDDKNGKIQNG
DMIKVAWPTSGTVKIEGYSKTVPLTVKGEQVGQAVITPDGATITFNDKVEKLSDVSGFAEFEVQGRNLTQ

TNTSDDKVATITSGNKSTNVTVHKSEAGTSSVFYYKTGDMLPEDTTHVRWFLNINNEKSYVSKDITIKDQ
IQGGQQLDLSTLNINVTGTHSNYYSGQSAITDFEKAFPGSKITVDNTKNTIDVTIPQGYGSYNSFSINYK
TKITNEQQKEFVNNSQAWYQEHGKEEVNGKSFNHTVHNINANAGIEGTVKGELKVLKQDKDTKAPIANVK
FKLSKKDGSVVKDNQKEIEIITDANGIANIKALPSGDYILKEIEAPRPYTFDKDKEYPFTMKDTDNQGYF
TTIENAKAIEKTKDVSAQKVWEGTQKVKPTIYFKLYKQDDNQNTTPVDKAEIKKLEDGTTKVTWSNLPEN
DKNGKAIKYLVKEVNAQGEDTTPEGYTKKENGLVVTNTEKPIETTSISGEKVWDDKDNQDGKRPEKVSVN
LLANGEKVKTLDVTSETNWKYEFKDLPKYDEGKKIEYTVTEDHVKDYTTDINGTTITNKYTPGETSATVT
KNWDDNNNQDGKRPTEIKVELYQDGKATGKTAILNESNNWTHTWTGLDEKAKGQQVKYTVEELTKVKGYT
THVDNNDMGNLIVTNKYTPETTSISGEKVWDDKDNQDGKRPEKVSVNLLADGEKVKTLDVTSETNWKYEF
KDLPKYDEGKKIEYTVTEDHVKDYTTDINGTTITNKYTPGETSATVTKNWDDNNNQDGKRPTEIKVELYQ
DGKATGKTAILNESNNWTHTWTGLDEKAKGQQVKYTVEELTKVKGYTTHVDNNDMGNLIVTNKYTPETTS
ISGEKVWDDKDNQDGKRPEKVSVNLLANGEKVKTLDVTSETNWKYEFKDLPKYDEGKKIEYTVTEDHVKD
YTTDINGTTITNKYTPGETSATVTKNWDDNNNQDGKRPTEIKVELYQDGKATGKTAILNESNNWTHTWTG
LDEKAKGQQVKYTVDELTKVNGYTTHVDNNDMGNLIVTNKYTPKKPNKPIYPEKPKDKTPPTKPDHSNKV
KPTPPDKPSKVDKDDQPKDNKTKPENPLKELPKTGMKIITSWITWVFIGILGLYLILRKRFNS

>CAA48133.1
MNNKMFLNKEAGFLAHTKRKRRFAVTLVGVFFMLLASAGAIGFGQVAYAADEKTVPHRVSQNPEFPWYGY
DFYKGPYTRYHNLQLNLNGSKTYQAYCFNLKRFEPKKEGSYFPNWYKRWDGSEETFVKYADNPRKDNESS
RVIDVELEKNILRVLYNGYPNNGNGIMEGLEPLNAILVTQNAVWYYSDNSSIFNTDNFFTTEAKDLNIKP
EQLSLMRVALKKLIDPKLSEESLKPVPSTFRLNIFESQDKLYQNLLSAEFVPENPPKPGETPEHGPKTPE
LDGTPIPEGPQRPNESLEPTLPPVMLDGQEVPEVPSESLEPALPPLMPELDGQEVPEVPSESLEPALPPL
MPELDGQEVPEKPSVDLPIEDPRYEFNNKDQSPLAGESGETEYITEVYGNQQNPVDIDKKLPNETGFSGN
MVETEDTKEPGVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPGVLMGGQSESVEFTKDTQTGMSG
QTASQVETEDTKEPGVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPGVLMGGQSESVEFTKDTQT
GMSGFSETVTIVEDTRPKLVFHFDNNEPKVEENREKPTKNITPILPATGDIENVLAFLGILILSVLPIFS
LLKKQTKQ

>AAA67928.1
MKTFEILKHLQADAIVLFMKVHNFHWNVKGTDFFNVHKATEEIYEEFADMFDDLAERLVQLGHHPLVTLS
EALKLTRVKDETKTSFHSKDIFKEILGDYKHLEKEFKELSNTAEKEGDKVTVTYADDQLAKLQKSIWMLE
AHLA

>CAA88138.1
MKLNRFGAAVGVLAAGALVLSACGNDDNVTGGGATTGQASAKVDCGGKKTLKASGSTAQANAMTRFVNVF
EQACPGQTLNYTANGSGAGISEFNGNQTDFGGSDVPLSKDEPSGAARCGSPAWNLPVVFGPIAVTYNLNS
VSSLNLDGPTLAKIFNGSITQWNNPAIQALNRDFTLPGERIHVVFRSDESGTTDNFQRYLQAASNGAWGK
GAGKSFQGGVGEGARGNDGTSAAAKNTPGSITYNEWSFAQAHDLTMANIVTSAGGDPVAITIDSVGQTIA
GATISGVGNDLVLDTDSFYRPKRPGSYPIVLATYEIVCSKYPDSQVGTAVKAFLQSTIGAGQSGLGDNGY
IPIPDEFKSRLSTAVNAIA

>AAC36000.1
MKKALATLIALALPAAALAEGASGFYVQADAAHAKASSSLGSAKGFSPRISAGYRINDLRFAVDYTRYKN
YKAPSTDFKLYSIGASAIYDFDTQSPVKPYLGARLSLNRASVDLGGSDSFSQTSIGLGVLTGVSYAVTPN
VDLDAGYRYNYIGKVNTVKNVRSGELSVGVRVKF

>AAB07068.1
MKKLLKSVLAFAVLGSASSLHALPVGNPAEPSLMIDGILWEGFGGDPCDPCTTWCDAISLRLGYYGDFVF
DRVLKTDVNKQFEMGAAPTGDADLTTAPTPASRENPAYGKHMQDAEMFTNAAYMALNIWDRFDVFCTLGA
TSGYLKGNSAAFNLVGLFGRDETAVAADDIPNVSLSQAVVELYTDTAFAWSVGARAALWECGCATLGASF
QYAQSKPKVEELNVLCNAAEFTINKPKGYVGQEFPLNIKAGTVSATDTKDASIDYHEWQASLALSYRLNM
FTPYIGVKWSRASFDADTIRIAQPKLETSILKMTTWNPTISGSGIDVDTKITDTLQIVSLQLNKMKSRKS
CGLAIGTTIVDADKYAVTVETRLIDERAAHVNAQFRF

>AAB09440.1
MKKLGTLLVLFLSAIILVACASGKKDTTSGQKLKVVATNSIIADITKNIAGDKIDLHSIVPIGQDPHEYE
PLPEDVKKTSEADLIFYNGINLETGGNAWFTKLVENAKKTENKDYFAVSDGVDVIYLEGQNEKGKEDPHA
WLNLENGIIFAKNIAKQLSAKDPNNKEFYEKNLKEYTDKLDKLDKESKDKFNKIPAEKKLIVTSEGAFKY
FSKAYGVPSAYIWEINTEEEGTPEQIKTLVEKLRQTKVPSLFVESSVDDRPMKTVSQDTNIPIYAQIFTD
SIAEQGKEGDSYYSMMKYNLDKIAEGLAK

>CAA63564.1
MSLISKEELIKLAYSIRPRENEYKTILTNLDEYNKLTTNNNENKYLQLKKLNESIDVFMNKYKTSSRNRA
LSNLKKDILKEVILIKNSNTSPVEKNLHFVWIGGEVSDIALEYIKQWADINAEYNIKLWYDSEAFLVNTL
KKAIVESSTTEALQLLEEEIQNPQFDNMKFYKKRMEFIYDRQKRFINYYKSQINKPTVPTIDDIIKSHLV
SEYNRDETVLESYRTNSLRKINSNHGIDIRANSLFTEQELLNIYSQELLNRGNLAAASDIVRLLALKNFG
GVYLDVDMLPGIHSDLFKTISRPSSIGLDRWEMIKLEAIMKYKKYINNYTSENFDKLDQQLKDNFKLIIE
SKSEKSEIFSKLENLNVSDLEIKIAFALGSVINQALISKQGSYLTNLVIEQVKNRYQFLNQHLNPAIESD
NNFTDTTKIFHDSLFNSATAENSMFLTKIAPYLQVGFMPEARSTISLSGPGAYASAYYDFINLQENTIEK
TLKASDLIEFKFPENNLSQLTEQEINSLWSFDQASAKYQFEKYVRDYTGGSLSEDNGVDFNKNTALDKNY
LLNNKIPSNNVEEAGSKNYVHYIIQLQGDDISYEATCNLFSKNPKNSIIIQRNMNESAKSYFLSDDGESI
LELNKYRIPERLKNKEKVKVTFIGHGKDEFNTSEFARLSVDSLSNEISSFLDTIKLDISPKNVEVNLLGC
NMFSYDFNVEETYPGKLLLSIMDKITSTLPDVNKNSITIGANQYEVRINSEGRKELLAHSGKWINKEEAI
MSDLSSKEYIFFDSIDNKLKAKSKNIPGLASISEDIKTLLLDASVSPDTKFILNNLKLNIESSIGDYIYY
EKLEPVKNIIHNSIDDLIDEFNLLENVSDELYELKKLNNLDEKYLISFEDISKNNSTYSVRFINKSNGES
VYVETEKEIFSKYSEHITKEISTIKNSIITDVNGNLLDNIQLDHTSQVNTLNAAFFIQSLIDYSSNKDVL
NDLSTSVKVQLYAQLFSTGLNTIYDSIQLVNLISNAVNDTINVLPTITEGIPIVSTILDGINLGAAIKEL
LDEHDPLLKKELEAKVGVLAINMSLSIAATVASIVGIGAEVTIFLLPIAGISAGIPSLVNNELILHDKAT
SVVNYFNHLSESKKYGPLKTEDDKILVPIDDLVISEIDFNNNSIKLGTCNILAMEGGSGHTVTGNIDHFF
SSPSISSHIPSLSIYSAIGIETENLDFSKKIMMLPNAPSRVFWWETGAVPGLRSLENDGTRLLDSIRDLY
PGKFYWRFYAFFDYAITTLKPVYEDTNIKIKLDKDTRNFIMPTITTNEIRNKLSYSFDGAGGTYSLLLSS
YPISTNINLSKDDLWIFNIDNEVREISIENGTIKKGKLIKDVLSKIDINKNKLIIGNQTIDFSGDIDNKD
RYIFLTCELDDKISLIIEINLVAKSYSLLLSGDKNYLISNLSNTIEKINTLGLDSKNIAYNYTDESNNKY
FGAISKTSQKSIIHYKKDSKNILEFYNDSTLEFNSKDFIAEDINVFMKDDINTITGKYYVDNNTDKSIDF
SISLVSKNQVKVNGLYLNESVYSSYLDFVKNSDGHHNTSNFMNLFLDNISFWKLFGFENINFVIDKYFTL
VGKTNLGYVEFICDNNKNIDIYFGEWKTSSSKSTIFSGNGRNVVVEPIYNPDTGEDISTSLDFSYEPLYG
IDRYINKVLIAPDLYTSLININTNYYSNEYYPEIIVLNPNTFHKKVNINLDSSSFEYKWSTEGSDFILVR
YLEESNKKILQKIRIKGILSNTQSFNKMSIDFKDIKKLSLGYIMSNFKSFNSENELDRDHLGFKIIDNKT
YYYDEDSKLVKGLININNSLFYFDPIEFNLVTGWQTINGKKYYFDINTGAALTSYKIINGKHFYFNNDGV
MQLGVFKGPDGFEYFAPANTQNNNIEGQAIVYQSKFLTLNGKKYYFDNNSKAVTGWRIINNEKYYFNPNN
AIAAVGLQVIDNNKYYFNPDTAIISKGWQTVNGSRYYFDTDTAIAFNGYKTIDGKHFYFDSDCVVKIGVF
STSNGFEYFAPANTYNNNIEGQAIVYQSKFLTLNGKKYYFDNNSKAVTGLQTIDSKKYYFNTNTAEAATG
WQTIDGKKYYFNTNTAEAATGWQTIDGKKYYFNTNTAIASTGYTIINGKHFYFNTDGIMQIGVFKGPNGF
EYFAPANTDANNIEGQAILYQNEFLTLNGKKYYFGSDSKAVTGWRIINNKKYYFNPNNAIAAIHLCTINN
DKYYFSYDGILQNGYITIERNNFYFDANNESKMVTGVFKGPNGFEYFAPANTHNNNIEGQAIVYQNKFLT
LNGKKYYFDNDSKAVTGWQTIDGKKYYFNLNTAEAATGWQTIDGKKYYFNLNTAEAATGWQTIDGKKYYF
NTNTFIASTGYTSINGKHFYFNTDGIMQIGVFKGPNGFEYFAPANTDANNIEGQAILYQNKFLTLNGKKY
YFGSDSKAVTGLRTIDGKKYYFNTNTAVAVTGWQTINGKKYYFNTNTSIASTGYTIISGKHFYFNTDGIM
QIGVFKGPDGFEYFAPANTDANNIEGQAIRYQNRFLYLHDNIYYFGNNSKAATGWVTIDGNRYYFEPNTA
MGANGYKTIDNKNFYFRNGLPQIGVFKGSNGFEYFAPANTDANNIEGQAIRYQNRFLHLLGKIYYFGNNS
KAVTGWQTINGKVYYFMPDTAMAAAGGLFEIDGVIYFFGVDGVKAPGIYG

>CAA71822.1
MKKMLLSIFTTFVAVFLAACGGNSDSGASNSLERIKQDGVVRIGVFGDKPPFGYVDEKGVNQGYDIVLAK
RIAKELLGDENKVQFVLVEAANRVEFLKSNKVDIILANFTQTPERAEQVDFCLPYMKVALGVAVPQDSNI
SSIEDLKDKTLLLNKGTTADAYFTKEYPDIKTLKYDQNTETFAALIDQRGDALSHDNTLLFAWVKEHPEF
KMAIKELGNKDVIAPAVKKGDKELKEFIDNLITKLGEEQFFHKAYDETLKSHFGDDVKADDVVIEGGKI

>CAA88137.1
MKFARSGAAVSLLAAGTLVLTACGGGTNSSSSGAGGTSGSVHCGGKKELHSSGSTAQENAMEQFVYAYVR
SCPGYTLDYNANGSGAGVTQFLNNETDFAGSDVPLNPSTGQPDRSAERCGSPAWDLPTVFGPIAITYNIK
GVSTLNLDGPTTAKIFNGTITVWNDPQIQALNSGTDLPPTPISVIFRSDKSGTSDNFQKYLDGASNGAWG
KGASETFNGGVGVGASGNNGTSALLQTTDGSITYNEWSFAVGKQLNMAQIITSAGPDPVAITTESVGKTI
AGAKIMGQGNDLVLDTSSFYRPTQPGSYPIVLATYEIVCSKYPDATTGTAVRAFMQAAIGPGQEGLDQYG
SIPLPKSFQAKLAAAVNAIS

>AAC16068.1
MVNKDVKQTTAFGAPVWDDNNVITAGPRGPVLLQSTWFLEKLAAFDRERIPERVVHAKGSGAYGTFTVTK
DITKYTKAKIFFKVGKKTECFFRFFTVAGERGSADAVRDPRGFAMKYYTEEGNWDLVGNNTPVFFIRDAI
KFPDFIHTQKRDPQTNLPNHDMVWDFWSNVPESLYQVTWVMSDRGIPKSFRHMDGFGSHTFSLINAKGER
FWVKFHFHTMQGVKHLTNEEAAEVRKYDPDSNQRDLFNAIARGDFPKWKLSIQVMPEEDAKKYRFHPFDV
TKIWYLQDYPLMEVGIVELNKNPENYFAEVEQVAFTPANVVPGIGYSPDRMLQGRLFSYGDTHRYRLGVN
YPQIPVNKPRCPFHSSSRDGYMQNGYYGSLQNYTPSSLPGYKEDKSARDPKFNLAHIEKEFEVWNWDYRA
DDSDYYTQPGDYYRSLPADEKERLHDTIGESLAHVTHKEIVDKQLEHFKKADPKYAEGVKKALEKHQKMM
KDMHGKDMHHTKKKK

>AAC28879.1
MSSGNILWGSQNPIVFKNSFGVSNADTGSQDDLSQQNPFAEGYGVLLILLMVIQAIANNKFIEVQKNAER
ARNTQEKSNEMDEVIAKAAKGDAKTKEEVPEDVIKYMRDNGILIDGMTIDDYMAKYGDHGKLDKGGLQAI
KAALDNDANRNTDLMSQGQITIQKMSQELNAVLTQLTGLISKWGEISSMIAQKTYS

>AAD04290.1
MEIQQTHRKINRPLVSLALVGALVSITPQQSHAAFFTTVIIPAIVGGIATGAAVGTVSGLLSWGLKQAEE
ANKTPDKPDKVWRIQAGRGFNNFPHKEYDLYKSLLSSKIDGGWDWGNAARHYWVKGGQWNKLEVDMKDAV
GTYKLSGLINFTGGDLDVNMQKATLRLGQFNGNSFTSYKDSADRTTRVDFNAKNILIDNFLEINNRVGSG
AGRKASSTVLTLQASEGITSSKNAEISLYDGATLNLASSSVKLMGNVWMGRLQYVGAYLAPSYSTINTSK
VTGEVNFNHLTVGDHNAAQAGIIASNKTHIGTLDLWQSAGLNIIAPPEGGYKDKPKDKPSNTTQNNANNN
QQNSAQNNNNTQVINPPNSAQKTEIQPTQVINGPFAGGKDTVVNINRINTNADGTIRVGGYKASLTTNAA
HLHIGKGGINLSNQASGRSLLVENLTGNITVDGPLRVNNQVGGYALAGSNANFEFKAGTDTKNGTATFNN
DISLGRFVNLKVDAHTANFKGIDTGNGGFNTLDFSGVTDKVNINKLITASTNVAIKNFNINELLVKTNGV
SVGEYTHFSEDIGSQSRINTVRLETGTRSIFSGGVKFKSGEKLVIDEFYYSPWNYFDARNIKNVEITRKF
ASSTPENPWGTSKLMFNNLTLGQNAVMDYSQFSNLTIQGDFINNQGTINYLVRGGKVATLNVGNAAAMMF
NNDIDSATGFYKPLIKINSAQDLIKNTEHVLLKAKIIGYGNVSTGTNGISNVNLEEQFKERLALYNNNNR
MDTCVVRNTDDIKACGMAIGNQSMVNNPDNYKYLIGKAWKNIGISKTANGSKISVYYLGNSTPTENGGNT
TNLPTNTTNNARSANYALVKNAPFAHSATPNLVAINQHDFGTIESVFELANRSKDIDTLYTHSGAKGRDL
LQTLLIDSHDAGYARQMIDNTSTGEITKQLNAATTTLNNIASLEHKTSSLQTLSLSNAMILNSRLVNLSR
KHTNNIDSFAKRLQALKDQRFASLESAAEVLYQFAPKYEKPTNVWANAIGGASLNNGSNASLYGTSAGVD
AYLNGQVEAIVGGFGSYGYSSFSNRANSLNSGANNTNFGVYSRIFANQHEFDFEAQGALGSDQSSLNFKS
ALLQDLNQSYNYLAYSAATRASYGYDFAFFKNALVLKPSVGVSYNHLGSTNFKSNSTNKVALSNGSSSQH
LFNASANVEARYYYGDTSYFYMNAGVLQEFANFGSSNAVSLNTFKVNAARNPLNTHARVMMGGELQLAKE
VFLNLGFVYLHNLISNIGHFASNLGMRYSF

>AAD23967.1
MKNIAKVTALALGIALASGYASAEEKIAFINAGYIFQHHPDRQAVADKLDAEFKPVAEKLAASKKEVDDK
IAAARKKVEAKVAALEKDAPRLRQADIQKRQEEINKLGAAEDAELQKLMQEQDKKVQEFQAQNEKRQAEE
RGKLLDSIQTATNNLARAKGYTYVLDANSVVFAVEGKDITEEVLKSIPASEKAQEKK

>AAD45811.1
MHTFRPYSLRHSDLLYEDIPLEIREQIILLIINTLGNCSSFYDMTLYCYHNSHSDEVYRRICKTLRKEYG
LFTL

>AAD51067.1
MKVKYLMLTLVGAIALNASAQENTVPATGQLPAKNVAFARNKAGRNWFVTLQGGVAAQFLNDNNNKDLMD
RLGAIGSLSVGKYHSPFFATRLQINGGQAHTFLGKNGEQEINTNFGAAHFDFMFDVVNYFAPYRENRFFH
LIPWVGVGYQHKFIGSEWSKDNVESLTANVGVMMAFRLGKRVDFVIEAQAAHSNLNLSRAYNAKKTPVFE
DPAGRYYNGFQGMATAGLNFRLGAVGFNAIEPMDYALINDLNGQINRLRSEVEELSKRPVSCPECPEVTP
VTKTENILTEKAVLFRFDSHVVDKDQLINLYDVAQFVKETNEPVTVVGYADPTGNTQYNEKLSERRAKAV
VDVLTGKYGVPSELISVEWKGDSTQPFSKKAWNRVVIVRSK

>AAD51068.1
MKAKSLLLALAGLACTFSATAQEATTQNKAGMHTAFQRDKASDHWFIDIAGGAGMALSGWNNDVDFVDRL
SIVPTFGIGKWHEPYFGTRLQFTGFDIYGFPQGSKERNHNYFGNAHLDFMFDLTNYFGVYRPNRVFHIIP
WAGIGFGYKFHSENANGEKVGSKDDMTGTVNVGLMLKFRLSRVVDFNIEGQAFAGKMNFIGTKRGKADFP
VMATAGLTFNLGKTEWTEIVPMDYALVNDLNNQINSLRGQVEELSRRPVSCPECPEPTQPTVTRVVVDNV
VYFRINSAKIDRNQEINVYNTAEYAKTNNAPIKVVGYADEKTGTAAYNMKLSERRAKAVAKMLEKYGVSA
DRITIEWKGSSEQIYEENAWNRIVVMTAAE

>AAD51621.1
MADLAKIVEDLSALTVLEAAELSKLLEEKWGVSAAAPVAVAAAGGAAPAAAAEEKTEFDVVLADGGANKI
NVIKEVRALTGLGLKEAKDLVEGAPKAVKEGASNDEAEKIKAQLEAAGAKVELK

>AAF17598.1
MTNETIDQQPQTEAAFNPQQFINNLQVAFLKVDNAVASYDPDQKPIVDKNDRDNRQAFDGISQLREEYSN
KAIKNPTKKNQYFSDFINKSNDLINKDNLIDIGSSIKSFQKFGTQRYRIFTSWVSHQNDPSKINTRSIRN
FMENIIQPPIPDDKEKAEFLKSAKQSFAGIIIGNQIRTDQKFMGVFDEFLKERQEAEKNGEPTGGDWLDI
FLSFVFNKEQSSDVKEAINQEPVPHVQPDIATTTTHIQGLPPESRDLLDERGNFSKFTLGDMEMLDVEGV
ADIDPNYKFNQLLIHNNALSSVLMGSHNGIEPEKVSLLYAGNGGFGAKHDWNATVGYKNQQGDNVATLIN
VHMKNGSGLVIAGGEKGINNPSFYLYKEDQLTGSQRALSQEEIRNKIDFMEFLAQNNAKLDNLSEKEKEK
FQNEIEDFQKDSKAYLDALGNDRIAFVSKKDPKHSALITEFGKGDLSYTLKDYGKKADRALDREKNVTLQ
GNLKHDSVMFVNYSNFKYTNASKSPDKGVGVTNGVSHLDAGFSKVAVFNLPDLNNLAITSFVRRNLENKL
VTEGLSLQEANKLIKDFLSSNKELVGKALNFNKAVADAKNTGNYDEVKKAQKDLEKSLRKREHLEKEVEK
KLESKSGNKNKMEAKAQANSQKDKIFALINKEANRDARAIAYSQNLKGIKRELSDKLEKINKDLKDFSKS
FDEFKNGKNKDFSKAEETLKALKGSVKDLGINPEWISKVENLNAALNEFKNGKNKDFSKVTQAKSDLENS
VKDVIINQKITDKVDNLNQAVSMAKATGDFSRVEQALADLKNFSKEQLAQQTQKNESFNVGKKSEIYQSV
KNGVNGTLVGNGLSGIEATALAKNFSDIKKELNEKFKNFNNNNNNGLENEPIYAKVNKKKTGQVASPEEP
IYAQVAKKVNAKIDRLNQAASGLGGVGQAGFPLKRHDKVDDLSKVGRSVSPEPIYATIDDLGGPFPLKRH
DKVDDLSKVGRSVSPEPIYATIDDLGGPFPLKRHDKVDDLSKVGRSVSPEPIYATIDDLGGPFPLKRHDK
VDDLSKVGLSRNQELAQKIDNLSQAVSEAKAGFFSNLEQTIDKLKDSTKYNSVNLWVESAKKVPASLSAK
LDNYATNSHTRINSNIQNGAINEKATGMLTQKNPEWLKLVNDKIVAHNVGSVPLSEYDKIGFNQKNMKDY
SDSFKFSTKLNNAVKDVKSSFTQFLANAFSTGYYSLARENAEHGIKNVNTKGGFQKS

>AAA99670.2
MLGRDGEARLCRRPTAAAWSSVAGTAPGQDVSSPIRRNHFAMAKTIAYDEEARRGLERGLNALADAVKVT
LGPKGRNVVLEKKWGSPTITNDGVSIAKEIELEDPYEKIGAELVKEVAKKTDDVAGDGTTTATVLAQALV
REGLRNVAAGANPLGLKRGIEKAVEKVTETLLKSAKEVETKDQIAATAAISAGDQSIGDLIAEAMDKVGN
EGVITVEESNTFGLQLELTEGMRFDKGYISGYFVTDAERQEAVLEDPFILLVSSKVSTVKDLLPLLEKVI
QAGKPLLIIAEDVEGEALSTLVVNKIRGTFKSVAVKAPGFGDRRKAMLQDMAILTGGQVISEEVGLSLES
ADISLLGKARKVVVTKDETTIVEGAGDSDAIAGRVAQIRTEIENSDSDYDREKLQERLAKLAGGVAVIKA
GAATEVELKERKHRIEDAVRNAKAAVEEGIVAGGGVALLHAIPALDELKLEGEEATGANIVRVALERPLK
QIAFNGGLEPGVVAEKVRNSPAGTGLNAATGKYEDLLKAGITEPVKVTRSALQNAASISGLFLTTEAVVA
DKPEKTAPPAGDPTGGMGGMDF

>AAF45140.1
MIDPSATSRYGSPRLVSNGFRHRRKVVYQRVGHRRFSLIFFFVVVLGRSPRLWAQVSFTPDIEGYAELAW
GIASDRGALKHGFKTTTDFKIVFPIVAKKDFKYRGEGNVYAEINVKALKLSLESNGGAKFDTKGSAKTIE
ATLHCYGAYLTIGKNPDFKSTFAVLWEPWTANGDYKSKGDKPVYEPGFEGAGGKLGYKQTDIAGTGLTFD

IAFKFASNTDWEGKDSKGNVPAGVTPSKYGLGGDILFGWERTREDGVQEYIKVELTGNSTLSSDYAQARA
PAAGAKVSMKLWGLCALAATDVGHKKNGAQGTVGADALLTLGYRWFSAGGYFASQASNVFGGVFLNMAMR
EHDCAAYIKLETKGSDPDTSFLEGLDLGVDVRTYMPVHYKVLKALPRADIHFPVYGKVWGSYRHDMGEYG
WVKVYANLYGGTNKKATPPAAPATKWSKEYCGYYECGVVVSPLEKVEIRLSWEQGKLQENSNVVIEKNVT
ERWQFVGACRLIW

>AAF81744.1
MQQQHLFRFNILCLSLMTALPAYAENVQAGQAQEKQLDTIQVKAKKQKTRRDNEVTGLGKLVKSSDTLSK
EQVLNIRDLTRYDPGIAVVEQGRGASSGYSIRGMDKNRVSLTVDGVSQIQSYTAQAALGGTRTAGSSGAI
NEIEYENVKAVEISKGSNSVEQGSGALAGSVAFQTKTADDVIGEGRQWGIQSKTAYSGKNRGLTQSIALA
GRIGGAEALLIHTGRRAGEIRAHEDAGRGVQSFNRLVPVEDSSNYAYFIVKEECKNGSYETCKANPKKDV
VGKDERQTVSTRDYTGPNRFLADPLSYESRSWLFRPGFRFENKRHYIGGILEHTQQTFDTRDMTVPAFLT
KAVFDANKKQAGSLPGNGKYAGNHKYGGLFTNGENGALVGAEYGTGVFYDETHTKSRYGLEYVYTNADKD
TWADYARLSYDRQGVGLDNHFQQTHCSADGSDKYCRPSADKPFSYYKSDRVIYGESHRLLQAAFKKSFDT
AKIRHNLSVNLGFDRFGSNLRHQDYYYQHANRAYSSNTPPQNNGKKISPNGSETSPYWVTIGRGNVVTGQ
ICRLGNNTYTDCTPRSINGKSYYAAVRDNVRLGRWADVGAGLRYDYRSTHSDDGSVSTGTHRTLSWNAGI
VLKPTDWLDLTYRTSTGFRLPSFAEMYGWRAGVQSKAVKIDPEKSFNKEAGIVFKGDFGNLEASWFNNAY
RDLIVRGYEAQIKDGKEEAKGDPAYLNAQSARITGINILGKIDWNGVWDKLPEGWYSTFAYNRVRVRDIK
KRADRTDIQSHLFDAIQPSRYVVGLGYDQPEGKWGVNGMLTYSKAKEITELLGSRALLNGNSRNTKATAR
RTRPWYIVDVSGYYTVKKHFTLRAGVYNLLNYRYVTWENVRQTAGGAVNQHKNVGVYNRYAAPGRNYTFS
LEMKF

>AAF81745.1
MNNPLVNQAAMVLPVFLLSACLGGGGSFDLDSVDTEAPRPAPKYQDVFSEKPQAQKDQGGYGFAMRLKRR
NWYPQAKEDEVKLDESDWEATGLPDEPKELPKRQKSVIEKVETDSDNNIYSSPYLKPSNHQNGNTGNGIN
QPKNQAKDYENFKYVYSGWFYKHAKREFNLKVEPKSAKNGDDGYIFYHGKEPSRQLPASGKITYKGVWHF
ATDTKKGQKFREIIQPSKSQGDRYSGFSGDDGEEYSNKNKSTLTDGQEGYGFTSNLEVDFHNKKLTGKLI
RNNANTDNNQATTTQYYSLEAQVTGNRFNGKATATDKPQQNSETKEHPFVSDSSSLSGGFFGPQGEELGF
RFLSDDQKVAVVGSAKTKDKPANGNTAAASGGTDAAASNGAAGTSSENGKLTTVLDAVELKLGDKKVQKL
DNFSNAAQLVVDGIMIPLLPEASESGNNQANQGTNGGTAFTRKFDHTPESDKKDAQAGTQTNGAQTASNT
AGDTNGKTKTYEVEVCCSNLNYLKYGMLTRKNSKSAMQAGESSSQADAKTEQVEQSMFLQGERTDEKEIP
SEQNIVYRGSWYGYIANDKSTSWSGNASNATSGNRAEFTVNFADKKITGTLTADNRQEATFTIDGNIKDN
GFEGTAKTAESGFDLDQSNTTRTPKAYITDAKVQGGFYGPKAEELGGWFAYPGDKQTKNATNASGNSSAT
VVFGAKRQQPVQ

>AAG18474.1
MKMNKKVLLTSTMAASLLSVASVQAQETDTTWTARTVSEVKADLVKQDNKSSYTVKYGDTLSVISEAMSI
DMNVLAKINNIADINLIYPETTLTVTYDQKSHTATSMKIETPATNAAGQTTATVDLKTNQVSVADQKVSL
NTISEGMTPEAATTIVSPMKTYSSAPALKSKEVLAQEQAVSQAAANEQVSTAPVKSITSEVPAAKEEVKP
TQTSVSQSTTVSPASVAAETPAPVAKVAPVRTVAAPRVASVKVVTPKVETGASPEHVSAPAVPVTTTSTA
TDSKLQATEVKSVPVAQKAPTATPVAQPASTTNAVAAHPENAGLQPHVAAYKEKVASTYGVNEFSTYRAG
DPGDHGKGLAVDFIVGKNQALGNEVAQYSTQNMAANNISYVIWQQKFYSNTNSIYGPANTWNAMPDRGGV
TANHYDHVHVSFNK

>AAG31201.1
MVKIQEVKRILTRWQPSSFSLYREVFTQYGGSINMHPDIVDYFMKRYNWHFKFFHYKEDDKIKGAYFICN
DQNIGILTRRTFPLSSDEILIPMAPDLRCFLPDRTNRLSALHQPQIRNAIWKLARKKQNCLVKETFSSKF
EKTRRNEYQRFLKKSGSVKSVADCSSDELTHIFIELFRSRFGNTSSCYPADNLANFFSQLHHLLFGHILY
IEGIPCAFDIVLKSESQMNIYFDVSNGAIKNECRPLSPGSILMWLNISRARHYCQERQKKLLFSIGILKP
EWEYKRMWSTPYFTGKSIC

>NP_045547.1
MKYHIITTIFVFLFLACRPDFNIDQKDIKYPPTEKSRPKTESSKQKESKPKTEEELKKKQQEEELKKKQQ
EEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKK
QQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELKKKQQEEELK
KKQQEEELKKKQQEEELKKKQQEEELKKKQQEEEKEELRKQQLKNTLSNDLKKQIESAYNFKEKYVKSME
KEPEDHYGMTSFRGLNWGPGTEDISDNTERSIRYRRHTYTVLSPLDPHELKEFANIIQDINKLASVASIF
NSFSAIGGALDIVSDHLYFKKDNLDKLDIADLEILKNSFEQILYIKGSVAGKAKKLLLDYKNLKTDINKL
KSYSNELVNGIKQQALEAENLEELIVSKYKL

>NP_045688.1
MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEMKVLVSKEKNKDGKYDLIATVDKLELKGTSDKN
NGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKSSTEEKFNEKGEVSEKIITRA
DGTRLEYTGIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVVKEGTVTLSKNISKSGEVSVELNDTDSSAAT
KKTAAWNSGTSTLTITVNSKKTKDLVFTKENTITVQQYDSNGTKLEGSAVEITKLDEIKNALK

>AAK19155.1
MKINKKYLVGSAAALILSVCSYELGLYQARTVKENNRVSYIDGKQATQKTENLTPDEVSKREGINAEQIV
IKITDQGYVTSHGDHYHYYNGKVPYDAIISEELLMKDPNYKLKDEDIVNEVKGGYVIKVDGKYYVYLKDA
AHADNVRTKEEINRQKQEHSQHREGGTPRNDGAVALARSQGRYTTDDGYIFNASDIIEDTGDAYIVPHGD
HYHYIPKNELSASELAAAEAFLSGRGNLSNSRTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQ
ASQSNDIDSLLKQLYKLPLSQRHVESDGLVFDPAQITSRTARGVAVPHGDHYHFIPYSQMSELEERIARI
IPLRYRSNHWVPDSRPEQPSPQPTPEPSPGPQPAPNLKIDSNSSLVSQLVRKVGEGYVFEEKGISRYVFA
KDLPSETVKNLESKLSKQESVSHTLTAKKENVAPRDQEFYDKAYNLLTEAHKALFXNKGRNSDFQALDKL
LERLNDESTNKEKLVDDLLAFLAPITHPERLGKPNSQIEYTEDEVRIAQLADKYTTSDGYIFDEHDIISD
EGDAYVTPHMGHSHWIGKDSLSDKEKVAAQAYTKEKGILPPSPDADVKANPTGDSAAAIYNRVKGEKRIP
LVRLPYMVEHTVEVKNGNLIIPHKDHYHNIKFAWFDDHTYKAPNGYTLEDLFATIKYYVEHPDERPHSND
GWGNASEHVLGKKDHSEDPNKNFKADEEPVEETPAEPEVPQVETEKVEAQLKEAEVLLAKVTDSSLKANA
TETLAGLRNNLTLQIMDNNSIMAEAEKLLALLKGSNPSSVSKEKIN

>AAK19156.1
MKKVRFIFLALLFFLASPEGAMASDGTWQGKQYLKEDGSQAANEWVXDTHYQSWFYIKADANYAENEWLK
QGDDYFYLKSGGYMAKSEWVEDKGAFYYLDQDGKMKRNAWVGTSYVGATGAKVIEDWVYDSQYDAWFYIK
ADGQHAEKEWLQIKGKDYYFKSGGYLLTSQWINQAYVNASGAKVQQGWLFDKQYQSWFYIKENGNYADKE
WIFENGHYYYLKSGGYMAANEWIWDKESWFYLKFDGKMAEKEWVYDSHSQAWYYFKSGGYMTANEWIWDK
ESWFYLKSDGKIAEKEWVYDSHSQAWYYFKSGGYMAKNETVDGYQLGSDGKWLGGKTTNENAAYYQVVPV
TANVYDSDGEKLSYISQGSVVWLDKDRKSDDKRLAITISGLSGYMKTEDLQALDASKDFIPYYESDGHRF
YHYVAQNASIPVASHLSDMEVGKKYYSADGLHFDGFKLENPFLFKDLTEATNYSAEELDKVFSLLNINNS
LLENKGATFKEAEEHYHINALYLLAHSALESNWGRSKIAKDKNNFFGITAYDTTPYLSAKTFDDVDKGIL
GATKWIKENYIDRGRTFLGNKASGMNVEYASDPYWGEKIASVMMKINEKLGGKD

>AAK19158.1
MFKRIRRVLVLAVFLFAGYKAYRVHQDVKQVMTYQPMVREILSEQDTPANEELVLAMIYTETKGKEGDVM
QSSESASGSTNTINDNASSIRQGIQTLTGNLYLAQKKGVDIWTAVQAYNFGPAYIDFIAQNGKENTLALA
KQYSRETVAPLLGNRTGKTYSYIHPISIFHGAELYVNGGNYYYSRQVRLNLYIIKCFTLFSTSG

>P17452.2
MKTKHFFNSDFTVKGKSADEIFRRLCTDHPDKQLNNVKWKEVFINRFGQMMLDTPNPRKIVEKIINEGLE
KQGLKNIDPETTYFNIFSSSDSSDGNVFHYNSLSESYRVTDACLMNIFVERYFDDWDLLNSLASNGIYSV
GKEGAYYPDHDYGPEYNPVWGPNEQIYHSRVIADILYARSVWDEFKKYFMEYWQKYAQLYTEMLSDTFLA
MAIQQYTRQTLTDEGFLMVCNTYYGNKEEVQITLLDIYGYPSTDIICIEQKGLPTPKVILYIPGGTQPFV
EFLNTDDLKQWIAWHLKDNKHMVAFRKHFSLKQRQEGETFTGIDKALQYIAEESPEWPANKYILYNPTHL
ETENLFNIMMKRTEQRMLEDSDVQIRSNSEATRDYALSLLETFISQLSAIDMLVPAVGIPINFALSATAL
GLSSDIVVNGDSYEKRKYGIGSLVQSALFTGINLIPVISETAEILSSFSRTEEDIPAFFTEEQALAQRFE
IVEEELHSISPDDPPREITDENLHKIRLVRLNNENQPLVVLRRLGGNKFIRIEPITFQEIKGSLVSEVIN
PVTNKTYYVSNAKLLGGSPYSPFRIGLEGVWTPEVLKARASVIGKPIGESYKRILAKLQRIHNSNILDER
QGLMHELMELIDLYEESQPSSERLNAFRELRTQLEKALYLPEMEALKKQILQIPNKGSGAARFLLRTAMN
EMAGKTSESTADLIRFALQDTVISAPFRGYAGAIPEAIDFPVKYVIEDISVFDKIQTNYWELPAYESWNE

GSNSALLPGLLRESQSKGMLSKCRIIENSLYIGHSYEEMFYSISPYSNQVGGPYELYPFTFFSMLQEVQG
DLGFEQAFATRNFFNTLVSDRLSLMENTMLLTESFDYTPWDAIYGDINYDEQFAAMSINERIEKCMNTYR
GVAFQNSSKSIDFFLNNLTTFIDNGLTEIAISDLPYDIVQQEISQFLQGSNEWKTLDAMLFNLDKGDING
AFRKLLQSAKDNNIKFRAIGHSDNSVPPFNNPYKSLYYKGNIIAEAIEKLDREGQKFVVFADSSLLNSTP
GTGRPMPGLVQYLKIPATVVDSDGAWQFLPDVASSRVPIEVTELENWQVLTPPQGKILGLKQFKLTAGFP
TEQSRLPLLENSVSEDLREELMQKIDAIKNDVKMNSLVCMEAGSCDSVSPKVAARLKDMGLEAGMGASIT
WWRREGGMEFSHQMHTTASFKFAGKEFAVDASHLQFVHDQLDTTILILPVDDWALEIAQRNRAINPFVEY
VSKTGNMLALFMPPLFTKPRLTRAL

>AAK26629.1
MKINKKYLVGSAAALILSVCSYELGLYQARTVKENNRVSYIDGKQATQKTENLTPDEVSKREGINAEQIV
IKITDQGYVTSHGDHYHYYNGKVPYDAIISEELLMKDPNYQLKDEDIISEIKGGYVIKVDGKYYVYLKDA
AHADNVRTKEEINRQKQEHSQHREGGTPRNDGAVALARSQGRYTTDDGYIFNASDIIEDTGDAYIVPHGD
HYHYIPKNELSASELAAAKAFLSGRGNLSNSRTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQ
ASQSNDIDSLLKQLYKLPLSQRHVESDGLIFDPAQITSRTANGVAVPHGDHYHFIPYSQLSPLEEKLARI
IPLRYRSNHWVPDSRPEQPSPQSTPEPSPSPQPAPNPQPAPSNPIDEKLVKEAVRKVGDGYVFEENGVPR
YIPAKDLSAETAAGIDSKLAKQESLSHKLGAKKTDLPSSDREFYNKAYDLLARIHQDLLDNKGRQVDFEA
LDNLLERLKDVSSDKVKLVDDILAFLAPIRHPERLGKPNAQITYTDDEIQVAKLAGKYTTEDGYIFDPRD
ITSDEGDAYVTPHMTHSHWIKKDSLSEAERAAAQAYAKEKGLTPPSTDHQDSGNTEAKGAEAIYNRVKAA
KKVPLDRMPYNLQYTVEVKNGSLIIPHYDHYHNIKFEWFDEGLYEAPKGYSLEDLLATVKYYVEHPNERP
HSDNGFGNASDHVQRNKNGQADTNQTEKPNEEKPQTEKPEEETPREEKPQSEKPESPKPTEEPEEESPEE
SPEESEEPQVETEKVKEKLREAEDLLGKIQNPIIKSNAKETLTGLKNNLLFGTQDNNTIMAEAEKLLALL
KESK

>Q9RB65.2
MKSQFSWLVLSSTLACFTSCSTVFAATAENIGPSDSFDGSTNTGTYTPKNTTTGIDYTLTGDITLQNLGD
SAALTKGCFSDTTESLSFAGKGYSLSFLNIKSSAEGAALSVTTDKNLSLTGFSSLTFLAAPSSVITTPSG
KGAVKCGGDLTFDNNGTILFKQDYCEENGGAISTKNLSLKNSTGSISFEGNKSSATGKKGGAICATGTVD
ITNNTAPTLFSNNIAEAAGGAINSTGNCTITGNTSLVFSENSVTATAGNGGALSGDADVTISGNQSVTFS
GNQAVANGGAIYAKKLTLASGGGGGISFSNNIVQGTTAGNGGAISILAAGECSLSAEAGDITFNGNAIVA
TTPQTTKRNSIDIGSTAKITNLRAISGHSIFFYDPITANTAADSTDTLNLNKADAGNSTDYSGSIVFSGE
KLSEDEAKVADNLTSTLKQPVTLTAGNLVLKRGVTLDTKGFTQTAGSSVIMDAGTTLKASTEEVTLTGLS
IPVDSLGEGKKVVIAASAASKNVALSGPILLLDNQGNAYENHDLGKTQDFSFVQLSALGTATTTDVPAVP
TVATPTHYGYQGTWGMTWVDDTASTPKTKTATLAWTNTGYLPNPERQGPLVPNSLWGSFSDIQAIQGVIE
RSALTLCSDRGFWAAGVANFLDKDKKGEKRKYRHKSGGYAIGGAAQTCSENLISFAFCQLFGSDKDFLVA
KNHTDTYAGAFYIQHITECSGFIGCLLDKLPGSWSHKPLVLEGQLAYSHVSNDLKTKYTAYPEVKGSWGN
NAFNMMLGASSHSYPEYLHCFDTYAPYIKLNLTYIRQDSFSEKGTEGRSFDDSNLFNLSLPIGVKFEKFS
DCNDFSYDLTLSYVPDLIRNDPKCTTALVISGASWETYANNLARQALQVRAGSHYAFSPMFEVLGQFVFE
VRGSSRIYNVDLGGKFQF

>BAB62098.1
MRRAENNKHSRYSIRKLSVGVTSIAIASLFLGKVAYAVDGIPPISLTQKTTATTSENWHHIDKDGLIPLG
ISLEAAKEEFKKEVEESRLSEAQKETYKQKIKTAPDKDKLLFTYHSEYMTAVKDLPASTESTTQPVEAPV
QETQASASDSMVTGDSTSVTTDSPEETPSSESPVAPALSEAPAQPAESEEPSVAASSEETPSPSTPAAPS
TPAAPETPEEPAAPSQPAESEESSVAATTSPSPSTPAESETQTPPAVTKDSDKPSSAAEKPAASSLVSEQ
TVQQPTSKRSSDKKEEQEQSYSPNRSLSRQVRAHESGKYLPSTGEKAQPLFIATMTLMSLFGSLLVTKRQ
KETKK

>AAK92094.1
MNMSLSRIVKAAPLRRTTLAMALGALGAAPAAHADWNNQSIVKTGERQHGIHIQGSDPGGVRTASGTTIK
VSGRQAQGILLENPAAELQFRNGSVTSSGQLSDDGIRRFLGTVTVKAGKLVADHATLANVGDTWDDDGIA
LYVAGEQAQASIADSTLQGAGGVQIERGANVTVQRSAIVDGGLHIGALQSLQPEDLPPSRVVLRDTNVTA
VPASGAPAAVSVLGASELTLDGGHITGGRAAGVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVP
GGFGPGGFGPVLDGWYGVDVSGSSVELAQSIVEAPELGAAIRVGRGARVTVSGGSLSAPHGNVIETGGAR
RFAPQAAPLSITLQAGAHAQGKALLYRVLPEPVKLTLTGGADAQGDIVATELPSIPGTSIGPLDVALASQ
ARWTGATRAVDSLSIDNATWVMTDNSNVGALRLASDGSVDFQQPAEAGRFKVLTVNTLAGSGLFRMNVFA
DLGLSDKLVVMQDASGQHRLWVRNSGSEPASANTLLLVQTPLGSAATFTLANKDGKVDIGTYRYRLAANG
NGQWSLVGAKAPPAPKPAPQPGPQPPQPPQPQPEAPAPQPPAGRELSAAANAAVNTGGVGLASTLWYAES
NALSKRLGELRLNPDAGGAWGRGFAQRQQLDNRAGRRFDQKVAGFELGADHAVAVAGGRWHLGGLAGYTR
GDRGFTGDGGGHTDSVHVGGYATYIADSGFYLDATLRASRLENDFKVAGSDGYAVKGKYRTHGVGASLEA
GRRFTHADGWFLEPQAELAVFRAGGGAYRAANGLRVRDEGGSSVLGRLGLEVGKRIELAGGRQVQPYIKA
SVLQEFDGAGTVHTNGIAHRTELRGTRAELGLGMAAALGRGHSLYASYEYSKGPKLAMPWTFHAGYRYSW

>NP_212737.1
MKSHILYKLIIFLTTSAAIFAADALKEKDIFKINPWMPTFGFENTSEFRLDMDELVPGFENKSKITIKLK
PFEANPELGKDDPFSAYIKVEDLALKAEGKKGDQFKIDVGDITAQINMYDFFIKISTMTDFDFNKESLFS
FAPMTGFKSTYYGFPSNDRAVRGTILARGTSKNIGTIQLGYKLPKLDLTFAIGGTGTGNRNQENDKDTPY
NKTYQGILYGIQATWKPIKNLLDQNEDTKSVIAETPFELNFGLSGAYGNETFNNSSITYSLKDKSVVGND
LLSPTLSNSAILASFGAKYKLGLTKINDKNTYLILQMGTDFGIDPFASDFSIFGHISKAANFKKETPSDP
NKKAEIFDPNGNALNFSKNTELGIAFSTGASIGFAWNKDTGEKESWAIKGSDSYSTRLFGEQDKKSGVAL
GISYGQNLYRSKDTEKRLKTISENAFQSLNVEISSYEDNKKGIINGLGWITSIGLYDILRQKSVENYPTTT
ISSTTENNQTEQSSTSTKTTTPNLTFEDAMKLGLALYLDYAIPIASISTEAYVVPYIGAYILGPSNKLSS
DATKIYLKTGLSLEKLIRFTTISLGWDSNNIIELANKNTNNAAIGSAFLQFKIAYSGS

>NP_250397.1
MEVRNLNAARELFLDELLAASAAPASAEQEELLALLRSERIVLAHAGQPLSEAQVLKALAWLLAANPSAP
PGQGLEVLREVLQARRQPGAQWDLREFLVSAYFSLHGRLDEDVIGVYKDVLQTQDGKRKALLDELKALTA
ELKVYSVIQSQINAALSAKQGIRIDAGGIDLVDPTLYGYAVGDPRWKDSPEYALLSNLDTFSGKLSIKDF
LSGSPKQSGELKGLSDEYPFEKDNNPVGNFATTVSDRSRPLNDKVNEKTTLLNDTSSRYNSAVEALNRFI
QKYDSVLRDILSAI

>NP_250468.1
MKLKNTLGVVIGSLVAASAMNAFAQGQNSVEIEAFGKRYFTDSVRNMKNADLYGGSIGYFLTDDVELALS
YGEYHDVRGTYETGNKKVHGNLTSLDAIYHFGTPGVGLRPYVSAGLAHQNITNINSDSQGRQQMTMANIG
AGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGLGVGFNFGGSKAAPAPEPVADVCSDSDNDGVCD
NVDKCPDTPANVTVDANGCPAVAEVVRVQLDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHT
DSVGTDAYNQKLSERRANAVRDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVEAEVEAEAK

>NP_214989.1
MAENSNIDDIKAPLLAALGAADLALATVNELITNLRERAEETRTDTRSRVEESRARLTKLQEDLPEQLTE
LREKFTAEELRKAAEGYLEAATSRYNELVERGEAALERLRSQQSFEEVSARAEGYVDQAVELTQEALGTV
ASQTRAVGERAAKLVGIELPKKAAPAKKAAPAKKAAPAKKAAAKKAPAKKAAAKKVTQK

>NP_215690.1
MRLSLTALSAGVGAVAMSLTVGAGVASADPVDAVINTTCNYGQVVAALNATDPGAAAQFNASPVAQSYLR
NFLAAPPPQRAAMAAQLQAVPGAAQYIGLVESVAGSCNNY

>NP_216402.1
MTDVSRKIRAWGRRLMIGTAAAVVLPGLVGLAGGAATAGAFSRPGLPVEYLQVPSPSMGRDIKVQFQSGG
NNSPAVYLLDGLRAQDDYNGWDINTPAFEWYYQSGLSIVMPVGGQSSFYSDWYSPACGKAGCQTYKWETF
LTSELPQWLSANRAVKPTGSAAIGLSMAGSSAMILAAYHPQQFIYAGSLSALLDPSQGMGPSLIGLAMGD
AGGYKAADMWGPSSDPAWERNDPTQQIPKLVANNTRLWVYCGNGTPNELGGANIPAEFLENFVRSSNLKF
QDAYNAAGGHNAVFNFPPNGTHSWEYWGAQLNAMKGDLQSSLGAG

>NP_216424.1
MPEQHPPITETTTGAASNGCPVVGHMKYPVEGGGNQDWWPNRLNLKVLHQNPAVADPMGAAFDYAAEVAT
IDVDALTRDIEEVMTTSQPWWPADYGHYGPLFIRMAWHAAGTYRIHDGRGGAGGGMQRFAPLNSWPDNAS
LDKARRLLWPVKKKYGKKLSWADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDEVYWGKEATWLGDERY

SGKRDLENPLAAVQMGLIYVNPEGPNGNPDPMAAAVDIRETFRRMAMNDVETAALIVGGHTFGKTHGAGP
ADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLEILYGYEWELTKSPAGA
WQYTAKDGAGAGTIPDPFGGPGRSPTMLATDLSLRVDPIYERITRRWLEHPEELADEFAKAWYKLIHRDM
GPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAEIASLKSQIRASGLTVSQLVSTAWAAASSFRGSDKRG
GANGGRIRLQPQVGWEVNDPDGDLRKVIRTLEEIQESFNSAAPGNIKVSFADLVVLGGCAAIEKAAKAAG
HNITVPFTPGRTDASQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLDKANLLTLSAPEMTVLVGG
LRVLGANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQGKDGSGKVKWTGSRVDLVFGS
NSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDVR

>NP_216442.1
MKLTTMIKTAVAVVAMAAIATFAAPVALAAYPITGKLGSELTMTDTVGQVVLGWKVSDLKSSTAVIPGYP
VAGQVWEATATVNAIRGSVTPAVSQFNARTADGINYRVLWQAAGPDTISGATIPQGEQSTGKIYFDVTGP
SPTIVAMNNGMEDLLIWEP

>NP_216496.1
MRIKIFMLVTAVVLLCCSGVATAAPKTYCEELKGTDTGQACQIQMSDPAYNINISLPSYYPDQKSLENYI
AQTRDKFLSAATSSTPREAPYELNITSATYQSAIPPRGTQAVVLKVYQNAGGTHPTTTYKAFDWDQAYRK
PITYDTLWQADTDPLPVVFPIVQGELSKQTGQQVSIAPNAGLDPVNYQNFAVTNDGVIFFFNPGELLPEA
AGPTQVLVPRSAIDSMLA

>NP_216547.1
MATTLPVQRHPRSLFPEFSELFAAFPSFAGLRPTFDTRLMRLEDEMKEGRYEVRAELPGVDPDKDVDIMV
RDGQLTIKAERTEQKDFDGRSEFAYGSFVRTVSLPVGADEDDIKATYDKGILTVSVAVSEGKPTEKHIQI
RSTN

>NP_217389.1
MINVQAKPAAAASLAAIAIAFLAGCSSTKPVSQDTSPKPATSPAAPVTTAAMADPAADLIGRGCAQYAAQ
NPTGPGSVAGMAQDPVATAASNNPMLSTLTSALSGKLNPDVNLVDTLNGGEYTVFAPTNAAFDKLPAATI
DQLKTDAKLLSSILTYHVIAGQASPSRIDGTHQTLQGADLTVIGARDDLMVNNAGLVCGGVHTANATVYM
IDTVLMPPAQ

>NP_218391.1
MAEMKTDAATLAQEAGNFERISGDLKTQIDQVESTAGSLQGQWRGAAGTAAQAAVVRFQEAANKQKQELD
EISTNIRQAGVQYSRADEEQQQALSSQMGF

>NP_222731.1
MKFQPLGERVLVERLEEENKTSSGIIIPDNAKEKPLMGVVKAVSHKISEGCKCVKEGDVIAFGKYKGAEI
VLDGTEYMVLELEDILGIVGSGSCCHTGNHDHKHAKEHEACCHDHKKH

>NP_222744.1
MSVTLINNENNARYEFETIECTRGPKAVDFSKLFETTGFFSYDPGYSSTAGCQSKISYINGKKGELYYRG
HRIEDLVAKYKYVDVCRLLLTGELPKNQDESLEFELELRHRSFVHESLLNMFSAFPSNAHPMAKLSSGVS
ILSTLYSTHQNMHTEEDYQTMARRIVAKIPTLAAICYRNEVGAPIIYPDIARSYVENILFMLRGYPYSRL
KHTTQGEVGITPLEVEAFDKILTLHADHGQNASSTTVRNVASTGVHPYAAISAGISALWGHLHGGANEKV
LLQLEEIGDVKNVDKYIARVKDKNDNFKLMGFGHRVYKSYDPRAKILKGLKDELHQKGVKMDERLSEIAA
KVEEIALKDEYFIERNLYPNVDFYSGTILRALKIPVRFFTPVFVIGRTVGWCAQLLEHVKSPQARITRPR
QVYVGD

>NP_224227.1
MKIPLRFLLISLVPTLSMSNLLGAATTEELSASNSFDGTTSTTSFSSKTSSATDGTNYVFKDSVVIENVP
KTGETQSTSCFKNDAAAGDLNFLGGGFSFTFSNIDATTASGAAIGSEAANKTVTLSGFSALSFLKSPAST
VTNGLGAINVKGNLSLLDNDKVLIQDNFSTGDGGAINCAGSLKIANNKSLSFIGNSSSTRGGAIHTKNLT
LSSGGETLFQGNTAPTAAGKGGAIAIADSGTLSISGDSGDIIFEGNTIGATGTVSHSAIDLGTSAKITAL
RAAQGHTIYFYDPITVTGSTSVADALNINSPDTGDNKEYTGTIVFSGEKLTEAEAKDEKNRTSKLLQNVA
FKNGTVVLKGDVVLSANGFSQDANSKLIMDLGTSLVANTESIELTNLEINIDSLRNGKKIKLSAATAQKD
IRIDRPVVLAISDESFYQNGFLNEDHSYDGILELDAGKDIVISADSRSIDAVQSPYGYQGKWTINWSTDD
KKATVSWAKQSFNPTAEQEAPLVPNLLWGSFIDVRSFQNFIELGTEGAPYEKRFWVAGISNVLHRSGREN
QRKFRHVSGGAVVGASTRMPGGDTLSLGFAQLFARDKDYFMNTNFAKTYAGSLRLQHDASLYSVVSILLG
EGGLREILLPYVSKTLPCSFYGQLSYGHTDHRMKTESLPPPPPTLSTDHTSWGGYVWAGELGTRVAVENT
SGRGFFQEYTPFVKVQAVYARQDSFVELGAISRDFSDSHLYNLAIPLGIKLEKRFAEQYYHVVAMYSPDV
CRSNPKCTTTLLSNQGSWKTKGSNLARQAGIVQASGFRSLGAAAELFGNFGFEWRGSSRSYNVDAGSKIK
F

>NP_224506.1
MKKLLFSTFLLVLGSTSAAHANLGYVNLKRCLEESDLGKKETEELEAMKQQFVKNAEKIEEELTSIYNKL
QDEDYMESLSDSASEELRKKFEDLSGEYNAYQSQYYQSINQSNVKRIQKLIQEVKIAAESVRSKEKLEAI
LNEEAVLAIAPGTDKTTEIIAILNESFKKQN

>NP_224995.1
MFEAVIADIQAREILDSRGYPTLHVKVTTSTGSVGEARVPSGASTGKKEALEFRDTDSPRYQGKGVLQAV
KNVKEILFPLVKGCSVYEQSLIDSLMMDSDGSPNKETLGANAILGVSLATAHAAAATLRRPLYRYLGGCF
ACSLPCPMMNLINGGMHADNGLEFQEFMIRPIGASSIKEAVNMGADVFHTLKKLLHERGLSTGVGDEGGF
APNLASNEEALELLLLAIEKAGFTPGKDISLALDCAASSFYNVKTGTYDGRHYEEQIAILSNLCDRYPID
SIEDGLAEEDYDGWALLTEVLGEKVQIVGDDLFVTNPELILEGISNGLANSVLIKPNQIGTLTETVYAIK
LAQMAGYTTIISHRSGETTDTTIADLAVAFNAGQIKTGSLSRSERVAKYNRLMEIEEELGSEAIFTDSNV
FSYEDSEE

>NP_218698.1
MRGTYCVTLWGGVFAALVAGCASERMIVAYRGAAGYVPEHTFASKVLAFAQGADYLQQDVVLSKDNQLIV
AQSHILDNMTDVAEKFPRRQRADGHFYVIDFTVEELSLLRATNSFYTRGKRHTPVYGQRFPLWKPGFRLH
TFEEELQFIRGLEQTTGKKIGIYSEIKVPWFHHQEGKDIAALTLALLKKYGYQSRSDLVYVQTYDFNELK
RIKRELLPKYEMNVKLIQRVAYTDQRETQEKDSRGKWINYNYNWMFEPGGMQKIAKYADGVGPDWRMLIE
NEWSKVGAVRLSPMVSAIQDAKLECHVHTVRKETLPSYARTMDEMFSILFKQTGANVVLTDFPDLGVKFL
GKPARY

>NP_219206.1
MKTRNFSLVSALYVLLGVPLFVSAASYDDNEFSRKSRAYSELAEKTYDAGEYDVSAEYARLAEDFAQKSS
VYIKETMARTTAEDAMNAARTRHAWAKNERIDRAYPTEYLLASEAIKTGGLAFDSKQYDVALTWARKALD
ALKNVKPESQLLAKAAKEEAARKAAEARKLEEQRIAAQKAQEERKRAEEEAARKAAEARKLEEQRIAAQK
AQEERKRAEEEAARKAAEEAARKAEELEKGRVLPAQYKVTTWSIDRECFWNIAKNPAVYGNPFLWKKLYE
ANKDKIPQSKNPNWVEPETVLVIPSLKGEEREGLYEPNVKYRPLP

>NP_219475.1
MNMCTDGKKYHSTATSAAVGASAPGVPDARAIAAICEQLRQHVADLGVLYIKLHNYHWHIYGIEFKQVHE
LLEEYYVSVTEAFDTIAERLLQLGAQAPASMAEYLALSGIAEETEKEITIVSALARVKRDFEYLSTRFSQ
TQVLAAESGDAVTDGIITDILRTLGKAIWMLGATLKA

>NP_273914.1
MKHTVSASVILLLTACAQLPQNNENLWQPSEHISSFAAEGRLAVKAEGKGSYANFDWTYQPPVETININT
PLGSTLGQLCQDRDGALAVDGKGNVYQAESAEELSRQLVGFKLPIQYLHIWADGRRVAGAPYRILPDGIL
EQYGWTVGRTADSGGQVRTLQLNNGNLNIRLVFTEIGMPSETETPERCAARTR

>NP_274190.1
MMNPKTLSRLSLCAAVLALTACGGNGQKSLYYYGGYPDTVYEGLKNDDTSLGKQTEKMEKYFVEAGNKKM
NAAPGAHAHLGLLLSRSGDKEGAFRQFEEEKRLFPESGVFMDFLMKTGKGGKR

>1DYQ_A
MSEKSEEINEKDLRKKSELQGTALGNLKQIYYYNEKAKTENKESHDQFRQHTILFKGFFTDHSWYNDLLV

RFDSKDIVDKYKGKKVDLYGAYAGYQCAGGTPNKTACMYGGVTLHDNNRLTEEKKVPINLWLDGKQNTVP
LETVKTNKKNVTVQELDLQARRYLQEKYNLYNSDVFDGKVQRGLIVFHTSTEPSVNYDLFGAQGQYSNTL
LRIYRDNKTINSENMHIDIYLYTS

>NP_345151.1
MKKSTVLSLTTAAVILAAYAPNEVVLADTSSSEDALNISDKEKVAENKEKHENIHSAMETSQDFKEKKTA
VIKEKEVVSKNPVIDNNTSNEEAKIKEENSNKSQGDYTDSFVNKNTENPKKEDKVVYIAEFKDKESGEKA
IKELSSLKNTKVLYTYDRIFNGSAIETTPDNLDKIKQIEGISSVERAQKVQPMMNHARKEIGVEEAIDYL
KSINAPFGKNFDGRGMVISNIDTGTDYRHKAMRIDDDAKASMRFKKEDLKGTDKNYWLSDKIPHAFNYYN
GGKITVEKYDDGRDYFDPHGMHIAGILAGNDTEQDIKNFNGIDGIAPNAQIFSYKMYSDAGSGFAGDETM
FHAIEDSIKHNVDVVSVSSGFTGTGLVGEKYWQAIRALRKAGIPMVVATGNYATSASSSSWDLVANNHLK
MTDTGNVTRTAAHEDAIAVASAKNQTVEFDKVNIGGESFKYRNIGAFFDKSKITTNEDGTKAPSKLKFVY
IGKGQDQDLIGLDLRGKIAVMDRIYTKDLKNAFKKAMDKGARAIMVVNTVNYYNRDNWTELPAMGYEADE
GTKSQVFSISGDDGVKLWNMINPDKKTEVKRNNKEDFKDKLEQYYPIDMESFNSNKPNVGDEKEIDFKFA
PDTDKELYKEDIIVPAGSTSWGPRIDLLLKPDVSAPGKNIKSTLNVINGKSTYGYMSGTSMATPIVAAST
VLIRPKLKEMLERPVLKNLKGDDKIDLTSLTKIALQNTARPMMDATSWKEKSQYFASPRQQGAGLINVAN
ALRNEVVATFKNTDSKGLVNSYGSISLKEIKGDKKYFTIKLHNTSNRPLTFKVSASAITTDSLTDRLKLD
ETYKDEKSPDGKQIVPEIHPEKVKGANITFEHDTFTIGANSSFDLNAVINVGEAKNKNKFVESFIHFESV
EEMEALNSNGKKINFQPSLSMPLMGFAGNWNHEPILDKWAWEEGSRSKTLGGYDDDGKPKIPGTLNKGIG
GEHGIDKFNPAGVIQNRKDKNTTSLDQNPELFAFNNEGINAPSSSGSKIANIYPLDSNGNPQDAQLERGL
TPSPLVLRSAEEGLISIVNTNKEGENQRDLKVISREHFIRGILNSKSNDAKGIKSSKLKVWGDLKWDGLI
YNPRGREENAPESKDNQDPATKIRGQFEPIAEGQYFYKFKYRLTKDYPWQVSYIPVKIDNTAPKIVSVDF
SNPEKIKLITKDTYHKVKDQYKNETLFARDQKEHPEKFDEIANEVWYAGAALVNEDGEVEKNLEVTYAGE
GQGRNRKLDKDGNTIYEIKGAGDLRGKIIEVIALDGSSNFTKIHRIKFANQADEKGMISYYLVDPDQDSS
KYQKLGEIAESKFKNLGNGKEGSLKKDTTGVEHHHQENEESIKEKSSFTIDRNISTIRDFENKDLKKLIK
KKFREVDDFTSETGKRMEEYDYKYDDKGNIIAYDDGTDLEYETEKLDEIKSKIYGVLSPSKDGHFEILGK
ISNVSKNAKVYYGNNYKSIEIKATKYDFHSKTMTFDLYANINDIVDGLAFAGDMRLFVKDNDQKKAEIKI
RMPEKIKETKSEYPYVSSYGNVIELGEGDLSKNKPDNLTKMESGKIYSDSEKQQYLLKDNIILRKGYALK
VTTYNPGKTDMLEGNGVYSKEDIAKIQKANPNLRALSETTIYADSRNVEDGRSTQSVLMSALDGFNIIRY
QVFTFKMNDKGEAIDKDGNLVTDSSKLVLFGKDDKEYTGEDKFNVEAIKEDGSMLFIDTKPVNLSMDKNY
FNPSKSNKIYVRNPEFYLRGKISDKGGFNWELRVNESVVDNYLIYGDLHIDNTRDFNIKLNVKDGDIMDW
GMKDYKANGFPDKVTDMDGNVYLQTGYSDLNAKAVGVHYQFLYDNVKPEVNIDPKGNTSIEYADGKSVVF
NINDKRNNGFDGEIQEQHIYINGKEYTSFNDIKQIIDKTLNIKIVVKDFARNTTVKEFILNKDTGEVSEL
KPHRVTVTIQNGKEMSSTIVSEEDFILPVYKGELEKGYQFDGWEISGFEGKKDAGYVINLSKDTFIKPVF
KKIEEKKEEENKPTFDVSKKKDNPQVNHSQLNESHRKEDLQREEHSQKSDSTKDVTATVLDKNNISSKST
TNNPNKLPKTGTASGAQTLLAAGIMFIVGIFLGLKKKNQD

>NP_357715.1
MNKKKMILTSLASVAILGAGFVASQPTVVRAEESPVASQSKAEKDYDAAKKDAKNAKKAVEDAQKALDDA
KAAQKKYDEDQKKTEEKAALEKAASEEMDKAVAAVQQAYLAYQQATDKAAKDAADKMIDEAKKREEEAKT
KFNTVRAMVVPEPEQLAETKKKSEEAKQKAPELTKKLEEAKAKLEEAEKKATEAKQKVDAEEVAPQAKIA
ELENQVHRLEQELKEIDESESEDYAKEGFRAPLQSKLDAKKAKLSKLEELSDKIDELDAEIAKLEDQLKA
AEENNNVEDYFKEGLEKTIAAKKAELEKTEADLKKAVNEPEKPAPAPETPAPEAPAEQPKPAPAPQPAPA
PKPEKPAEQPKPEKTDDQQAEEDYARRSEEEYNRLTQQQPPKAEKPAPAPKTGWKQENGMWYFYNTDGSM
ATGWLQNNGSWYYLNSNGAMATGWLQYNGSWYYLNANGAMATGWAKVNGSWYYLNANGAMATGWLQYNGS
WYYLNANGAMATGWAKVNGSWYYLNANGAMATGWLQYNGSWYYLNANGAMATGWAKVNGSWYYLNANGAM
ATGWVKDGDTWYYLEASGAMKASQWFKVSDKWYYVNGLGALAVNTTVDGYKVNANGEWV

>NP_373278.1
MKKIKIVPLILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIY
NSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQKDQ
SIHIENLKSERGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDT
FVPLKTVKKMDEYLSDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGK
KGLEKLYDKKLQHEDGYRVTIVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAI
HPQTGELLALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLD
DKTSYKIDGKGWQKDKSWGGYNVTRYEVVNGNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGE
DIPSDYPFYNAQISNKNLDNEILLADSGYGQGEILINPVQILSIYSALENNGNINAPHLLKDTKNKVWKK
NIISKENINLLTDGMQQVVNKTHKEDIYRSYANLIGKSGTAELKMKQGETGRQIGWFISYDKDNPNMMMA
INVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE

>NP_395165.1
MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKNIDISIKYDPRKDSEVFANRVITDDIEL
LKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMAVMHFSLTADRIDDDILKVIV
DSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSSSGTINIHDKSINLMDKNLYGYTDEEIFKAS
AEYKILEKMPQTTIQVDGSEKKIVSIKDFLGSENKRTGALGNLKNSYSYNKDNNELSHFATTCSDKSRPL
NDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDTSGK

>NP_395430.1
MKKISSVIAIALFGTIATANAADLTASTTATATLVEPARITLTYKEGAPITIMDNGNIDTELLVGTLTLG
GYKTGTTSTSVNFTDAAGDPMYLTFTSQDGNNHQFTTKVIGKDSRDFDISPKVNGENLVGDDVVLATGSQ
DFFVRSIGSKGGKLAAGKYTDAVTVTVSNQ

>NP_438542.1
MNKFVKSLLVAGSVAALAACSSSNNDAAGNGAAQTFGGYSVADLQQRYNTVYFGFDKYDITGEYVQILDA
HAAYLNATPAAKVLVEGNTDERGTPEYNIALGQRRADAVKGYLAGKGVDAGKLGTVSYGEEKPAVLGHDE
AAYSKNRRAVLAY

>NP_438563.1
MKKFNQSLLATAMLLAAGGANAAAFQLAEVSTSGLGRAYAGEAAIADNASVVATNPALMSLFKTAQFSTG
GVYVDSRINMNGDVTSHATIITSSSGIKAIEGGSASARNVVPGAFVPNLYFVAPVNDKFALGAGMNVNFG
LKSEYDDSYDAGIFGGKTDLSAINLNLSGAYRVTEGLSLGLGVNAVYAKAQVERNAGIIADSVKDNQVKT
ALTVQQEPLKFLDKYLPSKDTSVVSLQDRAAWGFGWNAGVMYQFNEANRIGLAYHSKVDIDFTDRTATSV
EANVIKAGKKGDLTLTLPDYLELSGFHQLTDKLAVHYSYKYTHWSRLTKLNASFEDGKKAFDKELQYSNN
SRVALGASYNLDEKLTLRAGIAYDQAASRHQRSAAIPDTDRTWYSLGATYKFTPNLSVDLGYAYLKGKKV
HFKEVKTIGDERSLTLNTTANYTSQAHANLYGLNLNYSF

>NP_539319.1
MQYSDRHWSPYWAQDFSGSLDVTASGFVGGVQAGYNWQLANGLVLGGEADFQGSTVKSKLVDNGDLSDIG
VAGNLSGDESFGLETKVQWFGTVRARLGFTPTERLMVYGTGGLAYGKVKTSLSAYDDGESFSAGNSKTKA
GWTLGAGVEYAVTNNWTLKSEYLYTDLGKRSFNYIDEENVNINMENKVNFHTVRLGLNYKF

>NP_539453.1
MNTRASNFLAASFSTIMLVGAFSLPAFAQENQMTTQPARIAVTGEGMMTASPDMAILNLSVLRQAKTARE
AMTANNEAMTKVLDAMKKAGIEDRDLQTGGINIQPIYVYPDDKNNLKEPTITGYSVSTSLTVRVRELANV
GKILDESVTLGVNQGGDLNLVNDNPSAVINEARKRAVANAIAKAKTLADAAGVGLGRVVEISELSRPPMP
MPIARGQFRTMLAAAPDNSVPIAAGENSYNVSVNVVFEIK

>NP_539986.1
MTRSEGLNMQVTETLNEGLKREIKVVVPAGDLEAKLAERLETARGRARINGFRPGKVPTAHLRKMYGKSF
MAEIVNEILNDSSRSILAERNEKSATQPEVIMSEDEKEAEKVLDGKADFVFSLNYEVLPAIEVKDFSKIA
VTREVVDISDEEVDEQVKRIASSTRTFETKKGKAENEDRVTIDYLGKLDGEPFEGGADNDAQLVLGSGQF
IPGFEEQLIGLKAGDEKVITVTFPAEYGAAHLAGKEATFDIKVKEVAKPNELVLDDETAKKLGIESLERL
RQVVREQIESQYGQITRQKVKRQILDALDGDYQFETPQKLVDAEFNNIWQQINFDLQQAGRTFEDEETTE
EAAREEYRKLAERRVRLGLVLSEIGEKAGVEVTEEELQRAVYDQVRRYPGQEKEIYDFLRRTPDAVANLR
APIFEEKVVDHLLANINVTDKKVSKEELTAEDEDAASEAKPAKKAAAKKKAAPKKKAEEGKSEEA

>NP_540166.1
MRTLKSLVIVSAALLPFSATAFAADAIQEQPPVPAPVEVAPQYSWAGGYTGLYLGYGWNKAKTSTVGSIK

PDDWKAGAFAGWNFQQDQIVYGVEGDAGYSWAKKSKDGLEVKQGFEGSLRARVGYDLNPVMPYLTAGIAG
SQIKLNNGLDDESKFRVGWTAGAGLEAKLTDNILGRVEYRYTQYGNKNYDLAGTTVRNKLDTQDFRVGIG
YKF

>NP_540501.1
MKNYRAIGLAFTFTALSSLSAFAASLPGGASTLQETYQDWTVSCQSQKDTTACVMRQEQSSAQAGQRVLT
AELRNVAGGKVDGVLLMPFGLDLAKGASLKIDDTAGPNLTFSTCLPQGCLAPVSFDAKQVAALKSGTNIN
VTTTALSPSQPVAFKISLKGFGAALDRIQALTK

>NP_541568.1
MAPVANAQEKQNVEVLHWWTSGGEASALEVLKKDLESKGISWTDMPVAGGGGTEAMTVLRARVTAGNAPT
AVQMLGFDIRDWAEQGALGNLDTVASKEGWEKVIPAPLQEFAKYDGHWIAAPVNIHSTNWMWINKAALDK
AGGKEPTNWDELIALLDNFKAQGITPIAHGGQPWQDATIFDAVVLSFGPDFYKKAFIDLDPEALGSDTMK
QAFDRMSKLRTYVDDNFSGRDWNLASAMVIEGKAGVQFMGDWAKGEFLKAGKKPGEDFVCMRYPGTQGAV
TFNSDMFAMFKVSEDKVPAQLEMASAIESPAFQSAFNVVKGSAPARTDVPDTAFDACGKKAIADVKEANS
KGTLLGSMAHGYANPAAVKNAIYDVVTRQFNGQLSSEDAVKELVVAVEAAK

>AAM99537.1
MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGI
ENKDGEVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQ
GLVVDALDSKSNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDK
DVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTID
EPTVDNQNTLKITFKPEKFKEIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE
NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKA
NTNKNYIAGEAVTGQPIKLKSHTDGTFEIKGLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVS
QTSYNTKPTDITVDSADATPDTIKNNKRPSIPNTGGIGTAIFVAIGAAVMAFAVKGMKRRTKDN

>AAM99541.1
MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETS
HETVEGSGEATFENIKPGDYTLREETAPIGYKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYP
KSAIYEDTKENYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNKYKIELTVEGK
TTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGEAVEKLIDKITSNKDNRVALVTYASTIFDGT
EATVSKGVADQNGKALNDSVSWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ
FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYISTSYQNQFNSFLNKIPDRSGI
LQEDFIINGDDYQIVKGDGESFKLFSDRKVPVTGGTTQAAYRVPQNQLSVMSNEGYAINSGYIYLYWRDY
NWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIRPKGYDIFTVGIGVNGDPGATPLEAEKFMQSISSKT
ENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMGEMIEFQLKNGQSFTHDDYVLVGNDGSQLKN
GVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTYDVRLKDNYISNKFYNTNNRTTLSPK
SEKEPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESLLGAKFQLQIEKDFSGYKQFV
PEGSDVTTKNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTNLKADPNANKNQIGYL
EGNGKHLITNTPKRPPGVFPKTGGIGTIVYILVGSTFMILTICSFRRKQL

>NP_688405.1
MKRINKYFAMFSALLLTLTSLLSVAPAFADEATTNTVTLHKILQTESNLNKSNFPGTTGLNGKDYKGGAI
SDLAGYFGEGSKEIEGAFFALALKEDKSGKVQYVKAKEGNKLTPALINKDGTPEITVNIDEAVSGLTPEG
DTGLVFNTKGLKGEFKIVEVKSKSTYNNNGSLLAASKAVPVNITLPLVNEDGVVADAHVYPKNTEEKPEI
DKNFAKTNDLTALTDVNRLLTAGANYGNYARDKATATAEIGKVVPYEVKTKIHKGSKYENLVWTDIMSNG
LTMGSTVSLKASGTTETFAKDTDYELSIDARGFTLKFTADGLGKLEKAAKTADIEFTLTYSATVNGQAII
DNPESNDIKLSYGNKPGKDLTELPVTPSKGEVTVAKTWSDGIAPDGVNVVYTLKDKDKTVASVSLTKTSK
GTIDLGNGIKFEVSGNFSGKFTGLENKSYMISERVSGYGSAINLENGKVTITNTKDSDNPTPLNPTEPKV
ETHGKKFVKTNEQGDRLAGAQFVVKNSAGKYLALKADQSEGQKTLAAKKIALDEAIAAYNKLSATDQKGE
KGITAKELIKTKQADYDAAFIEARTAYEWITDKARAITYTSNDQGQFEVTGLADGTYNLEETLAPAGFAK
LAGNIKFVVNQGSYITGGNIDYVANSNQKDATRVENKKVTIPQTGGIGTILFTIIGLSIMLGAVVIMKRR
QSKEA

>NP_688406.1
MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESK
IEKVTAELTGEATFDNLIPGDYTLSEETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELD
KQYPPTGIYEDTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKI
ELTVSGKTIVKPVDKQKPLDVVFVLDNSNSMNNDGPNFQRHNKAKKAAEALGTAVKDILGANSDNRVALV
TYGSDIFDGRSVDVVKGFKEDDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGL
TPEQQKEYYLSKVGETFTMKAFMEADDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQM
KKNGYLNKSNFLLTDKPEDIKGNGESYFLFPLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEH
GTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEEYKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKP
EYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTIEDPMGDKINLQLGNGQTLQPSDYTLQGNDG
SVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL
NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKLLLKGATFELQEFNEDYKLYL
PIKNNNSKVVTGENGKISYKDLKDGKYQLIEAVSPEDYQKITNKPILTFEVVKGSIKNIIAVNKQISEYH
EEGDKHLITNTHIPPKGIIPMTGGKGILSFILIGGAMMSIAGGIYIWKRYKKSSDMSIKKD

>AAN37923.1
MKKTVFRLNFLTACISLGIVSQAWAGHTYFGIDYQYYRDFAENKGKFSVGAKNIEVYNKEGTLVGTSMTK
APMIDFSVVSRNGVAALVGDQYIVSVAHNGGYNSVDFGAEGPNPDQHRFTYQIVKRNNYKPGKDNPYHGD
YHMPRLHKFVTDAEPAKMTDNMNGKNYADLSKYPDRVRIGTGEQWWRTDEEQKQGSKSSWLADAYLWRIA
GNTHSQSGAGNGTVNLSGDITKPNNYGPLPTGVSFGDSGSPMFIYDAIKQKWLINGVLQTGNPFSGAGNG
FQLIRKNWFYDNVFVEDLPITFLEPRSNGHYSFTSNNNGTGTVTQTNEKVSMPQFKVRTVQLFNEALKEK
DKEPVYAAGGVNAYKPRLNNGKNIYFGDRGTGTLTIENNINQGAGGLYFEGNFTVSSENNATWQGAGVHV
GEDSTVTWKVNGVEHDRLSKIGKGTLHIQAKGENLGSISVGDGKVILDQQADENNQKQAFKEVGIVSGRA
TVQLNSADQVDPNNIYFGFRGGRLDLNGHSLTFKRIQNTDEGAMIVNHNTTQVANITITGNESITAPSNK
NNINKLDYSKEIAYNGWFGETDENKHNGRLNLIYKPTTEDRTLLLSGGTNLKGNITQEGGTLVFSGRPTP
HAYNHLNRPNELGRPQGEVVIDDDWITRTFKAENFQIKGGSAVVSRNVSSIEGNWTVSNNANAAFGVVPN
QQNTICTRSDWTGLTTCKTVDLTDTKVINSIPTTQINGSINLTDNATVNIHGLAKLNGNVTLINHSQFTL
SNNATQTGNIQLSNHANATVDNANLNGNVHLTDSAQFSLKNSHFSHQIQGDKDTTVTLENATWTMPSDAT
LQNLTLNNSTVTLNSAYSASSNNAPRHRRSLETETTPTSAEHRFNTLTVNGKLSGQGTFQFTSSLFGYKS
DKLKLSNDAEGDYTLSVRNTGKEPEALEQLTLVESKDNKPLSDKLKFTLENDHVDAGALRYKLVKNNGEF
RLHNPIKEQELRNDLVRAEQAERTLEAKQVEQTAETQTSNARVRSKRAVFSDTLPDQSQLDVLQAEQVEP
TAEKQKNKAKKVRSKRAVFSDTLPDQSQLDVLQAEQVEPTAEKQKNKAKKVRSKRAAREFSDTPLDLSRL
KVLEVKLEVINAQQQVKKEPQDQEKQRKQKDLISRYSNSALSELSATVNSMLSVQDELDRLFVDQAQSAV
WTNIAQDKRRYDSDAFRAYQQKTNLRQIGVQKALANGRIGAVFSHSRSDNTFDEQVKNHATLTMMSGFAQ
YQWGDLQFGVNVGTGISASKMAEEQSRKIHRKAINYGVNASYQFRLGQLGIQPYFGVNRYFIERENYQSE
EVKVKTPSLAFNRYNAGIRVDYTFTPTDNISVKPYFFVNYVDVSNANVQTTVNSTVLQQPFGRYWQKEVG
LKAEILHFQLSAFISKSQGSQLGKQQNVGVKLGYRW

>AAN79749.1
MIHLFKTCMITAFILGLTWSAPLRAQDQRYISIRNTDTIWLPGNICAYQFRLDNGGNDEGFGPLTITLQL
KDKYGQTLVTRKMETEAFGDSNATRTTDAFLETECVENVATTEIIKATEESNGHRVSLPLSVFDPQDYHP
LLITVSGKNVN

>NP_757168.1
MKKSLLAVMLTGLFALVSLPALGNVNLEQLKQKAESGEAKAQLELGYRYFQGNETTKDLTQAMDWFRRAA
EQGYTPAEYVLGLRYMNGEGVPQDYAQAVIWYKKAALKGLPQAQQNLGVMYHEGNGVKVDKAESVKWFRL
AAEQGRDSGQQSMGDAYFEGDGVTRDYVMAREWYSKAAEQGNVWSCNQLGYMYSRGLGVERNDAISAQWY
RKSATSGDELGQLHLADMYYFGIGVTQDYTQSRVLFSQSAEQGNSIAQFRLGYILEQGLAGAKEPLKALE
WYRKSAEQGNSDGQYYLAHLYDKGAEGVAKNREQAISWYTKSAEQGDATAQANLGAIYFRLGSEEEHKKA
VEWFRKAAAKGEKAAQFNLGNALLQGKGVKKDEQQAAIWMRKAAEQGLSAAQVQLGEIYYYGLGVERDYV
QAWAWFDTASTNDMNLFGTENRNITEKKLTAKQLQQAELLSQQYIEKYAPEAWARMQKLKAQSAVKTGNK

>NP_783831.1

MPITINNFRYSDPVNNDTIIMMEPPYCKGLDIYYKAFKITDRIWIVPERYEFGTKPEDFNPPSSLIEGAS
EYYDPNYLRTDSDKDRFLQTMVKLFNRIKNNVAGEALLDKIINAIPYLGNSYSLLDKFDTNSNSVSFNLL
EQDPSGATTKSAMLTNLIIFGPGPVLNKNEVRGIVLRVDNKNYFPCRDGFGSIMQMAFCPEYVPTFDNVI
ENITSLTIGKSKYFQDPALLLMHELIHVLHGLYGMQVSSHEIIPSKQEIYMQHTYPISAEELFTFGGQDA
NLISIDIKNDLYEKTLNDYKAIANKLSQVTSCNDPNIDIDSYKQIYQQKYQFDKDSNGQYIVNEDKFQIL
YNSIMYGFTEIELGKKFNIKTRLSYFSMNHDPVKIPNLLDDTIYNDTEGFNIESKDLKSEYKGQNMRVNT
NAFRNVDGSGLVSKLIGLCKKIIPPTNIRENLYNRTASLTDLGGELCIKIKNEDLTFIAEKNSFSEEPFQ
DEIVSYNTKNKPLNFNYSLDKIIVDYNLQSKITLPNDRTTPVTKGIPYAPEYKSNAASTIEIHNIDDNTI
YQYLYAQKSPTTLQRITMTNSVDDALINSTKIYSYFPSVISKVNQGAQGILFLQWVRDIIDDFTNESSQK
TTIDKISDVSTIVPYIGPALNIVKQGYEGNFIGALETTGVVLLLEYIPEITLPVIAALSIAESSTQKEKI
IKTIDNFLEKRYEKWIEVYKLVKAKWLGTVNTQFQKRSYQMYRSLEYQVDAIKKIIDYEYKIYSGPDKEQ
IADEINNLKNKLEEKANKAMININIFMRESSRSFLVNQMINEAKKQLLEFDTQSKNILMQYIKANSKFIG
ITELKKLESKINKVFSTPIPFSYSKNLDCWVDNEEDIDVILKKSTILNLDINNDIISDISGFNSSVITYP
DAQLVPGINGKAIHLVNNESSEVIVHKAMDIEYNDMFNNFTVSFWLRVPKVSASHLEQYGTNEYSIISSM
KKHSLSIGSGWSVSLKGNNLIWTLKDSAGEVRQITFRDLPDKFNAYLANKWVFITITNDRLSSANLYING
VLMGSAEITGLGAIREDNNITLKLDRCNNNNQYVSIDKFRIFCKALNPKEIEKLYTSYLSITFLRDFWGN
PLRYDTEYYLIPVASSSKDVQLKNITDYMYLTNAPSYTNGKLNIYYRRLYNGLKFIIKRYTPNNEIDSFV
KSGDFIKLYVSYNNNEHIVGYPKDGNAFNNLDRILRVGYNAPGIPLYKKMEAVKLRDLKTYSVQLKLYDD
KNASLGLVGTHNGQIGNDPNRDILIASNWYFNHLKDKILGCDWYFVPTDEGWTND

>NP_854596.1
MDFGLLPPEVNSSRMYSGPGPESMLAAAAAWDGVAAELTSAAVSYGSVVSTLIVEPWMGPAAAAMAAAAT
PYVGWLAATAALAKETATQARAAAEAFGTAFAMTVPPSLVAANRSRLMSLVAANILGQNSAAIAATQAEY
AEMWAQDAAVMYSYEGASAAASALPPFTPPVQGTGPAGPAAAAAATQAAGAGAVADAQATLAQLPPGILS
DILSALAANADPLTSGLLGIASTLNPQVGSAQPIVIPTPIGELDVIALYIASIATGSIALAITNTARPWH
IGLYGNAGGLGPTQGHPLSSATDEPEPHWGPFGGAAPVSAGVGHAALVGALSVPHSWTTAAPEIQLAVQA
TPTFSSSAGADPTALNGMPAGLLSGMALASLAARGTTGGGGTRSGTSTDGQEDGRKPPVVVIREQPPPGN
PPR

>NP_857727.1
MSNFSGFTKGTDIADLDAVAQTLKKPADDANKAVNDSIAALKDKPDNPALLADLQHSINKWSVIYNINST
IVRSMKDLMQGILQKFP

>NP_880571.1
MNTNLYRLVFSHVRGMLVPVSEHCTVGNTFCGRTRGQARSGARATSLSVAPNALAWALMLACTGLPLVTH
AQGLVPQGQTQVLQGGNKVPVVNIADPNSGGVSHNKFQQFNVANPGVVFNNGLTDGVSRIGGALTKNPNL
TRQASAILAEVTDTSPSRLAGTLEVYGKGADLIIANPNGISVNGLSTLNASNLTLTTGRPSVNGGRIGLD
VQQGTVTIERGGVNATGLGYFDVVARLVKLQGAVSSKQGKPLADIAVVAGANRYDHATRRATPIAAGARG
AAAGAYAIDGTAAGAMYGKHITLVSSDSGLGVRQLGSLSSPSAITVSSQGEIALGDATVQRGPLSLKGAG
VVSAGKLASGGGAVNVAGGGAVKIASASSVGNLAVQGGGKVQATLLNAGGTLLVSGRQAVQLGAASSRQA
LSVNAGGALKADKLSATRRVDVDGKQAVALGSASSNALSVRAGGALKAGKLSATGRLDVDGKQAVTLGSV
ASDGALSVSAGGNLRAKQLVSSAQLEVRGQREVALDDASSARGMTVVAAGALAARNLQSKGAIGVQGGEA
VSVANANSDAELRVRGRGQVDLHDLSAARGADISGEGRVNIGRARSDSDVKVSAHGALSIDSMTALGAIG
VQAGGSVSAKDMRSRGAVTVSGGGAVNLGDVQSDGQVRATSAGAMTVRDVAAAADLALQAGDALQAGFLK
SAGAMTVNGRDAVRLDGAHAGGQLRVSSDGQAALGSLAAKGELTVSAARAATVAELKSLDNISVTGGERV
SVQSVNSASRVAISAHGALDVGKVSAKSGIGLEGWGAVGADSLGSDGAISVSGRDAVRVDQARSLADISL
GAEGGATLGAVEAAGSIDVRGGSTVAANSLHANRDVRVSGKDAVRVTAATSGGGLHVSSGRQLDLGAVQA
RGALALDGGAGVALQSAKASGTLHVQGGEHLDLGTLAAVGAVDVNGTGDVRVAKLVSDAGADLQAGRSMT
LGIVDTTGDLQARAQQKLELGSVKSDGGLQAAAGGALSLAAAEVAGALELSGQGVTVDRASASRARIDST
GSVGIGALKAGAVEAASPRRARRALRQDFFTPGSVVVRAQGNVTVGRGDPHQGVLAQGDIIMDAKGGTLL
LRNDALTENGTVTISADSAVLEHSTIESKISQSVLAAKGDKGKPAVSVKVAKKLFLNGTLRAVNDNNETM
SGRQIDVVDGRPQITDAVTGEARKDESVVSDAALVADGGPIVVEAGELVSHAGGIGNGRNKENGASVTVR
TTGNLVNKGYISAGKQGVLEVGGALTNEFLVGSDGTQRIEAQRIENRGTFQSQAPAGTAGALVVKAAEAI
VHDGVMATKGEMQIAGKGGGSPTVTAGAKATTSANKLSVDVASWDNAGSLDIKKGGAQVTVAGRYAEHGE
VSIQGDYTVSADAIALAAQVTQRGGAANLTSRHDTRFSNKIRLMGPLQVNAGGAVSNTGNLKVREGVTVT
AASFDNETGAEVMAKSATLTTSGAARNAGKMQVKEAATIVAASVSNPGTFTAGKDITVTSRGGFDNEGKM
ESNKDIVIKTEQFSNGRVLDAKHDLTVTASGQADNRGSLKAGHDFTVQAQRIDNSGTMAAGHDATLKAPH
LRNTGQVVAGHDIHIINSAKLENTGRVDARNDIALDVADFTNTGSLYAEHDATLTLAQGTQRDLVVDQDH
ILPVAEGTLRVKAKSLTTEIETGNPGSLIAEVQENIDNKQAIVVGKDLTLSSAHGNVANEANALLWAAGE
LTVKAQNITNKRAALIEAGGNARLTAAVALLNKLGRIRAGEDMHLDAPRIENTAKLSGEVQRKGVQDVGG
GEHGRWSGIGYVNYWLRAGNGKKAGTIAAPWYGGDLTAEQSLIEVGKDLYLNAGARKDEHRHLLNEGVIQ
AGGHGHIGGDVDNRSVVRTVSAMEYFKTPLPVSLTALDNRAGLSPATWNFQSTYELLDYLLDQNRYEYIW
GLYPTYTEWSVNTLKNLDLGYQAKPAPTAPPMPKAPELDLRGHTLESAEGRKIFGEYKKLQGEYEKAKMA
VQAVEAYGEATRRVHDQLGQRYGKALGGMDAETKEVDGIIQEFAADLRTVYAKQADQATIDAETDKVAQR
YKSQIDAVRLQAIQPGRVTLAKALSAALGADWRALGHSQLMQRWKDFKAGKRGAEIAFYPKEQTVLAAGA
GLTLSNGAIHNGENAAQNRGRPEGLKIGAHSATSVSGSFDALRDVGLEKRLDIDDALAAVLVNPHIFTRI
GAAQTSLADGAAGPALARQARQAPETDGMVDARGLGSADALASLASLDAAQGLEVSGRRNAQVADAGLAG
PSAVAAPAVGAADVGVEPVTGDQVDQPVVAVGLEQPVATVRVAPPAVALPRPLFETRIKFIDQSKFYGSR
YFFEQIGYKPDRAARVAGDNYFDTTLVREQVRRALGGYESRLPVRGVALVAKLMDSAGTVGKALGLKVGV
APTAQQLKQADRDFVWYVDTVIDGQKVLAPRLYLTEATRQGITDQYAGGGALIASGGDVTVNTDGHDVSS
VNGLIQGRSVKVDAGKGKVVVADSKGAGGGIEADDEVDVSGRDIGIEGGKLRGKDVRLKADTVKVATSMR
YDDKGRLAARGDGALDAQGGQLHIEAKRLETAGATLKGGKVKLDVDDVKLGGVYEAGSSYENKSSTPLGS
LFAILSSTTETNQSAHANHYGTRIEAGTLEGKMQNLEIEGGSVDAAHTDLSVARDARFKAAADFAHAEHE
KDVRQLSLGAKVGAGGYEAGFSLGSESGLEAHAGRGMTAGAEVKVGYRASHEQSSETEKSYRNANLNFGG
GSVEAGNVLDIGGADINRNRYGGAAKGNAGTEEALRMRAKKVESTKYVSEQTSQSSGWSVEVASTASARS
SLLTAATRLGDSVAQNVEDGREIRGELMAAQVAAEATQLVTADTAAVALSAGISADFDSSHSRSTSQNTQ
YLGGNLSIEATEGDATLVGAKFGGGDQVSLKAAKSVNLMAAESTFESYSESHNFHASADANLGANAVQGA
VGLGLTAGMGTSHQITNETGKTYAGTSVDAANVSIDAGKDLNLSGSRVRGKHVVLDVEGDINATSKQDER
NYNSSGGGWDASAGVAIQNRTLVAPVGSAGFNFNTEHDNSRLTNDGAAGVVASDGLTGHVKGDANLTGAT
IADLSGKGNLKVDGAVNAQNLKDYRDKDGGSGGLNVGISSTTLAPTVGVAFGRVAGEDYQAEQRATIDVG
QTKDPARLQVGGGVKGTLNQDAAQATVVQRNKHWAGGGSEFSVAGKSLKKKNQVRPVETPTPDVVDGPPS
RPTTPPASPQPIRATVEVSSPPPVSVATVEVVPRPKVETAQPLPPRPVAAQVVPVTPPKVEVAKVEVVPR
PKVETAQPLPPRPVVAEKVTTPAVQPQLAKVETVQPVKPETTKPLPKPLPVAKVTKAPPPVVETAQPLPP
VKPQKATPGPVAEVGKATVTTVQVQSAPPKPAPVAKQPAPAPKPKPKPKPKAERPKPGKTTPLSGRHVVQ
QQVQVLQRQASDINNTKSLPGGKLPKPVTVKLTDENGKPQTYTINRREDLMKLNGKVLSTKTTLGLEQTF
RLRVEDIGGKNYRVFYETNK

>NP_882282.1
MRCTRAIRQTARTGWLTWLAILAVTAPVTSPAWADDPPATVYRYDSRPPEDVFQNGFTAWGNNDNVLDHL
TGRSCQVGSSNSAFVSTSSSRRYTEVYLEHRMQEAVEAERAGRGTGHFIGYIYEVRADNNFYGAASSYFE
YVDTYGDNAGRILAGALATYQSEYLAHRRIPPENIRRVTRVYHNGITGETTTTEYSNARYVSQQTRANPN
PYTSRRSVASIVGTLVRMAPVIGACMARQAESSEAMAAWSERAGEAMVLVYYESIAYSF

>AAR88321.1
MDAMKRGLCCVLLLCGAVFVSPSEVKQENRLLNESESSSQGLLGYYFSDLNFQAPMVVTSSTTGDLSIPS
SELENIPSENQYFQSAIWSGFIKVKKSDEYTFATSADNHVTMWVDDQEVINKASNSNKIRLEKGRLYQIK
IQYQRENPTEKGLDFKLYWTDSQNKKEVISSDNLQLPELKQKSSNTSAGPTVPDRDNDGIPDSLEVEGYT
VDVKNKRTFLSPWISNIHEKKGLTKYKSSPEKWSTASDPYSDFEKVTGRIDKNVSPEARHPLVAAYPIVH
VDMENIILSKNEDQSTQNTDSETRTISKNTSTSRTHTSEVHGNAEVHASFFDIGGSVSAGFSNSNSSTVA
IDHSLSLAGERTWAETMGLNTADTARLNANIRYVNTGTAPIYNVLPTTSLVLGKNQTLATIKAKENQLSQ
ILAPNNYYPSKNLAPIALNAQDDFSSTPITMNYNQFLELEKTKQLRLDTDQVYGNIATYNFENGRVRVDT
GSNWSEVLPQIQETTARIIFNGKDLNLVERRIAAVNPSDPLETTKPDMTLKEALKIAFGFNEPNGNLQYQ
GKDITEFDFNFDQQTSQNIKNQLAELNATNIYTVLDKIKLNAKMNILIRDKRFHYDRNNIAVGADESVVK
EAHREVINSSTEGLLLNIDKDIRKILSGYIVEIEDTEGLKEVINDRYDMLNISSLRQDGKTFIDFKKYND
KLPLYISNPNYKVNVYAVTKENTIINPSENGDTSTNGIKKILIFSKKGYEIG

>AAR88322.1
MDAMKRGLCCVLLLCGAVFVSPSAGGHGDVGMHVKEKEKNKDENKRKDEERNKTQEEHLKEIMKHIVKIE

VKGEEAVKKEAAEKLLEKVPSDVLEMYKAIGGKIYIVDGDITKHISLEALSEDKKKIKDIYGKDALLHEH
YVYAKEGYEPVLVIQSSEDYVENTEKALNVYYEIGKILSRDILSKINQPYQKFLDVLNTIKNASDSDGQD
LLFTNQLKEHPTDFSVEFLEQNSNEVQEVFAKAFAYYIEPQHRDVLQLYAPEAFNYMDKFNEQEINLSLE
ELKDQRMLSRYEKWEKIKQHYQHWSDSLSEEGRGLLKKLQIPIEPKKDDIIHSLSQEEKELLKRIQIDSS
DFLSTEEKEFLKKLQIDIRDSLSEEEKELLNRIQVDSSNPLSEKEKEFLKKLKLDIQPYDINQRLQDTGG
LIDSPSINLDVRKQYKRDIQNIDALLHQSIGSTLYNKIYLYENMNINNLTATLGADLVDSTDNTKINRGI
FNEFKKNFKYSISSNYMIVDINERPALDNERLKWRIQLSPDTRAGYLENGKLILQRNIGLEIKDVQIIKQ
SEKEYIRIDAKVV

>YP_094511.1
MHPNYYLSPLAVAIALGIASPVKAADPIPLQKSSFSEVTQKFQLTLPGVMKGAVVSTNSLQFIRQHTDGN
KVTHVRMQQQYAGFPVFGGYAILHSKNATPSLATAKSDVKMNGVIYDGLQAELGQPKPSFVKNASLALQQ
FKDKYANKQISEDQVTPMIYIDEKHQAHWAYKVSVLVIHDDRIPERPTAIIDAETNKPFVQWDDVKTEKV
QAKGMGFGGNRKIGEYQFGKDLPLLEITRDSSVEMCFMENTDVKVVDMGHKYYSNNKPMQFTCKETPDTQ
STKTYYTGYSADGYDRDNGAASPTNDALYAGYVIKHMYHDWYGVEALTKSDGSPMQLVMRVHYGQGYENA
YWDGKQMTFGDGDTMMYPLVSLGVGGHEISHGFTEQHSGLEYFGQSGGMNESFSDMAAQAAEYYSVGKNS
WQIGPEIMKEDSGYDALRYMDKPSRDGMSIDVADDYYGGLDVHYSSGVYNHLFYILANQPNWNLRMAFDV
MVKANMDYWTPYSTFDEGGCGMLSAAKDLGYNLDDVKKSLSEVTINYQSCYVD

>YP_094724.1
MIMAKELRFGDDARLQMLAGVNALADAVQVTMGPRGRNVVLEKSYGAPTVTKDGVSVAKEIEFEHRFMNM
GAQMVKEVASKTSDTAGDGTTTATVLARSILVEGHKAVAAGMNPMDLKRGIDKAVLAVTKKLQAMSKPCK
DSKAIAQVGTISANSDEAIGAIIAEAMEKVGKEGVITVEDGNGLENELSVVEGMQFDRGYISPYFINNQQ
NMSCELEHPFILLVDKKVSSIREMLSVLEGVAKSGRPLLIIAEDVEGEALATLVVNNMRGIVKVCAVKAP
GFGDRRKAMLQDIAILTKGQVISEEIGKSLEGATLEDLGSAKRIVVTKENTTIIDGEGKATEINARITQI
RAQMEETTSDYDREKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGIVAGGGVAL
IRAQKALDSLKGDNDDQNMGINILRRAIESPMRQIVTNAGYEASVVVNKVAEHKDNYGFNAATGEYGDMV
EMGILDPTKVTRMALQNAASVASLMLTTECMVADLPKKEEGVGAGDMGGMGGMGGMGMM

>YP_096954.1
MFSLKKTAVAVLALGSGAVFAGTMGPVCTPGNVTVPCERTAWDIGITALYLQPIYDADWGYNGFTQVGGW
RHWHDVDHEWDWGFKLEGSYHFNTGNDINVNWYHFDNDSDHWFDFANWHNYNNKWDAVNAELGQFVDFSA
NKKMRFHGGVQYARIEADVNRYFNNFAFNGFNSKFNGFGPRTGLDMNYVFGNGFGIYAKGAAAILVGTSD
FYDGIGFVTGSKNAIVPELEAKLGADYTYAMAQGDLTLDVGYMWFNYFNAMHNTATNGLETDFAASGPYI
GLKYVGNV

>P69996.1
MKKISRKEYVSMYGPTTGDKVRLGDTDLIAEVEHDYTIYGEELKFGGGKTLREGMSQSNNPSKEELDLII
TNALIVDYTGIYKADIGIKDGKIAGIGKGGNKDMQDGVKNNLSVGPATEALAGEGLIVTAGGIDTHIHFI
SPQQIPTAFASGVTTMIGGGTGPADGTNATTITPGRRNLKWMLRAAEEYSMNLGFLAKGNASNDASLADQ
IEAGAIGFKIHEDWGTTPSAINHALDVADKYDVQVAIHTDTLNEAGCVEDTMAAIAGRTMHTFHTEGAGG
GHAPDIIKVAGEHNILPASTNPTIPFTVNTEAEHMDMLMVCHHLDKSIKEDVQFADSRIRPQTIAAEDTL
HDMGIFSITSSDSQAMGRVGEVITRTWQTADKNKKEFGRLKEEKGDNDNFRIKRYLSKYTINPAIAHGIS
EYVGSVEVGKVADLVLWSPAFFGVKPNMIIKGGFIALSQMGDANASIPTPQPVYYREMFAHHGKAKYDAN
ITFVSQAAYDKGIKEELGLERQVLPVKNCRNITKKDMQFNDTTAHIEVNPETYHVFVDGKEVTSKPANKV
SLAQLFSIF

>YP_177770.1
MKIRLHTLLAVLTAAPLLLAAAGCGSKPPSGSPETGAGAGTVATTPASSPVTLAETGSTLLYPLFNLWGP
AFHERYPNVTITAQGTGSGAGIAQAAAGTVNIGASDAYLSEGDMAAHKGLMNIALAISAQQVNYNLPGVS
EHLKLNGKVLAAMYQGTIKTWDDPQIAALNPGVNLPGTAVVPLHRSDGSGDTFLFTQYLSKQDPEGWGKS
PGFGTTVDFPAVPGALGENGNGGMVTGCAETPGCVAYIGISFLDQASQRGLGEAQLGNSSGNFLLPDAQS
IQAAAAGFASKTPANQAISMIDGPAPDGYPIINYEYAIVNNRQKDAATAQTLQAFLHWAITDGNKASFLD
QVHFQPLPPAVVKLSDALIATISS

>YP_178023.1
MTEQQWNFAGIEAAASAIQGNVTSIHSLLDEGKQSLTKLAAAWGGSGSEAYQGVQQKWDATATELNNALQ
NLARTISEAGQAMASTEGNVTGMFA

>P0A3S9.1
MRRIQSIARSPIAIALFMSLAVAGCASKKNLPNNAGDLGLGAGAATPGSSQDFTVNVGDRIFFDLDSSLI
RADAQQTLSKQAQWLQRYPQYSITIEGHADERGTREYNLALGQRRAAATRDFLASRGVPTNRMRTISYGN
ERPVAVCDADTCWSQNRRAVTVLNGAGR

>YP_248692.1
MKSVPLITGGLSFLLSACSGGGGSFDVDDVSNPSSSKPRYQDDTSSSRTKSNLEKLSIPSLGGGMKLVAQ
NLSGNKEPSFLNENGYISYFSSPSTIEDDVKNVKTENKIHTNPIGLEPNRALQDPNLQKYVYSGLYYIEN
WKDFSKLATEKKAYSGHYGYAFYYGNKTATDLPVSGVATYKGTWDFITATKYGQNYSLFSNARGQAYFRR
SATRGDIDLENNSKNGDIGLISEFSADFGTKKLTGQLSYTKRKTDIQQYEKEKLYDIDAHIYSNRFRGKV
TPTKSTSDEHPFTSEGTLEGGFYGPNAEELGGKFLARDKRVFGVFSAKETPETEKEKLSKETLIDGKLIT
FSTKTADATTSTTASTTADVKTDEKNFTTKDISSFGEADYLLIDNYPVPLFPEGDTDDFVTSKHHDIGNK
TYKVEACCKNLSYVKFGMYYEDKEKKNTNQTGQYHQFLLGLRTPSSQIPVTGNVKYLGSWFGYIGDDKTS
YSTTGNKQQDKNAPAEFDVNFDNKTLTGKLKRADSQNTVFNIEATFKNGSNAFEGKATANVVIDPKNTQA
TSKVNFTTTVNGAFYGPHATELGGYFTYNGNNPTATNSESSSTVPSPPNSPNARAAVVFGAKRQVEKTNK

>YP_248824.1
MKKTAIALVVAGLAAASVAQAAPQENTFYAGVKAGQGSFHDGINNNGAIKQNLSSANYGYRRNTFTYGVF
GGYQILNQDNFGLAAELGYDNFGRVKLREAGKPKAKHTNHGAHLSLKGSYEVLDGLDVYGKAGVALVRSD
YKFYEVANGTRDHKKGRHTARASGLFAVGAEYAVLPELAVRLEYQWLTRVGKYRPQDKPNTAINYNPWIG
SINAGISYRFGQGEAPVVAAPEMVSKTFSLNSDVTFAFGKANLKPQAQATLDSVYGEISQVKSAKVAVAG
YTDRIGSDAFNVKLSQERADSVANYFVAKGVAADAISATGYGKANPVTGATCDQVKGRKALIACLAPDRR
VEIAVNGTK

>YP_418725.1
MKSLFIASTMVLMAFPAFAESTTVKMYEALPTGPGKEVGTVVISEAPGGLHFKVNMEKLTPGYHGFHVHE
NPSCAPGEKDGKIVPALAAGGHYDPGNTHHHLGPEGDGHMGDLPRLSANADGKVSETVVAPHLKKLAEIK
QRSLMVHVGGDNYSDKPEPLGGGGARFACGVIE

>P0C109.1
MGISKASLLSLAAAGIVLAGCQSSRLGNLDNVSPPPPPAPVNAVPAGTVQKGNLDSPTQFPNAPSTDMSA
QSGTQVASLPPASAPDLTPGAVAGVWNASLGGQSCKIATPQTKYGQGYRAGPLRCPGELANLASWAVNGK
QLVLYDANGGTVASLYSSGQGRFDGQTTGGQAVTLSR

>YP_671699.1
MPTITTAQIKSTLQSAKQSAANKLHSAGQSTKDALKKAAEQTRNAGNRLILLIPKDYKGQGSSLNDLVRT
ADELGIEVQYDEKNGTAITKQVFGTAEKLIGLTERGVTIFAPQLDKLLQKYQKAGNKLGGSAENIGDNLG
KAGSVLSTFQNFLGTALSSMKIDELIKKQKSGSNVSSSELAKASIELINQLVDTAASINNNVNSFSQQLN
KLGSVLSNTKHLNGVGNKLQNLPNLDNIGAGLDTVSGILSVISASFILSNADADTGTKAAAGVELTTKVL
GNVGKGISQYIIAQRAAQGLSTSAAAAGLIASAVTLAISPLSFLSIADKFKRANKIEEYSQRFKKLGYDG
DSLLAAFHKETGAIDASLTTISTVLASVSSGISAAATTSLVGAPVSALVGAVTGIISGILEASKQAMFEH
VASKMADVIAEWEKKHGKNYFENGYDARHAAFLEDNFKILSQYNKEYSVERSVLITQQHWDTLIGELAGV
TRNGDKTLSGKSYIDYYEEGKRLEKKPDEFQKQVFDPLKGNIDLSDSKSSTLLKFVTPLLTPGEEIRERR
QSGKYEYITELLVKGVDKWTVKGVQDKGSVYDYSNLIQHASVGNNQYREIRIESHLGDGDDKVFLAAGSA
NIYAGKGHDVVYYDKTDTGYLTIDGTKATEAGNYTVTRVLGGDVKVLQEVVKEQEVSVGKRTEKTQYRSY
EFTHINGTDLTETDNLYSVEELIGTNRADKFFGSKFTDIFHGADGDDHIEGNDGNDRLYGDKGNDTLRGG
NGDDQLYGGDGNDKLTGGVGNNYLNGGDGDDELQVQGNSLAKNVLSGGKGNDKLYGSEGADLLDGGEGND
LLKGGYGNDIYRYLSGYGHHIIDDDGGKDDKLSLADIDFRDVAFKREGNDLIMYKAEGNVLSIGHKNGIT
FRNWFEKESGDISNHQIEQIFDKDGRVITPDSLKKAFEYQQSNNQANYVYGEYASTYADLDNLNPLINEI

SKIISAAGNFDVKEERSAASLLQLSGNASDFSYGRNSITLTASA

>YP_694509.1
MKRKICKALICAALATSLWAGASTKVYAWDGKIDGTGTHAMIVTQGVSILENDLSKNEPESVRKNLEILK
ENMHELQLGSTYPDYDKNAYDLYQDHFWDPDTDNNFSKDNSWYLAYSIPDTGESQIRKFSALARYEWQRG
NYKQATFYLGEAMHYFGDIDTPYHPANVTAVDSAGHVKFETFAEERKEQYKINTAGCKTNEAFYTDILKN
KDFNAWSKEYARGFAKTGKSIYYSHASMSHSWDDWDYAAKVTLANSQKGTAGYIYRFLHDVSEGNDPSVG
KNVKELVAYISTSGEKDAGTDDYMYFGIKTKDGKTQEWEMDNPGNDFMTGSKDTYTFKLKDENLKIDDIQ
NMWIRKRKYTAFSDAYKPENIKIIANGKVVVDKDINEWISGNSTYNIK

>YP_817402.1
MFASKSERKVHYSIRKFSIGVASVAVASLVMGSVVHATENEGSTQAATSSNMAKTEHRKAAKQVVDEYIE
KMLREIQLDRRKHTQNVALNIKLSAIKTKYLRELNVLEEKSKDELPSEIKAKLDAAFEKFKKDTLKPGEK
VAEAKKKVEEAKKKAEDQKEEDRRNYPTNTYKTLELEIAEFDVKVKEAELELVKEEAKESRNEGTIKQAK
EKVESKKAEATRLENIKTDRKKAEEEAKRKADAKLKEANVATSDQGKPKGRAKRGVPGELATPDKKENDA
KSSDSSVGEETLPSSSLKSGKKVAEAEKKVEEAEKKAKDQKEEDRRNYPTNTYKTLDLEIAESDVKVKEA
ELELVKEEAKEPRDEEKIKQAKAKVESKKAEATRLENIKTDRKKAEEEAKRKAAEEDKVKEKPAEQPQPA
PATQPEKPAPKPEKPAEQPKAEKTDDQQAEEDYARRSEEEYNRLTQQQPPKTEKPAQPSTPKTGWKQENG
MWYFYNTDGSMATGWLQNNGSWYYLNANGAMATGWLQNNGSWYYLNANGSMATGWLQNNGSWYYLNANGA
MATGWLQYNGSWYYLNSNGAMATGWLQYNGSWYYLNANGDMATGWLQNNGSWYYLNANGDMATGWLQYNG
SWYYLNANGDMATGWVKDGDTWYYLEASGAMKASQWFKVSDKWYYVNGSGALAVNTTVDGYGVNANGEWV
N

>YP_817150.1
MANKAVNDFILAMNYDKKKLLTHQGESIENRFIKEGNQLPDEFVVIERKKRSLSTNTSDISVTATNDSRL
YPGALLVVDETLLENNPTLLAVDRAPMTYSIDLPGLASSDSFLQVEDPSNSSVRGAVNDLLAKWHQDYGQ
VNNVPARMQYEKITAHSMEQLKVKFGSDFEKTGNSLDIDFNSVHSGEKQIQIVNFKQIYYTVSVDAVKNP
GDVFQDTVTVEDLKQRGISAERPLVYISSVAYGRQVYLKLETTSKSDEVEAAFEALIKGVKVAPQTEWKQ
ILDNTEVKAVILGGDPSSGARVVTGKVDMVEDLIQEGSRFTADHPGLPISYTTSFLRDNVVATFQNSTDY
VETKVTAYRNGDLLLDHSGAYVAQYYITWDELSYDHQGKEVLTPKAWDRNGQDLTAHFTTSIPLKGNVRN
LSVKIRECTGLAWEWWRTVYEKTDLPLVRKRTISIWGTTLYPQVEDKVEND

>YP_873931.1
MKTKIGLASICLLGLATSHVAANETEVAKTSQDTTTASSSSEQNQSSNKTQTSAEVQTNAAAHWDGDYYV
KDDGSKAQSEWIFDNYYKAWFYINSDGRYSQNEWHGNYYLKSGGYMAQNEWIYDSNYKSWFYLKSDGAYA
HQEWQLIGNKWYYFKKWGYMAKSQWQGSYFLNGQGAMMQNEWLYDPAYSAYFYLKSDGTYANQEWQKVGG
KWYYFKKWGYMARNEWQGNYYLTGSGAMATDEVIMDGTRYIFAASGELKEKKDLNVGWVHRDGKRYFFNN
REEQVGTEHAKKVIDISEHNGRINDWKKVIDENEVDGVIVRLGYSGKEDKELAHNIKELNRLGIPYGVYL
YTYAENETDAESDAKQTIELIKKYNMNLSYPIYYDVENWEYVNKSKRAPSDTGTWVKIINKYMDTMKQAG
YQNVYVYSYRSLLQTRLKHPDILKHVNWVAAYTNALEWENPHYSGKKGWQYTSSEYMKGIQGRVDVSVWY

>ABP89149.1
MTIKYPDGTIDLLSPVEVVKQADKTAPTVANDGKGNIVIVPSEKAVELVVSYVDNNGKSQTVVVTKGTDG
LWTASNTVVIVDPVTGQVIVPGSVIKPGTVVTAYSKDEVGNSSDSAEAEVVAVDENNSAAGVKVKSVTTN
ANNVEKKAKQLPNTGEEANSATSLGLVALGLGLALLAAKRRRDEEA

>ABP89196.1
MFEDPEVDIIYISTPHNTHINYLRKALKAGKHVLCEKSITLNSEELAEAIQLAEENQVVLAEAMTIFHMP
IYRQLSEVVASGKLGELKMIQMNFGSYKEYDMTNRFFNKNLAGGALLDIGVYALSFVRWFMIKKPNQVLS
QVKLAPTGVDEQVGILLSNDAGKMATIALTLHAKQPKRGTIAYDKGYIELYDYPRGQKAVITYTENGIQK
VIEAGETAKALSYEVADMEKAVAGIENTMHLAYTQDVMEIMTQLRKEWGLVYPEKM

>ABP90369.1
MRKSSHLILSSIVSLALVGVTPLSVLADSKQDINQYFQSLTYEPQEILTNEGEYIDNPPATTGMLENGRF
VVLRREKKNITNNSADIAVIDAKAANIYPGALLRADQNLLDNNPTLISIARGDLTLSLNLPGLANGDSHT
VVNSPTRSTVRTGVNNLLSKWNNTYAGEYGNTQAELQYDETMAYSMSQLKTKFGTSFEKIAVPLDINFDA
VNSGEKQVQIVNFKQIYYTVSVDEPESPSKLFAEGTTVEDLKRNGITDEVPPVYVSSVSYGRSMFIKLET
SSRSTQVQAAFKAAIKGVDISGNAEYQDILKNTSFSAYIFGGDAGSAATVVSGNIETLKKIIEEGARYGK
LNPGVPISYSTNFVKDNRPAQILSNSEYIETTSTVHNSSALTLDHSGAYVAKYNITWEEVSYNEAGEEVW
EPKAWDKNGVNLTSHWSETIQIPGNARNLHVNIQECTGLAWEWWRTVYDKDLPLVGQRKITIWGTTLYPQ
YADEVIE

>ABP90422.1
MDFGFDILAIMIPIIVVVLFVLLFVFGYVSAKPNEAIVITGLGKPRTLIGRSGFMIPFIEKRSYISIEQF
STDVQTTDFVPTLDFINVKADAVVKVKVGVSDELLNAAAQNFLNWKTADISASIQDVLEGNLREIIGQME
LRDMVNNRQAFAEKVQSNAAPDLAKMGLEIIAFTVQSFTDDNDVIKNLGIDNIVTIQKDAANARAKAERE
QAEVRAREDKAANDARVAADLEIAKKQNELAIEQANLKRRSDVQLAQANAAYGIEEQAQRKEIERATAEA
NIVKQQKEAEVKAEEVKVREQELSATIRKQAEAEKYARQQAAEADLIERQRKAEAELYETQREAEAQKAR
AEAAKYAAEQEAAGIEAKGRAEAEAIRLKLEAEAEGLSKKADAMAKFNDAAVTEMVVNVLPEIAKNIAAP
LGNVDKITMYGEGNASKMVGDLMTTMDKTTEGLGLNVRDLISATLTGRAMSTGLLTAQETKEVEGDKEK

>YP_001248444.1
MKKIMLIASAMSALSLPFSASAIELEDEVGLECGPYAKVGVVGGMITGAESTRLDSTDSEGKKHLSLTTG
LPFGGTLAAGMTIAPGFRAELGVMYLRNISAEVEVGKGEVDSKGEIKADSGGGTDAPIRKPFKLTPPQPT
MMPISIADRDFGIDIPNIPQAQAQAAQPPLNDQKRAAARIAWLKNCAGIDYMVKDPNNPGHMMVNPVLLN
IPQGNPNPVGQPPQRANQPANFAIHNHEQWRSLVVGLAALSNANKPSASPVKVLSDKIIQIYSDIKPFAD
IAGINVPDTGLPNSASIEQIQSKIQELGDTLEELRDSFDGYINNAFVNQIHLNFVMPPQAQQQQGQGQQQ
QAQATAQEAVAAAAVRLLNGSDQIAQLYKDLVKLQRHAGIRKAMEKLAAQQEEDAKNQGKGDCKQQQGAS
EKSKEGKVKETEFDLSMVVGQVKLYADLVTTESFSIYGGVGAGLAYTYGKIDNKDVKGYTGMVASGALGV
AINAAEGVCVDLEAGYMHSFSKVEDKYQVNAFIASACVRYNF

>YP_001331790.1
MNMKKKEKHAIRKKSIGVASVLVGTLIGFGLLSSKEADASENSVTQSDSASNESKSNDSSSVSAAPKTDD
TNVSDTKTSSNTNNGETSVAQNPAQQETTQSSSTNATTEETPVTGEATTTTTNQANTPATTQSSNTNAEE
LVNQTSNETTFNDTNTVSSVNSPQNSTNAENVSTTQDTSTEATPSNNESAPQSTDASNKDVVNQAVNTSA
PRMRAFSLAAVAADAPAAGTDITNQLTNVTVGIDSGTTVYPHQAGYVKLNYGFSVPNSAVKGDTFKITVP
KELNLNGVTSTAKVPPIMAGDQVLANGVIDSDGNVIYTFTDYVNTKDDVKATLTMPAYIDPENVKKTGNV
TLATGIGSTTANKTVLVDYEKYGKFYNLSIKGTIDQIDKTNNTYRQTIYVNPSGDNVIAPVLTGNLKPNT
DSNALIDQQNTSIKVYKVDNAADLSESYFVNPENFEDVTNSVNITFPNPNQYKVEFNTPDDQITTPYIVV
VNGHIDPNSKGDLALRSTLYGYNSNIIWRSMSWDNEVAFNNGSGSGDGIDKPVVPEQPDEPGEIEPIPED
SDSDPGSDSGSDSNSDSGSDSGSDSTSDSGSDSASDSDSASDSDSASDSDSASDSDSASDSDSDNDSDSD
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSESDSDSESDSDSDSDSDSDSDSDSDSDSDSASD
SDSGSDSDSSSDSDSESDSNSDSESGSNNNVVPPNSPKNGTNASNKNEAKDSKEPLPDTGSEDEANTSLI
WGLLASIGSLLLFRRKKENKDKK

>ACF35726.1
MSLRVRQIDRREWLLAQTATECQRHGREATLEYPTRQGMWVRLSDAEKRWSAWIKPGDWLEHVSPALAGA
AVSAGAEHLVVPWLAATERPFELPVPHLSCRRLCVENPVPGSALPEGKLLHIMSDRGGLWFEHLPELPAV
AGGRPKMLRWPLRFVIGSSDTQRSLLGRIGIGDVLLIRTSRAEVYCYAKKLGHFNRVEGGIIVETLDIQH
IEEENNTTETAETLPGLNQLPVKLEFVLYRKNVTLAELEAMGQQQLLSLPTNAELNVEIMANGVLLGNGE
LVQMNDTLGVEIHEWLSESGNGE

>ACF35754.1
MEKIMAQVINTNSLSLLTQNNLNKSQSALGTAIERLSSGLRINSAKDDAAGQAIANRFTANIKGLTQASR
NANDGISIAQTTEGALNEINNNLQRVRELAVQSANSTNSQSDLDSIQAEITQRLNEIDRVSGQTQFNGVK
VLAQDNTLTIQVGANDGETIDIDLKQINSQTLGLDTLNVQKKYDVKSEAVTPSATLSTTALDGAGLKTGT

GSTTDTGSIKDGKVYYNSTSKNYYVEVEFTDATDQTNKGGFYKVNVADDGAVTMTAATTKEATTPTGITE
VTQVQKPVAAPAAIQAQLTAAHVTGADTAEMVKMSYTDKNGKTIDGGFGVKVGADIYAATKNKDGSFSIN
TTEYTDKDGNTKTALNQLGGADGKTEVVSIDGKTYNASKAAGHNFKAQPELAEAAAATTENPLAKIDAAL
AQVDALRSDLGAVQNRFNSAITNLGNTVNNLSSARSRIEDSDYATEVSNMSRAQILQQAGTSVLAQANQV
PQNVLSLLR

>YP_002364847.1
MKKNTLSAILMTLFLFISCNNSGKDGNTSANSADESVKGPNLTEISKKITDSNAVLLAVKEVEALLSSID
ELAKAIGKKIKNDGSLGDEANHNESLLAGAYTISTLITQKLSKLNGSEGLKEKIAAAKKCSEEFSTKLKD
NHAQLGIQGVTDENAKKAILKANAAGKDKGVEELEKLSGSLESLSKAAKEMLANSVKELTSPVVVESPKK
P

>YP_002345675.1
MNNHKIWRLSAVAVALLISGNSYADQTLHFAQQPHNTGDHTVIAAPEVNNFSLADLTANAEGQAVDATNA
GALSLRVEPTQGDAQANGVLVTSGTFTNQAQGSITVSASSQDQHAVASGLQATDAYWSTLRNQGSIRATS
NSLHGSAEVYGIQAGLAFSDAAGDHYGDHMMLSNEGVMQVSATAEQDAFASGIYTNAGNGVVNNGRLDVS
AHASAGDARATGLHTLSANPGVDPDPWNPANPTHLMSNNGSLSVTASGNNATASAISTEGEGVYIYNNGT
LIVDAFSESAQGGASAYGIHVLNGSATINSSGRILATATGGSQQQAYEVMADGSVVNIERYTLALGNETP
WAVSNGGSIVLGNGTQGADLVLMAGQASEGFAYGKEYSLDNLAYDTTSGQQSTVGGSIAGLYGITPDIKV
IHSGSSSSSLGSAALVYAPDVSHAGVSALAQRSSMEQASNIISSQLHSQLVSNAGCNQKLESADSCVFIT
PYAGEYRRDPVTSGYSGQRYGVLLGQNQHFGDFQLGWHGGYESASTDFNGTSVGRKEEINTLMLGVQGGM
KLSESLFIAATSTWFSSDTDYSDSNTYYGVGSQSGSYDSSGLYTDVSVGNSWQLNRNYAMTPMVGLTHIW
QQRDGYTVSSNNKNYDLIDTRYSSYSDHAVALHAGVRLDGRYPLTNETLLKPFFNVGFQQMLYGDEITID
QSIPNSPVVGVSTKDKTTQGTFDLGMALVSDNGVSASLQLSGMVNSDRQDFTGWANLGWAF

>YP_002346905.1
MKDNNPADNLAWRVIWRQLISSVGSQARMLRRSMLALLLAAFMQGIAFACLYPIIDALLRGDAPQLLNWA
MAFSVAAIVTLVLRWYGLGFEYRGHLAQATHELRLRLGEQLRRVPLEKLQRGRAGEMNALLLGSVDENLN
YVIAIANILLLTIVTPLTASLATLWIDWRLGLVMLLIFPLLVPFYYWRRPAMRRQMQTLGEAHQRLSGDI
VEFAQGMMVLRTCGSDADKSRALLAHFNALENLQTRTHRQGAGATMLIASVVELGLQVVVLSGIVWVVTG
TLNLAFLIAAVAMIMRFAEPMAMFISYTSVVELIASALQRIERFMAIAPLPVAEQSEMPERYDIRFDNVS
YRYEEGDGHALNHVSLTFPAASMSALVGASGAGKTTVTKLLMRYADPQQGQISIGGVDIRRLTPEQLNSL
ISVVFQDVWLFDDTLLANIRIARPQATRQEVEEAARAAQCLEFISRLPQGWLTPMGEMGGQLSGGERQRI
SIARALLKNAPVVILDEPTAALDIESELAVQKAIDNLVHNRTVIIIAHRLSTIAGAGNILVMEEGQVVEQ
GTHAQLLSHHGRYQALWQAQMAARVWRDDGVSASGEWVHE

>YP_002646905.1
MKGRSALLRALWIAALSFGLGGVAVAAEPTAKAAPYENLMVPSPSMGRDIPVAFLAGGPHAVYLLDAFNA
GPDVSNWVTAGNAMNTLAGKGISVVAPAGGAYSMYTNWEQDGSKQWDTFLSAELPDWLAANRGLAPGGHA
AVGAAQGGYGAMALAAFHPDRFGFAGSMSGFLYPSNTTTNGAIAAGMQQFGGVDTNGMWGAPQLGRWKWH
DPWVHASLLAQNNTRVWVWSPTNPGASDPAAMIGQAAEAMGNSRMFYNQYRSVGGHNGHFDFPASGDNGW
GSWAPQLGAMSGDIVGAIR

>YP_003076587.1
MNKKIHSLALLVNLGIYGVAQAQEPTDTPVSHDDTIVVTAAEQNLQAPGVSTITADEIRKNPVARDVSEI
IRTMPGVNLTGNSTSGQRGNNRQIDIRGMGPENTLILIDGKPVSSRNSVRQGWRGERDTRGDTSWVPPEM
IERIEVLRGPAAARYGNGAAGGVVNIITKKGSGEWHGSWDAYFNAPEHKEEGATKRTNFSLTGPLGDEFS
FRLYGNLDKTQADAWDINQGHQSARAGTYATTLPAGREGVINKDINGVVRWDFAPLQSLELEAGYSRQGN
LYAGDTQNTNSDAYTRSKYGDETNRLYRQNYSLTWNGGWDNGVTTSNWVQYEHTRNSRIPEGLAGGTEGK
FNEKATQDFVDNDLDDVMLHSEVNLPIDFLVNQTLTLGTEWNQQRMKDLSSNTQALTGTNTGGAIDGVSA
TDRSPYSKAEIFSLFAENNMELTDSTIVTPGLRFDHHSIVGNNWSPALNISQGLGDDFTLKMGIARAYKA
PSLYQTNPNYILYSKGQGCYASAGGCYLQGNDDLKAETSINKEIGLEFKRDGWLAGITWFRNDYRNKIEA
GYVAVGQNAVGTDLYQWDNVPKAVVEGLEGSLNVPVSETVMWTNNITYMLKSENKTTGDRLSIIPEYTLN
STLSWQAREDLSMQTTFTWYGKQQPKKYNYKGQPAVGPETKEISPYSIVGLSATWDVTKNVSLTGGVDNL
FDKRLWRAGNAQTTGDLAGANYIAGAGAYTYNEPGRTWYMSVNTHF

>YP_003110629.1
MIKCNNKTFNNLLKLTILVNLLISCGLTGATKIKLESSAKAIVDEIDAIKKKAASMGVNFDAFKDKKTGS
GVSENPFILEAKVRATTVAEKFVIAIEEEATKLKETGSSGEFSAMYDLMFEVSKPLQELGIQEMTKTVSM
AAEENPPTTAQGVLEIAKKMREKLQRVHKKNQDTLKKKNTEDSTAKS

>YP_003293996.1
MNKVKCYVLFTALLSSLCAYGAPQSITELCSEYRNTQIYTINDKILSYTESMAGKREMVIITFKSGATFQ
VEVPGSQHIDSQKKAIERMKDTLRITYLTETKIDKLCVWNNKTPNSIAAISMEN

>ACZ17766.1
MKRVITLFAVLLMGWSVNAWSFACKTANGTAIPIGGGSANVYVNLAPAVNVGQNLVVDLSTQIFCHNDYP
ETITDYVTLQRGAAYGGVLSSFSGTVKYNGSSYPFPTTSETPRVVYNSRTDKPWPVALYLTPVSSAGGVA
IKAGSLIAVLILRQTNNYNSDDFQFVWNIYANNDVVVPTGGCDVSARDVTVTLPDYPGSVPIPLTVYCAK
SQNLGYYLSGTTADAGNSIFTNTASFSPAQGVGVQLTRNGTIIPANNTVSLGAVGTSAVSLGLTANYART
GGQVTAGNVQSIIGVTFVYQ

>ADE88959.1
MSRYKTDNKQPRFRYSVLARCVAWANISVQVLFPLAVTFTPVMAARAQHAVQPRLSMENTTVTADNNVEK
NVASLAANAGTFLSSQPDSDATRNFITGMATAKANQEIQEWLGKYGTARVKLNVDKNFSLKDSSLEMLYP
IYDTPTNMLFTQGAIHRTDDRTQSNIGFGWRHFSENDWMAGVNTFIDHDLSRSHTRIGVGAEYWRDYLKL
SANGYIRASGWKKSPDVEDYQERPANGWDIRAEGYLPAWPQLGASLMYEQYYGDEVGLFGKDKRQKDPHA
ITAEVNYTPVPLLTLSAGHKQGKSGENDTRFGLEVNYRIGEPLEKQLDTDSIRERRMLAGSRYDLVERNN
NIVLEYRKSEVIRIALPERIEGKGGQTVSLGLVVSKATHGLKNVQWEAPSLLAAGGKITGQGNQWQVTLP
AYQAGKDNYYAISAIAYDNKGNASKRVQTEVVISGAGMSADRTALTLDGQSRIQMLANGNEQKPLVLSLR
DAEGQPVTGMKDQIKTELTFKPAGNIVTRTLKATKSQAKPTLGEFTETEAGVYQSVFTTGTQSGEATITV
SVDDMSKTVTAELRATMMDVSNSTLSANEPSGDVVADGQQAYTLTLTAVDSEGNPVTGEASRLRLVPQDT
NGVTVGAISEIKPGVYSATVSSTRAGNVVVRAFSEQYQLGTLQQTLKFVAGPLDAAHSSITLNPDKPVVG
GTVTAIWTAKDANDNPVTGLNPDAPSLSGAAAAGSTASGWTDNGDGTWTAQISLGTTAGELDVMPKLNGQ
DAAANAAKVTVVADALSSNQSKVSVAEDHVKAGESTTVTLVAKDAHGNAISGLSLSASLTGTASEGATVS
SWTEKGDGSYVATLTTGGKTGELRVMPLFNGQPAATEAAQLTVIAGEMSSANSTLVADNKTPTVKTTTEL
TFTMKDAYGNPVTGLKPDAPVFSGAASTGSERPSAGNWTEKGNGVYVSTLTLGSAAGQLSVMPRVNGQNA
VAQPLVLNVAGDASKAEIRDMTVKVNNQLANGQSANQITLTVVDSYGNPLQGQEVTLTLPQGVTSKTGNT
VTTNAAGKVDIELMSTVAGELEIEASVKNSQKTVKVKFKADFSTGQASLEVDAAAQKVANGKDAFTLTAT
VKDQYGNLLPGAVVVFNLPRGVKPLADGNIMVNADKEGKAELKVVSVTAGTYEITASAGNDQPSNAQSVT
FVADKTTATISSIEVIGNRAVADGKTKQTYKVTVTDANNNLLKDSEVTLTASPENLVLTPNGTATTNEQG
QAIFTATTTVAATYTLTAKVEQADGQESTKTAESKFVADDKNAVLAASPERVDSLVADGKTTATLTVTLM
SGVNPVGGTMWVDIEAPEGVTEADYQFLPSKNDHFASGKITRTFSTNKPGTYTFTFNSLTYGGYEMKPVT
VTINAVPADTEGAEEK

>ADE89421.1
MNKKFKYKKSLLAAILSATLLAGCDGGGSGSSSDTPSVDSGSGTLPEVKPDPTPTPEPTPEPTPDPEPTP
DPTPDPEPTPEPEPEPVPTKTGYLTLGGSQRVTGATCNGESSDGFTFTPGNTVSCVVGSTTIATFNTQSE
AARSLRAVDKVSFSLEDAQELANSENKKTNAISLVTSSDSCPADAEQLCLTFSSVVDRARFEKLYKQIDL
ATDNFSKLVNEEVENNAATDKAPSTHTSTVVPVTTEGTKPDLNASFVSANAEQFYQYQPTEIILSEGQLV
DSLGNGVAGVDYYTNSGRGVTDENGKFSFSWGETISFGIDTFELGSVRGNKSTIALTELGDEVRGANIDQ
LIHRYSTTGQNNTRVVPDDVRKVFAEYPNVINEIINLSLSNGATLDEGDQNVVLPNEFIEQFKTGQAKEI
DTAICAKTDGCNEARWFSLTTRNVNDGQIQGVINKLWGVDTNYQSVSKFHVFHDSTNFYGSTGNARGQAV
VNISNSAFPILMARNDKNYWLAFGEKRAWDKNELAYITEAPSIVQPENVTRDTATFNLPFISLGQVGEGK
LMVIGNPHYNSILRCPNGYSWGGGVNSKGECTLSGDSDDMKHFMQNVLRYLSNDIWQPNTKSIMTVGTNL
ENVYFKKAGQVLGNSAPFAFHEDFTGITVKQLTSYGDLNPEEIPLLILNGFEYVTQWSGDPYAVPLRADT
SKPKLTQQDVTDLIAYLNKGGSVLIMENVMSNLKEESASSFVRLLDAAGLSMALNKSVVNNDPQGYPDRV
RQRRATGIWVYERYPAADGAQPPYTIDPNTGEVTWKYQQDNKPDDKPKLEVASWQEEVEGKQVTRYAFID

EAEYTTEESLEAAKAKIFEKFPGLQECKDSTYHYEINCLERRPGTDVPVTGGMYVPRYTQLNLDADTAKA
MVQAADLGTNIQRLYQHELYFRTKGSKGERLNSVDLERLYQNMSVWLWNDTKYRYEEGKEDELGFKTFTE
FLNCYANDAYAGGTKCSADLKKSLVDNNMIYGDGSSKAGMMNPSYPLNYMEKPLTRLMLGRSWWDLNIKV
DVEKYPGSVSAKGESVTENISLYSNPTKWFAGNMQSTGLWAPAQQDVTIKSSASVPVTVTVALADDLTGR
EKHEVALNRPPRVTKTYTLEANGEVTFKVPYGGLIYIKGDSKDDVSANFTFTGVVKAPFYKDGEWKNDLD
SPAPLGELESASFVYTTPKKNLEASNFTGGVAEFAKDLDTFASSMNDFYGRNDEDGKHRMFTYKNLTGHK
HRFTNDVQISIGDAHSGYPVMNSSFSTNSTTLPTTPLNDWLIWHEVGHNAAETPLNVPGATEVANNVLAL
YMQDRYLGKMNRVADDITVAPEYLDESNGQAWARGGAGDRLLMYAQLKEWAEENFDIKQWYPDGELPKFY
SDRKGMKGWNLFQLMHRKARGDDVGNSTFGGKNYCAESNGNAADTLMLCASWVAQADLSEFFKKWNPGAS
AYQLPGATEMSFQGGVSSSAYSTLASLKLPKPEKGPETINKVTEHKMSAE

>ADE91247.1
MAMFTPSFSGLKGRALFSLLFAAPMIHATDSVTTKDGETITVTADANTATEATDGYQPLSTSTATLTDMP
MLDIPQVVNTVSDQVLENQNATTLDEALYNVSNVVQTNTLGGTQDAFVRRGFGANRDGSIMTNGLRTVLP
RSFNAATERVEVLKGPASTLYGILDPGGLINVVTKRPEKTFHGSVSATSSSFGGGTGQLDITGPIEGTQL
AYRLTGEVQDEDYWRNFGKERSTFIAPSLTWFGDNATVTMLYSHRDYKTPFDRGTIFDLTTKQPVNVDRK
IRFDEPFNITDGQSDLAQLNAEYHLNSQWTARFDYSYSQDKYSDNQARVTAYDATTGTLTRRVDATQGST
QRMHATRADLQGNVDIAGFYNEILGGVSYEYYDLLRTDMIRCKKAKDFNIYNPVYGNTSKCTTVSASDSD
QTIKQENYSAYAQDALYLTDNWIAVAGIRYQYYTQYAGKGRPFNVNTDSRDEQWTPKLGLVYKLTPSVSL
FANYSQTFMPQSSIASYIGDLPPESSNAYEVGAKFELFDGITADIALFDIHKRNVLYTESIGDETIAKTA
GRVRSRGVEVDLAGALTENINIIASYGYTDAKVLEDPDYAGKPLPNVPRHTGSLFLTYDIHNMPGNNTLT
FGGGGHGVSRRSATNGADYYLPGYFVADAFAAYKMKLQYPVTLQLNVKNLFDKTYYTSSIATNNLGNQIG
DPREVQFTVKMEF

>ADE91828.1
MITSPPKRGMALVVVLVLLAVMMLVTITLSGRMQQQLGRTRSQQEYQQALWYSASAESLALSALSLSLKN
EKRVHLAQPWASGPRFFPLPQGQIAVTLRDAQACFNLNALAQPTTASRPLAVQQLIALISRLDVPAYRAE
LIAESLWEFIDEDRSVQTRLGREDSEYLARSVPFYAANQPLADISEMRVVQGMDAGLYQKLKPLVCALPM
ARQQININTLDVTQSVILEALFDPWLSPVQARALLQQRPAKGWEDVDQFLAQPLLADVDERTKKQLKTIL
SVDSNYFWLRSDITVNEIELTMNSLIVRMGPQHFSVLWHQTGESE

>NP_045689.2
MRLLIGFALALALIGCAQKGAESIGSQKENDLNLEDSSKKSHQNAKQDLPAVTEDSVSLFNGNKIFVSKE
KNSSGKYDLRATIDQVELKGTSDKNNGSGTLEGSKPDKSKVKLTVSADLNTVTLEAFDASNQKISSKVTK
KQGSITEETLKANKLDSKKLTRSNGTTLEYSQITDADNATKAVETLKNSIKLEGSLVGGKTTVEIKEGTV
TLKREIEKDGKVKVFLNDTAGSNKKTGKWEDSTSTLTISADSKKTKDLVFLTDGTITVQQYNTAGTSLEG
SASEIKNLSELKNALK

>ZP_05300691.1
MLVFITLILVSLPIAQQTEAKDASAFNKENSISSMAPPASPPASPKTPIEKKHADEIDKYIQGLDYNKNN
VLVYHGDAVTNVPPRKGYKDGNEYIVVEKKKKSINQNNADIQVVNAISSLTYPGALVKANSELVENQPDV
LPVKRDSLTLSIDLPGMTNQDNKIVVKNATKSNVNNAVNTLVERWNEKYAQAYPNVSAKIDYDDEMAYSE
SQLIAKFGTAFKAVNNSLNVNFGAISEGKMQEEVISFKQIYYNVNVNEPTRPSRFFGKAVTKEQLQALGV
NAENPPAYISSVAYGRQVYLKLSTNSHSTKVKAAFDAAVSGKSVSGDVELTNIIKNSSFKAVIYGGSAKD
EVQIIDGNLGDLRDILKKGATFNRETPGVPIAYTTNFLKDNELAVIKNNSEYIETTSKAYTDGKINIDHS
GGYVAQFNISWDEVNYDPEGNEIVQHKNWSENNKSKLAHFTSSIYLPGNARNINVYAKECTGLAWEWWRT
VIDDRNLPLVKNRNISIWGTTLYPKYSNKVDNPIE

>ZP_05821343.1
MMFARPLIASMLGAAVCLTALGTVAAPAFAAGESEVKVIVSGNAITNSDIKHRMAFLKLQRKSGNLNQLA
RNELTEEMLKRIEMKSRGINISDKEVDDAYAGFASRNKMTLAQLNQVMNQSGVTPEHFKKYIMVQMGWGR
LVSARFRATGMVSEQEAVQRMLKNGGKKPVATEYHLQQVIFVVPASKRSPALLAKRRQEANALRARFQNC
DSTRQQAKGILDVTVRDLGRIIEPQLPGEWSKDVKAAGVNRTTKPHDTEKGVEFLAVCSTRQVSDDRVAQ
LVFSMEGADSPAGQEKKAEELSKKYVQELREKATIVNR

>ZP_00785385.1
MKKKMIQSLLVASLAFGMAVSPVTPIAFAAETGTITVQDTQKGATYKAYKVFDAEIDNANVSDSNKDGAS
YLIPQGKEAEYKASTDFNSLFTTTTNGGRTYVTKKDTASANEIATWAKSISANTTPVSTVTESNNDGTEV
INVSQYGYYYVSSTVNNGAVIMVTSVTPNATIHEKNTDATWGDGGGKTVDQKTYSVGDTVKYTITYKNAV
NYHGTEKVYQYVIKDTMPSASVVDLNEGSYEVTITDGSGNITTLTQGSEKATGKYNLLEENNNFTITIPW
AATNTPTGNTQNGANDDFFYKGINTITVTYTGVLKSGAKPGSADLPENTNIATINPNTSNDDPGQKVTVR
DGQITIKKIDGSTKASLQGAIFVLKNATGQFLNFNDTNNVEWGTEANATEYTTGADGIITITGLKEGTYY
LVEKKAPLGYNLLDNSQKVILGDGATDTTNSDNLLVNPTVENNKGTELPSTGGIGTTIFYIIGAILVIGA
GIVLVARRRLRS

>ZP_00785378.1
MLKKCQTFIIESLKKKKHPKEWKIIMWSLMILTTFLTTYFLILPAITVEETKTDDVGITLENKNSSQVTS
STSSSQSSVEQSKPQTPASSVTETSSSEEAAYREEPLMFRGADYTVTVTLTKEAKIPKNADLKVTELKDN
SATFKDYKKKALTEVAKQDSEIKNFKLYDITIESNGKEAEPQAPVKVEVNYDKPLEASDENLKVVHFKDD
GQTEVLKSKDTAETKNTSSDVAFKTDSFSIYAIVQEDNTEVPRLTYHFQNNDGTDYDFLTASGMQVHHQI
IKDGESLGEVGIPTIKAGEHFNGWYTYDPTTGKYGDPVKFGEPITVTETKEICVRPFMSKVATVTLYDDS
AGKSILERYQVPLDSSGNGTADLSSFKVSPPTSTLLFVGWSKTQNGAPLSESEIQALPVSSDISLYPVFK
ESYGVEFNTGDLSTGVTYIAPRRVLTGQPASTIKPNDPTRPGYTFAGWYTAASGGAAFDFNQVLTKDTTL
YAHWSPAQTTYTINYWQQSATDNKNATDAQKTYEYAGQVTRSGLSLSNQTLTQQDINDKLPTGFKVNNTR
TETSVMIKDDGSSVVNVYYDRKLITIKFAKYGGYSLPEYYYSYNWSSDADTYTGLYGTTLAANGYQWKTG
AWGYLANVGNNQVGTYGMSYLGEFILPNDTVDSDVIKLFPKGNIVQTYRFFKQGLDGTYSLADTGGGAGA
DEFTFTEKYLGFNVKYYQRLYPDNYLFDQYASQTSAGVKVPISDEYYDRYGAYHKDYLNLVVWYERNSYK
IKYLDPLDNTELPNFPVKDVLYEQNLSSYAPDTTTVQPKPSRPGYVWDGKWYKDQAQTQVFDFNTTMPPH
DVKVYAGWQKVTYRVNIDPNGGRLSKTDDTYLDLHYGDRIPDYTDITRDYIQDPSGTYYYKYDSRDKDPD
STKDAYYTTDTSLSNVDTTTKYKYVKDAYKLVGWYYVNPDGSIRPYNFSGAVTQDINLRAIWRKAGDYHI
IYSNDAVGTDGKPALDASGQQLQTSNEPTDPDSYDDGSHSALLRRPTMPDGYRFRGWWYNGKIYNPYDSI
DIDAHLADANKNITIKPVIIPVGDIKLEDTSIKYNGNGGTRVENGNVVTQVETPRMELNSTTTIPENQYF
TRTGYNLIGWHHDKDLADTGRVEFTAGQSIGIDNNPDATNTLYAVWQPKEYTVRVSKTVVGLDEDKTKDF
LFNPSETLQQENFPLRDGQTKEFKVPYGTSISIDEQAYDEFKVSESITEKNLATGEADKTYDATGLQSLT
VSGDVDISFTNTRIKQKVRLQKVNVENDNNFLAGAVFDIYESDANGNKASHPMYSGLVTNDKGLLLVDAN
NYLSLPVGKYYLTETKAPPGYLLPKNDISVLVISTGVTFEQNGNNATPIKENLVDGSTVYTFKITNSKGT
ELPSTGGIGTHIYILVGLALALPSGLILYYRKKI

>ZP_05821969.1
MERNMAKVIGIDLGTTNSCVAVMDGKNAKVIENAEGARTTPSIIAFTDGDERLAGQPAKRQAVTNPEGTL
FAVKRLIGRRYDDPMVTKDKDLVPYKIVKGDNGDAWVEVHGKKYSPSQISAMILQKMKETAESYLGETVT
QAVITVPAYFNDAQRQATKDAGKIAGLEVLRIINEPTAAALAYGLDKSEGKTIAVYDLGGGTFDVSVLEI
GDGVFEVKSTNGDTFLGGEDFDIRLVEYLVAEFKKESGIDLKNDKLALQRLKEAAEKAKIELSSSQQTEI
NLPFITADQTGPKHLAIKLSRAKFESLVDDLVQRTVEPCKAALKDAGLKAGEIDEVVLVGGMTRMPKIQE
VVKAFFGKEPHKGVNPDEVVAMGAAIQGGVLQGDVKDVLLLDVTPLSLGIETLGGVFTRLIERNTTIPTK
KSQTFSTAEDNQSAVTIRVFQGEREMAADNKLLGQFDLVGIPPAPRGVPQIEVTFDIDANGIVNVSAKDK
GTGKEHQIRIQASGGLSDADIEKMVKDAEANAEADKKRRESVEAKNQAESLVHSTEKSLAEYGDKVSADD
KKAIEDAIAALKTSLEGEDAEDIKAKTQALAEVSMKLGQAMYEAAQAAEGAGAEGGEQASSSKDDVVDAD
YEEIDDNKKSS

>CAE55522.1
MDFGALPPEVNSARMYGGAGAADLLAAAAAWNGIAVEVSTAASSVGSVITRLSTEHWMGPASLSMAAAVQ
PYLVWLTCTAESSALAAAQAMASAAAFETAFALTVPPAEVVANRALLAELTATNILGQNVSAIAATEARY
GEMWAQDASAMYGYAAASAVAARLNPLTRPSHITNPAGLAHQAAAVGQAGASAFARQVGLSHLISDVADA
VLSFASPVMSAADTGLEAVRQFLNLDVPLFVESAFHGLGGVADFATAAIGNMTLLADAMGTVGGAAPGGG
AAAAVAHAVAPAGVGGTALTADLGNASVVGRLSVPASWSTAAPATAAGAALDGTGWAVPEEDGPIAVMPP
APGMVVAANSVGADSGPRYGVKPIVMPKHGLF

>AAA26522.1
MHNVSTTTTGFPLAKILTSTELGDNTIQAANDAANKLFSLTIADLTANQNINTTNAHSTSNILIPELKAP
KSLNASSQLTLLIGNLIQILGEKSLTALTNKITAWKSQQQARQQKNLEFSDKINTLLSETEGLTRDYEKQ
INKLKNADSKIKDLENKINQIQTRLSNLDPESPEKKKLSREEIQLTIKKDAAVKDRTLIEQKTLSIHSKL
TDKSMQLEKEIDSFSAFSNTASAEQLSTQQKSLTGLASVTQLMATFIQLVGKNNEESLKNDLALFQSLQE
SRKTEMERKSDEYAAEVRKAEELNRVMGCVGKILGALLTIVSVVAAAFSGGASLALAAVGLALMVTDAIV
QAATGNSFMEQALNPIMKAVIEPLIKLLSDAFTKMLEGLGVDSKKAKMIGSILGAIAGALVLVAAVVLVA
TVGKQAAAKLAENIGKIIGKTLTDLIPKFLKNFSSQLDDLITNAVARLNKFLGAAGDEVISKQIISTHLN
QAVLLGESVNSATQAGGSVASAVFQNSASTNLADLTLSKYQVEQLSKYISEAIEKFGQLQEVIADLLASM
SNSQANRTDVAKAILQQTTA

>AAA26524.1
MNITTLTNSISTSSFSPNNTNGSSTETVNSDIKTTTSSHPVSSLTMLNDTLHNIRTTNQALKKELSQKTL
TKTSLEEIALHSSQISMDVNKSAQLLDILSRNEYPINKDARELLHSAPKEAELDGDQMISHRELWAKIAN
SINDINEQYLKVYEHAVSSYTQMYQDFSAVLSSLAGWISPGGNDGNSVKLQVNSLKKALEELKEKYKDKP
LYPANNTVSQEQANKWLTELGGTIGKVSQKNGGYVVSINMTPIDNMLKSLDNLGGNGEVVLDNAKYQAWN
AGFSAEDETMKNNLQTLVQKYSNANSIFDNLVKVLSSTISSCTDTDKLFLHF

>AAC44468.1
MFRRSKNNSYDTLQTKQRFSIKKFKFGAASVLIGISFLGGFTQGQFNISTDTVFAAEVISGSAVTLNTNM
TKNVQNGRAYIDLYDVKNGKIDPLQLITLNSPDLKAQYVIRQGGNYFTQPSELTTVGAASINYTVLKTDG
SPHTKPDGQVDIINVSLTIYNSSALRDKIDEVKKKAEDPKWDEGSRDKVLISLDDIKTDIDNNPKTQSDI
ANKITEVTNLEKILVPRIPDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAFETPVDTATPGD
KPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAFET
PVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLP
KGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAED
SIGNLPDLPKGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVN
VGETPKAEDSIGNLPDLPKGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKND
PAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVD
PRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDT
VDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAFETPVDTATPGDKPAKVVV
TYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAFETPVDTATP
GDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPDLPKGTTVAF
ETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKAEDSIGNLPD
LPKGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQVNVGETPKA
EDSIGNLPDLPKGTTVAFETPVDTATPGDKPAKVVVTYPDGSKDTVDVTVKVVDPRTDADKNDPAGKDQQ
VNGKGNKLPATGENATPFFNVVALTIMSSVGLLSVSKKKED

>AAC02243.1
MKKTIVALAVAAVAATSANAATVYNQDGTKVDVNGSLRLILKKEKNERGDLVDNGSRVSFKASHDLGEGL
SALAYTELRFSKNVPVQVKDQQGEVVREYEVEKLGNNVHVKRLYAGFAYEGLGTLTFGNQLTIGDDVGLS
DYTYFNSGINNLLSSGEKAINFKSAEFNGFTFGGAYVFSADADKQALRDGRGFVVAGLYNRKMGDVGFAF
EAGYSQKYVKQEVEQNPPAAQKVFKDEKEKAFMVGAELSYAGLALGVDYAQSKVTNVDGKKRALEVGLNY
DLNDRAKVYTDFIWEKEGPKGDVTRNRTVAVGFGYKLHKQVETFVEAAWGREKDSDGVTTKNNVVGTGLR
VHF

>CAB46338.1
MFRQYNFEKGLNFSIRKFSVGIASVAIGSLLFILPQVLADETTSVTSATTPTGVTTTDANLVNPNNSTPT
STNRSATSTQGSNLSNTSEIIKPATLAATSPTTDNVAPSVDKRTYATSGDWTLQNPYADSVRNKNISPSV
RHESFKSAETTVVRHDNSTVKVTATITPVEGNDEGSGILTNGGNQSEYKATSEMFVGNVDPAKIPALGVY
TQPGRTEGGSKLSDKLNFNGKAPSTILTLKFDKAVTDPIIDLSGVGGNARLSFTETVMENGKIVEKFDYA
RGSYNSTFFDGITPGISLEKASSGVNLTVTANTVDVTDKNTFNESVVNPSDEDFVNGPDRTPDAVPAGTG
SIRLKGTFTDASFKLYHQAVPSTAFSKEKYHTGDGYYNSIANVRPTVVKPDSINGLNKHASDVIDYKISD
NNDISNDDLLRLSVRLQNPRGSVVVNYIDTEGNIIGTEYKDTTDAIPGTHYNTAESSGDLNSDATVERPS
TITKDGKVYDLVAENITVPVGKVNSDGTLATNGSSFNYGTDAASAEVAEGTKSVIYVISIKQESKGNVHA
RYVILGTETELASAKTVKSEAPIDEAYSDKAPATLEKDGKLYEFVHVRDNKGDAPADGKVTEQDQTITYE
YVEVPKGRVVVDYVIEGTATKPKDTYVDTPTAYIRDKEGQAIPYNTAENDSEKPLLLDKDGIKYELVSIQ
EGSAAEKGTLREGEQHVVYQYRKVVEVPSVKVGNVHARYVILGTETELASAKTVKSEAPIDEAYSDKAPA
TLEKDGKLYEFVHVRDNKGDAPADGKVTEQDQTITYEYKLKKDDADAVGNVVINYVDETGNVIKKPILDT
HESKVGTPYDTTDYKFAEIKFNGKIYKLVSAKTMGNEFGKVTEGTTEVTYVYRESVKSTLPKETGISIPS
ESESPKFISTQTTENRPNKGVLTSSKNPTNKSVLPTTGEESNRILGVVGITLVATTATLAASSLKRRKN

>AAF41790.1
MRKKLTALVLSALPLAAVADVSLYGEIKAGVEGRNYQLQLTEAQAANGGASGQVKVTKVTKAKSRIRTKI
SDFGSFIGFKGSEDLGDGLKAVWQLEQDVSVAGGGATQWGNRESFIGLAGEFGTLRAGRVANQFDDASQA
IDPWDSNNDVASQLGIFKRHDDMPVSVRYDSPEFSGFSGSVQFVPIQNSKSAYTPAYYTKNTNNNLTLVP
AVVGKPGSDVYYAGLNYKNGGFAGNYAFKYARHANVGRNAFELFLIGSGSDQAKGTDPLKNHQVHRLTGG
YEEGGLNLALAAQLDLSENGDKTKNSTTEIAATASYRFGNAVPRISYAHGFDFIERGKKGENTSYDQIIA
GVDYDFSKRTSAIVSGAWLKRNTGIGNYTQINAASVGLRHKF

>AAF42074.1
MNLKLVFESGDPVLIGVFVLMLLMSIVTWCLVVLRCIKLYRARKGNAAVKRHMRDTLSLNDAVEKVRAVD
APLSKLAQEALQSYRNYRRNEASELAQALPLNEYLVIQIRNSMAQIMRRFDYGMTALASIGATAPFIGLF
GTVWGIYHALINIGQSGQMSIAAVAGPIGEALVATAAGLFVAIPAVLAYNFLNRGTKILTQDLDAMAHDL
HVRLLNQKDS

>A46405
MFRRSKNNSYDTSQTKQRFSIKKFKFGAASVLIGLSFLGGVTQGNLNIFEESIVAASTIPGSAATLNTSI
TKNIQNGNAYIDLYDVKLGKIDPLQLIVLEQGFTAKYVFRQGTKYYGDVSQLQSTGRASLTYNIFGEDGL
PHVKTDGQIDIVSVALTIYDSTTLRDKIEEVRTNANDPKWTEESRTEVLTGLDTIKTDIDNNPKTQTDID
SKIVEVNELEKLLVLSVPDKDKYDPTGGETTVPQGTPVSDKEITDLVKIPDGSKGVPTVVGDRPDTNVPG
DHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDKYDPTGGETTVPQGTPVSDKEITDLVKIPDGSKGVPT
VVGDRPDTNVPGDHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDKYDPTGGETTVPQGTPVSDKEITDL
VKIPDGSKGVPTVVGDRPDTNVPGDHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDKYDPTGGETTVPQ
GTPVSDKEITDLVKIPDGSKGVPTVVGDRPDTNVPGDHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDK
YDPTGGETTVPQGTPVSDKEITDLVKIPDGSKGVPTVVGDRPDTNVPGDHKVTVEVTYPDGTKDTVEVTV
HVTPKPVPDKDKYDPTGGETTVPQGTPVSDKEITDLVKIPDGSKGVPTVVGDRPDTNVPGDHKVTVEVTY
PDGTKDTVEVTVHVTPKPVPDKDKYDPTGGETTVPQGTPVSDKEITDLVKIPDGSKGVPTVVGDRPDTNV
PGDHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDKYDPTGGETTVPQGTPVSDKEITDLVKIPDGSKGV
PTVVGDRPDTNVPGDHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDKYDPTGGETTVPQGTPVSDKEIT
DLVKIPDGSKGVPTVVGDRPDTNVPGDHKVTVEVTYPDGTKDTVEVTVHVTPKPVPDKDKYDPTGKAQQV
NGKGNKLPATGENATPFFNVAALTIISSVGLLSVSKKKED

>1DI0_A
MNQSCPNKTSFKIAFIQARWHADIVDEARKSFVAELAAKTGGSVEVEIFDVPGAYEIPLHAKTLARTGRY
AAIVGAAFVIDGGIYDHDFVATAVINGMMQVQLETEVPVLSVVLTPHHFHESKEHHDFFHAHFKVKGVEA
AHAALQIVSERSRIAALV

>AAK33158.1
MKKRILSAVLVSGVTLGAATTVGAEDLSTKIAKQDSIISNLTTEQKAAQNQVSALQAQVSSLQSEQDKLT
ARNTELEALSKRFEQEIKALTSQIVARNEKLKNQARSAYKNNETSGYINALLNSKSISDVVNRLVAINRA
VSANAKLLEQQKADKVSLEEKQAANQTAINTIAANMAMAEENQNTLRTQQAANLVAATANLALQLASATED
KANLVAQKEAAEKAAAEALAQEQAAKVKAQEQAAQQAASVEAAKSAITPAPQATPAAQSSNAIEPAALTA
PAAPSAGPQTSYDSSNTYPVGQCTWGAKSLAPWAGNNWGNGGQWAYSAQAAGYRTGSTPMVGAIAVWNDG
GYGHVAVVVEVQSASSIRVMESNYSGRQYIADHRGWFNPTGVTFIYPH

>AAK33267.1
MSNKKTFKKYSRVAGLLTAALIIGNLVTANAESNKQNTASTETTTTNEQPKPESSELTTEKAGQKTDDML

NSNDMIKLAPKEMPLESAEKEEKKSEDKKKSEEDHTEEINDKIYSLNYNELEVLAKNGETIENFVPKEGV
KKADKFIVIERKKKNINTTPVDISIIDSVTDRTYPAALQLANKGFTENKPDAVVTKRNPQKIHIDLPGMG
DKATVEVNDPTYANVSTAIDNLVNQWHDNYSGGNTLPARTQYTESMVYSKSQIEAALNVNSKILDGTLGI
DFKSISKGEKKVMIAAYKQIFYTVSANLPNNPADVFDKSVTFKELQRKGVSNEAPPLFVSNVAYGRTVFV
KLETSSKSNDVEAAFSAALKGTDVKTNGKYSDILENSSFTAVVLGGDAAEHNKVVTKDFDVIRNVIKDNA
TFSRKNPAYPISYTSVFLKNNKIAGVNNRTEYVETTSTEYTSGKINLSHQGAYVAQYEILWDEINYDDKG
KEVITKRRWDNNWYSKTSPFSTVIPLGANSRNIRIMARECTGLAWEWWRKVIDERDVKLSKEINVNISGS
TLSPYGSITYK

>AAK33444.1
MEKKQRFSLRKYKSGTFSVLIGSVFLVMTTTVAADELSTMSEPTITNHAQQQAQHLTNTELSSAESKSQD
TSQITLKTNREKEQSQDLVSEPTTTELADTDAASMANTGSDATQKSASLPPVNTDVHDWVKTKGAWDKGY
KGQGKVVAVIDTGIDPAHQSMRISDVSTAKVKSKEDMLARQKAAGINYGSWINDKVVFAHNYVENSDNIK
ENQFEDFDEDWENFEFDAEAEPKAIKKHKIYRPQSTQAPKETVIKTEETDGSHDIDWTQTDDDTKYESHG
MHVTGIVAGNSKEAAATGERFLGIAPEAQVMFMRVFANDIMGSAESLFIKAIEDAVALGADVINLSLGTA
NGAQLSGSKPLMEAIEKAKKAGVSVVVAAGNERVYGSDHDDPLATNPDYGLVGSPSTGRTPTSVAAINSK
WVIQRLMTVKELENRADLNHGKAIYSESVDFKDIKDSLGYDKSHQFAYVKESTDAGYNAQDVKGKIALIE
RDPNKTYDEMIALAKKHGALGVLIFNNKPGQSNRSMRLTANGMGIPSAFISHEFGKAMSQLNGNGTGSLE
FDSVVSKAPSQKGNEMNHFSNWGLTSDGYLKPDITAPGGDIYSTYNDNHYGSQTGTSMASPQIAGASLLV
KQYLEKTQPNLPKEKIADIVKNLLMSNAQIHVNPETKTTTSPRQQGAGLLNIDGAVTSGLYVTGKDNYGS
ISLGNITDTMTFDVTVHNLSNKDKTLRYDTELLTDHVDPQKGRFTLTSHSLKTYQGGEVTVPANGKVTVR
VTMDVSQFTKELTKQMPNGYYLEGFVRFRDSQDDQLNRVNIPFVGFKGQFENLAVAEESIYRLKSQGKTG
FYFDESGPKDDIYVGKHFTGLVTLGSETNVSTKTISDNGLHTLGTFKNADGKFILEKNAQGNPVLAISPN
GDNNQDFAAFKGVFLRKYQGLKASVYHASDKEHKNPLWVSPESFKGDKNFNSDIRFAKSTTLLGTAFSGK
SLTGAELPDGHYHYVVSYYPDVVGAKRQEMTFDMILDRQKPVLSQATFDPETNRFKPEPLKDRGLAGVRK
DSVFYLERKDNKPYTVTINDSYKYVSVEDNKTFVERQADGSFILPLDKAKLGDFYYMVEDFAGNVAIAKL
GDHLPQTLGKTPIKLKLTDGNYQTKETLKDNLEMTQSDTGLVTNQAQLAVVHRNQPQSQLTKMNQDFFIS
PNEDGNKDFVAFKGLKNNVYNDLTVNVYAKDDHQKQTPIWSSQAGASVSAIESTAWYGITARGSKVMPGD
YQYVVTYRDEHGKEHQKQYTISVNDKKPMITQGRFDTINGVDHFTPDKTKALDSSGIVREEVFYLAKKNG
RKFDVTEGKDGITVSDNKVYIPKNPDGSYTISKRDGVTLSDYYYLVEDRAGNVSFATLRDLKAVGKDKAV
VNFGLDLPVPEDKQIVNFTYLVRDADGKPIENLEYYNNSGNSLILPYGKYTVELLTYDTNAAKLESDKIV
SFTLSADNNFQQVTFKITMLATSQITAHFDHLLPEGSRVSLKTAQDQLIPLEQSLYVPKAYGKTVQEGTY
EVVVSLPKGYRIEGNTKVNTLPNEVHELSLRLVKVGDASDSTGDHKVMSKNNSQALTASATPTKSTTSAT
AKALPSTGEKMGLKLRIVGLVLLGLTCVFSRKKSTKD

>AAK33494.1
MRQIQSIRLIDVLELAFGVGYKEETTSQFSSDQPSQVVLYRGEANTVRFAYTNQMSLMKDIRIALDGSDK
SLTAQIVPGMGHVYEGFQTSARGIFTMSGVPESTVPVANPNVQTKYIRYFKVIDDMHNTMYKGTVFLVQP
QAWKYTMKSVDQLPVDDLNHIGVAGIERMTTLIKNAGALLTTGGSGAFPDNIKVSINPKGRQATITYGDG
STDIIPPAVLWKKGSVKEPTEADQSVGTPTPGIPGKFKRDQSLNEHEAMVNVEPLSHVVKDNIKVIDEKS
TGRFEPFRPNEDEKEKPASDVKVRPAEVGSWLEPATALPSVEMSAEDRLKS

>AAK33792.1
MKKYFILKSSVLSILTSFTLLVTDVQADQVDVQFLGVNDFHGALDNTGTAYTPSGKIPNAGTAAQLGAYM
DDAEIDFKQANQDGTSIRVQAGDMVGASPANSALLQDEPTVKVFNKMKFEYGTLGNHEFDEGLDEFNRIM
TGQAPDPESTINDITKQYEHEASHQTIVIANVIDKKTKDIPYGWKPYAIKDIAINDKIVKIGFIGVVTTE
IPNLVLKQNYEHYQFLDVAETIAKYAKELQEQHVHAIVVLAHVPATSKDGVVDHEMATVMEKVNQIYPEH
SIDIIFAGHNHQYTNGTIGKTRIVQALSQGKAYADVRGTLDTDTNDFIKTPSANVVAVAPGIKTENSDIK
AIINHANDIVKTVTERKIGTATNSSTISKTENIDKESPVGNLATTAQLTIAKKTFPTVDFAMTNNGGIRS
DLVVKNDRTITWGAAQAVQPFGNILQVIQMTGQHIYDVLNQQYDENQTYFLQMSGLTYTYTDNDPKNSDT
PFKIVKVYKDNGEEINLTTTYTVVVNDFLYGGGDGFSAFKKAKLIGAINTDTEAFITYITNLEASGKTVN
ATIKGVKNYVTSNLESSTKVNSAGKHSIISKVFRNRDGNTVSSEVISDLLTSTENTNNSLGKKETTTNKN
TISSSTLPITGDNYKMSPIMTILALISLGGLNAFIKKRKS

>AAK33814.1
MTNNQTLDILLDVYAYNHAFRIAKALPNIPKTALYLLEMLKERRELNLAFLAEHAAENRTIEDQYHCSLW
LNQSLEDEQIANYILDLEVKVKNGAIIDFVRSVSPILYRLFLRLITSEIPNFKAYIFDTKNDQYDTWHFQ
AMLESDHEVFKAYLSQKQSRNVTTKSLADMLTLTSLPQEIKDLVFLLRHFEKAVRNPLAHLIKPFDEEEL
HRTTHFSSQAFLENIITLATFSGVIYRREPFYFDDMNAIIKKELSLWRQSIV

>AAK34188.1
MKTKKVIILVGLLLSSQLTLIACQSRGNGTYPIKTKQSRKGMTSNKIKPIKKSKKTNKTHKGVAGVDFPT
DDGFILTKDSKILSKTDQGIVVDHDGHSHFIFYADLKGSPFEYLIPKGASLAKPAVAQRAASQGTSKVAD
PHHHYEFNPADIVAEDALGYTVRHDDHFHYILKSSLSGQTQAQAKQVATRLPQTSSLVSTATANGIPGLH
FPTSDGFQFNGQGIVGVTKDSILVDHDGHLHPISFADLRQGGWAHVADQYDPAKKAEKPAETHQTPELSE
REKEYQEKLAYLAEKLGIDPSTIKRVETQDGKLGLEYPHHDHAHVLMLSDIEIGKDIPDPHAIEHARELE
KHKVGMDTLRALGFDEEVILDIVRTHDAPTPFPSNEKDPNMMKEWLATVIKLDLGSRKDPLQRKGLSLLP
NLETLGIGFTPIKDISPVLQFKKLKQLLMTKTGVTDYRFLDNMPQLEGIDISQNNLKDISFLSKYKNLTL
VAAADNGIEDIRPLGQLPNLKFLVLSNNKISDLSPLASLHQLQELHIDNNQITDLSPVSHKESLTVVDLS
RNADVDLATLQAPKLETLMVNDTKVSHLDFLKNNPNLSSLSINRAQLQSLEGIEASSVIVRVEAEGNQIK
SLVLKDKQGSLTFLDVTGNQLTSLEGVNNFTALDILSVSKNQLTNVNLSKPNKTVTNIDISHNNISLADL
KLNEQHIPEAIAKNFPAVYEGSMVGNGTAEEKAAMATKAKESAQEASESHDYNHNHTYEDEEGHAHEHRD
KDDHDHEHEDENEAKDEQNHAD

>AAK34428.1
MKSFSLTFSFLNLLKYGTIKVMTKEFHHVTVLLHETVDMLDIKPDGIYVDATLGGSGHSAYLLSKLGEEG
HLYCFDQDQKAIDNAQVTLKSYIDKGQVTFIKDNFRHLKARLTALGVDEIDGILYDLGVSSPQLDERERG
FSYKQDAPLDMRMDRQSLLTAYEVVNTYPFNDLVKIFFKYGEDKFSKQIARKIEQARAIKPIETTTELAE
LIKAAKPAKELKKKGHPAKQIFQAIRIEVNDELGAADESIQDAMELLALDGRISVITFHSLEDRLTKQLF
KEASTVDVPKGLPLIPEDMKPKFELVSRKPILPSHSELTANKRAHSAKLRVAKKIRK

>AAK34472.1
MTTTEQELTLTPLRGKSGKAYKGTYPNGECVFIKLNTTPILPALAKEQIAPQLLWAKRMGNGDMMSAQEW
LNGRTLTKEDMNSKQIIHILLRLHKSKKLVNQLLQLNYKIENPYDLLVDFEQNAPLQIQQNSYLQAIVKE
LKRSLPEFKSEVATIVHGDIKHSNWVITTSGMIFLVDWDSVRLTDRMYDVAYLLSHYIPRSRWSEWLSYY
GYKNNDKVMQKIIWYGQFSHLTQILKCFDKRDMEHVNQEIYALRKFREIFRKK

>AAK34691.1
MRKKQKLPFDKLAIALMSTSILLNAQSDIKANTVTEDTPATEQAVETPQPTAVSEEAPSSKETKTPQTPD
DAEETIADDANDLAPQAPAKTADTPATSKATIRDLNDPSQVKTLQEKAGKGAGTVVAVIDAGFDKNHEAW
RLTDKTKARYQSKEDLEKAKKEHGITYGEWVNDKVAYYHDYSKDGKTAVDQEHGTHVSGILSGNAPSETK
EPYRLEGAMPEAQLLLMRVEIVNGLADYARNYAQAIIDAVNLGAKVINMSFGNAALAYANLPDETKKAFD
YAKSKGVSIVTSAGNDSSFGGKTRLPLADHPDYGVVGTPAAADSTLTVASYSPDKQLTETATVKTADQQD
KEMPVLSTNRFEPNKAYDYAYANRGMKEDDFKDVKGKIALIERGDIDFKDKIANAKKAGAVGVLIYDNQD
KGFPIELPNVDQMPAAFISRKDGLLLKENPQKTITFNATPKVLPTASGTKLSRFSSWGLTADGNIKPDIA
APGQDILSSVANNKYAKLSGTSMSAPLVAGIMGLLQKQYETQYPDMTPSERLDLAKKVLMSSATALYDED
EKAYFSPRQQGAGAVDAKKASAATMYVTDKDNTSSKVHLNNVSDKFEVTVTVHNKSDKPQELYYQATVQT
DKVDGKLFALAPKALYETSWQKITIPANSSKQVTIPIDVSQFSKDLLAPMKNGYFLEGFVRFKQDPTKEE
LMSIPYIGFRGDFGNLSALEKPIYDSKDGSSYYHEANSDAKDQLDGDGLQFYALKNNFTALTTESNPWTI
IKAVKEGVENIEDIESSEITETIFAGTFAKQDDDSHYYIHRHANGKPYAAISPNGDGNRDYVQFQGTFLR
NAKNLVAEVLDKEGNVVWTSEVTEQVVKNYNNDLASTLGSTRFEKTRWDGKDKDGKVVANGTYTYRVRYT
PISSGAKEQHTDFDVIVDNTTPEVATSATFSTEDRRLTLASKPKTSQPVYRERIAYTYMDEDLPTTEYIS
PNEDGTFTLPEEAETMEGATVPLKMSDFTYVVEDMAGNITYTPVTKLLEGHSNKPEQDGSDQAPDKKPET
KPEQDGSGQAPDKKPETKPEQDGSGQTPDKKPETKPEKDSSGQTPGKTPQKGQ
PSRTLEKRSSKRALATKASTKDQLPTTNDKDTNRLHLLKLVMTTFFLGLVAHIFKTKRTED

>AAK34694.1
MAKNNTNRHYSLRKLKTGTASVAVALTVLGAGFANQTEVKANGDGNPREVIEDLAANNPAIQNIRLRYEN

KDLKARLENAMEVAGRDFKRAEELEKAKQALEDQRKDLETKLKELQQDYDLAKESTSWDRQRLEKELEEK
KEALELAIDQASRDYHRATALEKELEEKKKALELAIDQASQDYNRANVLEKELETITREQEINRNLLGNA
KLELDQLSSEKEQLTIEKAKLEEEKQISDASRQSLRRDLDASREAKKQVEKDLANLTAELDKVKEDKQIS
DASRQGLRRDLDASREAKKQVEKDLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEA
NSKLAALEKLNKELEESKKLTEKEKAELQAKLEAEAKALKEQLAKQAEELAKLRAGKASDSQTPDTKPGN
KAVPGKGQAPQAGTKPNQNKAPMKETKRQLPSTGETANPFFTAAALTVMATAGVAAVVKRKEEN

>NP_250399.1
MNPITLERAGLPYGVADAGDIPALGRPVARDVESLRVERLAAPAAASASGTGVALTPPSAASQQRLEVAN
RAEIASLVQAVGEDVGLARQVVLAGASTLLSAGLMSPQAFEIELAKITGEVENQQKKLKLTEIEQARKQN
LQKMEDNQQKIRESEEAAKEAQKSGLAAKIFGWISAIASIIVGAIMVATGVGAAAGALMIAGGVMGVVSQ
SVQQAAADGLISKEVMEKLGPALMGIEMAVALLAAVVSFGGSAVGGLARLGAKIGGKAAEMTASLASKVA
DLGGKFGSLAGQSLSHSLKLGVQVSDLTLDVANGAAQATHSGFQAKAANRQADVQESRADLTTLQGVIER
LKEELSRMLEAFQEIMERIFAMLQAKGETLHNLSSRPAAI

>NP_219647.1
MTTPDNNTIDVSFPTFVRLNVATTDLADGNKSNAVTITETATANYVNVTQDLTSSTAKLECTQDLIAQGK
LIVTNPKSDISFGGRVNLADNTVNYSNGGAEVSFTNINSRQGKQYVPYGLYKNGEPKISMRSALSGGHVG
SGDTGGWGAEVLWDAYTEQLKDMTDGAVTLNSSNRGKLSFTASPEAPVLFRLSVFMRKNGDWLDNGVGGR
VMLYVNTTDSAGKTVRRLLGIAVCLGSTWYTTVPMFWCAATYYATSSGFFQLIVGERNFRVSSLSWSVVR
LPVVP

>NP_220344.1
MMKRLLCVLLSTSVFSSPMLGYSASKKDSKADICLAVSSGDQEVSQEDLLKEVSRGFSRVAAKATPGVVY
IENFPKTGNQAIASPGNKRGFQENPFDYFNDEFFNRFFGLPSHREQQRPQQRDAVRGTGFIVSEDGYVVT
NHHVVEDAGKIHVTLHDGQKYTAKIVGLDPKTDLAVIKIQAEKLPFLTFGNSDQLQIGDWAIAIGNPFGL
QATVTVGVISAKGRNQLHIVDFEDFIQTDAAINPGNSGGPLLNINGQVIGVNTAIVSGSGGYIGIGFAIP
SLMAKRVIDQLISDGQVTRGFLGVTLQPIDSELATCYKLEKVYGALVTDVVKGSPAEKAGLRQEDVIVAY
NGKEVESLSALRNAISLMMPGTRVVLKIVREGKTIEIPVTVTQIPTEDGVSALQKMGVRVQNITPEICKK
LGLAADTRGILVVAVEAGSPAASAGVAPGQLILAVNRQRVASVEELNQVLKNSKGENVLLMVSQGDVVRF
IVLKSDE

>NP_214802.1
MSQIMYNYPAMLGHAGDMAGYAGTLQSLGAEIAVEQAALQSAWQGDTGITYQAWQAQWNQAMEDLVRAYH
AMSSTHEANTMAMMARDTAEAAKWGG

>NP_268530.1
MDIIIGTSLLILVLAIFSLFNYKAPHGAKAMGALASAACASFLVEAFQDSFFGKVLGFQFLSEVGGANGS
LSGVAAAILVAIAIGVSPGYAVLIGLSVSGTGIIPGFVAGYVVSFLIKWMEKNIPGGLDLISIIIVGAPL
TRFLAQLITPVINSTLLTIGDILTSSANSNPIIMGMILGGTIVVVATAPLSSMALTAMLGLTGIPMAIGA
LSVFGSSFMNGVLFYRLKLGERKDNIAFAIEPLTQADVTSANPIPIYVTNFVGGAACGVLIALMKLVNDT
PGTATPIAGFAVMFAYNPVAKVLITALGCIIISLIVGYIGGSVFKNYRLVTKQELQARN

>NP_268623.1
MDLEQTKPNQVKQKIALTSTIALLSASVGVSHQVKADDRASGETKASNTHDDSLPKPETIQEAKATIDAV
EKTLSQQKAELTELATALTKTTAEINHLKEQQDNEQKALTSAQEIYTNTLASSEETLLAQGAEHQRELTA
TETELHNAQADQHSKETALSEQKASISAETTRAQDLVEQVKTSEQNIAKLNAMISNPDAITKAAQTANDN
TKALSSELEKAKADLENQKAKVKKQLTEELAAQKAALAEKEAELSRLKSSAPSTQDSIVGNNTMKAPQGY
PLEELKKLEASGYIGSASYNNYYKEHADQIIAKASPGNQLNQYQDIPADRNRFVDPDNLTPEVQNELAQF
AAHMINSVRRQLGLPPVTVTAGSQEFARLLSTSYKKTHGNTRPSFVYGQPGVSGHYGVGPHDKTIIEDSA
GASGLIRNDDNMYENIGAFNDVHTVNGIKRGIYDSIKYMLFTDHLHGNTYGHAINFLRVDKHNPNAPVYL
GFSTSNVGSLNEHFVMFPESNIANHQRFNKTPIKAVGSTKDYAQRVGTVSDTIAAIKGKVSSLENRLSAI
HQEADIMAAQAKVSQLQGKLASTLKQSDSLNLQVRQLNDTKGSLRTELLAAKAKQAQLEATRDQSLAKLA
SLKAALHQTEALAEQAAARVTALVAKKAHLQYLRDFKLNPNRLQVIRERIDNTKQDLAKTTSSLLNAQEA
LAALQAKQSSLEATIATTEHQLTLLKTLANEKEYRHLDEDIATVPDLQVAPPLTGVKPLSYSKIDTTPLV
QEMVKETKQLLEASARLAAENTSLVAEALVGQTSEMVASNAIVSKITSSITQPSSKTSYGSGSSTTSNLI
SDVDESTQRALKAGVVMLAAVGLTGFRFRKESK

>NP_268639.1
MIKRLISLVVIALFFAASTVSGEEYSVTAKHAIAVDLESGKVLYEKDAKEVVPVASVSKLLTTYLVYKEV
SKGKLNWDSPVTISNYPYELTTNYTISNVPLDKRKYTVKELLSALVVNNANSPAIALAEKIGGTEPKFVD
KMKKQLRQWGISDAKVVNSTGLTNHFLGANTYPNTEPDDENCFCATDLAIIARHLLLEFPEVLKLSSKSS
TIFAGQTIYSYNYMLKGMPCYREGVDGLFVGYSKKAGASFVATSVENQMRVITVVLNADQSHEDDLAIFK
TTNQLLQYLLINFQKVQLIENNKPVKTLYVLDSPEKTVKLVAQNSLFFIKPIHTKTKNTVHITKKSSTMI
APLSKGQVLGRATLQDKHLIGQGYLDTPPSINLILQKNISKSFFLKVWWNRFVRYVNTSL

>NP_273588.1
MALFLSIFPIVLLIWLMVKKNSMPSYVALPITAVLIYAIKLFYFDDAGMLLNATAASGLVKTLTPITVIF
GAIMFNRMMETTGCIDVIRKWLATISPNPVAQLMIIGWAFAFMIEGASGFGTPAAIAAPILMSLGFNPLK
VAIFTLVMNSVPVSFGAVGTPTWFGFAPLNLSAEDILAIGRQTGVMHFFAGFVIPVIGLGFIVPWSEIRK
NLGFVAIAVFSCTIPYVALAMVNEEFPSLVAGAIGLMVSVFAANQGWGLSKDHAKDPNAEKVPFAQVAKA
LAPLGMLIGMLVVTRIKQLGIKGILTSKEEWFSFQLPFDLSKITVSDSLTITFGNIFGQDVSASYQTLYV
PAWIPFVLTVWICILLYKTKFKDAWTIYSVTFNQTKKPLLALMGALIMVQLMLVGGDNSMVKIIGKEFAA
MAGEHWVYFSPYLGAIGAFFSGSNTVSNLTFGPIQQQIALDTGLSVTLILALQSVGGAMGNMVCLNNIIA
VCTVLDVKNSEGAIIKKTVIPMAIYGVIAVVAAMIFFL

>NP_346492.1
MSKDIRVRYAPSPTGLLHIGNARTALFNYLYARHHGGTFLIRIEDTDRKRHVEDGERSQLENLRWLGMDW
DESPESHENYRQSERLDLYQKYIDQLLAEGKAYKSYVTEEELAAERERQEVAGETPRYINEYLGMSEEEK
AAYIAEREAAGIIPTVRLAVNESGIYKWHDMVKGDIEFEGGNIGGDWVIQKKDGYPTYNFAVVIDDHDMQ
ISHVIRGDDHIANTPKQLMVYEALGWEAPEFGHMTLIINSETGKKLSKRDTNTLQFIEDYRKKGYLPEAV
FNFIALLGWNPGGEDEIFSREEFIKLFDENRLSKSPAAFDQKKLDWMSNDYIKNADLETIFEMAKPFLEE
AGRLTDKAEKLVELYKPQMKSVDEIIPLTDLFFSDFPELTEAEREVMTGETVPTVLEAFKAKLEAMTDDE
FVTENIFPQIKAVQKETGIKGKNLFMPIRIAVSGEMHGPELPDTIFLLGREKSIQHIENMLKEISK

>NP_395162.1
MTINIKTDSPIITTGSQLDAITTETVKQSGEIKKTEDTRHEAQAIKSSEASLSRSQVPELIKPSQGINVA
LLSKSQGDLNGTLSILLLLLELARKAREMGLQQRDIENKATITAQKEQVAEMVSGAKLMIAMAVVSGIMA
ATSTVASAFSIAKEVKIVKQEQILNSNIAGREQLIDTKMQQMSNIGDKAVSREDIGRIWKPEQVADQNKL
ALLDKEFRMTDSKANAFNAATQPLGQMANSAIQVHQGYSQAEVKEKEVNASIAANEKQKAEEAMNYNDNF
MKDVLRLIEQYVSSHTHAMKAAFGVV

>CAC98661.1
MKKATIAATAGIAVTAFAAPTIASASTVVVEAGDTLWGIAQSKGTTVDAIKKANNLTTDKIVPGQKLQVN
NEVAAAEKTEKSVSATWLNVRSGAGVDNSIITSIKGGTKVTVETTESNGWHKITYNDGKTGFVNGKYLTD
KAVSTPVAPTQEVKKETTTQQAAPAAETKTEVKQTTQATTPAPKVAETKETPVVDQNATTHAVKSGDTIW
ALSVKYGVSVQDIMSWNNLSSSSIYVGQKLAIKQTANTATPKAEVKTEAPAAEKQAAPVVKENTNTNTAT
TEKKETATQQQTAPKAPTEAAKPAPAPSTNTNANKTNTNTNTNTNTNNTNTNTPSKNTNTNSNTNTNTNS
NTNANQGSSNNNSNSSASAIIAEAQKHLGKAYSWGGNGPTTFDCSGYTKYVFAKAGISLPRTSGAQYAST
TRISESQAKPGDLVFFDYGSGISHVGIYVGNGQMINAQDNGVKYDNIHGSGWGKYLVGFGRV

>AAL18599.1
MKKLSILAISVALFASITACGAFGGLPSLKSSFVLSEDTIPGTNETVKTLLPYGSVINYYGYVKPGQAPD
GLVDGNKKAYYLYVWIPAVIAEMGVRMISPTGEIGEPGDGDLVSDAFKAATPEEKSMPHWFDTWIRVERM
SAIMPDQIAKAAKAKPVQKLDDDDDGDDTYKEERHNKYNSLTRIKIPNPPKSFDDLKNIDTKKLLVRGLY
RISFTTYKPGEVKGSFVASVGLLFPPGIPGVSPLIHSNPEELQKQAIAAEESLKKAASDATK

>NP_462896.1

MKSLNTLVILTSVISTSVFAGAYVENREAYNLASDQMEFMLRVGYNSDMGAGIMLTNTYTLQRDDELKHG
YNEIEGWYPLFKPTDKLTIQPGGLINDKSIGSGGAVYLDVNYKFTPWFNLTVRNRYNHNNYSSTDLNGEL
DNNDSYEIGNYWNFIITDKFSYTFEPHYFYNVNDFNSSNGTKHHWEITNTFRYRINEHWLPYFELRWLDR
NVGPYHREQNQIRIGAKYFF

>AAM22954.1
MTAQTPIHVYSEIGKLKKVLLHRPGKEIENLMPDYLERLLFDDIPFLEDAQKEHDAFAQALRDEGIEVLY
LETLAAESLVTPEIREAFIDEYLSEANIRGRATKKAIRELLMAIEDNQELIEKTMAGVQKSELPEIPASE
KGLTDLVESSYPFAIDPMPNLYFTRDPFATIGTGVSLNHMFSETRNRETLYGKYIFTHHPIYGGGKVPMV
YDRNETTRIEGGDELVLSKDVLAVGISQRTDAASIEKLLVNIFKQNLGFKKVLAFEFANNRKFMHLDTVF
TMVDYDKFTIHPEIEGDLRVYSVTYDNEELHIVEEKGDLADLLAANLGVEKVDLIRCGGDNLVAAGREQW
NDGSNTLTIAPGVVVVYNRNTITNAILESKGLKLIKIHGSELVRGRGGPRCMSMPFEREDI

>NP_665438.1
MSTSFENKATNRGVITFTISQDKIKPALDKAFNKIKKDLNAPGFRKGHMPRPVFNQKFGEEVLYEDALNI
VLPEAYEAAVTELGLDVVAQPKIDVVSMEKGKEWTLSAEVVTKPEVKLGDYKNLVVEVDASKEVSDEDVD
AKIERERQNLAELIIKDGEAAQGDTVVIDFVGSVDGVEFDGGKGDNFSLELGSGQFIPGFEDQLVGAKAG
DEVEVNVTFPESYQAEDLAGKAAKFMTTIHEVKTKEVPELDDELAKDIDEDVDTLEDLKVKYRKELEAAQ
ETAYDDAVEGAAIELAVANAEIVDLPEEMIHEEVNRSVNEFMGNMQRQGISPEMYFQLTGTTQEDLHNQY
SAEADKRVKTNLVIEAIAKAEGFEATDSEIEQEINDLATEYNMPADQVRSLLSADMLKHDIAMKKAVEVI
TSTASVK

>AAN47806.1
MNLSKHTAVLTGIVFFLAFSLSMYVSVIEKAGTKDEYPYTMKIYYPRLEGIHPGAPVRILGVEKGIVRSL
DVVPIDEVEDQRFLNKDQTKAIEIIVRLKEPITLWDNYKITFQTNTILSGRTIDIDPGSFDKEETSFFQP
TYLEEEQKSPDFLPSADYFEDFFAASTGVIRENREDIRTSFNNFYEISEKLKSNRGTIPQIINSPETYDN
VIELLTDARIFGNDARRYLEGNRKLERSAPIPLTINMYRRTTLIGNVSNRYYFGKL

>AAN48316.1
MKHPKMNLLGALLLSSLFLNPLNSDPTLGNCEVFPTNNIWNTPVDTLPLHPFSESYVRSIGAQKKLKADF
GSGLWEGMPIGIPFILTSGANPVPVSFEYTDESEPGPYPIPHNAPIEGGETSDGDRHVLVVEQKTCKLYE
LYSARKKGKSWTAVSGAVFDLKSNQLRPANWTSADAAGLPILPGLVRYEEIASGEIKHAIRFTAKKTQKA
YLWPARHYASKITDKNVPPMGTRFRLKASFNIDGFSKENQVILRALKKYGMILADNGSDWFLSGAPNEKW
NNDQLHKLGKVLGDQFEAVDSESLMISTDSGEAKQN

>AAN48653.1
MGKNNSFRPKNSFLKFQVVFGFFLLSFISLNSEERSSILDPTFFLIHKNLNPSLGGLFETRKKFTSYGAW
FELPVWKRERKHFWKTHFLYENHKTFSVVHQTYIGLEYSFISEKKDLHPSVFVGWERGEKDLGIFGIHLT
IPDKQTIQVFGKTGKDFKSGSIFLHSNFDTGLQLFLGFSRTWDKSYTEDQFTLGIRVSWEKIYSSFFWNQ
TQEEESFLTGKFGISNFQQDLKSKTLFESTDEFQSNSLFKSKNLSSKNTKSKSFETSLLEKDKTNYSYIS
SFIRSYTNISISVQELLSAGFTLSSALEISKASYNSKEEFLKLFHSLSVKEQTKIFVLLKKKNPKHILKN
SILPKDKR

>NP_754374.1
MLYNIPCRIYILSTLSLCISGIVSTATATSSETKISNEETLVVTTNRSASNLWESPATIQVIDQQTLQNS
TNASIADNLQDIPGVEITDNSLAGRKQIRIRGEASSRVLILIDGQEVTYQRAGDNYGVGLLIDESALERV
EVVKGPYSVLYGSQAIGGIVNFITKKGGDKLASGVVKAVYNSATAGWEESIAVQGSIGGFDYRINGSYSD
QGNRDTPDGRLPNTNYRNNSQGVWLGYNSGNHRFGLSLDRYRLATQTYYEDPDGSYEAFSVKIPKLEREK
VGVFYDTDVDGDYLKKIHFDAYEQTIQRQFANEVKTTQPVPSPMIQALTVHNKTDTHDKQYTQAVTLQSH
FSLPANNELVTGAQYKQDRVSQRSGGMTSSKSLTGFINKETRTRSYYESEQSTVSLFAQNDWRFADHWTW
TMGVRQYWLSSKLTRGDGVSYTAGIISDTSLARESASDHEMVTSTSLRYSGFDNLELRAAFAQGYVFPTL
SQLFMQTSAGGSVTYGNPDLKAEHSNNFELGARYNGNQWLIDSAVYYSEAKDYIASLICDGSIVCNGNTN
SSRSSYYYYDNIDRAKTWGLEISAEYNGWVFSPYISGNLIRRQYETSTLKTTNTGEPAINGRIGLKHTLV
MGQANIISDVFIRAASSAKDDSNGTETNVPGWATLNFAVNTEFGNEDQYRINLALNNLTDKRYRTAHETI
PAAGFNAAIGFVWNF

>NP_755498.1
MAQRQPEKTAAGGCCFNSLYNKYSGITMMRKKYMPRALGPLLLVVLSPAVAQQNDDNEIIVSASRSNRTV
AEMAQTTWVIENAELEQQIQGGKELKDALAQLIPGLDVSSQSRTNYGMNMRGRPLVVLIDGVRLNSSRSD
SRQLDSVDPFNIDHIEVISGATALYGGGSTGGLINIVTKKGQPETMMEFEAGTKSGFNSSKDHDERIAGA
VSGGNDHISGRLSVAYQKFGGWFDGNGDATLLDNTQTGLQHSNRLDIMGTGTLNIDESRQLQLITQYYKS
QGDDNYGLNLGKGFSAISGSSTPYVSKGLNSDRIPGTERHLISLQYSDSDFLRQELVGQVYYRDESLRFY
PFPTVNANKQATAFSSSQQDTDQYGMKLTLNSQLMDGWQITWGLDAEHERFTSNQMFFDLAQASASGGLN
NHKIYTTGRYPSYDITNLAAFLQSSYDINDIFTVSGGVRYQYTENRVDDFIDYTQQQKIAAGKAISADAI
PGGSVDYDNFLFNAGLLMHITERQQAWFNFSQGVALPDPGKYYGRGIYGAAVNGHLPLTKSVNVSDSKLE
GVKVDSYELGWRFTGDNLRTQIAAYYSLSNKSVERNKDLTISVKDDRRRIYGVEGAVDYLIPDTDWSTGV
NFNVLKTESKVNGQWQKYDVKESSPSKATAYINWAPEPWSLRVQSTTSFDVSDAEGNDINGYTTVDFISS
WQLPVGTLSFSVENLFDRDYTTVWGQRAPLYYSPGYGPASLYDYKGRGRTFGLNYSVLF

>NP_757022.1
MKNKYIIAPGIAVMCSAVISSGYASSDKKEDTLVVTASGFTQQLRNAPASVSVITSEQLQKKPVSDLVDA
VKDVEGISITGGNEKPDISIRGLSGDYTLILVDGRRQSGRESRPNGSGGFEAGFIPPVEAIERIEVIRGP
MSSLYGSDAIGGVINIITKPVNNQTWDGVLGLGGIIQEHGKFGNSTTNDFYLSGPLIKDKLGLQLYGGMN
YRKEDSISQGTPAKDNKNITATLQFTPTESQKFVFEYGKNNQVHTLTPGESLDAWTMRGNLKQPNSKRET
HNSRSHWVAAWNAQGEILHPEIAVYQEKVIREVKSGKKDKYNHWDLNYESRKPEITNTIIDAKVTAFLPE
NVLTIGGQFQHAELRDDSATGKKTTETQSVSIKQKAVFIENEYAATDSLALTGGLRLDNHEIYGSYWNPR
LYAVYNLTDNLTLKGGIAKAFRAPSIREVSPGFGTLTQGGASIMYGNRDLKPETSVTEEIGIIYSNDSGF
SASATLFNTDFKNKLTSYDIGTKDPVTGLNTFIYDNVGEANIRGVELATQIPVYDKWHVSANYTFTDSRR
KSDDESLNGKSLKGEPLERTPRHAANAKLEWDYTQDITFYSSLNYTGKQIWAAQRNGAKVPRVRNGFTSM
DIGLNYQILPDTLINFAVLNVTDRKSEDIDTIDGNWQVDEGRRYWANVRVSF

>YP_026537.1
MAKDIKFSEEARRSMLRGVDTLANAVKVTLGPKGRNVVLEKKFGSPLITNDGVTIAKEIELEDAFENMGA
KLVAEVASKTNDVAGDGTTTATVLAQAMIREGLKNVTAGANPMGLRKGIEKAVVAAVEELKTISKPIEGK
SSIAQVAAISAADEEVGQLIAEAMERVGNDGVITLEESKGFTTELDVVEGMQFDRGYASPYMITDSDKME
AVLDNPYILITDKKISNIQEILPVLEQVVQQGKPLLIIAEDVEGEALATLVVNKLRGTFNVVAVKAPGFG
DRRKAMLEDIAILTGGEVITEELGRDLKSATVESLGRAGKVVVTKENTTVVEGVGSTEQIEARIGQIRAQ
LEETTSEFDREKLQERLAKLVGGVAVIKVGAATETELKERKLRIEDALNSTRAAVEEGIVAGGGTSLMNV
YTKVASIVAEGDEATGINIVLRALEEPVRQIAINAGLEGSVVVERLKGEKVGVGFNAATGEWVNMLETGI
VDPAKVTRSALQNAASVAAMFLTTEAVVADKPEPNAPAMPDMGGMGMGGMGGMM

>YP_030461.1
MSKIIGIDLGTTNSCVAVMEGGEPKVIPNPEGNRTTPSVVAFKNEERQVGEVAKRQAITNPNTIMSVKRH
MGTDYKVEVEGKDYTPQEISAIILQNLKASAEAYLGETVTKAVITVPAYFNDAERQATKDAGRIAGLEVE
RIINEPTAAALAYGLEKQDEEQKILVYDLGGGTFDVSILELADGTFEVISTAGDNRLGGDDFDQVIIDHL
VAEFKKENNIDLSQDKMALQRLKDAAEKAKKDLSGVTQTQISLPFISAGAAGPLHLELTLTRAKFEELSA
GLVERTLEPTRRALKDAGFAPSELDKVILVGGSTRIPAVQEAIKRETGKEPYKGVNPDEVVALGAAVQGG
VLTGDVEGVLLLDVTPLSLGIETMGGVFTKLIERNTTIPTSKSQVFSTAADNQPAVDIHVLQGERPMSAD
NKTLGRFQLTDLPPAPRGIPQIEVTFDIDANGIVNVRAKDLGTSKEQAITIQSSSGLSDEEVERMVQEAE
ANADADQKRKEEVELRNEADQLVFQTDKVVKDLEGKVDAAEVAKATEAKEALQAAIEKNELEEIRAKKDA
LQEIVQQLTVKLYEQAQAAAGQAEGAEGAQDAGAKKDNVVDAEFEEVKEDK

>YP_177697.1
MSYVIAAPEMLATTAADVDGIGSAIRAASASAAGPTTGLLAAAADEVSSAAAALFSEYARECQEVLKQAA
AFHGEFTRALAAAGAAYAQAEASNTAAMSGTAGSSGALGSVGMLSGNPLTALMMGGTGEPILSDRVLAII
DSAYIRPIFGPNNPVAQYTPEQWWPFIGNLSLDQSIAQGVTLLNNGINAELQNGHDVVVFGYSQSAAVAT
NEIRALMALPPGQAPDPSRLAFTLIGNINNPNGGVLERYVGLYLPFLDMSFNGATPPDSPYQTYMYTGQY
DGYAHNPQYPLNILSDLNAFMGIRWVHNAYPFTAAEVANAVPLPTSPGYTGNTHYYMFLTQDLPLLQPIR

AIPFVGTPIAELIQPDLRVLVDLGYGYGYADVPTPASLFAPINPIAVASALATGTVQGPQAALVSIGLLP
QSALPNTYPYLPSANPGLMFNFGQSSVTELSVLSGALGSVARLIPPIA

>YP_328507.1
MKKLLKSVLVFAALSSASSLQALPVGNPAEPSLMIDGILWEGFGGDPCDPCTTWCDAISMRMGYYGDFVF
DRVLKTDVNKEFQMGAAPTTSDVAGLEKDPVANVARPNPAYGKHMQDAEMFTNAAYMALNIWDRFDVFCT
LGATTGYLKGNSASFNLVGLFGTKTQSSGFDTANIVPNTALNQAVVELYTDTTFAWSVGARAALWECGCA
TLGASFQYAQSKPKVEELNVLCNASEFTINKPKGYVGAEFPLDITAGTEAATGTKDASIDYHEWQASLAL
SYRLNMFTPYIGVKWSRVSFDADTIRIAQPKLAKPVLDTTTLNPTIAGKGTVVSSAENELADTMQIVSLQ
LNKMKSRKSCGIAVGTTVVDADKYAVTIETRLIDERAAHVNAQFRF

>YP_816547.1
MSQKNNKKKNKRKNLLTNILAGFLILLSLALIFNTQIRNIFIVWNTNKYQVSQVSKEKLEENQDTEGNFD
FDSVKAISSEAVLTSQWDAQKLPVIGGIAIPELEMNLPIFKGLDNVNLFYGAGTMKREQVMGEGNYSLAS
HHIFGVDNANKMLFSPLDNAKNGMKIYLTDKNKVYTYEIREVKRVTPDRVDEVDDRDGVNEITLVTCEDL
AATERIIVKGDLKETKDYSQTSDEILTAFNQPYKQFY

>ABN54438.1
MLFKPHRFVPKTVAAAALAAHGLAAHATAPFVVQDIKIEGLQRVEAGSVFAYLPIKQGDTFTDDKASEAI
RALYATGFFNDVRIATQGGVVIVQVQERPAIASIDFTGIKEFDKDNLNKALKAVGLSQGRYYDKALVDKA
EQELKRQYLTRGFYAAEVSTTVTPVDANRVSILFAVAEGPSAKIRQINFIGNKAFKTSTLRDEMQLSTPN
WFSWYTKNDLYSKEKLTGDLENVRSYYLNRGYLEFNIESTQVSISPDKKDMYLTVALHEGEPYTVSSVKL
AGNLLDRQAELEKLVKIKPGDRFSAEKLQQTTKAIVDKLGQYGYAFATVNAQPEIDQATHKVGLTLVVDP
SRRVYVRRINIVGNTRTRDEVVRREMRQLESSWFDSSRLALSKDRVNRLGYFTDVDVTTVPVEGTNDQVD
VNVKVAEKPTGAITLGAGFSSTDKVVLSAGISQDNVFGSGTSLAVNVNTAKSYRTLTVTQVDPYFTVDGI
KRITDVFYRTYQPLYYSTNSSFRIITAGGNLKFGIPFSETDTVYFGAGFEQNRLDVDSNTPQSYQDYVNE
FGRVSNTVPLTIGWSRDARDSALIPSRGYFTQANAEYGVPVGKIQYYKMDVQGQYYYSFARGFILGLNFQ
AGYGNGIGNPYPIFKNYYAGGIGSVRGYEPSSLGPRDTKTNDPIGGSKMVVGNIELTFPLPGTGYDRTLR
VFTFLDGGNVWGNAPGGTSTGANGLRYGYGIGLAWISPIGPLKLSLGFPLQKHEGDQYQKFQFQIGTAF

>EDK22348.1
MKIDTLTKNFSNYQTQINKDNDLVSNVHSDNVVKIQISDEARSLSQTNSKNEKEYEIKVSQEKIENHKDQ
NSIQNSKGGVNPALEALIAKLAEILAKIAELTQKMTNANEQMKTTFQKQIDVLKSQADVIQAQIQELQSQ
QA

>YP_001197700.1
MKKKAVVGSVAALVLGMSLCSYQLGRFQAMEEQKNRVSYIEDSQSVQTTVAEQLTPDQVSAKENIDAEQI
VVKITDQGYVTSHGDHFHYYNGKVPFDAIFSEELVMKDPNYVLQDSHIINEVQDGYVIKVDGKYYLYLKD
PSKHKNVRSKEEVERQKGISSADSKNQAAGNSKDGRYRTDDGYVFNPTDVIEDTGDGFIVPHGDHFHFIP
KKDLSAAELKAAQDYWNQKGSVSSASGSQYGDRNNRAQQTTISAGQGQDLASLLAQLDATPLSQRHVEAD
GLVFDPRTITKKTAAGVIVPHGDHYHFIPYSQMSPLEEKISRMIGVNGAGVSSGAQASHSQHTLTQPNRP
VTPIGTVTTQPVSPTQPVLPTQPKQSTGKVVSYKGRQIPAYGKGLDGKAYFTSDGYTFSKDSITSVDDQG
LIASHGDHFHYVGFGELEDFEIKQVEEWVNEKAGKQVPPKTSEQVGNDAKPTTPSQGNDSKPVKPIQEEN
RPAFEYKQVTAKRKLAGKVVYEMEVGGKTYTYGRDELDLMKISFAELTLAEKDKQYIFDIAPLAEGDLKP
AMLVGMDQIPMKGANATYDTGQSFIIPHIDHIHVLPYTWLSKEQIATIRYIMQHPEIRPSAWTTSGHGDG
EATDLVPPILNATPKANRLGLKNWQIIHTAEEVMDARAKGKFATNDGYIFSAEDVLDPASFVFSQAFSLP
RATGGSLRSISKKDLSKEELEAVQTLLDKRDAEELAKNVTPIEKRAGLKNWQIVHSAEELAEAKAAGKYT
TKDGYIFDPADLLDPKVKIGTDNYRIPRVITDGYRRINKSDLYYLSELIPAEAMVAQREKSNSSLPSTPA
PTETGASAGETTRPEQPQVAKETAEEIYNRVEAKKVVPFEALTYNAGYATEVRNGTLVIPHQDHYHYVSF
KWFDQGSARSPEGYSLEDFLATVKYYMTNPQERPVSDDGWGVFTPKYTI

>YP_001332107.1
MKTRIVSSVTTTLLLGSILMNPVANAADSDINIKTGTTDIGSNTTVKTGDLVTYDKENGMHKKVFYSFID
DKNHNKKLLVIRTKGTIAGQYRVYSEEGANKSGLAWPSAFKVQLQLPDNEVAQISDYYPRNSIDTKEYMS
TLTYGFNGNVTGDDTGKIGGLIGANVSIGHTLKYVQPDFKTILESPTDKKVGWKVIFNNMVNQNWGPYDR
DSWNPVYGNQLFMKTRNGSMKAADNFLDPNKASSLLSSGFSPDFATVITMDRKASKQQTNIDVIYERVRD
DYQLHWTSTNWKGTNTKDKWIDRSSERYKIDWEKEEMTN

>YP_002037642.1
MLEITLKSPYQFAHILFQPTIVPHGGHYHFIPESDLSAGELAATYVFNPNDIVRDTGDAYIVRHGDHYHY
IPKSSLNNPPSHSNTEEVGSSSSSVLSNTSLHVHHEEEDGHGFDANRIISEDSEGFVIPHGDHNHYIKVQ
TKGYEAALKNKIPSLQSNYQPGTFDEKAVLAKVDQLLADSRSIYKDRLS

>YP_002037641.1
MKFSKKYIAAGSAVXVSLSLCAYALNQHRSQENKDNNRVSYVDGSQSSQKSENLTPDQVSQKEGIQAEQI
VIKITDQGYVTSHGDHYHYYNGKVPYDALFSEELLMKDPNYQLKDADIVNEVKGGYIIKVDGKYYVYLKD
AAHADNVRTKDEINRQKQEHVKDNEKVNSNVAVARSQGRYTTNDGYVFNPADIIEDTGNAYIVPHGGHYH
YIPKSDLSASELAAAKAHLAGKNMQPSQLSYSSTASDNNTQSVAKGSTSKPANKSENLQSLLKELYDSPS
AQRYSESDGLVFDPAKIISRTPNGVAIPHGDHYHFIPYSKLSALEEKIARMVPISGTGSTVSTNAKPNEV
VSSLGSLSSNPSSLTTSKELSSASDGYIFNPKDIVEETATAYIVRHGDHFHYIPKSNQIGQPTLPNNSLA
TPSPSLPINPGTSHEKHEEDGYGFDANRIIAEDESGFVMSHGDHNHYFFKKDLTEEQIKAAQKHLEEVKT
SHNGLDSLSSHEQDYPSNAKEMKDLDKKIEEKIAGIMKQYGVKRESIVVNKEKNAIIYPHGDHHHADPID
EHKPVGIGHSHSNYELFKPEEGVAKKEGNKVYTGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPE
LEKKLGINMLVKLITPDGKVLEKVSGKVFGEGVGNIANFELDQPYLPGQTFKYTIASKDYPEVSYDGTFT
VPTSLAYKMASQTIFYPFHAGDTYLRVNPQFAVPKGTDALVRVFDEFHGNAYLENNYKVGEIKLPIPKLN
QGTTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTDKPSILPQFKRNKAQENSKLDEKVEEPKTS
EKVEKEKLSETGNSTSNSTLEEVPTVDPVQEKVAKFAESYGMKLENVLFNMDGTIELYLPSGEVIKKNMA
DFTGEAPQGNGENKPSENGKVSTGTVENQPTENKPADSLPEAPNEKPVKPENSTDNGMLNPEGNVGSDPM
LDPALEEAPAVDPVQEKLEKFTASYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA

>EET06073.1
MLGINSNINSLVAQQNLNGSQGALSQAITRLSSGKRINSAADDAAGLAIATRMQTQINGLNQGVSNANDG
VSILQTASSGLTSLTNSLQRIRQLAVQASNGPLSASDASALQQEVAQQISEVNRIASQTNYNGKNILDGS
AGTLSFQVGANVGQTVSVDLTQSMSAAKIGGGMVQTGQTLGTIKVAIDSSGAAWSSGSTGQETTQINVVS
DGKGGFTFTDQNNQALSSTAVTAVFGSSTAGTGTAASPSFQTLALSTSATSALSATDQANATAMVAQINA
VNKPQTVSNLDISTQTGAYQAMVSIDNALATVNNLQATLGAAQNRFTAIATTQQAGSNNLAQAQSQIQSA
DFAQETANLSRAQVLQQAGISVLAQANSLPQQVLKLLQ

>ACU44469.1
MNNNEKKVKYFLRKTAYGLASMSAAFAVCSGIVHADTSSGISDSIPHKKQVNLGAVTLKNLISKYRGNDK
AIAILLSRVDDFNRASQDTLPQLINSTEAEINNTLPQGRIIKQSIPVVRLKVERLGSGAIKAESINNIKA
ESINKIQGKSTNTIKAESINKIKVESINTIKAESINKIQAKPINTIKAESINTIKAESIHKIKPQSIKST
SATHVKVSDQELAKQSRRSQDIIKSLGFLSSDQKDILVKSISSSKDSQLILKFVTQATQLNNAESTKAKH
MAQNDVASIKNISLEVLEEYKEKIQRASTKSQVDELVAEAKKVVNSNKETLVNQANGKKQEIAKLENLSN
DEMLRYNTAIDNVVKQYNEGKLNITDAMNALNSIKQAAQEVAQKNLQKQYAKKIERISLKGLALSKKAKE
IYEKHKSILPTPGYYADSVGTYLNRFRDKRTFGNRSVWTGQSGLDEAKKMLDEVKKLLKELQDLTRGTKE
DKKPDVKPEAKPEAKPNIQVPKQAPTEAAKPALSPEALTRLTTWYNQAKDLLKDDQVKDKYVDILAVQKA
VDQAYDHVEEGKFITTDQANQLANKLRDALQSLELKDKKVAKPEAKPEVKPEAKPDVKPDVKPEAKPEAK
PEAKPEAKPEAKPEAKPEVKPDVKPEAKPDVKPEVKPDVKPEAKPDVKPEAKPDVKPEVKPEAKPE
VKPDVKPEARPEAKPEVKPDVKPEAKPEAKPEVKPDVKPEAKPEAKPATKKSVNTSGNLAVKKAIENKKY
SKKLPSTGEAASPLLAIVSLIVMLSAGLITIVLKHKKN

>AEB91475.1
MKANNHFKDFAWKKCLLGASVAALLVGCSPHIIETNEVALKLNYHPASEKVQALDEKILLLRPAFQYSDN
IAKEYENKFKNQTALKVEQILQNQGYKVISVDSSDKDDLSFSQKKEGYLAVAMNGEIVLRPDPKRTIQKK
SEPGLLFSTGLDKMEGVLIPAGFIKVTILEPMSGESLDSFTMDLSELDIQEKFLKTTHSSHSGGLVSTMV
KGTDNSNDAIKSALNKIFANIMQEIDKKLTQKNLESYQKDAKELKNKRNR

>AER01005.1
MVKKILNLILLGAIAFSFTLCSSAEKKEESAAPEPSTQEQSAAANRNVDVNSPEAIADSLNEKLKDFRYP
DGLTRPGFSYKKADVTPGDFSEWSKTNAPVIKEGLRKLPDSYALEITGHTDAIGPEQAEGAKKGNIFYSE
LRANAVKQALIKQGIPANRIVTKGAGSSEPVSGLDAKDAKNRRVTFRFATSAPQQ

>AER03729.1
MHTLLTLILSLLLFSGLQSENKNSSSKKLSDSASWIPKENFTQLASQREKFKNVSHNETLKLEIGYKLLT
GKILQFGKLLSTENAYVPIESNEITSELSQLNDKTVRILCSMKGSTCNPIRYEIYPFWYSKEIKPWTIKK
IPDYVNHNIFAFNPTVSPDGKYLFWTAYVKRGKSGTQKIWYSKLDEKGFWEDGKEMNAPLNNEMPSAVIS
ALPGGNELFVFGTFGEKELLDELSKDFETKADLAARSSKNSNVYRKKIEELRAEYDEKTKQISSRVPLYK
SFKEKDSWSKPSILNFPNFYNLYRKRNDSSQEIFGGSTLSSSGRILIYSSQHKDSKGKLDLYVSKMLNDG
TFPLGTNLGDVINTTHEEMAPFLASDDRTLYFSSDGHKGLSIYMTKRIGEGWDQWTKPVEVSENLKGVNF
FSIPANSDWAYISKDGRLFMAYLPKEMRPEKVVIINGKVLDTDGNPLSADIHYESLKSHEKIGSAKSDPS
NGNFSIILPFGENYGFYAQKKGYLPVSQNLNLSSKKKFSEKVEVILQLPPIRERGSIQINNLFFESKSFQ
IAPESAPELDRLAEIVKENPDIEIQIEGHTDNIGKKKDNLILSEKRAAAVAEYLFQKHSISKTRIQTKGF
GDSVPLSKNDSEEARKKNRRVNFTILKKSK

>AER04228.1
MSLKNKNYVLSKKTILILFLVYFVFIISFFSIYSQDLSINQNPKPEKLKGSINTSLNEFGISLTDDGNIL
YFYSKRQNSNYTDIYKSTRTKDEWTQGEEIEVLNSNFDDQSPFILNREEGILFSSNRDGATEFQFANGKI
GVSRDIYFSKKINSSWTEPVLLPRAVNTEEIEENPFLFNNRLYFTRYPFGQVSEADIFVSVYKNNTWEKA
MSLPDPINTVYSEIAATISKDGKTIYFSSNRPGGFGGYDLYKSTLLENGNYSEPINLGPEINTTGDEAFF
LETNDRKTFYFCRRKERDYDIYSIVSNPFQELEKGKSISLDSIHFSLGSYEILENSFSILDNLNSYLKEN
LNIKIKITGHTDLNGDSQDNLILSRNRANAVKDYLVKRGIDSQRIITDGKGSSEPIVPMKNPETDYKNRR
TEFQIISR

>2Y1V_C
GAGTTTTSVTVHKLLATDGDXDKIANELETGNYAGNKVGVLPANAKEIAGVXFVWTNTNNEIIDENGQTL
GVNIDPQTFKLSGAXPATAXKKLTEAEGAKFNTANLPAAKYKIYEIHSLSTYVGEDGATLTGSKAVPIEI
ELPLNDVVDAHVYPKNTEAKPKIDKDFKGKANPDTPRVDKDTPVNHQVGDVVEYEIVTKIPALANYATAN
WSDRXTEGLAFNKGTVKVTVDDVALEAGDYALTEVATGFDLKLTDAGLAKVNDQNAEKTVKITYSATLND
KAIVEVPESNDVTFNYGNNPDHGNTPKPNKPNENGDLTLTKTWVDATGAPIPAGAEATFDLVNAQTGKVV
QTVTLTTDKNTVTVNGLDKNTEYKFVERSIKGYSADYQEITTAGEIAVKNWKDENPKPLDPTEPKVVTYG
KKFVKVNDKDNRLAGAEFVIANADNAGQYLARKADKVSQEEKQLVVTTKDALDRAVAAYNALTAQQQTQQ
EKEKVDKAQAAYNAAVIAANNAFEWVADKDNENVVKLVSDAQGRFEITGLLAGTYYLEETKQPAGYALLT
SRQKFEVTATSYSATGQGIEYTAGSGKDDATKVVNKKITIPQTGG

>YP_005657787.1
MQINNSLNSLSQYVKVNSNEENQNSKNQEQNALAQDPAVEVNISKEAKEKSNTSNQNNSQAPAQALNAQN
NTQQDSSSDSEDKLTELTQKLAEIQAKIVELTAKMSKANEDQIKSIESQIATLNAQASTIQAQIQELQSQ
QA

>YP_007012122.1
MRLKSIVIATTVLLGSATASIAAGSDNIDTSANTNSATTQSSGFAANNFIAPFANTYSALTNKDNTWGPQ
DRTGQWYLGVDANGLAGTPNSPSGAGANFTIGYNINKYFAVQYNQLVGRVFAGLGEGVVNFSNNTMFTPY
AAGGAGWANLAGQATGAWDVGGGLKFELSRNVQASVDYRYIQTMAPSNISGANGRAGTNMIGAGLTWFFG
GKDTTNNDTGNIQDNGATTAAQTVAMPTIDESKYVLPAGIKQCEGNFNLTEDGVACYTVNGDEVTVYLDT
KFAYDKATLNAKGKKAIASFVNFIKDSNISSVTVKGYASQGQTGSEFDIYNQKLSEKRAQAVADYMKQLG
LDSEKIITKGFGYNDTLGGIHKSDPRNQRVEASVSAPLKEAN

>ENR47700.1
MKGFMIAAPASGSGKTTVTLGLLRALKRRGEVLAPVKAGPDYIDPAYHRAASGVDCFNLDPWAMRPELIS
ALSSRMTESGARVLVAEGMMGLFDGAIDGKGSSADLARLLDLPVVLVVDCARQSHSIAALVWGFSQFRKD
VLIEGVILNRVGSPRHEAMLRGALAPLGVPVLGALPRDPALSLPERHLGLVQADEHAGLESFLEQAADVM
EAHIDMDALQTIWLRPKRYDAMANVARLKPLGNRIAVARDDAFAFAYMHLFEGWRRRGAEISFFSPLADE
APKADADAIYLPGGYPELHAQRLAGASRFRTAIGDAAARGVTVYGECGGYMVLGKTLEDAAGVHHPMLGL
LPLETSFARRKLHLGYRLLEPLGGLPWDMPLKAHEFHYASIVREEKADRLFRVRDASGENLGEAGLRVGS
VSGSFMHVIDFSGEAA

>ENR50124.1
MEVILLERIGRLGQMGDTVKVKDGYARNFLLPQGKALRANEANKKKFEGQRAQLEAQNLERKNEAQAVAD
KLNGESFIVVRSAGETGQLYGSVSTRDIAEIITANGFTLHRNQVELNHPIKTIGLHEVSVSLHPEVQVKV
MVNIARSTEEAERQAKGEDLTSIEAIYGIEEQPLSEEVFDDEDEAEDQA

>WP_009872247.1
MKMNRIWLLLLTFSSAIHSPVQGESLVCKNALQDLSFLEHLLQVKYAPKTWKEQYLGWDLVQSSVSAQQK
LRTQENPSTSFCQQVLADFIGGLNDFHAGVTFFAIESAYLPYTVQKSSDGRFYFVDIMTFSSEIRVGDEL
LEVDGAPVQDVLATLYGSNHKGTAAEESAALRTLFSRMASLGHKVPSGRTTLKIRRPFGTTREVRVKWRY
VPEGVGDLATIAPSIRAPQLQKSMRSFFPKKDDAFHRSSSLFYSPMVPHFWAELRNHYATSGLKSGYNIG
STDGFLPVIGPVIWESEGLFRAYISSVTDGDGKSHKVGFLRIPTYSWQDMEDFDPSGPPPWEEFAKIIQV
FSSNTEALIIDQTNNPGGSVLYLYALLSMLTDRPLELPKHRMILTQDEVVDALDWLTLLENVDTNVESRL
ALGDNMEGYTVDLQVAEYLKSFGRQVLNCWSKGDIELSTPIPLFGFEKIHPHPRVQYSKPICVLINEQDF
SCADFFPVVLKDNDRALIVGTRTAGAGGFVFNVQFPNRTGIKTCSLTGSLAVREHGAFIENIGVEPHIDL
PFTANDIRYKGYSEYLDKVKKLVCQLINNDGTIILAEDGSF

>NP_049501.1
MKKMFMAVLFALASVNAMAADCAKGKIEFSKYNEDDTFTVKVDGKEYWTSRWNLQPLLQSAQLTGMTVTI
KSSTCESGSGFAEVQFNND
>YP_006626618.1
MMGYLSALITVLSMGTFLGIVWWAWSSHRQSANRESALLPFALPDEDGPAQQDGAMQP

>YP_006625715.1
MKPLPLAYLAALLPWYAGVIQAQSAPAAGDDASITLEAVRVEASADASAGGLAPAFAGGQVATGAKVGIL
GTRDNLETPFSITAYTNELIQDRQAKGVGDVLQNDPGVRVARGFGNFQESYFIRGFILSSDDIAYNGLYG
LLPRQYISTQLFERVEVLRGASAFLTGAPPSGGGIGGVINLVPKRAPNEPLTRFSAGYGSDSVLEASADI
GRRFGPDDSVGIRINAAQRGGETAIDGERTRTTVFALGLDWRGERARLSADIGYQDNRLKRARPNVTLAG
DAAKVPGAPDAGSNYAQPWSYSNERDVFGTLRGEYDFNGRITGWVAYGMRQSKEENSLANLNNVNGAGQG
KFYRFDNAREDTVNTGEIGLRAKARTGPVGHELVASASYFDLEKKNAYVMDFFNQFDTSIYDPVSYAKPA
ISSTAFRGNDMDDPAKQGVIRLASYALGDTMSFFDDKVLLTAGIRHQRLYQRDYSYDTGIGGTPYEQSHN
SPAAGLVVRVTPQVSLYANYIEALSAGDTAPQTANGLPVVNHGESLAPYVSKQKEVGVKFEHDGLGGGLA
LFSTDKPRGFVGDDQVFRASGKDRHRGVELTTYGELTRSVRVLGGLTWLDAKQLSTGNAATDGKRVIGVP
RFQANLGVEWDIPGVQGLTVDGRVVYTGSSYADAANTLEVPGWTRLDAGLRYMTDIGGHLVTWRARVENI
ANRDYWSSVGGYPGNGYLVLGGPRTFTLSASMEF

>YP_006626823.1
MPAATDPDIAALVARAAAVAPERRSGTLRDLDHIVILMQENRGFDHYYGALPGARGRADPHRAPTPDGDV
FVQAHAGQRLAPYPLAPALPPGQPLGHLTPHTWDDAQRAWNDGRMDQWLAAKGRLGMGYYRPEELPLQTA
LARAFTLCEAYHCSMHAGTNPNRLFLWTGTNDPHGQAGGPALVNTHDRPGPAHEGYAWTTYPERLQAAGV
DWAIYQDMADNYHDNPLAGFRQYRAELAGGAARAPLRERALSTRTLAALAEDAANGRLPQVAWIIAPAAD
SEHPEVSSPAQGAAFSARVLDILTRAPALWRRCALLLTYDENDCFFDHMPPPAPPGAAGGGSTADTAGEY
HQARGGPSAGTPDDPRALHGRAYGLGPRVPMLVVSPFSRGGWLDARVYDHTSVIRLLESRFGVAEPNISP
WRRAVCGDLSHAFDFTGAQDQAGAPGPRSRPSPYACHVEAWAADGRQRVRMANPGHATLVLHVYDCLRLA
QGPRRYTIEPGRQWEDSWPDAGADLACDLWILGPDGFHRHIRRHGAAAPLAAAWRDQPPALLLENRGAQA
LQARIESAYGEAPALLRLAPGEQAAWPYEPASRGWYDLTASAAGQSLRLAGRMRAAGPGRPDPRLG

>NP_212396.1
MIIPKKKQRVEKRKKNFLFNSKKSVNFELKDFANISNIGKRRKKVFKIKNFFKKKINFFKKVSLFFYKLK
IQNINHYEYKYYYKSLRDKVFDIFSVRFDYKLVFKLNAIIFIFILTFYINIFSYYGSYVFLNRLSLPKDY

FIDTFLYYSDQDIAQISSYLPESNVSANVPGFKKNFVLKVFDHKIKPGETLSHVAARYQITSETLISFNE
IKDVRNIKPNSVIKVPNMKGIVYIVKKNDSISSIASAYNVPKVDILDSNNLDNEVLFLGQKLFIPGGRLP
KDFLKEVLGETFIYPVQGVITSGYGYRPDPFTGVISFHNGIDIANLANTPIKASREGVVVTAGFNAGGYG
KYIVISHSNGFQTLYAHLNSFAVKVGKKVSRGAVIGYMGSTGYSTGNHLHFTIFKNGKTENPMKYLR

>NP_212370.1
MLIFGFIGLFFLNIFSLHAQGIVTNKDAQEEFKWALNSYNNGIYDDALLSFKKILSFDPNNLDYHFWTGN
VYYRLGYVEEALMEWRNLKDQGYKVPYLRHLISTIEQRRGIFSNYELNFKKLVKVASLDNSIYKRPHGYQ
ITSLRADKYGGYYAANFVGNEILYFDVNNNVNALVKDGFSYLKSPYDVIEANNLLYVTLYSSDEIGVYDK
VLGVKRKSIGNKGTKDGELLAPQYMAIDKRNYIYVSEWGNKRVSKFGLEGDFILHFGSRTSGYKGLLGPT
GVTYLNENIYVADSLRNTIEVFDTSGNHLYSVFTSIEGIEGLSSDFVGNNVIVSSKDGVYKYSIAKKTIT
KILKADKMNSKISSSILDANNQMIVSDFNNAKVSVYKSDASLYDSLNVDVRRIIRLGGPKIYVELNVSSK
SGLPVVGLKSENFSISNENYYIVNPKVAYNVNASKDINIAVVFDKSSYMKKYDTDQIVGLNALMELSKNK
NFSFINATSVPIIDNIESLTNSIRNTSSLGPYSTDAVKTDVSLKLAGSGLMSKSSRRAVVYFSGGILNRK
AFEKYSLDTIVSYYKNNDIRFYLILFGNDPINSKLQYLVNETGGAVIPFSSYEGVSKVYDLILEQKTGTY
LLEYYYPGPQEPNKYFNLSVEANINQQTGRGEFAYFIN

>NP_212734.1
MAIFLKNKYFYLSLIFIIFLFLFVFSGFLFYSKPIIYDISPIPTSHKDIIVIKGNNLGYSTGEININNNY
LVKSSIISWNNTEIVFKITDEVNSGLIFVKGERGTSNELFLVISRQVPVKLNRKNIPFIFSEDKIILNAN
SSTLLQGMNLFSPFSTITIFLETKDKLYTILPQNILNVSENRVEFVSPKTLNSSGKLYVLLDNIQSNKVP
FSVKNDFFKWTLSDFKEFVIIEEIYFSQDVSSNFDSNPQDINFNIFYLRPIENERQKITERNSEHLDFNI
DNLFFENLKTNKFIFKTRVKTYKLNLEFLDAKYLESIEVNRDINNQEYKKYVQDKKKDYLSYSYVDLMSL
DSLILSKTSGSNSVYKLAKAIIDVLTSNFKIVENNLSLKDSIEEKKISSGNLIVLTNLLFLKYDIPLRNI
VGLYYDSNSLKLKEHFWFEFFLAGVGFVYFDIINAVLFKDSSKYFLNISDNYIQYGCKEDYDKNEFFDGY
LDSGFLKYKSLTNGSYSLMHRFVLEDNF

>NP_212856.1
MKINKTFILLFLFTKFSFVQAQANQILTEISPLSILSKNGKGSVYLKVSKSSDYILTLDKSSNSDFVFKI
YDISNKKYITDKVKRRDFKIRLDKNSLYAIIYVGTKNENIKFSLTDLDFSILSSDSLKAKTSKIEKEDLF
FTLKDLPVLNLTAKLKKYVLRIYKSNIYIAYQLENSDDIKVAEFIEDVGWFNLDSSVNRNITNIVNFDFS
INSKGNLYIAFVTKSGADFASELIVKKFNSRKWIDISPGHIENFGSLLNISIDLKDRLYLAYLREIRGEY
KINLISNMGYGSIWTDVIHAYLSKGDSNVNSSNIGLISEPFLGIFYNYKSNNEIKSEFIVNNENAWVNAN
IPSVYMANFIKGFFDSNFNQIIMSFVSENRPIVNICPLKSSRWINISPNVEMEGLSADIGLYKNNLFLAF
EDNNNVRLIYFKNKNWYFLNKLENFKSNVKSPQIGIYGNQGLVISTLSSNSNELFFTLICQ

>NP_212213.1
MNIDYFIKNKILHNSGINYQVKTKNLNPSNKLIEKIKAVNPKIKAKTWNEYNKEFYKILKIERNTMLIIL
ASIFIVIAVNTYYLQKRIIINKNKAILILLAMGLRIKKIKQIFFIHSIIICTVGGLLGLTLGISISLNIN
EILKIIDNLVNTLINFLNQILALKIDGIKIQIVKNTITPKLFLSDLTFTFCFACFSTMYSSMKATKKIGS
QKNIETINGQ

>NP_212767.1
MNKILEKIQNNTTILIEASAGTGKTHILENVVINLIKTKLYSINEILVLTFTKKATEEMHTRILKVIENA
YSNSKTNEILKEAYEQSKKLFISTINKFALHALNNFQIETENYSKYKPKEKFSKEIDEIVYDFLRKSDSL
IQALDIKDYELKVFKSDAKKTEEIVLKIKKAYERDTTQELGDWLKTQTAFENILLKKEELIKDYNKIIED
LDKMTKDEILSFYNKHIQTGKLEIEYSKENDIFKIAETLLKNKFFSTLIEKETKKNSKLSPKELKIKNDL
ICLGINIKHEKYKSEDNRNKNRNNLKQYVILKVEYKILKYIEKELKKTIKSTNTIDQNYIISNLKNYLKS
EDKKLLNAIKNRYKIILIDEAQDLSLIQIEIFKILKTAGIKLIFIADPKQIIYSFRKADISFYNKEIKNK
INTDARIVLKINHRSSKKLIGPLNKIFNNIYNNAIADEIEKIDFTNSLPNQKNDNNKIVINGQEIEGINI
ITTNTESEEDIYQKTALTIKYLLAYGKIAENNKIRNIKMQDIKVLCRGKNEINLIDKALKKEQIQTNKTQ
EKFLKTKEFSEIFYIIKCLDRKQSFKTLNYILSSKILNVPWNLQRILIKQDKICLIEEFIENIIVLLEKN
EITLINAINKITFEKNLWIKIANITKDQKIIEWAKNKINYKGLLIKEGKLENLKTYETTLEIISKIYHKE
QNIQSLISTLESLIINEEPEEIEEKINNINNDNESIELMTIHKSKGLGMNIVFLLNTTPIENSNFFSKKN
QFYKFYQDGKIEYDFFKLEENKKYARLKILSEEKNIFYVGATRAKFALFIIKINSITSKLLEIAKIFTID
DIKHDFNIHEFIGQKRFNKKKYNTNVNTKLIPPKPIIKNMFKKEYTSSFSSLTAQAHHKEFYENYDFKNI
NYEKETELDYEPGLEETLPKGKDIGNILHAAMEEIIFSTAKDTFDNFKKNNIEIIEKQIQKINSNLNTIE
IQNSLAKMIYNILTYNIRAINTRLCDIEELQKEMEFLIKINPEFQKQKYLFDKHFEDLHIKLSDGYLKGI
VDLIFKANNKIYILDYKTNYLGKNKEDYNITNLENTIKKEYYDLQYKIYALGIKKILFKNKKEYNQKFGG
IIYLFTRAFEDNIECLKSKFENGIYFNLPKFNDVDLDKIILELGIKRHL

>YP_222268.1
MIQSALFLMLGVLVAVFVVVLLGPSVWRRAFFLARRQVQAELPITLAEIRADRDGLRAEHAVAVSKLEQL
LKLERGKTAEQKVTLARQQDELKRIPLLEQNIARLEKRLGEEEKGAAEARQARDTALEKAEILQAELERV
QGHYVALESLADTLRIEISAHEAEISRLMSEITEMRHDRKDATARYNELSTQLTAAQTKLKSETRRNGEL
QQKLEKLISELSDAQEKLERQNRHGDGALMSADQIEIARQNAALREEMATLAARMVAITAEREGPDSPIH
KLLKGTALTEKGKGKKNKAPKSLATRIRDMQ

>YP_221386.1
MRYLARFCCKVRFSDAAAGDGHDSLVQPGRP

>YP_220816.1
MEQFVLDKRLHADTFYVARLGLCELRLMNDRRWPWLILVPRRPGLTEIHQMTPLDQTMLTFEAGIVAHAL
KTVTACQKINTGALGNVVRQLHVHVIARNEGDAGWPGPVWGHGVRETYDEKDAQKLIAEVRAAL

>YP_221262.1
MKQKFERILEPTDTWAIFDTTADVPAMMGTRLLIGLTEKEAEELLAILNAPPKGRNKRSAA

>YP_222757.1
MFRWGVLSTAKIGVTQVIPALAASDNGVVHAIASRDHARARAVAGRFGAPLAFGSYEELLASKEVDGVYI
PLPTSQHIEWTLKAAEAGKHVLCEKPIALCAGDIDQLIEARDRHGVTIAEAFMVYYHPQWIKLRALLAEG
AIGNLSHVQGVFSYFNVDPGNMRNQPGLGGGALPDIGVYPTVVTRMVTGREPLSVRASVEYDPGFGTDRY
ANVSARFDGFDLTFYVATQLAGRQCMVFHGDKGFIEVYAPFNTGKYGHGRITLHDASHMQATEWSFGDAN
HYQLQAQSFVRAARGENVPLFSLENSRANQKFIDAIYSAGRSGVSETV

>YP_222058.1
MLDIIILVVIGGAFGAMTREFIMLMVPPLTDGFPLDILVANVVACFLLGTVTALYARKIHSRDVHTIIGT
GMMGGVSTFSSFAYGSVVLASASMSAFLIAAAYVTVSVVAGYVAVLAGMKFGEKSADILHRYPPMASIID
SGLVTVESRHSVAETIERVAAKAKSMGMNVFTRVDHGAGAKEAGLGLPPTELIIFGNPQNGTVLMQDKRT
IGLDLPIRALAWEDGSGKVWLTVNDPAWLAQRHSLGLSSDVAIKAMVTGTGTVTKYAAGD

>YP_002733631.1
MSLISIGTLGGTIAMVEGGSGGGVVPTLTADALIAAVPQIRNYAQIKAQSLFQLPSPSLKFRHGLEVIGW
AENEVKSGADGIVLTQGTDTLEEMAFLLDLLWEHPQPLVLTGAMRSPRAPGADGPANLVAAVLTVASRLS
RERGVLVAMNDTIHAARWITKSHALSVQTFVSSDSGPLGYVYEETPVYINTLSRLPKIDRTQFNADVKIG
FYESLLDGDAQMLRSMFESGRYDGIVVAGFGAGHVSGDEADIIERYASRIPVVVASRTYGGRTATKTYGF
HGSEIDLQTKGAMLAGWLSPPQGAVIAVGTTGSG

>YP_002733166.1
MSAGTTRQKAIAKSLTLLLPAVPYSDSEPIRAAALAPHMKTLPPSTAVWLATVAHVRHTHTDYDALRDDG
YDKDSARFFVLNAINAKLTEWRATRLLSPEDDEAVEM

>YP_002731823.1
MLLLAAHFSSFALSPALAFEIFGIHLWGKDKKQDPDIIDPKTYSVDVTTTGDRKNADGKEADLKSVIEGA
SGLVSDADKPASGSAGLLAKARGDYRRILSALYGEGRYGGTISIKVDGREANDIPPDTEIPNNAKVAITV
DPGPQFLFSRTAISNIAPPPGNRRDKVQTPEEAGFAPGQEAKSGTILKAERLAVEAWRQEGYAKARVTGE
DVVADHADNRVSADIALDPGRKAYYGPVSVVGTARMDPQFVAWMTGLKPGQEYDPDDIENAKKRLGRMEV
FRAMTFEEADKIELDGSLPITLNVQERKPRRFGFGAEYSTIDGFGVTSYWMHRNLLGRGERLRFDAKVSG

IGGSQDNSFDPKNYTYLLGASFAKPGVYTPDTDFVATLDAKREVLDAYTETSINAKTGFTQIFSDELSGA
LYANASQGHFVDDVFGKRDFTTAGLEGNLLYDSRNNKPDPSSGFYLVGNIQPFYEFHYGNFATRFTAEGR
TYHGFGQTDRVVLAGRLKVGSIVGGSIADLPPSQLFLAGGGGSVRGYGYRNIGVSAGNGNIIGGRSLVEA
NGEVRTRITDSIGAVAFVDAGYVGEKSFPDFSEQMRVGVGGGLRYLTSLGPIRLDVAVPLNRRSGDPNYG
FYVGIGQAF

>YP_002731824.1
MTRLFAIIAFFVFVVAAAVFAWPEQKVQMPAQAQAQTDTQAQGAAQPADDQPQAEEEKSYFLSFVENKLS
APNRRISISGISGVLSSEATVGSITVADREGVWLRIENVKLDWSRTALFLGRLSVQSLTAERIDVIRKPL
PDKGLPSPESSSFSLPELPISINLEQLDVARLHFGPTIFGLESDVSLKGSLSLADGSLNSNFDIKRLDGP
GGNFALRAAYANADQKLDLDLKVSEPANGIVANLLKIDGRPPVDLTLTGAGTLDDLKVALALDTNGSRTL
TGEFTLDRQGKGPDGGRVFAARFNGPIAVLVPPVFRDFFGAETVLSADGLVKDGGGFRLDDLALTSAALK
LKASAETDNDHFLDKLRIDAAIADESKGRVLLPVKGAETYIDNARFQVAYGDRPTNDWTGMLSINGLKNA
TLSASAIALDMGGLAENLNDPASRHITFSVNGGMDGISSPRADIAEALGGKITLAVNGAWRAGSPVELAQ
ADIEGNGLSLDLKGKINVFAFNGDIAAKVASLAPFGALANRNVAGSIDVKATGMVRPISGAFQLDFDGTA
ENVRTGTEAVDRILDGAVNLSGALGRSSEGFSARNFRIGNALSEITANGSLASDKADFDFGVMLSDLKLL
SDKASGPMTVKGSARGDDKLIALALRADVPQGTLAGRKLSNGGLDFNALLDGNATTGVALSGKLDGGALL
NGERVDLGSLIDIAHDEKRLTNLMFNAGGAHLSGNVTQSAKGLLSGALKLDAPDISTAAALFLADATGAA
NADITLDASDGRQNATVTANASGIRINGNHIASADIALTAKDLLGVPVVDGNAAARDIMVGTFGIDSFDA
SANASGSKTDFTANTRLKIGTVANAAGSLEPKDGGFELALATADVKQGALSAHLAAPATISMKGDEITFG
DIVVNTGDGQVRINGLVDGRLDLSVALTNLPLALANTFKPELGLGGTVNGTARITGTKEAPDIAFDMTAR
DVMAAELKKQGLGALNAEAKGTSANERLNIDARVTGGAASGTGIDVRANGAVPLGKGDLALDVNLTNLPL
VLLNGVVKGQDLAGNVTGTARITGPLTNPAANFNLRGSGLAAKPLRESGLAPLTLQAAGKYAAKAVDLSS
LQLDGPGGLNVTANGKIPFSGQGLGVNVNGNVPLALANRFLADRGAQASGTLSLTANVSGGFQKPQLRGM
FSTNGATFVDPETNLRFVNINVMGSMEGETITLRQVNANLSSGGSLSASGTISTNAVANFPANIEVKFNR
ARYADGKLVVATIDGGVSITGPLMRDPLISGRIDVDKAEITVPENLGGGAAYVPVHHKNTPKNVQVTLDR
ARVETRKNKVPTPVARPTVPRLDVVVSAPNQIFVRGRGLDTELGGRVRLTGPITNIRPVGSFDLIRGRIS
ILGQRITFDEGHVTLVGDLNPQLNFVARSEGDGITAIVTVTGTVDDLNIVFSSSPELPQDEVLAQLIFKR
SIGKLSAFQIAQLAAAAAELAGGSNNSLINKLRQGTGLDDLDVVTNSKGETSVRAGRYIRDNIYLGVEAG
ARGTTKGTVNLDISRNLKVKGAFGAEDDSSGGIFYEKDY

>YP_002732480.1
MKFTSNRKHHRYFRLFSKAMLSYQPSIQCAESQILDLTVNTTAHEIIPKSVTSSDKKRTIPARAFRLFQL
MKFGNWERAHFCRHPILMDTSPGSFAYEYDCLLKMAFWSRSSLSQFLCSFHHCIEYPSDAIPSKEYFGI

>YP_002733724.1
MSADIRFDSLKTIVPAVSRETADRLIAFEDLFRKWSKAINLASPSTLADLWNRHILDSAQLFPLAKEATR
WLDIGSGGGFPGIVTACFLAERSGGCIDLVESAGKKAAFLRTAAGHLHVPARVHSARIESMWEKIETPQV
VTARALASLGDLFTLAEPWLSDGAKALFQKGRDYQREIDESRVGWSFDLVKHPSAIDQASVILEISNLRR
KTD

>YP_008746387.1
MQVSYKTISSYEYDAISGQYKQVDKQVEDYSSTSDSDFMDILSKLEENSGISSTEENLQNNTQSSNSNSS
QYAQMSSMYAYRFRQNEGELSLKAQSASVHNDLLNQNANGQNKDSSLLGDLLNAI

>YP_179748.1
MQNFKELNEKIAWQKVDNLLPVIIQDAKTCEVLMLGFMNNEALEKSLESGKVVFFSRTKQRLWMKGEESG
NFLNIIDLSLDCDNDTLLILANPVGPTCHTGDISCFEKISKNADFVFLARLEKLINARKNADENTSYTAK
LFKSGTKRIAQKVGEEGVETALAATVKDKEELICEAADLMYHLSVLLADANLSFSDVISKLKERHKA

>NP_296643.1
MLVMPFSLRSTSFCFLACLCSYSYGLASSPQVLTPNVIIPFKGDDIYLNGDCVFASIYAGAEQGSIISAN
GQNLTIVGQNHTLSFTDSQGPALQNCAFISAEEKISLRDFSSLLFSKNVSCGEKGMISGKTVSISGGDSI
VFKDNSVGYSSLPSVGQTPTTPIVGDVLKGSIFCVETGLEISGVKKELVFDNTAGNFGAVFCSRAAQGDT
TFTVKDCKGKILFQDNVGSCGGGVIYKGEVLFQDNEGEMLFRGNSAHDDLGILDANPQPPTEVGGGGGVI
CTPEKTVTFKGNKGPITFDYNFAKGRGGAIQSQTFSLVADSAVVFSNNTAEKGGGAIYALEVNVSTNGGS
ILFEGNRASEGGAICVSEPIAANNGGLTLHAADGDIIFSKNMTSDRPGERSAIRILDSGTNVSLNASGAS
KMIFYDPVVQNNPATPPTGTSGEIKINESGSGSVVFTAETLTPSEKLNVINATSNFPGNLTVSSGELVVT
KGATLTVGNITATSGRVTLGSGASLSAVAGTAGTCTVSKLGIDLESFLVPTYETAKLGADTTVAVNNNPT
LDLVMANETEMYDNPLFMNAVTIPFVTLVSLQTTGGVTTSAVTLNNADTAHYGYQGSWSADWRRPPLAPD
PSGMTPLDKSNTLYVTWRPSSNYGVYKLDPQRRGELVPNSLWVSGSALRTFTNGLKEHYVSRDVGFIASV
QALGDYVLNYKQGNRDGFLARYGGFQAVAASHYENGGIFGVAFGQLYGQTKSRLYDSKDAGNITILSCFG
RSYIDVKGTETVVYWETAYGYSVHRMHTQYFNGKTNKFDHSKCRWHNNSYYAFVGAEHNFLEYCIPTRQL
ARDYDLTGFMRFEMSGGWSSGAKETGALPRHFDRGTGHNMSLPIGVVAHAVSNGRRSPPSKLTINMGYRP
DIWRVTPHCNMKIIANGVKTPIQGSPLARHAFFLEVHDTLYVRHLGRAYMNYSLDARHRQTTHFVSLGLN
RIF

>NP_445596.1
MSTRRPIQLLDPLTINQIAAGEVIENSVSVVKELIENSLDAGADEIEIETLGGGQGAIIIRDNGCGFRAE
DIPIALQRHATSKIREFSDIFSLNSFGFRGEALPSIASISKMEIQSSIEGDEGVRTVIHGGDIVSCEPCA
RQLGTTVIVNSLFYNVPVRRGFQKSMQSDRLGIRKLIENRILSTANIGWSWISEGHHEIQIAKQQGFQER
VAYVMGDHFMQDALTIDKEANGVRIVGVLGSPSFHRPTRQGQKIFINDRPIESLFISKKVGDAYALLLPL
HRYPVFVLKLYLPSSWCDFNVHPQKIEARILKEELVGDCIKEAIVETLACPPGILCRTHQEIEESDSVPL
PMFRMLETSDVQEEESVEFDQNLFAYSSEDVSLEKQEYTSRGPKSQMDWIYSSDVRFLTSLGRVVLAEDL
EGVHIIFTAAARKHLFFLSLMQENSRMYQSQALLIPLRLQVTPEEAFFFSHHGRTLCDLGIEISQVGPCV
FSIESTPTVIGEEELKEWLLLLAARGSTDINSEALTALMKETLTQATFSKHQHVFDVSWLKLLWSVGKPE
KGFDGARIRRLILDSDFMEG

>NP_445074.1
MPGWQWFFYLYSQYTRHLKALILEKKQGLCFWAKFFSFC

>NP_444680.1
MEQTLSIIKPDSVSKAHIGEILSIFEQSGLRIAAMKMMHLSQTEAEGFYFVHRERPFFQELVDFMVSGPV
VVLVLEGANAVSRNRELMGATNPAEAASGTIRAKFGESIGVNAVHGSDTLENAAVEIAYFFSKIEVVNAS
KPLV

>NP_444832.1
MASGIGGSSGLGKIPPKDNGDRSRSPSPKGELGSHEISLPPQEHGEEGASGSSHIHSSSSFLPEDQESQS
SSSAASSPGFFSRVRSGVDRALKSFGNFFSAESTSQARETRQAFVRLSKTITADERRDVDSSSAAATEAR
VAEDASVSGENPSQGVPETSSGPEPQRLFSLPSVKKQSGLGRLVQTVRDRIVLPSGAPPTDSEPLSLYEL
NLRLSSLRQELSDIQSNDQLTPEEKAEATVTIQQLIQITEFQCGYMEATQSSVSLAEARFKGVETSDEIN
SLCSELTDPELQELMSDGDSLQNLLDETADDLEAALSHARLSFSLDDNPTPIDNNPTLISQEEPIYEEIG
GAADPQRTRENWSTRLWNQIREALVSLLGMILSILGSILHRLRIARHAAAEAVGRCCTCRGEECTSSEED
SMSVGSPSEIDETERTGSPHDVPRRNGSPREDSPLMNALVGWAHKHGAKTKESSESSTPEISISAPIVRG
WSQDSSVSFIVMEDDHIFYDVPRRKDGIYDVPSSPRWSPARELEEDVFGDYEVPITSAEPSKDKNIYMTP
RLATPAIYDLPSRPGSSGSSRSPSSDRVRSSSPNRRGVPLPPVPSPAMSEEGSIYEDMSGASGAGESDYE
DMSRSPSPRGDLDEPIYANTPEDNPFTQRNIDRILQERSGGASASPVEPIYDEIPWIHGRPPATLPRPEN
TLTNVSLRVSPGFGPEVRAALLSESVSAVMVEAESIVPPTEPGDGESEYLEPLGGLVATTKILLQKGWPR
GESNA

>YP_006096033.1
MDTIDLGNNESLVYGVFPNQDGTFTAMTYTKSKTFKTENGARRWLERNSGE

>YP_006096653.1
MQELSMLRMTPLASAIVALLLGIEAYAAEETFDTHFMIGGMKDQQVANIRLDDNQPLPGQYDIDIYVNKQ
WRGKYEIIVKDNPQETCLSREVIKRLGINSDNFASGKQCLTFEQLVQGGSYSWDIGVFRLDFSVPQAWVE
ELESGYVPPENWERGINAFYTSYYVSQYYSDYKASGNNKSTYVRFNSGLNLLGWQLHSDASFSKTNNNPG
VWKSNTLYLERGFAQLLGTLRVGDMYTSSDIFDSVRFSGVRLFRDMQMLPNSKQNFTPRVQGIAQSNALV

TIEQNGFVVYQKEVPPGPFAITDLQLAGGGADLDVSVKEADGSVTTYLVPYAAVPNMLQPGVSKYDFAAG
RSHIEGASKQSDFVQAGYQYGFNNLLTLYGGSMVANNYYAFTLGTSWNTRIGAISVDATKSHSKQDNGDV
FDGQSYQIAYNKFVSQTSTRFGLAAWRYSSRDYRTFNDHVWANNKDNYRRDENDIYDIADYYQNDFGRKN
SFSANMSQSLPEGWGSVSLSTLWRDYWGRSGSSKDYQLSYSNNLRRISYTLAASQAYDENHHEEKRFNIF
ISIPFDWGDDVTTPRRQIYMSNSTTFDDQGFASNNTGLSGTVGNRDQFNYGVNLSHQHQGNETTAGANLT
WNAPVATVNGSYSQSSTYRQAGASVSGGIVAWSGGVNLANRLSETFAVMNAPGIKDAYVNGQKYRTTNRN
GVVVYDGMTPYRENHLMLDVSQSDSEAELRGNRKIAAPYRGAVVLVNFDTDQRKPWFIKALRTDGQPLTF
GYEVNDIHGHNIGVVGQGSQLFIRTNEVPPSVNVAIDKQQGLSCTITFGKEIDESRNYICQ

>YP_006097351.1
MKYISLFLFILLLCGCKQQELLNHLDQQQANDVLAVLQRHNINAEKKDQGKTGFSIYVEPTDFASAVDWL
KIYNLPGKPDIQISQMFPADALVSSPRAEKARLYSAIEQRLEQSLKIMDGIVSSRVHVSYDVDTGDSGKT
ALPIHISVLAVYEKDINPEIKINDIKRFIVNSFASVQYENISVVLSKRRDIIEQAPTYEISEPVFAYDKT
MPVSILLALMSIATCWLLWKYRAILTNLIRLKNK

>YP_006095340.1
MNTARLNQGTPLLLNAVSKHYAENIVLNQLDLHIPAGQFVAVVGRSGGGKSTLLRLLAGLETPTAGDVLA
GSTPLAQIQEDTRMMFQDARLLPWKSVIDNVGLGLQGQWRDAARRALAAVGLENRAREWPAALSGGQKQR
VALARALIHRPGLLLLDEPLGALDALTRLEMQDLIVSLWQEHGFTVLLVTHDVSEAVAMADRVLLIEEGK
IGLDLTVDIPRPRRLGSVRLAELEAEVLQRVMQRGHSEQPIRRHG

>YP_006095913.1
MTNINTACVKNNASYQLNNALPNKETISSNFCERLAQWGNKSLNNGEERAIAVERIKEAYNSNMASLDLS
YLDLSELPPIPSTVNTLNLEKNCLTCLDFTDNASLVNINLSFNKIKTITFPNESKLENIYIDHNNLENLD
LKNQYSLVNLEAQNNNLTKINISDSYKLKFLNLDYNKLASLDLSRQESLIELSAHHNMITDLILHNHPRM
KKITLNDNHIAHLNAKTTTKLEYLNLSNNNLLPTDDIDQLISSKHLWHVLVNGINNDPLAQMQYWTAVRN
IIDDTNEVTIELSYNLAITNIDTSDEHLVEVSENSEGNHIKDNDSMSIRYRSKYYSREYALIEEETIFSD
AELKAILPMHRMYGVGDYKSNSSSLPSHSGLKDPTGTPVCYYIHNEDKPSLGYGSTPNNWLSQSFSTEL

>YP_006094651.1
MPLRRFSPGLKAQFAFGMVFLFVQPDASAADISAQQIGGVIIPQAFSQALQDGMSVPLYIHLAGSQGRQD
DQRIGSAFIWLDDGQLRIRKIQLEESEDNASVSEQTRQQLMALANAPFNEALTIPLTDNAQLDLSLRQLL
LQLVVKREALGTVLRSRSEDIGQSSVNTLSSNLSYNFGIYNNQLRNGGNNTSSYLSLNNVTALREHHVVL
DGSLYGIGSGQQDSELYKAMYERDFAGHRFAGGMLDTWNLQSLGPMTAISAGKIYGLSWGNQASSTIFDS
SQSATPVIAFLPAAGEVHLTRDGRLLSVQNFTMGNHEVDTRGLPYGIYDVEVEVIVNGRVISKRTQRVNK
LFSRGRGVGAPLAWQIWGGSFHMDRWSENGKKTRPAKESWLAGASTSGSLSTLSWAATGYGYDNQAVGET
RLTLPLGGAINVNLQNMLASDSSWSNIASISATLPGGFSSLWVNQEKTRIGNQLRRSDADNRAIGGTLNL
NSLWSKLGTFSISYNDDRRYNSHYYTADYYQNVYSGPFGSLGLRAGIQRYNNGDSNANTGKYIALDLSLP
LGNWFSAGMTHQNGYTMANLSARKQFDEGTIRTVGANLSRAISGDTGDDKTLSGGAYAQFDARYASGTLN
INSAADGYINTNLTANGSVGWQGKNIAASGRTDGNAGVIFDTGLENDGQISAKINGRIFPLNGKRNYLPL
SPYGRYEVELQNSKNSLDSYDIVSGRKSHLTLYPGNVAVIEPEVKQMVTVSGRIRAEDGTLLANARINNH
IGRTRTDENGEFVMDVDKKYPTIDFRYSGNKTCEVALELNQARGAVWVGDVVCSGLSSWAAVTQTGEENE
S

>YP_006094524.1
MVKKAIVTAMAVISLFTLMGCNNRAEVDTLSPAQAAELKPMPQSWRGVLPCADCEGIETSLFLEKDGTWV
MNERYLGAREEPSSFASYGTWARTADKLVLTDSKGEKSYYRAKGDALEMLDREGNPIESQFNYTLEPAQS
SLPMTPMTLRGMYFYMADAATFTDCATGKRFMVANNAELERGYLAARGHSEKPMLLSVEGHFTLEANPDT
GAPTKVLAPDTAGKFYPNQDCSSLGQ

>YP_006094626.1
MSPAVFYMKAAGKTKGIVMFIGHLPAGYLITCTLMKRLPLIARHARWAMVIGLLGAVAPDFDLFWCYLVD
NGQRHHHLYPSHWPLLWFALLAATLVWAGIAKNKLPAWLGVVFCLNSIAHLLLDTLVGDIWWLMPFVNQP
FAFFHITAVHHPWWLNFLLHWSFLLELALVVAAIAMWCRRNAKMVNA

>YP_006094558.1
MPTPCYISITGQTQGNITAGAFTADSVGNIYVQGHEDEMLVQEFLHNVTVPTDPQSGQPSGQRAHKPFIF
TVALNKAVPLLYNALASGEMLPKVELHWWRTSVEGKQEHYFTTRLTDATIVDMNLHMPHCQDPAQREFTQ
LLAVSLAYRKVEWEHIKSGTSGADDWRAPLEA

>YP_006097604.1
MSNTLIINGAKKFAHSNGQLNDTLTEVADGTLRDLGHDVRIVRADSDYDVKAEVQNFLWADVVIWQMPGW
WMGAPWTVKKYIDDVFTEGHGTLYASDGRTRKDPSKKYGSGGLVQGKKYMLSLTWNAPMEAFTEKDQFFH
GVGVDGVYLPFHKANQFLGMEPLPTFIANDVIKMPDVPRYTEEYRKHLVEIFG

>YP_005179212.1
MTTENPAIPTKKKKSFWKKMKPLFGLTVLIPTAFSAVYFGLFASDIYVSESSFVVRSPRSQSSLSGVGAL
LQSTGFSRSQDDTYSVQEYMRSRTALSALEQGLPLRTFYSEKGDLLSRFNGFGLNDTQEAFYRYFKERLS
VDVDSISGIATLRVHAFDAEEGYQINERLLKEGESLINRLNERARKDTIEFAEQAVKDAEKNVNETAQAL
SQYRIKNKIFDLPAQSGVQLSLISSLKSELIRVETQLAQLVSITPDNPQVPALQMRQKSLKKEIDEQTRQ
LSGNGNSAATQTADYQRLMLANELAQQQLAAAMTSLQNTRGEADRQQLYLEVISQPSKPDWALEPSRIYN
IIATFIIGLMLYGVLNLLIASIREHKN

>YP_005179356.1
MKKSLLAVIVGAFAFASVADANIYAEGDIGLSQTKANGSNNTRVEPRVSVGYKVGNTRVAGDYTHHGKVD
GTKIQGLGASVLYDFDTNSKVQPYVGARVATNQFKYTNRAEQKFKSSSDIKLGYGVVAGAKYKLDGNWYA
NGGVEYNRLGNFDSTKVNNYGAKVGVGYGF

>YP_005178261.1
MQAINPNWNVPKNIHAFTTTREGGVSLAPYLSFNLGDHVGDDKSAVKTNRTLLVEKFGLPQKPIFLTQTH
STRVLQLPYSEQNLEADAVYTNVPNQVCVVMTADCLPVLFTTTSGNEVAAAHAGWRGLCDGVLEETVKYF
QAKPEDIIAWFGPAIGPKAFQVGIDVVEKFVAVDEKAKLAFQPDAIEDGKYLSNLYQIATQRLNNLGITQ
IYGGNHCTFNEKEKFFSYRRDNQTGRMASVIWFE

>YP_005178684.1
MKILGIETSCDETGVAIYDEEKGLIANQLYTQIALHADYGGVVPELASRDHIRKTAPLIKAALEEANLTA
SDIDGIAYTSGPGLVGALLVGATIARSLAYAWNVPAIGVHHMEGHLLAPMLDENSPHFPFVALLVSGGHT
QLVRVDGVGKYEVIGESIDDAAGEAFDKTAKLLGLDYPGGAALSRLAEKGTPNRFTFPRPMTDRAGLDFS
FSGLKTFAANTVNQAIKNEGELIEQTKADIAYAFQDAVVDTLAIKCKRALKETGYKRLVIAGGVSANKKL
RETLAHLMQNLGGEVFYPQPQFCTDNGAMIAYTGFLRLKQGQHSDLAIDVKPRWAMAELPAI

>YP_005179346.1
MKKTAIALVVAGLAAASVAQAAPQENTFYAGVKAGQASFHDGLRALAREKNVGYHRNSFTYGVFGGYQIL
NQNNLGLAVELGYDDFGRAKGREKGKTVAKHTNHGAHLSLKGSYEVLDGLDVYGKAGVALVRSDYKFYED
ANGTRDHKKGRHTARASGLFAVGAEYAVLPELAVRLEYQWLTRVGKYRPQDKPNTAINYNPWIGSINAGI
SYRFGQGAAPVVAAPEVVSKTFSLNSDVTFAFGKANLKPQAQATLDSIYGEMSQVKSAKVAVAGYTDRIG
SDAFNVKLSQERADSVANYFVAKGVAADAISATGYGKANPVTGATCDQVKGRKALIACLAPDRRVEIAVN
GTK

>YP_005178875.1
MKTKAILTALLGAIALTGCANNNDAKQVSERNDSLEDFNRTMWKFNYNVIDRYVLEPAAKGWNNYVPKPI
SSGLAGIANNLDEPVSFINRLIEGEPKKAFVHFNRFWINTVFGLGGFIDFASASKELRIDNQRGFGETLG
SYGVDAGTYIVLPIYNATTPRQLTGAVVDAAYMYPFWQWVGGPWALVKYGVQAVDARAKNLNNAELLRQA
QDPYITFREAYYQNLQFKVNDGKLVESKESLPDDILKEID

>NP_208154.1
MDHLKHLQQLQNIERIVLSGIVLANHKIEEVHSVLEPSDFYYPPNGLFFEIALKLHEEDCPIDENFIRQK
MPKDKQIKEEDLVAIFAASPIDNIEAYVEEIKNASIKRKLFGLANTIREQALESAQKSSDILGAVEREVY

ALLNGSTIEGFRNIKEVLESAMDLITENQRKGSLEVTGIPTGFVQLDNYTSGFNKGSLVIIGARPSMGKT
SLMMNMVLSALNDDRGVAVFSLEMSAEQLALRALSDLTSINMHDLESGRLDDDQWENLAKCFDHLSQKKL
FFYDKSYVRIEQIRLQLRKLKSQHKELGIAFIDYLQLMSGSKATKERHEQIAEISRELKTLARELEIPII
ALVQLNRSLENRDDKRPILSDIKDSGGIEQDADIVLFLYRGYIYQMRAEDNKIDKLKKEGKIEEAQELYL
KVNEERRIHKQNGSIEEAEIIVAKNRNGATGTVYTRFNAPFTRYEDMPIDSHLEEGQETKVDYDIVTT

>NP_207004.1
MSELVTEYANATNNLLFKELIKHVSGNSEGIKNFCQCVKEIKKCNTPNKKYNSDEFFIMGKHKQNQLAKI
YSYFKKLSEGEIKPQNEDILKKLKSLDEIFKTTDFTKFTPETEVKDIIKEIDEKYPINENFKQQFRTFRL
NIGNLKKKIKNSLKYLEKTRKNFERKKESLIREIEYYCKNQKTLEFDYDVLLDNIQQICKKYIASHVVND
ASKDIKSMMCQFYLEKIDLLFNSEIEQYRYSDFLESARKFLWEDIKTLDEKSGVHLFPKNIGEIKDKFET
NKEKFKQSKNYSEFAEYCRECNPYTAFQNLRNKVQFPLSGGLSYKSYKLVPTMKEYKEPKITDNDLKTAL
FTLFDYSSPSEFDQWDWFFRNSLFRKMDFNPYNIWKNLNLDDKKDFAEILEYNMQLKINSLITKEFNKLL
AIAEDSSQDSYQLKIRVRHNNKFYDYSKKSTAYEIKLEIHDCRKSHDQNEPIILSQQSTGFQWAFNFMFG
FLYNVGSDFSLNKNIIYVMDEPATHLSVPARKEFRRFLKEYAHKNHVTFVLATHDPFLVDTDHLDEIRIV
EKETEGSVIKNHFNYPLNNAGKDSDALDKIKRSLGVGQHVFHNPKKHQIIFVEGITDYCYLSAFKLYFNE
REFKENPIPFTFLPISGLKNNPNEMKETIQKLCELDNHPIVLTDDDRKDGSDPQRAKSEQFKNANEEMHD
PIRILQLSDCDRHFKQIEDCFSANDRKKYAKNKQMELAMAFKTRLLYGEKDDVMSEETKKNFLKLFEWIK
KECNNLTIKKEYIKFDYNTPQML

>NP_206845.1
MKEKIALITGVTGQDGSYLAEYLLNLGYEVHGLKRRSSSINTSRIDHLYEDLHSDHKRRFFLHYGDMTDS
SNLIHLIATTKPTEIYNLAAQSHVKVSFETPEYTANADGIGTLRILEAMRILGLEKKTRFYQASTSELYG
EVLETPQNENTPFNPRSPYAVAKMYAFYITKNYREAYNLFAVNGILFNHESRVRGETFVTRKITRAASAI
AYNLTDCLYLGNLDAKRDWGHAKDYVKMMHLMLQAPIPQDYVIATGKTTSVRDFVKMSFEFIGINLEFQN
TGIKEIGLIKSVDEKRANALKLNLSHLKKGQIVVRIDERYFRPTEVDLLLGDPTKAEKELDWVREYDLKE
LVKDMLEYDLKECQKNLYLQDGGYILRNFYE

>NP_207326.1
MLGKKNEEVLIDENLVGGVIALDRLAKLNKANRTFKRAFYLSMALNVAAVTSIVMMMPLKKTDIFVYGID
RYTGEFKIVKRSDARQIVNSEAVVDSATSKFVSLLFGYSKNSLRDRKDQLMQYCDVSFQTQAMRMFNENI
RQFVDKVRAEAIISSNIQREKVKNSPLTRLTFFITIKITPDTMENYEYITKKQVTIYYDFARGNSSQENL
IINPFGFKVFDIQITDLQNEQTVSEILRKIREVESKNKALNK

>NP_206910.1
MKDEHNQEHDHLSQKEPEFCEKACKEQQYEEKQEAGEKEGEIKEDFELKYKEMHEKYLRVHADFENVKKR
LERDKSMALEYAYEKIALDLLPVIDALLGAHKSAAEEDKESALTKGLELTMEKLHEVLARHGIEGIECLE
EFDPHFHNAIMQVKSEEKENGKIVQVLQQGYKYKGRVLRPAMVSIAKND

>NP_207134.1
MVGGGTVKKDLKKAIQYYVKACELNEMFGCLSLVSNSQINKQKLFQYLSKACELNSGNGCRFLGDFYENG
KYVKKDLRKAAQYYSKACGLNDQDGCLILGYKQYAGKGVVKNEKQAVKTFEKACRLGSEDACGILNNY

>NP_207706.1
MRQEKYFLTSSLSLLSFLLCPAEAFDYRFSGRVENFSKIGFNNSQINTKKGIYPTESFIDIVTLAQVKVN
LLPKGTENHRLSVSLGGAIAAIPYDKTKYDINQANGKIFGSIVENFIGGYHGYFFNKYLGPAYAGTSQSA
SYHARPYVVDTAFLRYDYKDVFGFKAGRYEANIDFMSGSNQGWEVYYQPYKTETQRLRFWWWSSFGRGLA
FNSWIYEFFATVPYLKKGGNPNNSNDFINYGWHGITTTYSYKGLDAQFFYYFAPKTYNAPGFKLVYDTNR
NFQNVGFRSQSMIMTTFPLYYRGWYNPETNTYSLEDSTPHGSLLGRNGVTLNIRQVFWWDNFNWSIGFYN
TFGNSDAFLGSHTMPRGNNTSYIGSEISITTRHAGMIGYDFWDNTAYDGLADAITNANTFTFYTSVGGIH
KRFAWHVFGRVSHANKNALGQVGRANEYSLQFNASYAFTESILLNFRITYYGARINKGYQAGYFGAPKFN
NPDGDFSANYQDRSYMMTNLTLKF

>YP_123028.1
MAKLYFYYAAMNAGKSTVLLQSSYNYRERGMQTLLFTPAIDTRFQYGTICSRIGLSEQAYAFNNSDNLYV
LTQEFQLQTQKYSCVLIDEAQFLTREQVYQLTEITDQMSIPVLAYGLRTDFRGELFPGSQFLLAWADELI
ELKTICHCGRKAIMNMRIDENGQAVIEGEQVLIGGNESYVATCRLHYKRGEAGKTFPRNKLFNKDTNTF

>YP_123067.1
MNKRPIQNTAVIAAALGAALASIYTYTDWLSSDEITVKRERCYNVARAGKNDCATSQHSCAAQSTADRDP
EAFIMLPKGLCERIVGGRSG

>YP_123605.1
MLNVTGLAFADAGDSLFKNEKLTYDGHLFFNAANSAIQGEKYLLNQQLSNIKMGSELQVTDWNKMKGLLI
YNTLPTPVAPQFYFEQFYDELQIEPSNIFLAFGKKWLTFGNYKSDLIYKPLTKALGQTNEVTAIIGYDSL
YYANASFFKPYSRINTSSLPLYYNLNTGVHNQSMDAGVSYLYSLAESQLFQYNKGFGGLLSQSLKTRVPG
FATYFNLKYKKTSTYLTYVTAITPFALEDMSYNNRRASPKALSIQNSYDANIKKFSFKIIGFYDYTFQAL
ALGIPKQRLGLGLSATPFPYLGIQFQYSRDYSYGNNAIASGINHFVNGRNTKTNNLALQTILNF

>YP_001600056.1
MTRLTRAFAAALIGLCCTAGAHADTFQKIGFINTERIYLESKQARKIQKTLDSEFSARQDELQKLQREGL
DLERQLAEGKLRNAKKAQAEEKWRGLVAAFRKKQAQFEEDYNLRRNEEFASLQQNANRVIVKIAKQEGYD
VILQNVIYVNTQYDVTDSVIKEMNAR

>YP_001600126.1
MHNRSPYLINFGLSCLIGLEKAAAAPRDVPRDKKCSAF

>YP_001598536.1
MAQTYNYAIVMNERNQLDVRRDGQYQKSTLKDKDRERKFIYTSQRNNLGQQNNFISFDNTDTLVSQQRGT
AVFGTATYLPPYGKVSGFDADGLEKRNNAVDWIRTTRIALAGYRYDNVVCRSTTDCPKLVYKTRFSFDNP
DLAKTGVGLDKHTEPSRDNSPIYKLKDHPWLGVSFNLGAEGTAKNGKTINKLVSSFNEKNSNNNLVYTTE
GRDISLGDWQRETTAMAYYLNAKLHLLDKKGIKDIAQGKTVDLGTLRPRVETTGRSWLNFWAKWDIKDNG
QIPVKLGLPEVKAGRCINKANPNPKAQALSPALTAPALWFGAGQDGKAEMYSASVSTYPDSSSSRIFLQN
LKRKTDTSRPGRYSLADLSASDIQSKEPTFTSRQTIIRLDGGVREIKLDRSNEATGLNGNDGKNETFGIV
SEGSFMPDDSEWKKVLLPWTVRGSADDGRFKSINQESSKYSQRYRIRENGNNSKRDLGDIVNSPIVAVGE
YLATSANDGMVHIFKQSGGDKRSYNLKLSYIPGTMPRKDIENKDSTLAKELRAFAEKGYVGDRYGVDGGF
VLRRITDDQDKQKHFFMFGAMGLGGRGAYALDLTKAENGDPTAVSLFDVKHDNNGKNSNNSVQLGYTVGT
PQIGKTHDGKYAAFLASGYATKQIDSGENTTALYVYDLENNNGTLIRKIEVTSGKGGLSSPTLVDKDLDG
IVDIAYAGDRGGNMYRFDLSGQDPNQWSVRTIFSGNKPITSAPAISQLKDKRVVIFGTGSDLSEEDVDSK
EIQHVYGIFDNDTDTGTAQDGQGKGLLEQKLEKDKDGKTLFLSDYKRSDGSGDKGWVVKLEAGQRVTVKP
TVVLRTAFVTIHKYTGNDKCGAETAILGINTADGGKLTKKSARPIVPADNTAVAQYSGHKKTSSGKSIPI
GCMEKDNGIVCPNGYVYDKPVNVRYLDEKKTDGFSTTADGDAGGSGTFKEGKKPARNNRCFSRKGVRTLL
MNDLDSLDITGPMCGMKRISWREVFY

>YP_001598607.1
MCAPARMRSMIYGSMPSEKLTIFQTAFVMQVNIQIPCMLYRRGSVKPPLFEAPRLLPSFTDPVVPKLSAP
GGYIVDIPKGNLKTEIEKLAKQPEYAYLKQLQVAKNVNWNQVQLAYDKWDYKQEGLTRAGAAIIALAVTV
VTAGAGVGAALGLNGAAAAAADAAFASLASQASVSLINNKGDVGKTLKELGRSSTVKNLVVAAATAGVAD
KIGASALNNVSDKQWINNLTVNLANAGSAALINTAVNGGSLKDNLEANILAALVNTAHGEAASKIKQLDQ
HYIVHKIAHAIAGCAAAAANKGKCQDGAIGAAVGEIVGEALVKNTDFSDMTPEQLDLEVKKITAYAKLAA
GTVAGVTGGDVNTAAQTAQNAVENNAVKAVVTAAKVVYKVARKGLKNGKINVRDLKQTLKDEGYNLADNL
TTLFDETLDWNDAKAVIDIVVGTELNRANKEEAAQKVKEVLEKNRPYIPNKGAVPNMSTYMKNNPFGKQL
AQISEKTTLPTQQGQSVFLVKRNQGLLKTGDRFYLDGQHKNHLEVFDKNGNFKFVLNMDGSLNQMKTGAA
KGRKLNLK

>YP_001599887.1
MPLQSDSRQPIQIEADQGSLNQANQSTTFSGNVVIRQGTLNISAARVNVTRNDKGEQFMKADGSPVRFSQ
TLDGGKGTVRGQANNVAYSSAGSTVVLTGNAKVQRGGDVAEGAVITYNTKTEVYTISGSTKSGAKSASKS

GRVSVIIQPSSTQKSE

>YP_001248137.1
MMNFQNQGNFTRGSQLFAHKLRMFGQGSANVFTIGLGLSIFWIICRLYQKVFLSSLYYFAIERYVQLKLT
IGEHFYDIDQIGIVFYSLRFKKWIHLNAQDFLHEFYTGQHGFKIQQLWEFLINSALLEGLIVFAIGVIIS
IVFFTAQGKKTIIKAKIRGADFVRSRNLAKMLKSAKKASKICFGGLPLVKNSERLHILITGTTGTGKTNM
LNELLPQIRSQQDRAIIIDLTGAFTDRFFDSKCDKLLNPLEKNSEQWLPWNDCFEAADFHDIASSFSNYT
HRIDDFFAKNAELVLSEALKLYQDDKDIIKLIHTIIYSDNRQFVKAFRNTAVSGIISESAPETSAGIQST
LGKNITSLQYLKPGGNFSIKEWFSNSAETGWLFITATPSQRATLQPLISAWISIAIKALMCRNPNHDNKN
IWFILDELPALQKVSSLPVALAESRKYGGCFVAGLQNIHQLEAIYGSAECASMLDLFNSKFIFRVSDQVT
AYKSALTLGEEEIIETQENLSYGSNTMRDGVNMNNVERKKLLVMPSEIMNLPDLTCYVKLVGNFPITKLT
MKLQNLNIAYVWGYTLLKKLKLVNY

>NP_246737.1
MCKRLANIDFHSIINDQCTGLNKELIMLKSLSLISLITLLSACSMSSYVPFMNDKKAVIDLDKTQIDQKS
YATAYEATLATYKGRVNQDYDVHSFSSGANDWYLNRILLPIDKIKENLYQGGHDSNIHAYYSGVVFASAL
QSNFNKLNPNCWSYLDAPSVTQGIYDAMKDLQKGKQRAEDDPYIVQGSEQLLKRCAQ

>YP_004510868.1
MKLSFLRQINACICLDRYEGVEGFLCNGLIVMKPIRRREIIRSSEKDK

>YP_004510562.1
MGAKKNKNNTMKKALLCFACGALCLATITGCATLFGKQTHAVEMVSNPQGAEVYINGDKMGQTPLQLSLK
ADKSYTVEFRIPGKQPIIRIINNRVEAKWVVLDVLGGLIPVAIDAITGNWYGLDQNKLNVDFTMQEPTK

>YP_007708566.1
MSDPLAPRPPLTGLILSGGGARAAYQVGVLAAIADLLPDSAGNPFPVIVGTSAGAINAVGLACGALQFRQ
AIQRLTAVWQGFRTEHVYRSDWPGVTRQASRFLGHSLLGLGRRVPVALLDNRPLHELLARELDFSGIHAA
VRQRQLRAVAVTAFGYESGQAVTFYQGRATIDPWVRHRRLGIPTRLELRHLLASAAIPLLFPPVRINREY
FGDGAVRSQAPISPALHLGATRVLVIGVSGNPATGGDSAMTGLPPSDRPPSLAQIGAHLLNSTFIDSLEG
DIEQLQRMNQIGQLVAPEVRRQAGGLEPVEVLLVSPSRPLDEIAARHYRELSRVLRLFLRGPGATRASGA
GVLSYLLFERGYCSELIELGYHDAMGKRAELERFLRIER

>YP_007711124.1
MLIGFALPILAAVGIVVLHEAHTSRLQARELARYAATLDYEVRQGPSEAILFPADGPFDRRLGYQLLPSF
MQRLYDRDYAITRQAHFSPALMRYAEHRLFPPYAEKAQAGLDIADCRGVPIYDFRYPQRRYANFASLPPL
AVQSLLFIENRDLLDHERPYLNPAVDWGRFTQAALSQIGKMLGFSAHSSGGSTLATQIEKYRHSAEGRTG
SIGDKLRQMLSASVRSYREGPANFAARQDIVLTYLNSVPLSAAPGYGEVTGLADGLWVWYGADYRQVGEA
LDGKAGLAAQGLALRQVLALMIAHRRPSFYLAPRGRDELDRMTDSHLRVLTQAGVIEAPLRDAALAQKLA
FRDPRSQPTVQPLPTNKGVTLARTRLAGMLGVPLYDLDRLDLRANTTLQHELQESVTAYLQKLADPDYAA
QLGLIGERLLTPTSTRSVRYSFTLFERTPSGNQVRVQTDNTDQPFDINEGSKLELGSTAKLRVLASYLET
VAQIHRDYGGMSVAELRKVEVEPLDFILRWGIDHLVASRDRDLSAMLQAAMERRYSASPYESFFTGGGLH
TFNNFRKEDNGRRPMLLEALRESINLPFVRLLRDLIRHDIYQNAGSKVQLLADDKDPRRADYLDRFVDKE
SQVYLRRFWVKYRDKDANQRLETFLDGLRPWPVRLAAIHRYLQPQADLASFSAFLRERLPRGSLTDKRAA
ELYERYGPGKFNLNDQGYVARVHPLELWLLGYLQKQPQATFGEAVAASGAERKEVYGWLFKSRHKNARDK
RIRILLEVEAFLDLHQQWKKLGYPFDHLVPSYATALGSSGDRPAALAELMGIIVNDGVRMPTLRLEHLDF
ALGTPYETRFAPEATLGQRVMTSEVATTLRNALSQVVDAGTARRLQGTFSRPDGAPLVMGGKTGTGDNRL
QTFSAGGHVRSSKALNRTATFVFYIGPRHFGTLTAYVDGSESSQFKFTSALPVQVLKGMAPLLRDYLDPR
TATRCQVPLSPQQIGRL

>YP_007710813.1
MLSALWVSKTGLSAQDMNLTTISNNLANVSTTGFKRDRAEFQDLLYQIRRQPGGQSTQDSELPSGLQLGT
GVRVVGTQKIFTPGSLQTTEQPLDMAVNGRGFFQVLLPDGTVSYTRDGSFHLNSDGQIVTSNGFALEPAI
VVPNETQTFTVGQDGTVSVTTTGNAQPQVIGNIQTADFINPAGLQAIGNNLFLETGSSGAPQVGTPGLNG
LGTVAQNTLENSNVNVVEELVNMITTQRAYEMNSKVISTADQMLSFVTQNL

>YP_001494432.1
MTKKIALLLLPFILISCNGLGPTRVKNIVELTPKLAIQTHEPIYLDSNANIYAFNANMLKNKQYSFTRSK
TITEPVFIGDMIYALDIRSNISAFSIEKNKIIWSYNLSRHKKDNYIGGGILHHNGKLYVTYGSRLLVVLD
AKSGYEIIRKELPDIIRIKPIVLNDNTVLVQTISNQTIALNAETLKTVWEHESLAEVLSASYFMTPIVQH
DNVIVTYNSGQILALNITNGEVKWNFEFTNLNDRTAIPNFDESSILCTPVHDNMNLYIATGLGKLIKLNV
ATGSVIWQVNAEDIQSMSLIGNSLFVTNNARQIAAFNPETGKVKFVADLNDGQDPKKLKSAAFLVPFVGV
NNNNKRSLNVISVNGVLYSFDVDNNGLNMIPHVVKIIKNIRYYGLRANNNLYFSTDRKIIFGSK

>YP_001494827.1
MWLAKICCGLKYEVEGLEKLPKTIAIVVSNHQSFWDQMFMQLIIPKHSWVLKRELFNIPLLGWGLRMVKP
IAVDRGTNSSVAQILREGQEKIKEGLWLIIFPESTKVPPDRTVKFKPSAVKLASITKVPIVMIAHNAGLF
WPRGFWFKQPGTIKIKIIGVIEKEEIEQTDVRILNDKVEQIINSEKQKLLN

>YP_002144929.1
MALLQVFLLSFCYRLATQLLPKSYLFAIKWGRGYYCEDSLIFPIIRRRVIQNINWHEREENRIISMSETA
TLLKLKGQ

>YP_002144974.1
MAKEVAKENDAKLHPVAAFITNSINVVMTGKGGVGKSYVAQTLSQYFLHYKRIATSTSDSDPVNASTFRI
KALNPEFIKIMENNTILQSNFDQVIESFVANEEITFVLDTGASTYVPLMQYFYDNDLVDFFSSLGRPVFL
HTIIMAGQELPDTLNGFKSLCEMVKGTNVKVVAWINELKGTPIVDVKGVSTPLIETPFFENFTDSLGGVV
VIEDRKSDAFTADIKALTSKNLTLEEAKQSSEFNIMQKTRLNKVYKAVYDQLDEIYTA

>YP_002144968.1
MNSKGDFMNIVTKTTQNYAKARQLILDSDFCDENKQPFIWVCVDDDSSEPMFSLFANDDGSFSYKGNIWL
SDATREEIPAFIRDEKHLRSVLAFVAQDMKPQIAECFS

>YP_002144944.1
MSQRKDKFVTDWFYVPCASQKKLQRTGFLGLGDPKIVKDYVNNVPDLDDYAAKLAESYNKLDSAGYDVIN
IVPINIGSSDQCFQDNKNYVGDVGYSITRGAVVVGKLREVEG

>YP_002144942.1
MKNKLKYKLLHIRLLDFLLSCTVILASCYYSIASLFGVFNPIMWLSSFLIDSLIGKKGSFPQSIHEYSSW
WDRLEFSFPEIMQFFMAGLFLCVIVYATFHATVNIAGYIAELLERNYIKYIFGARFLRLYDKMQKRKGKI
ITRQNKKKCEKDDLNDATFEHYTKWKTFYKSDLSFDEWKNKVLNINSKS

>YP_001197594.1
MCYNKNMNTVVLKASAQQKQAMMAHYDRYRQTSKNPYIEAFFKLTGASLSIYTSGKVVFQGEMAEQEASR
WGYEPESSGQATIPGQNLPMIGTDEVGNGSYFGGLAVVAGFVRPEDHSFLKSLGVDDSKKMTDQKICQIA
PLLKEKIPHQALLLSPKKYNEVIEQGYNAVSVKVALHNQAIFLLLQKGVQPEKIVIDAFTSSQNYQKYLK
KEANQFANPVTLEEKAEGRYLAVAVSSIIARSMFLENLVQLGQLVGMHLPSGAGSKSDQVAASILKQYGM
AGLNQTAKLHFANTQKAQKLLK

>YP_001933695.1
MRRLLACSAGVLCFSQLGALELFLSPKIGITSVYQFGSNGGSDGTSSGKGVSFDRLIGRVDLGLILVNGL
TISASAESSLTNVFVRAQALIGYAVRVGGLRAIVSSGVNICGDSCATSEGKSSAWYSKLLYSVPLNLEVQ
YYLTSFAGVAVAASTAVGVRDFNFKEFTLPLSLTIGPTFRV

>YP_001933316.1
MAAICPSSTAKAGGGGIPVDPLNAVIVEGNVVPSASARAPEAAVCAFCIQTQRRVQGEGRVHTEVSYFEV
EAWDALARVCAQQVRPGVGLRVVGRLKQDRWQQEDGVRVQRVKIVAEHVEFQTPFVW

>YP_001933301.1
MGSGIFVADIGTSSLKAAIISQDGKVLQYQRVFFPQPVKAQDWVRSFFTVFERLRAVHHVIAITISGNGP
SVVAVHKKSHAEDQLILWNQAGASDPRCGVSLFLPKVLLLLQRLHFCARDVQFFLSSHEYLIYRLTGCAV
TVLPERRYMPTYWTSESLRACALPETLFAPFVAPGSIVASYRGIPVVCGAPDFAAALIGTNTLHAGSGCD
RAGSSEGLNVCVRLPPAIRSADARVRVLPSLRADFWNVSFLIADSGSRFASYMRRTHAQGFGERMGQIMA
LPFQLHDAYPPTVVGEGRQLVEDLAFEVCAGLEYLESVTQLQPVYTVSGGQAKDSRWLQLKADVSGRCFV
LPEIHDAELTGNAALACVALDFDADMQTAAQRLCRLKREFIPNRARHEQYAQKRLLRAAREAQG

>YP_001933610.1
MNTQQQRQLLQPRRRTYVSDIKRGIYDVKDAVTYTYSTGKGLNAQVVEKISRRKREPQWMLDLRLRSLRY
FMKRPMPEWGADISDLDIQEIVHYIVSDFKPIAESWDDVPEEIKKTFDRLGIPEAERRSLAGVGAQYDSE
VVYHNLRADLEQQGVVYLDMESAVHKHEDIVRAHFMHLIKPNEHKFAALHGAVWSGGSFVYVPKGVQVDL
PLQSYFRLNANQSGQFEHTLIIVDEGASLHFIEGCSAPKYYKNALHAGAVELYVKKNARLRYSTIENWSR
NLYNLNTKRAIVDEDGVIEWISGSFGSRVTMLYPMSILRGDRSRSEFTGITFASAGQYLDTGTKTVHLGR
NTVSEVHARSISKNGGTANYRGLLSIGPKADGAKAVAECESLMLDNQSHTDTIPIIDVRTDNVDIGHEAK
IGRISDRVVFYLMQRGLDEQTAISLIVRGFVEPVSKELPLEYAVELNNLISIELEGAIG

>YP_001607072.1
MCGLLLFQAYPVNAVSKTANITVTATLLPTCLARFLVSGSTSFGTLDFGSTVALTQPISVAGQTNSGAIT
VQCSNGTSFNVLLSSGQSGNTNNRYLSGGPSAQQVSYNLYTNATYSVIWDDVVGVSQVATGQVVTIPVYG
LVPAQSTPAVGTYTDTVQVTVSW

>YP_001607100.1
MVFGAAISGQVNASCTLHLVSNSESVAGVMPLTFPISSFTFTIDADAPNDSTVPILEKVAQPQGLAVIYY
CTSVDRYGKNVGPVLGQDLGNGLFATNIDGIAIKPAWNNGAAYGYFNSSGIMPAFEREGVPTEEGFWTYP
ATSHFRFELYKIKDTLNLTDTNGERVLPGGTIAYTWATNNSLANYAQRLEIGEIKVISTPSCTFDGPQKV
DFGIVTSSNLNNGGIERDLDFNITCKTDYGHYSATAAIFTQTSSADNNYIKVKDSQNQEDRLLIKISDTN
GQQMKVNGSTTEQQVNIASGVPAEFKWKAKLEAAPAANKPAIGNFSAMAEIILQIK

>YP_001606551.1
MKMTRLYPLALGGLLLPAIANAQTSQQDESTLVVTASKQSSRSASANNVSSTVVSAPELSDAGVTASDKL
PRVLPGLNIENSGNMLFSTISLRGVSSAQDFYNPAVTLYVDGVPQLSTNTIQALTDVQSVELLRGPQGTL
YGKSAQGGIINIVTQQPDSTPRGYIEGGVSSRDSYRSKFNLSGPIQDGLLYGSVTLLRQVDDGDMINPAT
GSDDLGGTRASIGNVKLRLAPDDQPWEMGFAASRECTRATQDAYVGWNDIKGRKLSISDGSPDPYMRRCT
DSQTLSGKYTTDDWVFNLISAWQQQHYSRTFPSGSLIVNMPQRWNQDVQELRAATLGDARTVDMVFGLYR
QNTREKLNSAYDMPTMPYLSSTGYTTAETLAAYSDLTWHLTDRFDIGGGVRFSHDKSSTQYHGSMLGNPF
GDQGKSNDDQVLGQLSAGYMLTDDWRVYTRVAQGYKPSGYNIVPTAGLDAKPFVAEKSINYELGTRYETA
DVTLQAATFYTHTKDMQLYSGPVGMQTLSNAGKADATGVELEAKWRFAPGWSWDINGNVIRSEFTNDSEL
YHGNRVPFVPRYGAGSSVNGVIDTRYGALMPRLAVNLVGPHYFDGDNQLRQGTYATLDSSLGWQATERMN
ISVYVDNLFDRRYRTYGYMNGSSAVAQVNMGRTVGINTRIDFF

>YP_001605456.1
MFENLTDRLSRTLRNISGRGRLTEENIKETLREVRMALLEADVALPVVRDFINRVKERAVGHEVNKSLTP
GQEFVKIVKNELISAMGEVNNELNLAAQPPAVVLMAGLQGAGKTTSVAKLGKFLKEKQKKKVLVVSADVY
RPAAIKQLETLAQGVGIDFFPSDAQEKPIDIVNRALQQAKLKFYDVLIVDTAGRLHVDEAMMDEIKQVHA
AIKPVETLFVVDAMTGQDAANTAKAFNEALPLTGVVLTKVDGDARGGAALSIRHITGKPIKFLGVGEKSD
ALEPFHPDRIASRILGMGDVLSLIEDIESKVDRAQAEKLATKLKKGDGFDLNDFLDQLKQMRNMGGMASM
LSKMPGAGQLPENVKSQMDDKVTVRMEAIINSMTLKERAKPEIIKGSRKRRIATGSGVQVQDVNRLLKQF
DEMQRMMKKMKNGGLAKMMRGMKGMMPPGFPGR

>YP_001605885.1
MPQVNNISTNNIHSAGFNNSNSIQKYTGAVSSISDDLRINNEKCKSDIGTISGDIKINRHSAVYGNVNSV
SGDITVKNSIVDKDITTVSGDVNAVNSTIGKNIKTVSGSIEVEQSTVSGNLETTSGGIDIDTTKINGNVH
TTSGSISMNDSTIDGSVTCKAGSVTIVNSTIKESLNVTSEKIIVGTASCIGKINISPPESVNFNIMNFGN
DSIVMGMRNFCISGEVNFTITNGKVFVNEQRVGHTASQSTSKKVEEVTINIAKNASVNDIVFYTKKCHII
LEGNAKYNGEKKDGMQFTHVNAPKSHAYA

>YP_220963.1
MQSGRQNTTSSTDALIGLLEAAPILAKLEAREIDALATLDIIENNFAADAIIIEQGARVRSIHLVRSGWG
CIYRDLSSGQRQIIDFPMRCDFVGLRTSNGYSYNTIAAITPMSIFEIPLNSLENAIKHAPRLSFILIELL
SRQRSFLIEHLTNVGCRNAFVRTAHLLLELSDRVKSCGMGEPDSFYCPLTQYQLADALGLTPIHLNRMLR
ELREEGLVLFRSNRVEILNREQLAALAEYDGEFMRMAVFARHE

>YP_221849.1
MPLLPVDEALAYILNSAAPHGTEDVSLTDAGGRVTASDITAQLLQPPFDCSAMDGYALIAPEEITYPLEL
TVIGESAAGKRFEGTLRKGEAIRIFTGAPMPENADSIIIQEHTEREGNRLIILHGLDKGRHIRRAGLDFA
PGKVVVPAGRELDAPALSLAAASGHARLPVFTRPRVAILATGDELVPPGAIPGPDQIVASNSIGIAEITR
RAGGSPEDLGIIADDPARIEKAISDALEDGIDMLVTIGGASVGDRDFVHGALRNCGVALDFWKIAMRPGK
PLMYGRKAVNGKTVHVLGLPGNPVSSLVCSLLFLRPLVAKLSGLALKADIRAAKLGVAMRANDHRRDFIR
VTVESEPDGTLVATPFPMQDSSMLSALVCSDALLIREENAPEALPGDPCRILML

>YP_221497.1
MHRYRSHTCAALRKTDVGSNVRLSGWVHRVRDHGGILFIDLRDHYGITQIVADPDSPAFKVAETVRGEWV
IRVDGEVKARADDAVNTNLPTGEVEIFATEIEVLSPAKELPLPVFGEPDYPEDIRLKYRFLDLRRETLHK
NIMSRTKIIAAMRRRMTEIGFNEFSTPILTASSPEGARDFLVPSRIHPGKFYALPQAPQQYKQLLMVAGF
DRYFQIAPCFRDEDPRADRLPGEFYQLDLEMSFVTQEEVWETMEPVMRGIFEEFAEGKPVTKVFRRIAYD
DAIRTYGSDKPDLRNPIEMQAVTDHFAGSGFKVFANMIANDAKVEVWAIPAKTGGSRAFCDRMNSWAQSE
GQPGLGYIFWRKEGDKLEGAGPIAKNIGEERTEAIRKQMGLEDGDACFFVAGLPSKFYKFAGDARTRAGE
ELNLVDRDRFELAWIIDFPFYEWDEDNKKIDFAHNPFSLPQGGMDALENMDPLEIKAYQYDLVCNGFEIA
SGSIRNQLPEVMVKAFEKVGLSQQDVEERFGGLYRAFQYGAPPHGGMAAGIDRVIMLLVGAKNLREISLF
PMNQQALDLLMGAPSEVSPAQLRDLHVRLAPVQKS

>YP_001065842.1
MTRNARRTWQWISGLLAICAAALAALPARADVAAGPGAWSSQQTWGADTVNGGNLTGYFYWPATQPTTPN
GKRALVLVLHGCLQSASGDVIDNSRGAGFNWKSVADRYGAIILAPNATGNVYGNHCWDYANTSPNRASGH
VGVLLDLVSRFVGDAQYAIDPNQVYVAGLSSGGGMTMVLGCIAPDVFAGIGINAGPPPGTTTLQIGYVPS
GYTAQTAANQCKAWAGSKADQFSTQIAGAVWGTSDYTVAQAYGPLDTAAFRLIYGGTFAQGSKVSIAGGG
TNTPYSDSNGKVRTHEIVVSGMAHAWPAGAGGDNANYVDATHVNYPAFVMDYWVKNNLRANGAPVQAGTP
PTGLTVTNATQTSISLAWNPVANASSYNVYRNGNKVGSSTSTAYTDAGLIAGTAYSYTVTEIDPSLGESA
PSPAVSATTQPSFACSATTATNYAHVQAGRAHDSLGIAYATGSNQNMGLDNVFYTNTLAQTSAGYYIIGD
CP

>YP_001066635.1
MTTLRLRSPHRCRPPALAAAATLAALSGPAHAQSTLTLYGVADAGVQYLSRADGRHAAWRLQNYGILPSQ
LGIKGEEDLGGGWRARFQLEQGINLNDGTATVPGYAFFRGAYVGMGGPAGTVTLGRQFSTLFDKTLFYDP
LWYASYSGQGVLVPLSANFVDHSIKFQSATFAGFDVEALAAMAGIAGNTRAGRVLELGGQFTSRGLSASA
VLHRSHGTAQGGADRSAQRRDIGTFAARYAFASLPLTVHAGVQRLTGELDPARTIVWGGARYQASGRFGF
AGGIYHTDSPTPQVGHPTLFIASTTCSLSKRTVAYLNLGYAKNSGRSAQTVYEYDPTPLAGASQFGAMLG
MYHVF

>YP_179430.1
MINLLFGNAKLYIALALMAILTGYFYLRLDSTQAKLEKSQSDLALALKINENNQEKLKELNQIHKTELKA
LNEANNQKNQVQERVQYVKEYIYKSNENNITKLFNDVVDRLWDANSTSSNQNRNSKSKNSARTTNIKSS

>YP_178110.1
MKKTKILGTALIGALLFSGCAQIAYTDGKASQIKKGDALTLGLDRQDFESAAETMINSMLSDPAFANIKP
GTRKVIAIGRVVNDTPQRIDTEKLTAKITSALRKSGKFVLTSAVAAGGALDSMSEDVRELRDNDEFNQKT
IAKKGTLVSPDFSLAGKIRQDNVKLSNGKTQVEYFFLLRLTDLTSGLVYWEDEQTIDKTGSSKSVTW

>YP_001654124.1
MGIRLVIDKGPLSGTVLILENGTSWSLGSDGKASDILLQDEKLAPSQIRITLKDGEYYLENLDASRPVSV
DGTVITAPVLLKDGVSFVMGSCQVSFFKGEEVEGDIELSFQTEGGNEGEPAAQGSSSVSSEAPKKETGNP
SLPSETKASGEVSSSAIAKEQELAASFLASVEKEPGTPKEVSEPKVSSQEGQTPSVTGEKKDLELPLASQ
EQPKQTTPSGSGEPTQSQNASMEENRTSPDQNQQPQLSSASESGSQSPENQEQQPSQTPPPSPETPEPSG
EPNSATEENSPSPMEKASVTEEGSSGTSEEEKEGEEDTAESAANEELKAEASQEEEKKEEDKGEVLAPFN
VQDLFRFDQGIFPAEIEDLAQKQVAVDLTQPSRFLLKVLAGANIGAEFHLDSGKTYIVGSDPQVADIVLS
DMSISRQHAKIIIGNDNSVLIEDLGSKNGVIVEGRKIEHQSTLSANQVVALGTTLFLLVDYTAPSDTVMA
TISSEDYGLFGRPQSPEEIAARAAEEEEEKRKRATLPTGAFILTLFIGGLALLFGIGTTSLFHTKEVVSI
DQIDLIHDIEHVIQQFPTVRFTFNKNNGQLFLIGHVRNSIDKSELLYKVDALSFVKSVDDNVIDDEAVWQ
EMNILLSKNPEFKGISMQSPEPGIFVISGYLKTEEQAACLADYLNLHFNYLSLLDNKVIIESQVMKALAG
HLVQSGFANVHVSFTNGEAVLTGYINNKDADKFRTVVQELQDIAGIRAVKNFVVLLPAEEGVIDLNMRYP
GRYRVTGFSKCGDISINVVVNGRILTRGDILDGMTVTSIQPHCIFLEREGLKYKIEYNK

>YP_001654964.1
MKTLIDNNIVRFKNISKTKQGIFVNFQVKGERGGASFTASIAVDIDAADVSAGDSLETIIERCALIGIRE
FQKCEFQFEGIICL

>YP_001654127.1
MADLDVFKEDFALLFEAGMVAIKQGDEASAKALFQALQVLDPEHTAHELGSGLLHLHKMELTKAEVLFRA
IVEKDPENWSAKAFLSLTLMMIVLQQGSSFEVRRESLERCLQLADQVLESCEVESTRALAKSVLDWHDGL
VAKSGGPLN

>YP_001654874.1
MKNILSWMLMFAVALPIVGCDNGGGSQTSATEKSMVEDSALTDNQKLSRTFGHLLARQLSRTEDFSLDLV
EVIKGMQSEIDGQSAPLTDTEYEKQMAEVQKASFEAKCSENLASAEEFLKENKEKAGVIELEPNKLQYRV
VKEGTGRVLSGKPTALLHYTGSFIDGKVFDSSEKNKEPILLPLTKVIPGFSQGMQGMKEGEVRVLYIHPD
LAYGTAGQLPPNSLLIFEVKLIEANDDNVSVTE

>YP_006095491.1
MNKSMLAGIGIGVAAALGVAAVASLNVFERGPQYAQVVSATPIKETVKTPRQECRNVTVTHRRPVQDENR
ITGSVLGAVAGGVIGHQFGGGRGKDVATVVGALGGGYAGNQIQGSLQESDTYTTTQQRCKTVYDKSEKML
GYDVTYKIGDQQGKIRMDRDPGTQIPLDSNGQLILNNKV

>YP_006094393.1
MKKRIPTLLATMIATALYSQQGLAADLASQCMLGVPSYDRPLVQGDTNDLPVTINADHAKGDYPDDAVFT
GSVDIMQGNSRLQADEVQLHQKEAPGQPEPVRTVDALGNVHYDDNQVILKGPKGWANLNTKDTNVWEGDY
QMVGRQGRGKADLMKQRGENRYTILDNGSFTSCLPGSDTWSVVGSEIIHDREEQVAEIWNARFKVGPVPI
FYSPYLQLPVGDKRRSGFLIPNAKYTTTNYFEFYLPYYWNIAPNMDATITPHYMHRRGNIMWENEFRYLS
QAGAGLMELDYLPSDKVYEDEHPNDDSSRRWLFYWQHSGVMDQVWRFNVDYTKVSDPSYFNDFDNKYGSS
TDGYATQKFSVGYAVQNFNATVSTKQFQVFSEQNTSSYSAEPQLDVNYYQNDVGPFDTRIYGQAVHFVNT
RDDMPEATRVHLEPTINLPLSNNWGSINTEAKLLATHYQQTNLDWYNSRNTTKLDESVNRVMPQFKVDGK
MVFERDMEMLAPGYTQTLEPRAQYLYVPYRDQSDIYNYDSSLLQSDYSGLFRDRTYGGLDRIASANQVTT
GVTSRIYDDAAVERFNISVGQIYYFTESRTGDDNITWENDDKTGSLVWAGDTYWRISERWGLRGGIQYDT
RLDNVATSNSSIEYRRDEDRLVQLNYRYASPEYIQATLPKYYSTAEQYKNGISQVGAVASWPIADRWSIV
GAYYYDTNANKQADSMLGVQYSSCCYAIRVGYERKLNGWDNDKQHAVYDNAIGFNIELRGLSSNYGLGTQ
EMLRSNILPYQNTL

>YP_006097512.1
MSDCHPVLLPEGPFSRKQAMAATTAYRNVLIEDDHGTHFLLVIRNAEGQLRWRCWNFESDAGKQLNSYLA
SEGILRQ

>YP_763859.1
MKKIPLIISLTASMSFGIALAAEAKAPVTQVQDSSESSPIAQGEGKCASGKCGSLKKFGVVDVDADEQDG
KLVRARDGNCGTKVCKAYGKQDKKEIAQLGKCSNGVCGQ

>NP_207182.1
MQKNILKMTLLLVFLFLRNAVGLEDKKATTQPESVQNTPKDLPPIQLRLNQVHEELIEMLENMGKGTQYE
FPKIKEILEQSEEEWLKVAHEECVALVMLISPKASIENSPIYKNCYEAYVKQRIHDLYDFYIESKKVKRK
IKKAHKQETAINQSQPLTKESPKNENKKNLVKPNLKDASIPKGYYLQIGAXLNAPSKDFLQTLKTFPYQI
KKKDSLTHYFIGPYKTKEEALKQLENAIKSFKNKPVLVEK

>YP_000389.1
MIQELKADLNGKGQKHCVIVSRFNEFITESLLKGALESFRMHGVEDVTVVRVPGAYEMPVVVSKAAASKK
YDSIVCLGAVIRGATAHFDLVAGESAKIGSIGVQHSIPVIFGVLTTDTIEQAIERAGTKAGNKGAEAAAT
AVEMVNLLSLL

>YP_002830.1
MSNPIQNRYEFVYLFDVKDGNPNGDPDAGNQPRVDPETGNGLITDVSLKRKIRNYVTIVKSATPPNDIYI
KEKAVLIETHEKAYVAVGAKLETSKKEEKEKRTGGDQVGKAREWMCKNFYDVRTFGAVMALKVNAGVVKG
PIQFTFARSIDPVINLEHSITRMAVATKKEAEDQDGDNRTMGRKHTISYGLYRAHGFISAHFANDTGFSE
EDLELFWSSLQNMFDHDRSAARGEMNCRGLYVFKHVGDGKNTNQAKLGVAPAHKLFNLISVSKKDNSTPA
RDFSDYSVKIQESDLPAGVELIKKVS

>YP_000391.1
MDKEIRKNRLFLEGIEEEEEFYFPENKEPVRIRRSYNKLLIFWILMGLLVLGGIGFAVYHQFFRNSSSGSE
FAGAFNKDLIQNKSDINRLLERPYIPDGNANPALTKCINLYKERFTRQAFDTCNEFLDSTGTQEEKSIAL
TVLGVIHDESGRYPQAIERLQKAIQYDPKNFYAYYNLTLSYKHAGRFADARMAALKAKEIAPSDPRVSLL
AGNLFNELNDPDAAIDAYKEGLSTSPDDMHLTYNLGVSYFKKGEIPQAEEEFKKVVIKTPSGRLAALSHS
YLGNIAYNKQDYKNAEYHFRQASNLSPNEAKYLYNLAIVLQKNGNKEEALKYLELARDAGANDPEIYRLI
AEGFSNLNQGEMSISALQKSLKYNPTDVDSLFQLAEAYYNKGDLLSAEETYRRIVSSTPGDSFTETALIN
LGVVLDQMERYGEAVTTLNRVIELNPKNAKAYHTLGIVYKHSGNGTLAIENWRKSTAIEPENIQSREALG
DYLLENKFFREAVEEYIGLVKHKDDAYKVYLKMAEAYMGMQDDSNAEKILLKVLNSSRDGADLKNAHKKL
ALLYNKSKDPDLKNRAKDEAFRSAHMDPNDMEGRLVLAKILIDSNSILDREKAIDELTAIVRSDVRPKTA
ATAYNYLGICYYKNGEFKRAVRAFQSSIDLDPSLSEAYENKRAASAALEENTRREGYF

>YP_002979.1
MGDCKRFSSAKQAAYYAGLVPRVDISGDTVRYGRIINRGCHSIRRVIVQAAWSLVRCQHGGKVKEFYQRL
YLKKGAKKSIIATSRKMIEVLYVMIRTGKLFDSMPENILNRKLTQYGLM

>YP_003043.1
MEQTHRIISMKINSKFKQINLYRLEINKHINILIYSFKKITRKFKIFPEKFYKMNLKFDDSRKMKLSCKN
LIGQLFFLDFRRIDFKNNLQLKNFFCYQIYLLSLYKFWDEFSNTHRIFSVKKENKFKFFSNFGNFWNIKS
EKAISLKKLKSKIMYTFSINTFINNYKLIFYNINSLLFQKHKVYYKKKPNAKKVLKLQNVFFTFISFFIL
VFTSECTDDSGIRTGDGIVEPRLEEVTGITTEKIKSILPEQELTIDELYAFAVERTERIALKEEAIQQAD
AQKAAAFASFFPSLSLVYNKFYRIPGPNSHFNPFYTEPNPLTGSSSTTSSLPPTVGPGTRLLLSIPILNG
VSQYTTYKASGALTNVRLNEARYESGRLYLEIAQAYYNVLQLKEIILLEEKKTDLVRKTIQERRRLFSLG
KITRADLSGAEADFSRSEAYLEDFKYQLKQAEMALESLIGAGEGGLRLAIPKEVISIPKNLLPEERIAKR
YDVIAAKENLKAAEFNLKKAWGGHLPSVTLNNYYTIPEHNTTPNKDITMQLSINVPLLSAGIVTAGVKQA
ESALRQAELQLSQTKRIATDEIRKAYESSLNSARLLSLYSKARNSAESNLSSQRRGFSFKTVSRLELLVS
ETSFLDSEIAYRKAYYQHSLNTIWYSVAIGELPKLKKLKEEDKTTN

>YP_000859.1
MSRYSARTGNPNLKMEESEHAEIGWDQKIGNFWNIKIEGYQNTFSNLSIADPYAMDPFSRNRDLMRESLD
PNADLSLVRRSNLNYSNSMTGYSRGVEVFIKKEASAESGLYGWISYTKSITKRNRNLPELTKQEYSSWLA
ESSAKDLIHQENTDYYYANFYRDGSYDVLFKNSKEELYDFDRTHMFNMVIGWKFGEKAQIGLRGTYLTNY
AYTPVVGSKSITSEQFNSQLLPSAAPPPPSSSSNSSSSLFSIYQPVYSDMLRSARLPHYHQFDLRFDRFI

PTSWGRMTAYLELVNITGSRIAVSADTFVPIFPFVPGANPETQYIYLNGLQSLRTEKNKIPYLNFGIEFR
F

>YP_001880.1
MMRIKTINLVASVFCLLVSSQLLSQTGKKPAADTVAPAPPPSQTIELEEKKWYDQVEFSGFADVYYMYNL
NPKQGNDIDSTRAFETSNKNFAVNAVALTIQKAAEKSSPWGFRIDIQNGQNNAFQEAPYSQSNSIYNYNM
LKQAYVSLYFPVLKGMTLDVGKMATHIGYEVLESMSNPNYSIGAIFQNTIPFIHTGARLTTQFTDKWAGT
FYVYNSGAGTGYNSPATVTTAPFNVITDPTYSGNAAGKSYFVEGQTERKAIGTQIKGQLIEDKLSVTWNT
LYSSDGAYARVDPSKAAMATELSSALGDPNIAKYNVTNPARAKYNKDYWFMNHAILSITPTDKITIDLDY
TWSEKSGALANNSLDQKRYNVDANTAEVFNKDTLFWGLNNKRDLKTTYKAYGIFAKFKINENWGVNVRAE
YIDDKHNNGALTIFNPFMGPNTYLSEFKKLTDAAIADAVVASLAADPNLGPLGITKEQILEVMSDDYKNY
GGTRNAGQYKTFTVTPVWNFTENLLIKLDLRRDWATGKQFVDQKGDKTDHQYGMTLGIVAKF

>YP_001599782.1
MNTRNMRYILLTGLLPTVSAFGETALQCAALTDNVTRLACYDRIFAAQLPSSAGQEGQESKAVLNLTETV
RSSLDKGEAVIVVEKGGDALPADSAGETADIYTPLSLMYDLDKNDLRGLLGVREHNPMYLMPLWYNNSPN
YAPGSPTRGTTVQEKFGQQKRAETKLQVSFKSKIAEDLFKTRADLWFGYTQRSDWQIYNQGRKSAPFRNT
DYKPEIFLTQPVKADLPFGGRLRMLGAGFVHQSNGQSRPESRSWNRIYAMAGMEWGKLTVIPRVWVRAFD
QSGDKNDNPDIADYMGYGDVKLQYRLNDRQNVYSVLRYNPKTGYGAIEAAYTFPIKGKLKGVVRGFHGYG
ESLIDYNHKQNGIGIGLMFNDLDGI

>YP_001598909.1
MGNTFKSILVWVALGIGLMAAFNALDGKKEDNGQIEYSQFIQQVNNGEVSGVNIEGSVVSGYLIKGERTD
KSTFFTNAPLDDNLIKTLLDKNVRVKVTPEEKPSALAALFYSLLPVLLLIGAWFYFMRMQTGGGKGGAF
SFGKSRARLLDKDANKVTFADVAGCDEAKEEVQEIVDYLKAPNRYQSLGGRVPRGILLAGSPGTGKTLLA
KAIAGEAGVPFFSISGSDFVEMFVGVGASRVRDMFEQAKKNAPCIIFIDEIDAVGRQRGAGLGGGNDERE
QTLNQLLVEMDGFESNQTVIVIAATNRPDVLDPALQRPGRFDRQVVVPLPDIRGREQILNVHSKKVPLDE
SVDLLSLARGTPGFSGADLANLVNEAALFAGRRNKVKVDQSDFEDAKDKIYMGPERRSMVMHEDEKRATA
YHESGHAIVAESLPFTDPVHKVTIMPRGRALGLTWQLPERDRISMYKDQMLSQLSILFGGRIAEDIFVGR
ISTGASNDFERATQMAREMVTRYGMSDKMGVMVYAENEGEVFLGRSVTRSQNISEKTQQDIDAEIRRILD
EQYQVAYKILDENRDKMETMCKALMEWETIDRDQVLEIMAGKQPSPPKDYSHNLRENADAAEDNAPHAPT
REETEAPAPADTASSESEQRSENKA

>YP_001598413.1
MKKNILEFWVGLFVLIGAAAVGFLAFRVAGGAAFGGSDKTYAVYADFGDIGGLKVNAPVKSAGVLVGRVG
AIGLDPKSYQARVRLDLDGKYQFSSDVSAQILTSGLLGEQYIGLQQGGDTENLAAGDTISVTSSAMVLEN
LIGKFMTSFAEKNADGGNAEKAAE

>NP_245935.1
MYLISEFIFQFILYEVCMKNKSKLLACCLMALPISSFSIGNNNLIGVGVSAGNSIYQVKKKTAVEPFLML
DLSFGNFYMRGAAGLSELGYQHVFTPSFSTSLFLSPFDGAPIKRKDLKPGYDSIQDRKTQVAVGLGLDYD
LSDLFNLPNTNISLEMKKGRRGFNSDITLTRTFMLTDKLSISPSFGLSYYSAKYTNYYFGIKKAELNKTK
LKSVYHPKKAYSGHIALNSHYAITDHIGMGLSFSWETYSKAIKKSPIVKRSGEISSALNFYYMF

>YP_007710775.1
MAIMKKTVAPKAAKPAPAVETRESIEAQVAAFLQAGGEIQKIAKGVSGQVYGPSRQITISKKR

>YP_002144969.1
MATTNIDSLSKLRSVAEKGVIKRADSYKVLYSELNIKPGHNVRGIYLSQDEYWNSEEVKDYIKGLAYSYA
HNIKVPALSVIVEDGKIWVNNGEHRYRAIALAKELFGVDIEYVDVVEDDDQLTSNSARNHTPIEMAVLYE
RYNTDKGLSVQEIAERAGKSVPHVYKYLTAVRKWPSDLLAKVQSGELSFTAAQTLYEETRRKKPDAPENV
SGTDLPNDDNGESGTVLPKENDQNSPDNNVQDVTNAAPAASSGGESGTDLPKGGTDKGATKSKKSITKSI
SDSLVNLIFGFEAVETDSGYQLNLTKDQYEAIIALQKEIEAEIKGK

>NP_837945.1
MAITKSTPAPLTGGTLWCVTIALSLATFMQMLDSTISNVAIPTISGFLGASTDEGTWVITSFGVANAIAI
PVTGRLAQRIGELRLFLLSVTFFSLSSLMCSLSTNLDVLIFFRVVQGLMAGPLIPLSQSLLLRNYPPEKR
TFALALWSMTVIIAPICGPILGGYICDNFSWGWIFLINVPMGIIVLTLCLTLLKGRETETSPVKMNLPRL
TLLVLGVGGLQIMLDKGRDLDWFNSSTIIILTVVSVISLISLVIWESTSENPILDLSLFKSRNFTIGIVS
ITCAYLFYSGAIVLMPQLLQKTMGYNAIWAGLAYAPIGIMPLLISPLIGRYGNKIDMRVLVTFSFLMYAV
CYYWRSVTFMPTIDFTGIILPQFFQGFAVACFFLPLTTISFSGLPDNKFANASSMSNFFRTLSGSVGTSL
TMTLWGRRESLHHSQLTATIDQFNPVFNSSSQIMDKYYGSLSGVLNEINNEITQQSLSISANEIFRMAAI
AFILLTVLVWFAKPPFTAKGVG

>NP_836897.1
MGIKQHNGNTKADRLAELKIRSPSIQLIKFGAIGLNAIIFSPLLIAADTGSQYGTNITINDGDRITGDTA
DPSGNLYGVMTPAGNTPGNINLGNDVTVNVNDASGYAKGIIIQGKNSSLTANRLTVDVVGQTSAIGINLI
GDYTHADLGTGSTIKSNDDGIIIGHSSTLTATQFTIENSNGIGLTINDYGTSVDLGSGSKIKTDGSTGVY
IGGLNGNNANGAARFTATDLTIDVQGYSAMGINVQKNSVVDLGTNSTIKTNGDNAHGLWSFGQVSANALT
VDVTGAAANGVEVRGGTTTIGADSHISSAQGGGLVTSSSDATINFSGTAAQRNSIFSGGSYGASAQTATA
VINMQNTDITVDRNGSLALGLWALSGGRITGDSLAITGAAGARGIYAMTNSQIDLTSDLVIDMSTPDQMA
IATQHDDGYAASRINASGRMLINGSVLSKGGLINLDMHPGSVWTGSSLSDNVNGGKLDVAMNNSVWNVTS
NSNLDTLALSHSTVDFASHGSTAGTFTTLNVENLSGNSTFIMRADVVGEGNGVNNRGDLLNISGSSAGNH
VLAIRNQGSEATTGNEVLTVVKTTDGAASFSASSQVELGGYLYDVRKNGTNWELYASGTVPEPTPNPEPT
PAPAQPPIVNPDPTPEPAPTPKPTTTADAGGNYLNVGYLLNYVENRTLMQRMGDLRNQSKDGNIWLRSYG
GSLDSFASGKLSGFDMGYSGIQFGGDKRLSDVMPLYVGLYIDSTHASPDYSGGDGTARSDYMGMYASYMA
QNGFYSDLVIKASRQKNSFHVLDSQNNGVNANGTANGMSISLEAGQRFNLSPTGYGFYIEPQTQLTYSHQ
NEMAMKASNGLNIHLNHYESLLGRASMILGYDITAGNSQLNVYVKTGAIREFSGDTEYLLNDSREKYSFK
GNGWNNGVGVSAQYNKQHTFYLEADYTQGNLFDQKQVNGGYRFSF

>YP_001606647.1
MGFALPNGAGIYLAKTYETEVAVTAVSNAVDAVLTVATGHDIAEGDIVQLTSSWGALNDLAAKVTASTTT
SLTLGSIDTSNTDRFAVGGGVGTVKKIASWIEIPQITEVANSGGDQQMIQIQFLSDTRQRNLNTFKAAQS
QTLTLAHDYSQPVYPVLRAADESEQTLATYMYVPKAKENRYSTVKVSFNDIPTTAINAIETVAVVFNLQS
QAMTFYKAGIAVPVTGVTLNKTTTILAVAATETLTATVAPANATNKSGAWSSSAPTKATVDPVTGVVTGV
AAGSVNIIYTTADGAKTATCAVTVTA

>YP_006095865.1
MDRRRFIKGSMAMAAVCGTSGIASLFSQAAFAADSDIADGQTQRFDFSILQSMAHDLAQTAWRGAPRPLP
DTLATMTPQAYNSIQYDAEKSLWHNVENRQLDAQFFHMGMGFRRRVRMFSVDPATHLAREIHFRPELFKY
NDAGVDTKQLEGQSDLGFAGFRVFKAPELARRDVVSFLGASYFRAVDDTYQYGLSARGLAIDTYTDSKEE
FPDFTAFWFDTVKPGATTFTVYALLDSASITGAYKFTIHCEKSQVIMDVENHLYARKDIKQLGIAPMTSM
FSCGTNERRMCDTIHPQIHDSDRLSMWRGNGEWICRPLNNPQKLQFNAYTDNNPKGFGLLQLDRDFSHYQ
DIMGWYNKRPSLWVEPRNKWGKGTIGLMEIPTTGETLDNIVCFWQPEKAVKAGDEFAFQYRLYWSAQPPV
HCPLARVMATRTGMGGFPEGWAPGEHYPEKWARRFAVDFVGGDLKAAAPKGIEPVITLSSGEAKQIEILY
IEPIEGYRIQFDWYPTSDSTDPVDMRMYLRCQGDAISETWLYQYFPPAPDKRQYVDDRVMS

>YP_002814629.1
MVRAFDIAGHEFTHAVTSSESNLEFFGESGAINEALSDIMGTAIEKYINNGKFNWTIGEQSGSVLRDMKN
PSSVKFFDGVPYPDDYSKYSDLNGEDNEGVHFNSSIINKVAYLIAQGGTHNGVTVNGIGEDKMFDIFYYA
NTDELNMTSNFSELRLACLKVATNKYGANSIEVETVQKAFDAAKIKGPVKEDEKTEVNQVPELTVPLTIT
LRVGDTFDPISNVKAIDKEDGDLTTRVKHKGDVDTSKPGKYIVDYSVIDSQGGNATATQTVIVEGNGETA
DLKPTLTVPVATTITVGDSFDPMAKVKAIDKEDSNLTSKVKIDGEVDTSKAGTYVLTYKVTDSKGNEGTA
KQTVTVKVREEVKNEIPILQVPATTTITKGDKFDLMVGVSATDKEDGDLTSKVMYEGTVDTSRTGTFEIK
YSVRDSVGNEVHTIQEVFVKDKDTGKTNSLANNSNSNKNESTYKELPNTGGQYN

>YP_002813347.1
MSSKKFILKLFICSTIFITFVFALHDKRVVAASSVNELENWSRWMQPIPDNIPLARISIPGTHDSGTFKL

QNPIKQVWGMTQEHDFRYQMEHGARIFDIRGRLTDDNTIVLHHGPLYLYVTLHEFINEAKQFLKDNPSET
IIMSLKKEYEDMKGAEGSFSSTFEKSYFVDPIFLKTEGNIKLGDARGKIVLLKRYSGSNESGGYNNFYWP
DNETFTTTVNQNVNVTVQDKYKVSYDEKVKSIKDTINETMNNSEDLHHLYINFASLSSGGTAWNSPYYYA
SYINPEIANYIKQKSPTRVGWIIQDYINEKWSPLLYQEVIRANKSLIRD

>YP_001088774.1
MSVLKTQEKLKQPRDLYFCCSIFTFVGLAYYGMKLVLLLFLAEQVSKGGLGLTPAESASMLASFLAWTYF
SPVIGGWVSDRFLGPKNCIILGMLMVSIGYFLGYIANSKMDINIMIFLSGLGIGFYKGNLHTMVGNLYTN
DDPRKDGAFSITYMATNLGTLFGPLICGLVANEWFAMKSGLSISMYGYRYVFLFSSITVLIGLFFFIFGN
RKFYKPLNNDPINGRESIRKSLEYKKMLASRPLTLIEKKRIIVILVLAFFTVFFWIAYNQASMSIALYTK
EHINLTVGNFQIPTSWIDSYNGLLCVILGPISALVWVKLSQTKKGDLSVPKKMSLGFILLAIAFLFMIVA
VIQTGTDPNSIHKASVFWVVGFLTFQSIGEICFSPVGYGMVNKLSPEKYISLLMGVWFLGKFAANKLSGY
TQAIIDQLGMLQVFIVIPSFLLIFGIILLIMNKKLVELSN

>YP_696046.1
MNSLQRFRLSKNLSRSEIAKLLGISESYYTKIELGIRNPSYNFLKKFKSKFRCTLDEIFFAN

>YP_695964.1
MQDKNPLSTFGPDLNEFSRDVNFLTLAKNSDFIYLRASGSGTGKLRIDNKFLEFAKECRRLGIPCGAYHF
AKPSKDLDSAVIQADQFIDVLQQGFGDGDYGDLFPVLDVETPTDKSLTTTELVNWIDRFRDRFEEKTRRR
LMLYTGLFFIGLYDDFKVPGKGYPLSDMPLWIAMYTRIPSNPRIPPNVGGWKRWTMWQFTDEGKLDGVGS
PVDLNWGPNSIDSLMPPSAVTGLNAYISGNKIFVNWTANKEDDLNGYNVFVNDNYAGTLPRKATKIVIDK
SRFYLPKGKPIRISIEAFDITGDFSKERTEYVLDNNGEFLG

>NP_781237.1
MGRVDMVDKNSFIKNSLENELKDIYVTEDLKLKTLSRIKKEDNVVKPFKRKNRFLYKSMKGLALVSSILI
MFFTFSFLKYNRFMNDNSPTIDNDNNGKNVVLDNKEKEEKLDNPKEKIDEKKEVKDTKNENIKEKDPNEG
QPKKPKDNKIVNGGDNTPTNKEKTQPNKHIKNPKKENPKPSQGKEKPKLSEEQMLAKAENLWGGKINLPS
YIPKGYDVTNIGMDTAYGSKVLKITYKNPNSGNFLELKVLEGDKTAFEDGNGSKNPKENEKNEDDNRQDN
PKENVDKPTETVKSVNSNKNGVEYNIEGNIPAKTLEKIAESIEE

>YP_005844337.1
MGAAYEYLLKRFVDDAGQKVGEFFTPRSVVHLITRLLKPQENETVYDPTCSTGGMLFEPVAAVDANGGDT
RT

>YP_003412342.1
MKLKKVAMGITVVMASSLLLVGCGSSDDSSKDKKSTDTKQTETKKTAKTDGTMTDGTYKLEEKNFDDKGW
KGFMSIEVKDGKITKANYDYKNKDGKLKSEDADYEKAMKDKVGTGPQEYLKQLSDSLVKNQSAASVEVVS
GATHSSDAFINYANQLIQAAQKADTTTISINNLAKMEDGTYKLEEQNYAHGYRVVFSMDVKDGKITKSDY
NYVDKDGKLKSDDADYEKNMKAKSGTGPKEYIPALNKSLVEKQDVAAVDTVSGATNSSNQFKIYAAQLQN
AAQNGNTDTIKVYNLVEAE

>YP_884378.1
MHMGLRGEAAIVGYVELPPERLSKASPAPFVLEQWAELGAAALQDAGLPGEVVNGIVASHLAESEIFVPS
TIAEYLGVGARFAEHVDLGGASAAAMVWRAAAAIELGICDAVLCALPARYITPSSKKKPRPMVDAMFFGS
SSNQYGSPQAEFEIPYGNLGQNGPYGQVAQRYAAVYGYDERAMAKIVVDQRVNANHTDGAIWRDTPLTVE
DVLASPVIADPLHMLEIVMPCVGGAAVVVANADLAKRARHRPVWVKGFGEHVPFKTPTYAEDLLRTPIAA
AADTAFAMTGLSRAQMDMVSIYDCYTITVLLSLEDAGFCEKGRGMEFVADHDLTFRGDFPLNTAGGQLGF
GQAGLAGGMHHVCDATRQIMGRAGAAQVPDCNRAFVSGNGGILSEQTTLILEGD

>NP_856124.1
MKNARTTLIAAAIAGTLVTTSPAGIANADDAGLDPNAAAGPDAVGFDPNLPPAPDAAPVDTPPAPEDAGF
DPNLPPPLAPDFLSPPAEEAPPVPVAYSVNWDAIAQCESGGNWSINTGNGYYGGLQFTAGTWRANGGSGS
AANASREEQIRVAENVLRSQGIRAWPVCGRRG

>YP_177962.1
MVMSLMVAPELVAAAAADLTGIGQAISAANAAAAGPTTQVLAAAGDEVSAAIAALFGTHAQEYQALSARV
ATFHEQFVRSLTAAGSAYATAEAANASPLQALEQQVLGAINAPTQLWLGRPLIGDGVHGAPGTGQPGGAG
GLLWGNGGNGGSGAAGQVGGPGGAAGLFGNGGSGGSGGAGAAGGVGGSGGWLNGNGGAGGAGGTGANGGA
GGNAWLFGAGGSGGAGTNGGVGGSGGFVYGNGGAGGIGGIGGNGGDAGLFGNGGAGGAGAAGLPGAA
GLNGGDGSDGGNGGTGGNGGRGGLLVGNGGAGGAGGVGGDGGKGGAGDPSFAVNNGAGGNGGHGGNPGVG
GAGGAGGLLAGAHGAAGATPTSGGNGGDGGIGATANSPLQAGGAGGNGGHGGLVGNGGTGGAGGAGHAGS
TGATGTALQPTGGNGTNGGAGGHGGNGGNGGAQHGDGGVGGKGGAGGSGGAGGNGFDAATLGSPGADGGM
GGNGGKGGDGGKAGDGGAGAAGDVTLAVNQGAGGDGGNGGEVGVGGKGGAGGVSANPALNGSAGANGTAP
TSGGNGGNGGAGATPTVAGENGGAGGNGGHGGSVGNGGAGGAGGNGVAGTGLALNGGNGGNGGIGGNGGS
AAGTGGDGGKGGNGGAGANGQDFSASANGANGGQGGNGGNGGIGGKGGDAFATFAKAGNGGAGGNGGNVG
VAGQGGAGGKGAIPAMKGATGADGTAPTSGGDGGNGGNGASPTVAGGNGGDGGKGGSGGNVGNGGNGGAG
GNGAAGQAGTPGPTSGDSGTSGTDGGAGGNGGAGGAGGTLAGHGGNGGKGGNGGQGGIGGAGERGADGAG
PNANGANGENGGSGGNGGDGGAGGNGGAGGKAQAAGYTDGATGTGGDGGNGGDGGKAGDGGAGENGLNSG
AMLPGGGTVGNPGTGGNGGNGGNAGVGGTGGKAGTGSLTGLDGTDGITPNGGNGGNGGNGGKGGTAGNGS
GAAGGNGGNGGSGLNGGDAGNGGNGGGALNQAGFFGTGGKGGNGGNGGAGMINGGLGGFGGAGGGGAVDV
AATTGGAGGNGGAGGFASTGLGGPGGAGGPGGAGDFASGVGGVGGAGGDGGAGGVGGFGGQGGIGGEGRT
GGNGGSGGDGGGGISLGGNGGLGGNGGVSETGFGGAGGNGGYGGPGGPEGNGGLGGNGGAGGNGGVSTTG
GDGGAGGKGGNGGDGGNVGLGGDAGSGGAGGNGGIGTDAGGAGGAGGAGGNGGSSKSTTTGNAGSGGAGG
NGGTGLNGAGGAGGAGGNAGVAGVSFGNAVGGDGGNGGNGGHGGDGTTGGAGGKGGNGSSGAASGSGVVN
VTAGHGGNGGNGGNGGNGSAGAGGQGGAGGSAGNGGHGGGATGGDGGNGGNGGNSGNSTGVAGLAGGAAG
AGGNGGGTSSAAGHGGSGGSGGSGTTGGAGAAGGNGGAGAGGGSLSTGQSGGPRRQRWCRWQRRRWLGRQ
RRRRWCRWQRRCRRQRWRWRCRQRRLRRQWRQGRRRCRPWLHRRRGRQGRRWRQRRFQQRQRSRWQRR

>YP_177739.1
MSFVIATPEMLTTAATDLAKIGSTITAANTAAAAVAKVLPASADEVSVAVAALFGTHAQEYQTVSAQVAT
FHDRFVQTLSAAASSYVAAEAVNVEQSLLAAVNAPTQALFGRPLIGNGADGSPGTGQAGGPGGILYGNGG
NGGSGAPGQRGGAGGAAGLIGNGGNGGAGGVGTTGGAGGHGGAGGWLYGNGGAGGFGGAGAVGGNGGAGG
TAGLFGVGGAGGAGGNGIAGVTGTSASTPGGSGTAGGAGGIGGNGGAGGAGGVLMGNGGNGGAGGEGGPG
GAGGAGASGAHATNLGADGQAGGNGGNGGAGGTGGVGGPGGGHGLLGLGGSHGAGGAGGSGGDGGAPGDG
GNGATGTWGHNLGAGGTGGNGGNPGAGGAGGAGGASVGGSAHGANGAPGTTSTSGGNGGDGGKGADAISS
GQTGANGGRGGDGGQVGNGGAGGAGGRGGAGGLGFGSEAPGRPGGAGGTGGAGGNGGTQAGDGGTGGAGG
AGGDGGSGGAGSIGFNASAPGAAGSPGGNGGNGGPGGAGGEGGAGGLALAASGQNGSQGAGGDGGAGGNG
GTPGNGGHGAAGALGVNGGVGGAGGHGGDPGVGGAGGQGGSGSTPGANGAPGNTPTSGGNGGNGGRGADA
TGFGQTGASGGRGGDGGLVGNGGAGGAGGNGSKGLPGLGRLGNPGLDGGTGGNGGAGGSGGAWAGNGGTG
GAGGTGGVGGTGGSGSDGVNGSSAGADGHPGGTGGVGGTGGKGGDGGDGGAAPNGVAGSQGPGGAGGDGG
TGGVGGNGGRGIDGADGATAGARGQDGGAGGAGGKGGRGGTGGPGGAGPAGTTGSQGAGGNGGSGGTGGD
PGDGGNGANGSVFTNNGIGGNGGNGGNAGPSGAGGSGGAGSTFGATGSSSSIHVNGGNGGNGGNGDHALS
GNGAAGGNGGNGGNGSLRGSGGAGGHGGNGGNASRGMGGDGGTGGAGGNAGQIGNGGAGGNGGDGGTGSD
GNPGAITGSGGRGGDGGVGGQGGSVAGDGADGGRGGAGGTGGTGLRGTTGATGATGTFDAGADGHGGNGG
TGGVGGTGGAGGGGNGGAGGKALSPTGNNGSQGAGGDGGAGGAGGTGGTGGDGGRGAHGTLFSSLAGTG
GTGGNGGTGGTGGTGGAGGAGGTGSTLGATGATGAAGRAGNGGVGGSGGLGSAFGPGGTGGMGGAGGTST
VSAGGDGGRGGFGGDGLDASSGGNGGDGGHGGDGFRTAGAGGRGGDGGKGADPGGLFPIPGAGGKGGTGG
TGGTAHLGPLAIIGQSGQPGQFGSPGADGRGGAGGAGGGGAGGSF

>NP_214765.1
MNNLALWSRPVWDVEPWDRWLRDFFGPAATTDWYRPVAGDFTPAAEIVKDGDDAVVRLELPGIDVDKDVN
VELDPGQPVSRLVIRGEHRDEHTQDAGDKDGRTLREIRYGSFRRSFRLPAHVTSEAIAASYDAGVLTVRV
AGAYKAPAETQAQRIAITK

>NP_214754.1
MLSIDTNILLYAQNRDCPEHDAAAAFLVECAGRADVAVCELVLMELYQLLRNPTVVTRPLEGPEAAEVCQ
TFRRNRRWALLENAPVMNEVWVLAATPRIARRRLFDARLALTLRHHGVDEFATRNINGFTDFGFSRVWDP

ITSDG

>NP_217552.1
MRYLIATAVLVAVVLVGWPAAGAPPSCAGLGGTVQAGQICHVHASGPKYMLDMTFPVDYPDQQALTDYIT
QNRDGFVNVAQGSPLRDQPYQMDATSEQHSSGQPPQATRSVVLKFFQDLGGAHPSTWYKAFNYNLATSQP
ITFDTLFVPGTTPLDSIYPIVQRELARQTGFGAAILPSTGLDPAHYQNFAITDDSLIFYFAQGELLPSFV
GACQAQVPRSAIPPLAI

>NP_216438.1
MDSTVTASIRRMLGLLAATLLLGGCTGQHTTRTAASTTYTPHIKASSQDVLDGAINADEPGCSAAVGVEG
KVIWSGVRGIADLASGAKITTDTVFDIASVSKQFTATAILLLVEAGKLTLDDPISQYVPELPDWAQTVTV
EQLMHQTSGIPDYVALLAARGYQVSDRTIEAEARQALAAAPELQFKPGTRFDYSNSNYLLLGEIVHRASG
QPLPEFLSAEIFQPLGLAMVVDPVGKVPNKAVSYEKGTGGNRSEYRVGNPAWEQIGDGGIQTTPSQLARW
ADNYRTGSVGGLKLLEAQLAGAVETEPGGGDRYGAGIVSRADGTLDHAGAWAGFVTAFHISSDRRTSVAI
SCNTDKPDPVAMADALGRLWM

>NP_215002.1
MMTLKVAIGPQNAFVLRQGIRREYVLVIVALCGIADGALIAAGVGGFAALIHAHPNMTLVARFGGAAFLI
GYALLAARNAWRPSGLVPSESGPAALIGVVQMCLVVTFLNPHVYLDTVVLIGALANEESDLRWFFGAGAW
AASVVWFAVLGFSAGRLQPFFATPAAWRILDALVAVTMIGVAVVVLVTSPSVPTANVALII

>NP_217394.1
MSLRLVSPIKAFADGIVAVAIAVVLMFGLANTPRAVAADERLQFTATTLSGAPFDGASLQGKPAVLWFWT
PWCPFCNAEAPSLSQVAAANPAVTFVGIATRADVGAMQSFVSKYNLNFTNLNDADGVIWARYNVPWQPAF
VFYRADGTSTFVNNPTAAMSQDELSGRVAALTS

>NP_216684.1
MSGGSSRRYPPELRERAVRMVAEIRGQHDSEWAAISEVARLLGVGCAETVRKWVRQAQVDAGARPGTTTE
ESAELKRLRRDNAELRRANAILKTASAFFAAELDRPAR

>NP_215994.1
MRHTRFHPIKLAWITAVVAGLMVGVATPADAEPGQWDPTLPALVSAGAPGDPLAVANASLQATAQATQTT
LDLGRQFLGGLGINLGGPAASAPSAATTGASRIPRANARQAVEYVIRRAGSQMGVPYSWGGGSLQGPSKG
VDSGANTVGFDCSGLVRYAFAGVGVLIPRFSGDQYNAGRHVPPAEAKRGDLIFYGPGGGQHVTLYLGNGQ
MLEASGSAGKVTVSPVRKAGMTPFVTRIIEY

>NP_856026.1
MVNFSVLPPEINSGRMFFGAGSGPMLAAAAAWDGLAAELGLAAESFGLVTSGLAGGSGQAWQGAAAAAMV
VAAAPYAGWLAAAAARAGGAAVQAKAVAGAFEAARAAMVDPVVVAANRSAFVQLVLSNVFGQNAPAIAAA
EATYEQMWAADVAAMVGYHGGASAAAALAPWQQAVPGLSGLLGGAANAPAAAAQGAAQGLAELTLNLGVG
NIGSLNLGSGNIGGTNVGSGNVGGTNLGSGNYGSLNWGSGNTGTGNAGSGNTGDYNPGSGNFGSGNFGSG
NIGSLNVGSGNFGTLNLANGNNGDVNFGGGNTGDFNFGGGNNGTLNFGFGNTGSGNFGFGNTGNNNIGIG
LTGDGQIGIGGLNSGTGNIGFGNSGNNNIGFFNSGDGNIGFFNSGDGNTGFGNAGNINTGFWNAGNLNTG
FGSAGNGNVGIFDGGNSNSGSFNVGFQNTGFGNSGAGNTGFFNAGDSNTGFANAGNVNTGFFNGGDINTG
GFNGGNVNTGFGSALTQAGANSGFGNLGTGNSGWGNSDPSGTGNSGFFNTGNGNSGFSNAGPAMLPGFNS
GFANIGSFNAGIANSGNNLAGISNSGDDSSGAVNSGSQNSGAFNAGVGLSGFFR

>NP_218183.1
MVRQMRAALAALATGLLVLAPVAGCGGGVLSPDVVLVNGGEPPNPLIPTGTNDSNGGRIIDRLFAGLMSY
DAVGKPSLEVAQSIESADNVNYRITVKPGWKFTDGSPVTAHSFVDAWNYGALSTNAQLQQHFFSPIEGFD
DVAGAPGDKSRTTMSGLRVVNDLEFTVRLKAPTIDFTLRLGHSSFYPLPDSAFRDMAAFGRNPIGNGPYK
LADGPAGPAWEHNVRIDLVPNPDYHGNRKPRNKGLRFEFYANLDTAYADLLSGNLDVLDTIPPSALTVYQ
RDLGDHATSGPAAINQTLDTPLRLPHFGGEEGRLRRLALSAAINRPQICQQIFAGTRSPARDFTARSLPG
FDPNLPGNEVLDYDPQRARRLWAQADAISPWSGRYAIAYNADAGHRDWVDAVANSIKNVLGIDAVAAPQP
TFAGFRTQITNRAIDSAFRAGWRGDYPSMIEFLAPLFTAGAGSNDVGYINPEFDAALAAAEAAPTLTESH
ELVNDAQRILFHDMPVVPLWDYISVVGWSSQVSNVTVTWNGLPDYENIVKA

>NP_217661.1
MNVYIPILVLAALAAAFAVVSVVIASLVGPSRFNRSKQAAYECGIEPASTGARTSIGPGAASGQRFPIKY
YLTAMLFIVFDIEIVFLYPWAVSYDSLGTFALVEMAIFMLTVFVAYAYVWRRGGLTWD

>NP_856888.1
MPVRAPAAVRGAGLIVAVQGGAALVVAAALLVRGLAGADQHIVNGLGTAGWFVLVGGAVLAAGCRLAVGK
LWGRGLAVFAQLLLLPVAWYLIVGSHQPAIGIPVGIIALGVLVLLFSPPSIRWAAGRDQRGAASAANRGP
DSR

>NP_215288.2
MMARMPELSRRAVLGLGAGTVLGATSAYAIDMLLQPRTSHAAPAAAIGTNVPLAPTPALDPAPPAQAAPT
MSTGSFVSAARAGKMTNWAIARPPGQTQALRPVIALHGLGGSASAVMDGGVEQGLAQAVNAGLPPFAVVS
VDGGSSYWHQRASGEDAGAMVLNELIPLLDTQRLDTSRVAFLGWSMGGYGALLLGSRLGPARTAAICAVS
PALWLSAGSVAPGSFDGPDDWSANSVFGLPALGSIPIRVDCGNSDPFYAATKQFVAQLPHPPAGGFSPGG
HNGGFWSAQLPAELTWFAPLLTG

>NP_217065.1
MIFVDTSFWAALGNAGDARHGTAKRLWASKPPVVMTSNHVLGETWTLLNRRCGHRAAVAAAAIRLSTVVR
VEHVTADLEEQAWEWLVRHDEREYSFVDATSFAVMRKKGIQNAYAFDGDFSAAGFVEVRPE

>YP_008168548.1
MDDKTKNDQQESNEDKDELELFTRNTSKKRRQRKRSKATHFSNQNKDDTSQQADFDEEIYLINKDFKKEE
SNDKNNDSASSHANDNNIDDSTDSNIENEDYRYNQEIDDQNESNVISVDNEQPQSAPKEQNSDSIDEETV
TKKERKSKVTQLKPLTLEEKRKLRRKRQKRIQYSVITILVLLIAVILIYMFSPLSKIAHVNINGNNHVST
SKINKVLGVKNDSRMYTFSKKNAINDLEENPLIKSVEIHKQLPNTLNVDITENEIIALVKYKGKYLPLLE
NGKLLKGSNDVKINDAPVMDGFKGTKEDDMIKALSEMTPEVRRYIAEVTYAPSKNKQSRIELFTTDGLQV
IGDISTISKKMKYYPQMSQSLSRDSSGKLKTRGYIDLSVGASFIPYRGNTSSQSESDKNVTKSSQEENQA
KEELQSVLNKINKQSSKNN

>YP_008167623.1
MKKGFIHLIGMSDEGASMITYDLIGNTPLVLLEHYSDDKVKIYAKLEQWNPGGSVKDRLGKYLVEKAIQE
GRVRAGQTIVEATAGNTGIGLAIAANRHHLKCKIFAPYGFSEEKINIMIALGAEVSRTSQSEGMHGAQLA
ARSYAEKYGAVYMNQFESEHNPDTYFHTLGPELTSALQQIDYFVAGIGSGGTFTGTARYLKQHHVQCYAV
EPEGSVLNGGPAHAHDTEGIGSEKWPIFLERRLVDGIFTIKDQDAFRNVKSLAINEGLLVGSSSGAALQG
ALNLKAQLSEGTIVVVFPDGSDRYMSKQIFNYEENNNEQEN

>YP_008168905.1
MNLLKKNKYSIRKYKVGIFSTLIGTVLLLSNPNGAQALTTDNNVQSDTNQATPVNSQDKDVANNRGLANS
AQNTPNQSATTNQATNQALVNHNNGSIVNQATPTSVQSSTPSAQNNNHTDGNTTATETVSNANNNDVVSN
NTALNVPTKTNENGSGGHLTLKEIQEDVRHSSNKPELVAIAEPASNRPKKRSRRAAPADPNATPADPAAA
AVGNGGAPVAITAPYTPTTDPNANNAGQNAPNEVLSFDDNGIRPSTNRSVPTVNVVNNLPGFTLINGGKV
GVFSHAMVRTSMFDSGDNKNYQAQGNVIALGRIHGTDTNDHGDFNGIEKALTVNPNSELIFEFNTMTTKN
GQGATNVIIKNADTNDTIAEKTVEGGPTLRLFKVPDNVRNLKIQFVPKNDAITDARGIYQLKDGYKYYSF
VDSIGLHSGSHVFVERRTMDPTATNNKEFTVTTSLKNNGNSGASLDTNDFVYQVQLPEGVEYVNNSLTKD
FPSNNSGVDVNDMNVTYDAANRVITIKSTGGGTANSPARLMPDKILDLRYKLRVNNVPTPRTVTFNETLT
YKTYTQDFINSAAESHTVSTNPYTIDIIMNKDALQAEVDRRIQQADYTFASLDIFNGLKRRAQTILDENR
NNVPLNKRVSQAYIDSLTNQMQHTLIRSVDAENAVNKKVDQMEDLVNQNDELTDEEKQAAIQVIEEHKNE
IIGNIGDQTTDDGVTRIKDQGIQTLSGDTATPVVKPNAKKAIRDKATKQREIINATPDVTEDEIQDALNQ
LATDETDAIDNVTNATTNADVETAKNNGINTIGAVVPQVTHKKAARDAINQATATKRQQINSNREATQEE
KNAALNELTQATNHALEQINQATTNADVDNAKGDGLNAINPIAPVTVVKQAARDAVSHDAQQHIAEINAN
PDATQEERQAAIDKVNAAVTAANTNILNANTNADVEQVKTNAIQGIQAITPATKVKTDAKNAIDKSAETQ

HNTIFNNNDATLEEQQAAQQLLDQAVATAKQNINAADTNQEVAQAKDQGTQNIVVIQPATQVKTDTRNVV
NDKAREAITNINATTGATREEKQEAINRVNTLKNRALTDIGVTSTTAMVNSIRDDAVNQIGAVQPHVTKK
QTATGVLNDLATAKKQEINQNTNATTEEKQVALNQVDQELATAINNINQADTNAEVDQAQQLGTKAINAI
QPNIVKKPAALAQINQHYNAKLAEINATPDATNDEKNAAINTLNQDRQQAIESIKQANTNAEVDQAATVA
ENNIDAVQVDVVKKQAARDKITAEVAKRIEAVKQTPNATDEEKQAAVNQINQLKDQAINQINQNQTNDQV
DTTTNQAVNAIDNVEAEVVIKPKAIADIEKAVKEKQQQIDNSLDSTDNEKEVASQALAKEKEKALAAIDQ
AQTNSQVNQAATNGVSAIKIIQPETKVKPAAREKINQKANELRAKINQDKEATAEERQVALDKINEFVNQ
AMTDITNNRTNQQVDDTTSQALDSIALVTPDHIVRAAARDAVKQQYEAKKREIEQAEHATDEEKQVALNQ
LANNEKRALQNIDQAIANNDVKRVETNGIATLKGVQPHIVIKPEAQQAIKASAENQVESIKDTPHATVDE
LDEANQLISDTLKQAQQEIENTNQDAAVTDVRNQTIKAIEQIKPKVRRKRAALDSIEENDKNQLDAIRNT
LDTTQDERDVAIDTLNKIVNTIKNDIAQNKTNAEVDRTETDGNDNIKVILPKVQVKPAARQSVGVKAEAQ
NALIDQSDLSTEEERLAAKHLVEQALNQAIDQINHADKTAQVNQDSINAQNIISKIKPATTVKATALQQI
QNIATNKINLIKANNEATDEEQNIAIAQVEKELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRK
ASAREQLTTLFNDKKQAIEANIQATVEERNSILAQLQNIYDTAIGQIDQDRSNAQVDKTASLNLQTIHDL
DVHPIKKPDAEKTINDDLARVTALVQNYRKVSNRNKADALKAITALKLQMDEELKTARTNADVDAVLKRF
NVALSDIEAVITEKENSLLRIDNIAQQTYAKFKAIATPEQLAKVKVLIDQYVADGNRMIDEDATLNDIKQ
HTQFIVDEILAIKLPAEATKVSPKEIQPAPKVCTPIKKEETHESRKVEKELPNTGSEGMDLPLKEFALIT
GAALLARRRTKNEKES

>YP_008168566.1
MSKNITKNIILTTTLLLLGTVLPQNQKPVFSFYSEAKAYSIGQDETNINELIKYYTQPHFSFSNKWLYQY
DNGNIYVELKRYSWSAHISLWGAESWGNINQLKDRYVDVFGLKDKDTDQLWWSYRETFTGGVTPAAKPSD
KTYNLFVQYKDKLQTIIGAHKIYQGNKPVLTLKEIDFRAREALIKNKILYNENLNKGKLKITGGGNNYTI
DLSKRLHSDLANVYVKNPNKITVDVLFD

>YP_007968655.1
MKLIGIVGTNSNKSTNRQLLQYMQQHFADKAEIELIEVKDLPLFNKPADKNVPQVILDIAAKIEEADGVI
IGTPEYDHSIPSALMSVLAWLSYGIYPLLNKPVMITGASYGTLGSSRAQLQLRQILNAPELKASVLPDEF
LLSHSLQAFDKDGNLHDIETSQKLDAIFDDFRLFVKITGKLSNARDLLQKEAENFDWESL

>YP_007967992.1
MKISQYNKWSIRRLKVGAASVMITSGSIVALGQSHIVSADEMPQHKTTITAPTANTSTNVESSTDKALSK
VTTMETSSEMPKMQNMAKVEKTSDKPMMVATSVRKMMATPTPVAMTKTTSVDEVKESTDTAFKQTVDVPA
HYVNAAKGNGPFLAGVNQTIPYEAFGGDGMLTRLILKSSEGAKWSDNGVDKNSPLLPLKGLTKGKYFYQV
SLNGNTTGKEGQALLDQIKANDKHSYQATIRVYGAKDGKVDLKNMISQKMVTINIPHITTDMEVKNSLKM
AFKEKVDVPAKYVSAAKAKGPFLAGVNETIPYEAFGGDGMLTRLILKASEGAKWSDNGVDKNSPLLPLKD
LTKGKYFYQVSLNGNTAGKKGQALLDQIKANGSHTYQATITIYGTKDGKVDMNTILGKKTVMIHINVAKK
DMNSTSMMMKKDKMTMPMKKEMTSSKINTGMMMSNNKMSVNMQMSSQAKSNDKAGKKMSMISKNLPNTGE
TKQQNVGVLGVLSLAFATGLTALGLKKSKQR

>YP_007969093.1
MSKQKVMATLLLSTLVLSLSSPLVTLAETINPETSLTMATASTESSSEAEKQEKIQPTDSETVSPSAEGS
ISTEKTEVGTTETSSSNESPSSSSHQSSSNEDAKTSDSASTASTPSTNTTNSSQADSKPGQSTKTELKPE
PTLPLVEPKITPAPSQIESVQTNQNASVPALSFDDNLLSTSISPVTATPFYVEHWSGQNAYSHYLLSHRY
GIKAEQLDGYLKSLGIQYDSNRINGAKLLQWEKDSGLDVRAIVAIAVLESSLGTQGVAKMPGANMFGYGA
FDHDSSHASAYNDEEAIMLLTKNTIIKNNNSSFEIQDLKAQKLSSGQLNTVTEGGVYYTDNSGTGKRRAQ
IMEDLDRWIDQHGGTPEIPAALKALSTASLADLPSGFSLSTAVNTASYIASTYPWGECTWYVFNRAKELG
YTFDPFMGNGGDWQHKAGFETTHSPKVGYAVSFSPGQAGADGTYGHVAIVEEVKKDGSVLISESNAMGRG
IVSYRTFSSAQAAQLTYVIGHK

>YP_007968918.1
MKVKNKILTMVALTVLTCATYSSIGYADTSDKNTDTSVVTTTLSEEKRSDELNQSSTGSSSENESSSSSE
PETNPSTNPPTTEPSQPSPSEENKPDGSTKTEIGNNKDISSGTKVLISEDSIKNFSKASSDQEEVDRDES
SSSKASDEKKGHSKPKKELPKTGDSHSDTVIASTGGIILLSLSLYNKKMKLYLNIKRRTNLLFIFK

>YP_007968429.1
MNKSFNTKLGLVTVAALSGIVLTSQLPVNAKAETPVMAASAQQGFRFVATVVDSQTRVLPGKLVTLSEVT
SGQPKIIASVKTNDAGQAIFNNLPIKHNLSVSVDGQVKGYTIRTDVAGSSKAASFTATGVGTNEPTYSKK
TIDITVRDQNAEPVSGRTVTLKTQAGREIASLVSGDNGLTRFTDRLLDGTFYQYFVDGKKIGDIVPGESR
SAYVNVAPKKDYFTFTVTALDKDGLVVKGKEVTLTDITDGKAVALASLKTNDNGQAIFTNLPLSRNISVS
IDGKSKGYTLRTDVAGSHKAAAFYVDGKGTKAPQFSAEAAVVTVYDANGNTIANQEVTLTNSNGTIVAKG
LTEKDGKARFANKLMAGTLYNIFVNGIEMPKTALVGSDVSVFLTDKQIKKEKPVTPDKLDKEMNQEKEKE
LLDTPKPVPPTSKPTAPIKNKKDDSPKKFMKSQSSSLKSSVKAIGAKPKAMTALAKKLPKTGDQAISILT
LVGFIVTGLAIMLGFLKKSRSK

>YP_007969319.1
MNNNEKKVKYFLRKTAYGLASMSAAFIVCSGIVNTPTVSAESSNTLKVEKLGGIKNVESVHTLKPVSIPK
EQLNNELTKQKRRGELTIKSLGFITPNEQAKLIQDMNKIEDPNEVLTFVQKATLRDKDYGTKPKQITQKL
IEMLPNISDVLLKKYEAMIQDATSQADLNNIVDQAKKEVLEGDSEKYRHEIDNLSVTGLTMMAEADEIYQ
KHKDLLHSHYKDSIGTYTNLFTNRGGVTPWVGAEGLKDAQEAFRKAKVLLANLKALQEQIEKTQKKPVAK
GTYDVKYVDTEGKEVAGYKLVRTEGDVLNVFTAGAQVRTYVYEKVKPEAKPATKKSVNTSGNLAAKKAIE
NKKYSKKLPSTGEAASPLLAIVSLIVMLSAGLITIVLKHKKN

>YP_007969275.1
MKQNTTLNFKTNFQELSPEQLNNITGGGWIDDIKKIINLDKLNFLRL

>YP_002741380.1
MDETDEVSSKHGFEVVDETDEVSSKHGFEVVDETDEVSNHTYGKATLTWFEEIFEEY

>YP_002739733.1
MNKGLFEKRCKYSIRKFSLGVASVMIGAAFFGTSPVLADSVQSGSTANLPADLATALATAKENDGRDFEA
PKVEEDQGSPEVTDGPKTEEELLALEKEKPAEEKPKEDKPAAAKPETPKTVTPEWQTVEKKEQKGTVTIR
EEKGVRYNQLSSTAQNDNAGKPALFEKKGLTVDANGNATVDLTFKDDSEKGKSRFGVFLKFKDTNNNVFV
GYDKDGWFWEYKSPTTSTWYRGSRVVAPETGSTNRLSITLKSDGQLNASNNDVNLFDTVTLPAAVNDHLK
NEKKILLKAGSYGNDRTVVSVKTDNQEGVKADDTPAQKETGPVVDDSKVTYDTIQSKVLKAVIDQAFPRV
KEYTLNGHTLPGQVQQFNQVFINNHRITPEVTYKKINETTAEYLMKIRDDAHLINAEMTVRLQVVDNQLH
FDVTKIVNHNQVTPGQKIDDERKLLSSISFLGNALVSVSSDQTGAKFDGATMSNNTHVSGDDHIDVTNPM
KDLAKGYMYGFVSTDKLAAGVWSNSQNSYGGGSNDWTRLTAYKETVGNANYVGIHSSEWQWEKAYKGIVF
PEYTKELPSAKVVITEDANADKKVDWQDGAIAYRSIMNNPQGWEKVKDITAYRIAMNFGSQAQNPFLMTL
DGIKKINLHTDGLGQGVLLKGYGSEGHDSGHLNYADIGKRIGGVEDFKTLIEKAKKYGAHLGIHVNASET
YPESKYFNEKILRKNPDGSYSYGWNWLDQGINIDAAYDLAHGRLARWEDLKKKLGDGLDFIYVDVWGNGQ
SGDNGAWATHVLAKEINKQGWRFAIEWGHGGEYDSTFHHWAADLTYGGYTNKGINSAITRFIRNHQKDAW
VGDYRSYGGAANYPLLGGYSMKDFEGWQGRSDYNGYVTNLFAHDVMTKYFQHFTVSKWENGTPVTMTDNG
STYKWTPEMRVELVDADNNKVVVTRKSNDVNSPQYRERTVTLNGRVIQDGSAYLTPWNWDANGKKLPTDK
EKMYYFNTQTGATTWTLPSDWAKSKVYLYKLTDQGKTEEQELTVKDGKITLDLLANQPYVLYRSKQTNPE
MSWSEGMHIYDQGFNSGTLKHWTISGDASKAEIVKSQGANDMLRIQGNKEKVSLTQKLTGLKPNTKYAVY
VGVDNRSNAKASITVNTGEKEVTTYTNKSLALNYVKAYAHNTRRDNATVDDTSYFQNMYAFFTTGSDVSN
VTLTLSREAGDQATYFDEIRTFENNSSMYGDKHDTGKGTFKQDFENVAQGIFPFVVGGVEGVEDNRTHLS
EKHDPYTQRGWNGKKVDDVIEGNWSLKTNGLVSRRNLVYQTIPQNFRFEAGKTYRVTFEYEAGSDNTYAF
VVGKGEFQSGRRGTQASNLEMHELPNTWTDSKKAKKATFLVTGAETGDTWVGIYSTGNASNTRGDSGGNA
NFRGYNDFMMDNLQIEEITLTGKMLTENALKNYLPTVAMTNYTKESMDALKEAVFNLSQADDDISVEEAR
AEIAKIEALKNALVQKKTALVADDFASLTAPAQAQEGLANAFDGNVSSLWHTSWNGGDVGKPATMVLKEP
TEITGLRYVPRGSGSNGNLCDVKLVVTDESGKEHTFTATDWPDNNKPKDIDFGKTIKAKKIVLTGTKTYG
DGGDKYQSAAELIFTRPQVAETPLDLSGYEAALAKAQKLTDKDNQEEVASVQASMKYATDNHLLTERMVE
YFADYLNQLKDSATKPDAPTVEKPEFKLSSLASEQGKTPDYKQEIARPETPEQILPATGESQSDTALFLA
GVSLALSALFVVKTKKD

>YP_002740618.1

MLRLVYYQFLHNKKQWLGVSPVIFVSSLVMGLAVNGVINVENNSQVFVGLPDPKPIFMFPIVFGGVTLFF
VLSNIINMLVEIFRDDYELLEVLGASRLQLSFLVGGQIFIISSIISFIAYLCSIFVTSNYYYFLQYFFGE
NILPDIQFQMSAVGCIITVVLISFLAFLSGCFYTFKKIRNRKSSKIRHVLSIVKRILLLAGFSVIWLLSL
QQIFQDSTILAKAQIIFNIVILDIVIIYQLSPFIQSCFIKLLSIIIFRNNFMFIVSKWNLLYRKPYIKSI
SAAITGAILLISSFQMISQNILSQFQDDSDLELKVAFIVYVGAPILIVLANIISIAFLSSHQERIEIQQF
EILGTSNYQMVKIKVGEAIFLTFVTSLIAFLLNIKIIALIYYSLEDILIDDMNLLGLILPNFIVSIILFI
LIFITKSSYFIFKNAKIIS

>YP_002740442.1
MKIDKYSAILGNTVGFHDMSTLTDHRPVASLPFDGKYRLIDFPLSSLANAGVRSVFGIFQQDNISSVFDH
IRSGREWGLSTLLSHYYLGIYNTRVESSTVGKEYYQQLLTYLKRSGSNQTVALNCDVLINIDLNQVFHLH
STTKEPITVVYKKLAKKDISEVNAILDVDETDHVLSHKLFDSKSTAETFNMSTDIFVVDTPWLIEHLEEE
AKKEHPEKLRYVLRDLAVKEGAFAYEYTGYLANIHSVKSYYQANIDMLESQKFYSLFSPNQKIYTKVKNE
EPTYYANTSKVSTSQFASGSIIEGQVANSVLSRNIHVHKDSLVKDSLLFPRVVIGEGAQVEYAILDKGVE
VEPGVVIRGTAEHPVVVKKGAKVTEDIHS

>YP_002739906.1
MSRVHVQIMNQFERKSHEYKAIKRYWKLIQQDSRKLSDKRFYRPTFRMHLTNKEILDKILSYSEDLKHHY
QIYQLLLFHFQNKDPEKFFGLIEDNLKQVHPIFQTVFKTFLKNKEKIVNALQLPYSNAKLEATNNLIKLI
KRNAFGFRNFENFKKRIFIALNIKKERTKFVLSRA

>YP_002740339.1
MGLAFGYFCHRLTLLYDSLTNAPPMERFAYLLGEGLNQVFNPLWLFSFTRKSLLAFILGVITMSLVYLYV
STGQKVYREGEEYGSARFGTSKEKRFFYSKNPFNDTILTRDVRLTLLEKKKPQFDRNKNLIVIGGSGAGK
TFRFVKPNLIQLNCSNIVVDPKDHLAEKTGKLFLENGYQVKVLDLVNMTNSDGFNPFRYVETENDLNRML
TVYFNNTKGNGSRSDPFWDEASMTLVRAIASYLVDFYNPPGSSKQEQEARRKRGRYPAFSEIGKLIKLLS
KGDNQDKSVLEVLFEDYAKKYGHENFTMRNWADFQNYKDKTLDSVIAVTTAKFALFNIQSVIDLTQRDTM
DLKTWGTQKTMVYLVIPDNDTTFRFLSALFFSTVFSTLTRQADVDFKGQLPIHVRSYLDEFANVGEIPDF
AEQTSTVRSRNMSLVPILQNIAQLQGLYKEKEAWKTILGNCDSLLYLGGNDEETFKFMSGLLGKQTIDVR
STSRSFGQTGSSSTSHQKIARDLMTADEVGNMKRDECLVRIAGVLVFRTKKYFPLKHKNWKWLADKESDE
RWWHYHINPLITEEEIDLSGHTIRDLSTEMSIY

>YP_002740734.1
MQKKYVKILYSSPIGILSLVADDHYLYGIWVQEQKHFERGLGDETIEEVVSHPILDPVIACLDDYFKGKP
QDLSNLLLAPIGTNFEKRVWDYLQGIPYGQTVTYGQIAQDLQVASAQAIGGAVGRNPWSILVPCHRVLGA
GKRLTGYAAGVEKKAWLLEHEGVDFKDRNNRRRSTC

>YP_002740617.1
MAEEVLNLQLVSVQVDETDEVDGMRFSTFSTNRCGNWSAFSWENC

>YP_002741366.1
MSFKVLHRGYQHIRLSSSFSLTLDIQDYLRSLARDEKGIESIQFYMDQQHFTLRIKEGFSVLDNAEAFLK
RIDKGKVSELMTLPIRREESAYSIVSGAAVKRVLFRSFVPYPIRYIWTCYQALGYIREAYQTLARKELTM
EVLDCSAILLSLFMNQSKTASNIMFMLDLGNHLDQWSLKKTATDLEQSLLAKESDVFLVQGDTVVSIKSS
DVQIGDVLILSQGNEILFDGQVVSGLGMVNESSLTGESFPVEKRESDLVCANTVLETGELRIRVTDNQMN
SRILQLIELMKKSEENKKTKQRYFIKMADKVVKYNFLGAGLTYLLTGSFSKAISFLLVDFSCALKISTPV
AYLTVIKEGLNREMVIKDGDVLEKYLEVDTFLFDKTGTITTSYPIVEKVLPFGDYSEEDILRISACLEEH
IYHPIANAIVKQAEIEGIEHEEMHGKLQYIASKGIKSHIDGQPVLIGNYVLMQDEQIHISSEQNALIEEY
KSHYNLLFLAYQNELIGMFCIHTPLRKEAKTALDKLKAQGKKLILATGDTLIRTEELVKDLPFDQVYTDL
KPDGKFELVEKLQKAGHTILMVGDGLNDSAALTLSDIGVVMNESADISKQMSDILLLDNRLDFFQELDSL
SSSLQTLIKKNIQDTVVVNSSLIGFGLFNWLSPSNLSILHNLTTLRIVLRSLSIKNR

>YP_002740654.1
MKLIVSVMPRSLEEAQALDATRYLDADIIEWRADYLPKEAILQVAPAIFEKFAGRELVFTLRTRSEGGEI
DLSPEEYIHLIKEVAQLYQPDYIDFEYYSYKDVFEEMLDFPNLVLSYHNFQETPENMMEILSELTILNPK
LVKVAVMAHTEQDVLDLMNYTRGFKTLNPEQEYVTISMGKVGKVSRITADVTGSSWSFASLDEVSAPGQI
SLASMKKIREILDEA

>YP_008242889.1
MINKKYIIPVSLLTLAITLTSVEEVTSRQNLTYANEIVTQRPKRESVISDKSNFPVISPYLASVDFGERK
TPLPTPDKGVKVTTEQSIAQVRKGPEERTYTVTGKITSVINGWGGYGFYIQDSEGIGLYVYPQKDLGYSK
GDIVQLTGTLTRFKGDLQLQQVTAHKKLELSFPTSVKEAVISELETTTPSTLVKLSHVTVGELSTDQYNN
TSFLVRDDSGKSIVVHIDHRTGVKGADVVTKISQGDLINLTAILSIVDGQLQLRPFSLEQLEVVKKVTSS
NSDASSRNIVKIGEIQGASHTSPLLKKAVTVEQVVVTYLDDSTHFYVQDLNGDGDLATSDGIRVFAKNAK
VQVGDVLTISGEVEEFFGRGYEERKQTDLTITQIVAKAVTKTGTAQVPSPLVLGKDRIAPANIIDNDGLR
VFDPEEDAIDYWESMEGMLVAVDDAKILGPMKNKEIYVLPGSSTRPLNNSGGVLLPANSYNTDVIPVLFK
KGKQIIKAGDSYKGRLAGPVSYSYGNYKVFVDDSKNMPSLMDGHLKPEKTNLQKDLSKLSIASYNIENFS
ANPSSTKDEKVKRIAESFIHDLNAPDIIGLIEVQDNNGPTDDGTTDATQSAQRLIDAIKKLGGPTYRYVD
IAPENNVDGGQPGGNIRTGFLYQPERVSLSDKPKGGARDALTWVNGELNLSVGRIDPTNVAWKDVRKSLA
AEFIFQGRKVVVVANHLNSKRGDNALYGRVQPVTFKSEQRRHVLANMLAQFAKEGAKHQANIVMLGDFND
FEFTKTIQLIEEGDMVNLVSRHDISDRYSYFHQGNNQTLDNILVSRHLLDHYEFDMVHVNSPFMEAHGRA
SDHDPLLLQLSFSKENDKAESSKQSVKAKKTSKGKLLPKTGDSLVYVITLLGTASLLVSILLLTKGKKES

>YP_008243726.1
MKMKKISKCAFVAISALVLIQATQTVKSQEPLVQSQLVTTVALTQDNRLLVEEIGPYASQSAGKEYYKHI
EKVIVDNDVYEKSLEGERTFDINYQGIKINANLIKDGKHELTIVNKKDGDILITFIKKGDKVTFISAQKL
GTTDHQDSLKKDVLSDKTVPQNQGTQKVVKSGKNTANLSLITKLSQEDGAILFPEIDRYSDNKKIKALTQ
QITKVTVNGTVYKDLISDSVKDTNGWVSNMTGLHLGTKAFKDGENTIVISSKGFEDVTITVTKKDGQIHF
VSAKQKQHVTAEDRQSTKLDVTTLEKAIKEADAIIAKESNKDAVKDLAEKLQVIKDSYKEIKDSKLLDDT
HRLLKDTIESYQAGEVSINNLTEGTYTLNFKANKENSEESSMLQGAFDKRAKLVVKADGTMEISMLNTAL
GQFLIDFSIESKGTYPAAVRKQVGQKDINGSYIRSEFTMPIDDLDKLHKGAVLVSAMGGQESDLNHYDKY
TKLDMTFSKTVTKGWSGYQVETDDKEKGVGTERLEKVLVKLGKDLDGDGKLSKTELEQIRGELRLDHYEL
TDISLLKHAKNITELHLDGNQITEIPKELFSQMKQLRFLNLRSNHLTYLDKDTFKSNAQLRELYLSSNFI
HSLEGGLFQSLHHLEQLDLSKNRIGRLCDNPFEGLSRLTSLGFAENSLEEIPEKALEPLTSLNFIDLSQN
NLALLPKTIEKLRALSTIVASRNHITRIDNISFKNLPKLSVLDLSTNEISNLPNGIFKQNNQLTKLDFFN
NLLTQVEESVFPDVETLNLDVKFNQIKSVSPKVRALIGQHKLTPQKHIAKLEASLDGEKIKYHQAFSLLD
LYYWEQKTNSAIDKELVSIEEYQQLLQEKGSDTVSLLNDMQVDWSIVIQLQKKASNGQYVTVNEKLLSND
PKDDLTGEFSLKDPGTYRIRKALITKKFATQKEHIYLTSNDILVAKGPHSHQKDLVEKGLRALNQKQLRD
GIYYLNASMLKTDLASESMSNKAINHRVTLVVKKGVSYLEVEFRGIKVGKMLGYLGELSYFVDGYQRDLA
GKPVGRTKKAEVVSYFTDVTGLPLADRYGKNYPKVLRMKLIEQAKKDGLVPLQVFVPIMDAISKGSGLQT
VFMRLDWASLTTEKAKVVKETNNPQENSHLTSTDQLKGPQNRQQEKTPTSPPSAATGIANLTDLLAKKAT
GQSTQETSKTDDTDKAEKLKQLVRDHQTSIEGKTAKDTKTKKSDKKHRSNQQSNGEESSSRYHLIAGLSS
FMIVALGFIIGRKTLFK

>YP_001199090.1
MEFLGELLVEFLTGLADFDEKKHPPFGIRYWLGWLGVLIHVLLLALLTCVTVFFFKFFLVGKGLINVVVA
VVFLLFALFWLWKFGKTILKMWQATIYYLAIH

>YP_001198726.1
MKKILIVDDEKPISDIIKFNMTREGYEVVTAFDGREALEVFEAEFPDIVILDLMLPELDGLEVARTIRKT
SNVPILMLSAKDSEFDKVIGLEIGADDYVTKPFSNRELQARVKALLRRSELPETQTNIESTGTPELVIGD
LVILPDAFVAKQAW

>YP_001197582.1
MNKKIFLKSAIILLSVTCLINAIKALDSSGVSSSTGEIVNQDRIYNITLKVKTDGVYNDLGGNVISENNV
LSGISGTKINLKVSTNSGYRFVGYQSFLESGNSTDGLLPIENDSFEISDKTGNVTVVAMFEKIPIDETIY
FSDEFSTENINNYLLSENLIGKIAVSNGKLNIHAPGVSSIKPILSRQIDESSTLGYEIQFSIQQIGEVKQ
WNTFRVVFKENSDGSVYALEFTGKAVSIKKLSSIDAPNQTGEKYAETGHNLGSEEHRIRLVVRGDTVTVS

DNEIPLLSYSSPENWEGATASIVFTPISNRSVSLDDIIIRQTRALRSLLVVSRIDGQEVTDIQPGSIRGN
TSQVFVGDSLPLEVIEKPGYQFIGFKDEFGNVVDLSTFSVPNDESDLVIYADFQTAEVVNRETKTFYIDS
IEGNDTNSGESETNAWKTLEQLRKNTLIAGDRVLLKRGSRFVGEDAALTFKGSGLEDAPILISSYGEGEL
PLLEAQGKIESVIKLYNQEYITIENLEITNLDPNFSTSFELNSNNNRSKILRGVHVIAEDYGVVHDIVLR
NMYLHDINGNLNSKWNGGIFFDVYGTTVPTKYDGILIENNYLERVDRSGIKLVGSTWANQNLKNNKNIPL
NWYPSTNVVVRGNRIEKAGGDSITVRDTDGALIEYNISADARYQDTGYNAGIWPFQASNTVIQYNESFRT
HGVQDGQGLDLDHVSNNSVMQYNYSHDNEGGFMLIMNWYEQTSPTVRYNISQNDKDKILELARGGAQGTA
IYNNTIFSDSKLTGRAGVIDMPSTSGGTGVKDIFLFNNIFYFTDGEKMFVEASDAGKYKDSIHFYNNAYV
GVEVPDDPKAITEGITLKGVGTGPTENKSMIANAGKFLTGQLDGYRLPENSSLAKLGVSKEEAISYFYKK
LGVQPTIDFQDNGSLTMSPTEAFALARQTNSVDAIARVYPAIEGVTYDTDFFGETLSSENLSVGAAQEKK
SVFEEKISDIEISDVKTGIGMRYKRSNQLADVNLSIRLINKDELQYTINANNPTFSYEIQLQKNFGEYMD
LSESVTITIPIEADKKILNVLKINKNEQTESVKELPYRIENNKIILETDELGSFIIEYQMEKIKKENNTN
ATISESGKSSLGLQNKENSDVQAVGDKDYFSGLELSNKGEELSREDSKLMGNSINNLQTIKESESIFEGN
LGRNKELSDMYQKSLPKTGESNQAYIVLIGLFSLFVVLICQIFKKSID

>YP_002815839.1
MKVFCLGGAGKICREAILDLVQFSSFETITVADFNEEESLKIVEWLNDPRVNFVKVDVTNHEDTVAKMKG
YDIVMDGTTIKLNGLSTRCIAEAGCHGVNLNGFGEENDSHSIFVQNGKTCLPGFGMTPGVTQMMAMHAAN
QLHTVESVRVSHGSYRPIAFSASITETTTYEYDPHLPSRTVYENGEFKQVPPFARPREIELPAPYGKATQ
YIIPHSETITLAKALENKGVKLIETRGTWPKQNMQLVRALYDYGILRNDQVEINGEAIGIMDCVSKYLLQ
SKEGQETEVYGYALHVEVIGMKNNEKQRHVLYHTHPLSDGSVEGWEKLRAYTRNVGIPFGIATELIAKGD
VNKVGVVTPEEAFVNPQVIFDELKKRGIHIHEEISTYKESYNFV

>YP_002817680.1
MTEKELELLACLEKSSRLSVDTLAKLLNIEVEEVKKMVEKLEAEKIIVDYVTHIDWTKVKEHTGLTAMID
VKVTPKHGVGFDAVAEQIYRYSEVKSVYLMSGTYDLSITLEGKTMGEVAMFVSEKLATIESVVSTTTHFI
LKKYKHEGIIYEKNDDDKRIVVTP

>YP_003413400.1
MGVIASVVRQQYLIMTAPEPDPAFHSKEQNFLNELSPRAQEIQEKHGILTSITLAQAILESDWGQSGLAQ
KGNNLFGVKGKSPQPMVTMTTKEFVDGKWIEINANFRKYKDWNESLDSHAELFLNGTSWNKDKYNGVIAA
DDYKKAAQELQSAGYATDPDYAEKLINIIEKYDLALYDRIEDKIYYDTKSTGFGNVKKDVSGAIWTKPYG
LSGALKVEEINYYKREDLKLLREAKTDSGTWYQIAVDTEPIGWVKQELIDKK

>NP_218377.1
MYERDEFLRDRIRPHQPGTPRGYSPRPPSGDRCPAPPPGRHAAAATPPGPPRLPSAPLRPLPDPAWPRQP
EAPPPSTWADPALAPIRSRTRPGERGWRRMVRLVTFGLVGLGRSGMQRQEAQFEATIRTVLHGNHKVAVL
GKGGVGKTSVAACVGSILAELRQQDRIVGIDADTAFGRLSSRIDPRAAGSFWELTTDTNLRSFTDITARL
GRNSAGLYVLAGQPASGPRRVLDPAIYREAALRLDHHFAISVIDCGSSMEAAVTQEVLRDVDALIVVSSP
WADGASAAANTIEWLSDYGLTGLLRRSIVVLNDSDGHADKRTKSLLAQEFIDHGQPVVEVPFDPHLRPGG
VIDMSHEMAPTTRLKILQVAATVTAYFASRPADAHGSPPR

>NP_215959.1
MVGYAEPVLIERQSVVAAPAEQVWQRVVTPEGINDELRPWMTMSVPRGAKGMTVDTVPIGAPIGRAWLRL
FGVLPFDYDRLSIAELEPGRRFREDSTMLSMRQWQHERTVTPEGDTKTIVRDRITFQTRAGLRFAAPLIA
AGLRALFGHRHRRLQRHFAQG

>YP_008169312.1
MTYRIKKWQKLSTITLLMAGVITLNGGEFRSVDKHQIAVADTNVQTPDYEKLRNTWLDVNYGYDKYDENN
PDMKKKFDATEKEATNLLKEMKTESGRKYLWSGAETLETNSSHMTRTYRNIEKIAEAMRNPKTTLNTDEN
KKKVKDALEWLHKNAYGKEPDKKVKELSENFTKTTGKNTNLNWWDYEIGTPKSLTNTLILLNDQFSNEEK
KKFTAPIKTFAPDSDKILSSVGKAELAKGGNLVDISKVKLLECIIEEDKDMMKKSIDSFNKVFTYVQDSA
TGKERNGFYKDGSYIDHQDVPYTGAYGVVLLEGISQMMPMIKETPFNDKTQNDTTLKSWIDDGFMPLIYK
GEMMDLSRGRAISRENETSHSASATVMKSLLRLSDAMDDSTKAKYKKIVKSSVESDSSYKQNDYLNSYSD
IDKMKSLMTDNSISKNGLTQQLKIYNDMDRVTYHNKDLDFAFGLSMTSKNVARYESINGENLKGWHTGAG
MSYLYNSDVKHYHDNFWVTADMKRLSGTTTLDNEILKDTDDKKSSKTFVGGTKVDDQHASIGMDFENQDK
TLTAKKSYFILNDKIVFLGTGIKSTDSSKNPVTTIENRKANGYTLYTDDKQTTNSDNQENNSVFLESTDT
KKNIGYHFLNKPKITVKKESHTGKWKEINKSQKDTQKTDEYYEVTQKHSNSDNKYGYVLYPGLSKDVFKT
KKDEVTVVKQEDDFHVVKDNESVWAGVNYSNSTQTFDINNTKVEVKAKGMFILKKKDDNTYECSFYNPES
TNSASDIESKISMTGYSITNKNTSTSNESGVHFELTK

>YP_008243986.1
MKTIRRYDVNEDRGHTGLVEAGDFYYLNYCVGNVGQDIESQINGAFDEMERRLALVGLTLDAVVQMDCLF
RDVWNIPVMEKMIKERFNGRYPARKSIQTEFAHHGGPQGLLFQVDGVAYSKHISMT

>YP_008243740.1
MTNRLSWQDYFMANAELISKRSTCDRAFVGAVLVKDNHIIATGYNGGVSATDNCNEAGHYMEDGHCIRTV
HAEMNALIQCAKEGISTDGTEIYVTHFPCINCTKALLQAGITKITYKAHYRPHPFAIELMEKKGVAYVQH
DVPQIVLGGDKEQ

>YP_007968868.1
MKRQISSDKLSQELDRVTYQKRFWSVIKNTIYILMAVASTAILIAVLWLPVLRIYGHSMNKTLSAGDVVF
TVKGSNFKTGDVVAFYYNNKVLVKRVIAESGDWVNIDSQGDVYVNQHKLKEPYVIHKALGNSNIKYPYQV
PDKKIFVLGDNRKTSIDSRSTSVGDVSEEQIVGKISFRIWPLGKISSIN

>YP_007968919.1
MKRKYFILNTVTVLTLAAAMNTSSIYANSTETSASVAPTTNTIVQTNDSNPTAKFASESGQSVIGQVKPA
NSAALTTVDTPHISAPDALKTTQSSPVVESPSTKLTEETYKQKDGQDLANMVRSGQVTSEELVNMAYDII
AKENPSLNAVITTRRQEAIEEARKLKDTNQPFLCIPLLVKGLGHSIKGGETNNGLIYADGKISTFDSSYV
KKYKDLGFIILGQTNFPEYGWRNITDSKLYGPTHNPWNLAHNAGGSSGGSAAAIASGMTPIASGSDAGGS
IRIPSSWTGLVGLKPTRGLVSNEKPDSYSTAVHFPLTKSSRDAETLLTYLKKSDQTLVSVNDLKSLPIAY
TLKSPMGTEVSQDAKNAIMDNVTFLRKQGFKVTEIDLPIDGRALMRDYSTLAIGMGGAFSTIEKDLKKHG
FTKEDVDPITWAVHVIYQNSDKAELKKSIMEAQKHMDDYRKAMEKLHKQFPIFLSPTTASLAPLNTDPYV
TEEDKRAIYNMENLSQEERIALFNRQWEPMLRRTPFTQIANMTGLPAISIPTYLSESGLPIGTMLMAGAN
YDMVLIKFATFFEKHHGFNVKWQRIIDKEVKPSTGLIQPTNSLFKAHSSLVNLEENSQVTQVSISKKWMK
SSVKNKPSVMAYQKALPKTGDTESSLSPVLVVTLLLACFSFVTKKNQKS

>YP_007968207.1
MTKKHLKTLALALTTVSVVTYSQEVYGLEREESVKQEQTQSASEDDWFEEDNERKTNVSKENSTVDETVS
DLFSDGNSNNSSSKTESVVNDPKQVPKAKPEVTQEASNSSNDTSKVEVPKQDTVSKKETIETSTWEAKDF
VTRGDTLVGFSKSGINKLSQTSHLVLPSHAADGTQLTQVASFAFTPDKKTAIAEYTSRLGENGKPSRLDI
DQKEIIDEGEIFNAYQLTKLTIPNGYKSIGQDAFVDNKNIAEVNLPESLETISDYAFAHMSLKQVKLPDN
LKVIGELAFFDNQIGGKLYLPRHLIKLAERAFKSNRIQTVEFLGSKLKVIGEASFQDNNLSNVMLPDGLE
KIESEAFTGNPGDEHYNNQVILRTRTGQNPHQLATENTYVNPDKSLWRATPDMDYTKWLEEDFTYQKNSV
TGFSNKGLQKVRRNKNLEIPKQHNGITITEIGDNAFRNVDFQSKTLRKYDLEEVKLPSTIREEAFKRNHL
KEVKGSSTLSQIAFNAFDQNDGDKRFGKKVVVRTHNNSHMLADGERFIIDPDKLSSTMVDLEKVLKIIEG
LDYSTLRQTTQTQFREMTTAGKALLSKSNLRQGEKQKFLQEAQFFLGRVDLDKAIAKAEKALVTKKATKN
GHLLERSINKAVLAYNNSAIKKANVKRLEKELDLLTDLVEGKGPLAQATMVQGVYLLKTPLPLPEYYIGL
NVYFDKSGKLIYALDMSDTIGEGQKDAYGNPILNVDEDNEGYHALAVATLADYEGLYIKDILNSSLDKIK
AIRQIPLAKYHRLGFFQAIRNAAAEADRLIPKTPKGYLNAVPNYRKKQVEKNSKPVDYKTPIFDKALPNE
KVDGDRVAKGHNINAETNNSVAVTPIRSEQQLHKSQSDVNLPQTSSKNNFIYEILGYVSLCLLFLVTAGK
KGKRARK

>YP_007968202.1
MDKHHSKKAILKLTLITTSILLMHSNQVNAEEQELKNQEQSPVIANVAQQPSPSVTTNIVEKTSVTAASA
SNTAKEMGDTSVKNDKTEDELLEELSKNLDTSNLGADLEEEYPSKPEITNNKENNVVTNASTAIAQKVPS
AYEEVKPESKSSLAVLDTSKITKLQATTQRGKGNVVAIIDTGFDINHDIFRLDSPKDDKHSFKTKAEFEE
LKAKHNITYGKWVNDKIVFAHNYANNTETVADIAAAMKDGYGSEAKNISHGTHVAGIFVGNSKRPAINGL

LLEGAAPNAQVLLMRIPDKIDSDKFGEAYAKAITDAVNLGAKTINMSLGKTADSLIALNDKVKLALKLAS
EKGVAVVVAAGNEGAFGMDYSKPLSTNPDYGTVNSPAISEDTLSVASYESLKTISEVVETTIEGKLVKLP
IVTSKPFDKGKAYDVVYANYGAKKDFEGKDFKGKIALIERGGGLDFMTKITHATNAGVVGIVIFNDQEKR
GNFLIPYRELPVGVISKVDGERIKNTSSQLTFNQSFEVVDSQGGNRMLEQSSWGVTAEGAIKPDVTASGF
EIYSSTYNNQYQTMSGTSMASPHVAGLMTMLQSHLAEKYKGMNLDSKKLLELSKNILMSSATALYSEEDK
AFYSPRQQGAGVVDAEKAIQAQYYVTGNDGKAKINLKRVGDKFDITVTIHKLVEGVKELYYQANVATEQV
NKGEFALKPQALLDTNWQKVILRDKETQVRFTIDASQFSQKLKEQMANGYFLEGFVRFKEAKDSNQELMS
IPFVGFNGDFANLQALETPIYKTLSKGSFYYNPNDTIHKDQLEYNESAPFESNNYTALLTQSASWGYVDY
VKNGGELELAPESPKRIILGTFENKVEDKTIHLLERDAANNPYFAISPNKDGNRDEITPQATFLRNVKDI
SAQVLDQNGNVIWQSKVLPSYRKNFHNNPKQSDGHYRMDALQWSGLDKDGKVVADGFYTYRLRYTPVAEG
ANSQESDFKVQVSTKSPNLPSRAQFDETNRTLSLAMPKESSYIPTYRLQLVLSHVVKDEEYGDETSYHYF
HIDQEGKVTLPKTVKIGESEVAVDPKALTLVVEDKAGNFATVKLSDLLNKAVVSEKENAIVISNSFKYFD
NLKKEPMFISKEGKVVNKNLEETTLVKPQTTVTTQSLSKEITKSGNEKVLTSTNNNSSRVAKIISPKHNG
DSVNHTLPSTSDRATNGLFVGTLALLSSLLLYLKPKKTKNNSK

>YP_002740124.1
MVKVAVMLAQGFEEIEALTVVDVLRRANITCDMVGFEEQVTGSHAIQVRADHVFDGDLSDYDMIVLPGGM
PGSAHLRDNQTLIQELQSFEQEGKKLAAICAAPIALNQAEILKNKRYTCYDGVQEQILDGHYVKETVVVD
GQLTTSRGPSTALAFAYELVEQLGGDAESLRTGMLYRDVFGKNQ

>YP_002740034.1
MKLFKPLLTVLALAFALIFITACSSGGNAGPSSGKTTAKARTIDENKKSGELRIAVFGDKKPFDYVDNDG
SYQGYDIELGNQLAQDLGVKVKYISVDAANRAEYLISNKVDITLANFTVTDERKKQVDFALPYMKVSLGV
VSPKTGLITDVKQLEGKTLIVTKGTTAETYFEKNHPEIKLQKYDQYSDSYQALLDGRGDAFSTDNTEVLA
WALENKGFEVGITSLGDPDTIAAAVQKGNQELLDFINKDIEKLGKENFFHKAYEKTLHPTYGDAAKADDL
VVEGGH

>YP_002739327.1
MSKNIVQLNNSFIQNEYQRRRYLMKERQKRNRFMGGVLILIMLLFILPTFNLAQSYQQLLQRRQQLADLQ
TQYQTLSDEKDKETAFATKLKDEDYAAKYTRAKYYYSKSREKVYTIPDLLQR

>YP_002740557.1
MTDVRGTSCFVIKFGKAGEQLAAKLWEEGKMVYASYASMTKRLKLAMRARCNGVYGR

>YP_002741520.1
MKHLKTFYKKWFQLLVVIVISFFSGALGSFSITQLTQKSSVNNSNNNSTITQTAYKNENSTTQAVNKVKD
AVVSVITYSANRQNSVFGNDDTDTDSQRISSEGSGVIYKKNDKEAYIVTNNHVINGASKVDIRLSDGTKV
PGEIVGADTFSDIAVVKISSEKVTTVAEFGDSSKLTVGETAIAIGSPLGSEYANTVTQGIVSSLNRNVSL
KSEDGQAISTKAIQTDTAINPGNSGGPLINIQGQVIGITSSKIATNGGTSVEGLGFAIPANDAINIIEQL
EKNGKVTRPALGIQMVNLSNVSTSDIRRLNIPSNVTSGVVVRSVQSNMPANGHLEKYDVITKVDDKEIAS
STDLQSALYNHSIGDTIKITYYRNGKEETTSIKLNKSSGDLES

>YP_008242460.1
MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNI
LQDGEHKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSG
NLKKQIAKVLTEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLI
DTTDIDKYEDVQFDLFVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLY
RQLPGEKEVAVDDAELKQINSEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGL
TVTNTYVKPTSGYYDIEVTFGNGHIDITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMP
GEEDGTNSNKYEEVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSG
KTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEIATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKP
TKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDLIIGGQGQIVETTEDTQTGMHGDSGCKTEVEDTKLV
QSFHFDNKESESNSEIPKKDKPKSNTSLPATGEKQHNMFFWMVTSCSLISSVFVISLKSKKRLSSC

>YP_008243496.1
MTSKKACLSSIIVLASLTCGNDTVSANHLSATGDKFDDCSTLVEKDVAPKDELEMLAWSSSQTTDDADRD
YEDFLDDDSFISQNETDKMFENLTDDRLLNELDELDELDEEDTIEPEQNVIMPSDDELFDLTDAVETRLT
VSSAPHLEAELPKPHLRSLSDTALRSGEIRGHLDNKLDTLSVTATKLALTMAQKFDLTTHVYSIGESFSE
VLAAHYEDRKAESAFSKKKRFHLPIATPDVVIEELRRLVSSIGSSKEDVSVPYNRKLGMAVAKRKIALPQ
TGEEFSYYPALLGLMILGLTPIMIPKKINN

>YP_008243141.1
MFRLLKRACSFLLLFVIYQSFVIHHNVQRVLAYKPMVEKTLAENDTKANVDLVLAMIYTETKGGEADVMQ
SSESSSGQKNSITDSQASIEHGVNLLSHNLALAEEAGVDSWTAVQAYNFGTAYIDYIAKHGGQNTVDLAT
TYSKTVVAPSLGNTSGQTYFYYHPLALISGGKLYKNGGNIYYSREVHFNLYLIELMSLF

>YP_008243242.1
MKPKHLLCLSTVVAGLALFSTMTHSVLADDASNPDTILMNNNQANLQRDALVQKLDEGHQQLEAIKHEAK
GTDIEATVNKAIDAVDHMKSSIRFNTETIYDFSSIGARVEALSDAIKAIVFSTTQLTHKVEKAHTDMGFA
ITKLVIRIIDPFASVDAIKAQVQEIKALEEKVINYPDLQPTDRATIYTKAKLNKAIWNTRLERDKKVLGI
KPFDVYNRLNKAITHAVGVQLNPTTTVQQVDDEVIAVQNALETALKS

>YP_008243171.1
MVTNKLGRDIPQPYADQYGVFEGELVNIKQYDESSRRIKPVKPGDSKLLGSVREAIEKTGLTDGMTISFH
HHFREGDFIMNMVLEEIAKMGIKNLSIAPSSIANVHEPLIDHIKNGVVTNITSSGLRDKVGAAISEGLME
NPVVIRSHGGRARAIASGDIHIDVAFLGAPSSDAYGNVNGTKGKATCGSLGYAMIDAKYADQVVILTDNL
VPYPNTPISIPQTDVDYVVTVDAIGDPQGIAKGATRFTKNPKELLIAEYAAKVITNSPYFKEGFSFQTGT
GGASLAVTRFMREAMIKENIKASFALGGITNAMVELLEEELVEKILDVQDFDHPSAVSLGKHAEHYEIDA
NMYASPLSKGAVINQLDTCILSALEVDTNFNVNVMTGSDGVIRGASGGHCDTAFAAKMSLVISPLIRGRI
PTFVDEVNTVITPGTSVDVIVTEVGIAINPNRQDLVDHFKSLNVPKFSIEELKEKAYAIVGTPERIQYGD
KVVALIEYRDGSLMDVVYNV

>YP_008243931.1
MLTSKHHNLNKLVWRYGLTSAAAVLLAFGGGVSSVKADHQKESRAELLQKLTELQSQHQDNVPQTSGIEE
EWTFLGGTSYDSVELYKYLQKMEQYLEKQKKHEEEWKKEIIAGLESDALRGEKGEAGPKGEKGPKGEKGP
KGEKGAQGAQGVRGPKGAKGETGAQGPVGPQGEKGETGATGAQGPQGEAGKDGAQGPVGPQGEKGETGAT
GAQGPQGEAGKDGAQGPVGPQGEKGETGAQGPAGPRGERGPEGKQGPAGEKGEKGETGAPGKDGAPGKDG
VQGPKGDKGETGDRGEKGDTGAQGPVGPQGEAGKPGEKAPEKSPEVTPTPEMPEQPGEKAPEKSKEVTPA
PEKPADKEANQTPERRNDNMAKTPVANNHRRLPATGEQANPFFTAAVAVMTTAGVLAVTKRKENN

>YP_008242584.1
MRMIMEHNYEKFASVYDAVMDDSLYDLWTDFSLRHLPKSKGRNRLLELACGTGIQSVRFAQAGFDVTGLD
LSQDMLTIAEKRAQSAKKKIDFIQGNMLDLSQVGQFDFVTCYSDSICYMQDEVDVGDVFKEVYDVLANDG
IFIFDVHSTYQTDECFPGYSYHENADDFAMVWDTYAGEAPHSVVHELTFFIQEDDGRFSRFDEVHEERTY
ELLTYDILLEQAGFKSFKVYADFEDKEPTKTSKRWFFVAYK

>YP_008242934.1
MKVSIIPEKCIACGLCQTYSSLFDYHDNGIVTFSSSSETSQSICPSDKDAILAVKSCPTKALTLE

>YP_008243117.1
MTKPIIGITANQRLNMALDNLPWSYAPTGFVQAVTQSGGLPLLLPIGDEAAAKTYVSMVDKIILIGGQNV
DPKYYQEEKAAFDDDFSPERDTFELAIIKEAITLKKPILGICRGTQLMNVALGGNLNQHIDSHWQEAPSD
FLSHEMIIEPDSILYPIYGHKTLINSFHRQSLKTVAKDLKVIARDPRDGTIEAVISTNDAIPFLGVQWHP
ELLQGVRDEDLQLFRLFVNDF

>YP_008243078.1
MNSQDLKKRQEKIRNFSIIAHIDHGKSTLADRILEKTETVSSREMQAQLLDSMDLERERGITIKLNAIEL
NYTAKDGETYIFHLIDTPGHVDFTYEVSRSLAACEGAILVVDAAQGIEAQTLANVYLALDNDLEILPVIN

KIDLPAADPERVRHEVEDVIGLDASEAVLASAKAGIGIEEILEQIVEKVPAPTGDVDAPLQALIFDSVYD
AYRGVILQVRIVNGIVKPGDKIQMMSNGKTFDVTEVGIFTPKAVGRDFLATGDVGYVAASIKTVADTRVG
DTVTLANNPAKEALHGYKQMNPMVFAGIYPIESNKYNDLREALEKLQLNDASLQFEPETSQALGFGFRCG
FLGLLHMDVIQERLEREFNIDLIMTAPSVVYHVHTTDGDMIEVSNPSEFPDPTRVASIEEPYVKAQIMVP
QEFVGAVMELSQRKRGDFVTMDYIDDNRVNVIYQIPLAEIVFDFFDKLKSSTRGYASFDYDMSEYRRSQL
VKMDILLNGDKVDALSFIVHKEFAYERGKIIVEKLKKIIPRQQFEVPIQAAIGQKIVARSDIKALRKNVL
AKCYGGDVSRKRKLLEKQKAGKKRMKAIGSVEVPQEAFLSVLSMDDDTKK

>YP_008242573.1
MTFMKKSKWLAAVSVAILSVSALAACGNKNASGGSEATKTYKYVFVNDPKSLDYILTNGGGTTDVITQMV
DGLLENDEYGNLVPSLAKDWKVSKDGLTYTYTLRDGVSWYTADGEEYAPVTAEDFVTGLKHAVDDKSDAL
YVVEDSIKNLKAYQNGEVDFKEVGVKALDDKTVQYTLNKPESYWNSKTTYSVLFPVNAKFLKSKGKDFGT
TDPSSILVNGAYFLSAFTSKSSMEFHKNENYWDAKNVGIESVKLTYSDGSDPGSFYKNFDKGEFSVARLY
PNDPTYKSAKKNYADNITYGMLTGDIRHLTWNLNRTSFKNTKKDPAQQDAGKKALNNKDFRQAIQFAFDR
ASFQAQTAGQDAKTKALRNMLVPPTFVTIGESDFGSEVEKEMAKLGDEWKDVNLADAQDGFYNPEKAKAE
FAKAKEALTAEGVTFPVQLDYPVDQANAATVQEAQSFKQSVEASLGKENVIVNVLETETSTHEAQGFYAE
TPEQQDYDIISSWWGPDYQDPRTYLDIMSPVGGGSVIQKLGIKAGQNKDVVAAAGLDTYQTLLDEAAAIT
DDNDARYKAYAKAQAYLTDSAVDIPVVALGGTPRVTKAVPFSGGFSWAGSKGPLAYKGMKLQDKPVTAKQ
YENAKEKWMKAKAKSNAKYAEKLADHVEK

>YP_008242894.1
MEEAKIPMLKLGPITFNLTLLAVCIVTIAVIFAFVFWASRQMKLKPEGKQTALEYLISFVDGIGEEHLDH
NLQKSYSLLLFTIFLFVAVANNLGLFTKLETVNGYNLWTSPTANLAFDLALSLFITLMVHIEGVRRRGLV
AHLKRLATPWPMTPMNLLEEFTNFLSLAIRLFGNIFAGEVVTGLIVQLANYRVYWWPIAFLVNMAWTAFS
VFISCIQAFVFTKLTATYLGKKVNESEE

>YP_008243928.1
MKNYLSFGMFALLFALTFGTVKPVQAIAGPEWLLGRPSVNNSQLVVSVAGTVEGTNQEISLKFFEIDLTS
RPAQGGKTEQGLRPKSKPLATDKGAMSHKLEKADLLKAIQEQLIANVHSNDGYFEVIDFASDATITDRNG
KVYFADRDDSVTLPTQPVQEFLLSGHVRVRPYRPKAVHNSAERVNVNYEVSFVSETGNLDFTPSLKEQYH
LTTLAVGDSLSSQELAAIAQFILSKKHPDYIITKRDSSIVTHDNDIFRTILPMDQEFTYHIKDREQAYKA
NSKTGIEEKTNNTDLISEKYYILKKGEKPYDPFDRSHLKLFTIKYVDVDTKALLKSEQLLTASERNLDFR
DLYDPRDKAKLLYNNLDAFGIMGYTLTGKVEDNHDDTNRIITVYMGKRPEGENASYHLAYDKDRYTEEER
EVYSYLRDTGTPIPDNPKDK

>YP_008243165.1
MQFLQKWFKHSKKEKLKESSPLSTEIEPSENWEKIPAYIPADKSDYKKVTLITSAIAAGDRPNSQFKVKR
ILKRNPEAITVSLIASSIAAGVYPESQFRVTSIYCKR

>YP_001197641.1
MSKQKVVTNLLLSTAVLGGFLICHAPSVSAETGNLQEIPTESATVSSENQTATAVDGISPTDNQLPTSEE
GDKQAPEEKLVSSDSTREESATQVATDPDAPSQSLSVSGAEISTADSNIQQAEEKLIQESNYYRFDTHQG
EQIRKAVSIEKISADNDEIKWKVTFDRKNWTFSGEGSGYYFILPKGLQLTKIIDSQSEEDIFNKFPKDVN
SEANSGGQPYRFFSREKNANNDERSFESQ

Appendix J: Full list of annotation features and bioinformatics protein annotation tools used to annotate proteins in BPAD200+N+B+AF

.

Tool	Refference	Tool Description	Annotation Feature	Feature Number
Length		Protein Length in (AA)	Length	1
DictyOGlyc	1	The DictyOGlyc server produces neural network predictions for O-glycosylation	DictOGlyc-MaxScore	2
			DictOGlyc-AvgScore	3
			DictOGlyc-No_Sites_Score	4
			DictOGlyc-No_Sites	5
			DictOGlyc-AvgDiff	6
			DictOGlycMaxPosDiff	7
			DictOGlyc_Largest_Diff	8
			DictOGlyc-No_Score_Sites_Length	9
			DictOGlyc-No_Sites_length	10
			DictOGlyc-Average_Length	11
			DictOGlyc-AvgDiff_Length	12
			DictOGlyc-Max_Threshold	13
			DictOGlyc_Average_Threshold	14
			DictOGlyc_Average_Threshold_Length	15
			DictOGlyc_Thr-MaxScore	16
			DictOGlyc_Thr-AvgScore	17
			DictOGlyc_Thr-No_Sites_Score	18
			DictOGlyc_Thr-No_Sites	19
			DictOGlyc_Thr-AvgDiff	20
			DictOGlyc_ThrMaxPosDiff	21
			DictOGlyc_Thr_Largest_Diff	22
			DictOGlyc_Thr-No_Score_Sites_Length	23
			DictOGlyc_Thr-No_Sites_length	24
			DictOGlyc_Thr-Average_Length	25
			DictOGlyc_Thr-AvgDiff_Length	26
			DictOGlyc_Thr-Max_Threshold	27
			DictOGlyc_Thr_Average_Threshold	28
			DictOGlyc_Thr_Average_Threshold_Length	29
			DictOGlyc_Ser-MaxScore	30
			DictOGlyc_Ser-AvgScore	31
			DictOGlyc_Ser-No_Sites_Score	32
			DictOGlyc_Ser-No_Sites	33
			DictOGlyc_Ser-AvgDiff	34
			DictOGlyc_SerMaxPosDiff	35
			DictOGlyc_Ser_Largess_Diff	36
			DictOGlyc_Ser-No_Score_Sites_Length	37
			DictOGlyc_Ser-No_Sites_length	38
			DictOGlyc_Ser-Average_Length	39
			DictOGlyc_Ser-AvgDiff_Length	40
			DictOGlyc_Ser-Max_Sereshold	41
			DictOGlyc_Ser_Average_Sereshold	42
			DictOGlyc_Ser_Average_Sereshold_Length	43
NetAcet	2	NetAcet 1.0 server predicts substrates of N-acetyltransferase A (NatA). The method was trained on yeast data but, as mentioned in the article describing the method, it obtains similar performance values on mammalian substrates acetylated by NatA orthologs.	NetAcet-MaxScore	44
			NetAcet-Average	45
			NetAcet-Count_Score	46
			NetAcet-Count	47
			NetAcet-Average_Length	48
			NetAcet-Count_Score_Length	49
			NetAcet-Count_Length	50
			NetAcet-A-MaxScore	51
			NetAcet-A-Average	52
			NetAcet-A-Count_Score	53
			NetAcet-A-Count	54
			NetAcet-A-Average_Length	55
			NetAcet-A-Count_Score_Length	56
			NetAcet-A-Count_Length	57
			NetAcet-T-MaxScore	58
			NetAcet-T-Average	59
			NetAcet-T-Count_Score	60
			NetAcet-T-Count	61
			NetAcet-T-Average_Length	62
			NetAcet-T-Count_Score_Length	63
			NetAcet-T-Count_Length	64
			NetAcet-G-MaxScore	65
			NetAcet-G-Average	66
			NetAcet-G-Count_Score	67
			NetAcet-G-Count	68
			NetAcet-G-Average_Length	69
			NetAcet-G-Count_Score_Length	70
			NetAcet-G-Count_Length	71
			NetAcet-S-MaxScore	72
			NetAcet-S-Average	73
			NetAcet-S-Count_Score	74
			NetAcet-S-Count	75
			NetAcet-S-Average_Length	76
			NetAcet-S-Count_Score_Length	77
			NetAcet-S-Count_Length	78

NetGlycate	3	NetGlycate 1.0 server predicts glycation of ε amino groups of lysines in mammalian proteins.	NetGlycate-MaxScore	79
			NetGlycate-AvgScore	80
			NetGlycate-Count_score	81
			NetGlycate-Count	82
			NetGlycate-Count_Score_Length	83
			NetGlycate-Count_Length	84
			NetGlycate-Average_Length	85
NetPhosBac	4	PhosBac 1.0 server predicts serine and threonine phosphorylation sites in bacterial proteins	NetPhosBac-MaxScore	86
			NetPhosBac-AvgScore	87
			NetPhosBac-Count_Score	88
			NetPhosBac-Count	89
			NetPhosBac-Count_Score_Length	90
			NetPhosBac-Count_Length	91
			NetPhosBac-Average-Length	92
			NetPhosBac_T-MaxScore	93
			NetPhosBac_T-AvgScore	94
			NetPhosBac_T-Count_Score	95
			NetPhosBac_T-Count	96
			NetPhosBac_T-Count_Score_Length	97
			NetPhosBac_T-Count_Length	98
			NetPhosBac_T-Average-Length	99
			NetPhosBac_S-MaxScore	100
			NetPhosBac_S-AvgScore	101
			NetPhosBac_S-Count_Score	102
			NetPhosBac_S-Count	103
			NetPhosBac_S-Count_Score_Length	104
			NetPhosBac_S-Count_Length	105
			NetPhosBac_S-Average-Length	106
NetPhosYeast	5	NetPhosYeast 1.0 server predicts serine and threonine phosphorylation sites in yeast proteins	NetPhosYeast-MaxScore	107
			NetPhosYeast-AvgScore	108
			NetPhosYeast-Count_Score	109
			NetPhosYeast-Count	110
			NetPhosYeast-Count_Score_Length	111
			NetPhosYeast-Count_Length	112
			NetPhosYeast-Average-Length	113
			NetPhosYeast_T-MaxScore	114
			NetPhosYeast_T-AvgScore	115
			NetPhosYeast_T-Count_Score	116
			NetPhosYeast_T-Count	117
			NetPhosYeast_T-Count_Score_Length	118
			NetPhosYeast_T-Count_Length	119
			NetPhosYeast_T-Average-Length	120
			NetPhosYeast_S-MaxScore	121
			NetPhosYeast_S-AvgScore	122
			NetPhosYeast_S-Count_Score	123
			NetPhosYeast_S-Count	124
			NetPhosYeast_S-Count_Score_Length	125
			NetPhosYeast_S-Count_Length	126
			NetPhosYeast_S-Average-Length	127
ProtParam	6	ProtParam is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence.	ProtParam-Isoelectric	128
			ProtParam-Instability	129
			ProtParam-MolecWeight	130
			ProtParam-Aromaticity	131
			ProtParam-GRAVY	132
			ProtParam-PercHelix	133
			ProtParam-PercTurn	134
			ProtParam-PercSheet	135
			ProtParam-PercAlanine	136
			ProtParam-PercCysteine	137
			ProtParam-PercAsparticAcid	138
			ProtParam-PercGlutamicAcid	139
			ProtParam-PercPhenylalanine	140
			ProtParam-PercGlycine	141
			ProtParam-PercHistidine	142
			ProtParam-PercIsoleucine	143
			ProtParam-PercLysine	144
			ProtParam-PercLeucine	145
			ProtParam-PercMethionine	146
			ProtParam-PercAsparagine	147
			ProtParam-PercProline	148
			ProtParam-PercGlutamine	149
			ProtParam-PercArginine	150
			ProtParam-PercSerine	151
			ProtParam-PercThreonine	152
			ProtParam-PercValine	153
			ProtParam-PercTryptophan	154
			ProtParam-PercTyrosine	155
NetPhosK	7	The NetPhosK 1.0 server produces	NetPhosK-MaxScore	156
			NetPhosK-AvgScore	157

		neural network predictions of kinase specific eukaryotic protein phosphorylation sites	NetPhosK-Count	158
			NetPhosK-Count_Length	159
			NetPhosK-Average_Length	160
			NetPhosK-T-MaxScore	161
			NetPhosK-T-AvgScore	162
			NetPhosK-T-Count	163
			NetPhosK-T-Count_Length	164
			NetPhosK-T-Average_Length	165
			NetPhosK-S-MaxScore	166
			NetPhosK-S-AvgScore	167
			NetPhosK-S-Count	168
			NetPhosK-S-Count_Length	169
			NetPhosK-S-Average_Length	170
			NetPhosK-Y-MaxScore	171
			NetPhosK-Y-AvgScore	172
			NetPhosK-Y-Count	173
			NetPhosK-Y-Count_Length	174
			NetPhosK-Y-Average_Length	175
YinOYang	8	The YinOYang WWW server produces neural network predictions for O-ß-GlcNAc attachment sites in eukaryotic protein sequences	YinOYang-MaxScore	176
			YinOYang-AvgScore	177
			YinOYang-MaxDiff1	178
			YinOYang-AvgDiff1	179
			YinOYang-MaxDiff2	180
			YinOYang-AvgDiff2	181
			YinOYang-Count	182
			YinOYang-Count_Length	183
			YinOYang-Average_Length	184
			YinOYang-abs_diff_thresh1	185
			YinOYang-abs_diff_thresh2	186
			YinOYang-Average-Difference1_Length	187
			YinOYang-Average-Difference2_Length	188
			YinOYang-T-MaxScore	189
			YinOYang-T-AvgScore	190
			YinOYang-T-MaxDiff1	191
			YinOYang-T-AvgDiff1	192
			YinOYang-T-MaxDiff2	193
			YinOYang-T-AvgDiff2	194
			YinOYang-T-Count	195
			YinOYang-T-Count_Length	196
			YinOYang-T-Average_Length	197
			YinOYang-T-abs_diff_thresh1	198
			YinOYang-T-abs_diff_thresh2	199
			YinOYang-T-Average-Difference1_Length	200
			YinOYang-T-Average-Difference2_Length	201
			YinOYang-S-MaxScore	202
			YinOYang-S-AvgScore	203
			YinOYang-S-MaxDiff1	204
			YinOYang-S-AvgDiff1	205
			YinOYang-S-MaxDiff2	206
			YinOYang-S-AvgDiff2	207
			YinOYang-S-Count	208
			YinOYang-S-Count_Length	209
			YinOYang-S-Average_Length	210
			YinOYang-S-abs_diff_thresh1	211
			YinOYang-S-abs_diff_thresh2	212
			YinOYang-S-Average-Difference1_Length	213
			YinOYang-S-Average-Difference2_Length	214
LipoP	9	The LipoP 1.0 server produces predictions of lipoproteins and discriminates between lipoprotein signal peptides, other signal peptides and n-terminal membrane helices in Gram-negative bacteria.Note: Although LipoP 1.0 has been trained on sequences from Gram-negative bacteria only, the following paper reports that it has a good performance on sequences from Gram-positive bacteria also	LipoP_Signal_MaxScore	215
			LipoP_Signal_AvgScore	216
			LipoP_Signal_Count	217
			LipoP_Signal_Avg_Length	218
			LipoP_Signal_Count_Length	219
			LipoP_CleavI_MaxScore	220
			LipoP_CleavI_AvgScore	221
			LipoP_CleavI_Count	222
			LipoP_CleavI_Avg_Length	223
			LipoP_CleavI_Count_Length	224
			LipoP_CleavII_MaxScore	225
			LipoP_CleavII_AvgScore	226
			LipoP_CleavII_Count	227
			LipoP_CleavII_Avg_Length	228
			LipoP_CleavII_Count_Length	229
			LipoP_TMH_MaxScore	230
			LipoP_TMH_AvgScore	231
			LipoP_TMH_Count	232
			LipoP_TMH_Avg_Length	233
			LipoP_TMH_Count_Length	234
			LipoP_SPI_MaxScore	235
			LipoP_SPI_AvgScore	236

			LipoP_SPI_Count	237
			LipoP_SPI_Avg_Length	238
			LipoP_SPI_Count_Length	239
			LipoP_CYT_MaxScore	240
			LipoP_CYT_AvgScore	241
			LipoP_CYT_Count	242
			LipoP_CYT_Avg_Length	243
			LipoP_CYT_Count_Length	244
TargetP	10		TargetP-SecretScore	245
		TargetP 1.1 predicts the subcellular location of eukaryotic proteins.	TargetP-MitoScore	246
			TargetP-OtherScore	247
			TargetP-SecretFlag	248
			TargetP-MitoFlag	249
			TargetP-OtherFlag	250
			TargetP_RC_Score	251
NetNGlyc	Website	The NetNglyc server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons. http://www.cbs.dtu.dk/services/NetNG	NetNGlyc-MaxScore	252
			NetNGlyc-AvgScore	253
			NetNGlyc-Count	254
			NetNGlyc-Count_Score	255
			NetNGlyc-Count_Length	256
			NetNGlyc-Count_Score_Length	257
			NetNGlyc-Average_Length	258
NetOGlyc	11		NetOGlyc-Max-G	259
		The NetOglyc server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins.	NetOGlyc-Max-I	260
			NetOGlyc-Avg-G	261
			NetOGlyc-Avg-I	262
			NetOGlyc-Count_Score	263
			NetOGlyc-Count	264
			NetOGlyc-Average_G_Length	265
			NetOGlyc_Average_I_Length	266
			NetOGlyc_Count_Score_Length	267
			NetOGlyc_Count_Length	268
			NetOGlyc-T-Max-G	269
			NetOGlyc-T-Max-I	270
			NetOGlyc-T-Avg-G	271
			NetOGlyc-T-Avg-I	272
			NetOGlyc-T-Count_Score	273
			NetOGlyc-T-Count	274
			NetOGlyc-T-Average_G_Length	275
			NetOGlyc-T_Average_I_Length	276
			NetOGlyc-T_Count_Score_Length	277
			NetOGlyc-T_Count_Length	278
			NetOGlyc-S-Max-G	279
			NetOGlyc-S-Max-I	280
			NetOGlyc-S-Avg-G	281
			NetOGlyc-S-Avg-I	282
			NetOGlyc-S-Count_Score	283
			NetOGlyc-S-Count	284
			NetOGlyc-S-Average_G_Length	285
			NetOGlyc-S_Average_I_Length	286
			NetOGlyc-S_Count_Score_Length	287
			NetOGlyc-S_Count_Length	288
ProP	12	ProP 1.0 server predicts arginine and lysine propeptide cleavage sites in eukaryotic protein sequences using an ensemble of neural networks.	PropFurin-MaxScore	289
			PropFurin-AvgScore	290
			PropFurin-Count_Score	291
			PropFurin-Count_Score_Length	292
			PropFurin-Average_Length	293
			PropFurin-Count	294
			PropFurin-Count_Length	295
			PropGeneral-MaxScore	296
			PropGeneral-AvgScore	297
			PropGeneral-Count_Score	298
			PropGeneral-Count_Score_Length	299
			PropGeneral-Average_Length	300
			PropGeneral-Count	301
			PropFurinGeneral-Count_Length	302
Bepipred	13	BepiPred 1.0 server predicts the location of linear B-cell epitopes using a combination of a hidden Markov model and a propensity scale	Bepipred-Max_Score	303
			Bepipred-Average_Score	304
			Bepipred-Count	305
			Bepipred-Count_Length	306
			Bepipred-Average_Score_Length	307
TMHMM	Website	Prediction of transmembrane helices in proteins. http://www.cbs.dtu.dk/services/TMHM	TMHMM-No_aa_in_Helices	308
			TMHMM-StartAAcount	309
			TMHMM-Count	310
			TMHMM-No_aa_in_Helix_Length	311
			TMHMM-Count_Length	312
HMMTOP	14	HMMTOP is an automatic server for predicting transmembrane helices	HMMTOP-Count	313
			HMMTOP-Final_Localization	314

		and topology of proteins	HMMTOP-Count_Length	315
PSORTb	15	Bacterial Subcellular Localization predictor	PSORTb-ProbCytoMem	316
			PSORTb-ProbCytoplasm	317
			PSORTb-ProbPeriplasm	318
			PSORTb-ProbExtraCell	319
			PSORTb-ProbOuterMem	320
			PSORTb-ProbCellWall	321
SignalP	16	SignalP server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms	SignalP-NN-MaxC	322
			SignalP-NN-MaxY	323
			SignalP-NN-MaxS	324
			SignalP-NN-AvgS	325
			SignalP-NN_LengthAvgS	326
			SignalP-D	327
			SignalP-HMM-MaxC	328
			SignalP-HMM-ProbS	329
CBTOPE	17	B Cell epitope predictor	CBTOPE-MaxScore	330
			CBTOPE-AvgScore	331
			CBTOPE-Count	332
			CBTOPE-Count_Length	333
			CBTOPE-Average_Length	334
CSS-Palm	18	Post translational modification predictor of S-palmitoylation (also called as thoacylation or S-acylation).	CSS-Max_Score	335
			CSS-Number_of_Sites	336
			CSS-Average_Score	337
			CSS-CorCount	338
			CSS-Average_Score_Length	339
			CSS-Max_difference	340
			CSS-AverageDiff	341
			CSS_Length_AvDiff	342
GPS-ARM	19	D box and Ken box protein degradation motif predictor	GPS-ARM_Dbox-MaxScore	343
			GPS-ARM_Dbox-AvgScore	344
			GPS-ARM_Dbox-Count	345
			GPS-ARM_Dbox_MaxDiff	346
			GPS-ARM_Dbox_AverageDiff	347
			GPS-ARM_Dbox-Count_Length	348
			GPS-ARM_Dbox-Average_Length	349
			GPS-ARM_Dbox-AverageDiff_Length	350
			GPS-ARM_KENBox-MaxScore	351
			GPS-ARM_KENBox-AvgScore	352
			GPS-ARM_KENBox-Count	353
			GPS-ARM_KENBox_MaxDiff	354
			GPS-ARM_KENBox_AverageDiff	355
			GPS-ARM_KENBox-Count_Length	356
			GPS-ARM_KENBox-Average_Length	357
			GPS-ARM_KenBox-AverageDiff_Length	358
			GPS-ARM_DorKen-1ifmoreKENsites1ifmoreDsites	359
GPS-CCD	20	Calpains Ca2+ dependent cysteine proteases control numerous biological processes, gene expression, cell death, cell cycle progression. Predicts calpain cleavage motifs in proteins.	CCDmax_score	360
			CCD_count	361
			CCD_max_cutoff	362
			CCD_av_score	363
			CCD_av_diff	364
			CCD_CorCount	365
			CCD_CorAvg	366
			CCD_CorAvgDiff	367
GPS-MBA	21	Predicts HLA-DQ8 and I-Ag7 epitopes. (MHC binding of 2 specific MHC class 2 haplotypes that have been linked to causing auto immune	MBAAgI7_max_score	368
			MBAAgI7_count	369
			MBAAgI7_max_cutoff	370
			MBAAgI7_av_score	371
			MBAAgI7_av_diff	372
			MBAAgI7_CorCount	373
			MBAAgI7_CorAvg	374
			MBAAgI7_Cor_avg_diff	375
			MBADQ8_max_score	376
			MBADQ8_count	377
			MBADQ8_max_cutoff	378
			MBADQ8_av_score	379
			MBADQ8_av_diff	380
			MBADQ8_CorCount	381
			MBADQ8_CorAvg	382
			MBADQ8_Cor_avg_diff	383
GPS-Polo	22	Predicts phosphoprotein-binding domains (PPBDs) which have been shown important in facilitating a lot of reversible phosphorylation post translational modifications.	PoloBind_max_score	384
			PoloBind_count	385
			PoloBind_max_diff_cutoff	386
			PoloBind_av_score	387
			PoloBind_av_diff	388
			PoloBind_CorCount	389
			PoloBind_CorAvg	390
			PoloBind_Cor_avG_diff	391
			PoloPhosphorylation_max_score	392
			PoloPhosphorylation_count	393

			PoloPhosphorylation_max_cutoff	394
			PoloPhosphorylation_av_score	395
			PoloPhosphorylation_av_diff	396
			PoloPhosphorylation_CorCount	397
			PoloPhosphorylation_CorAvg	398
			PoloPhosphorylation_Cor_avG_diff	399
GPS-PUP	23	Prokaryotic ubiquitin-like proteint (PUP) identified as a tagging (ubiquitin like) system in prokaryotes. Predicts PUP binding sites.	PUP_max_score	400
			PUP_count	401
			PUP_max_cutoff_diff	402
			PUP_av_score	403
			PUP_av_diff	404
			PUP_CorCount	405
			PUP_CorAvg	406
			PUP_Cor_av_diff	407
GPS-SNO	24	Predicts Nitrosylation sites, S-nitrosylation has been proposed to modulate a protein's stability, activity and trafficking.	SNO_max_score	408
			SNO_count	409
			SNO_max_cutoffdiff	410
			SNO_av_score	411
			SNO_av_diff	412
			SNO_Length_Count	413
			SNO_Length_Average	414
			SNO_Length_AvDiff	415
GPS SUMO	25	Small ubiquitin-like modifiers play an essential role in the regulation of a variety of biological processes, including gene expression, DNA repai, chromosome assembly and cellular signaling. Predicts	GPS_SUMO_interaction_Max_Score	416
			GPS_SUMO_interaction_Number_of_Sites	417
			GPS_SUMO_interaction_Max_Diff_Cutoff	418
			GPS_SUMO_interaction_average_Diff_Cutoff	419
			GPS_SUMO_interaction_Average_Score	420
			GPS_SUMO_interaction_Count_length	421
			GPS-SUMO_interaction_Average_length	422
			GPS-SUMO_interaction_DiffAverage_length	423
			GPS_SUMO_Sumoylation_Max_Score	424
			GPS_SUMO_Sumoylation_Number_of_Sites	425
			GPS_SUMO_Sumoylation_Max_Diff_Cutoff	426
			GPS_SUMO_Sumoylation_average_Diff_Cutoff	427
			GPS_SUMO_Sumoylation_Average_Score	428
			GPS_SUMO_Sumoylation_Count_length	429
			GPS-SUMO_Sumoylation_Average_length	430
			GPS-SUMO_Sumoylation_DiffAverage_length	431
Net Chop	26	The NetChop server produces neural network predictions for cleavage of the human proteasome.	Net_Chop_MaxScore	432
			Net_Chop_Average	433
			Net_Chop_Number_of_Sites	434
			Net_Chop_CorCount	435
			Net_Chop_CorAverage	436
NetsurfP	27	NetSurfP predicts the surface accessibility and seondary structure of amino acids in an amino acid	NetsurfP_ASA_Exposed-MaxScore	437
			NetsurfP_ASA_Exposed-AvgScore	438
			NetsurfP_ASA_Exposed-Count	439
			NetsurfP_ASA_Exposed-Count_Length	440
			NetsurfP_ASA_Exposed-Average_Length	441
			NetsurfP_ASA_Exposed_MaxDiff	442
			NetsurfP_ASA_Exposed_AverageDiff	443
			NetsurfP_ASA_Exposed-AverageDiff_Length	444
			NetsurfP_ASA_Buried-MaxScore	445
			NetsurfP_ASA_Buried-AvgScore	446
			NetsurfP_ASA_Buried-Count	447
			NetsurfP_ASA_Buried-Count_Length	448
			NetsurfP_ASA_Buried-Average_Length	449
			NetsurfP_ASA_Buried_MaxDiff	450
			NetsurfP_ASA_Buried_AverageDiff	451
			NetsurfP_ASA_AverageDiff_Length	452
			NetsurfP_ASA_Buried_or_Exposed-1ifmoreExposedsites1ifmorel	453
			NetsurfP_RSA_Exposed-MaxScore	454
			NetsurfP_RSA_Exposed-AvgScore	455
			NetsurfP_RSA_Exposed-Count	456
			NetsurfP_RSA_Exposed-Count_Length	457
			NetsurfP_RSA_Exposed-Average_Length	458
			NetsurfP_RSA_Exposed_MaxDiff	459
			NetsurfP_RSA_Exposed_AverageDiff	460
			NetsurfP_RSA_Exposed-AverageDiff_Length	461
			NetsurfP_RSA_Buried-MaxScore	462
			NetsurfP_RSA_Buried-AvgScore	463
			NetsurfP_RSA_Buried-Count	464
			NetsurfP_RSA_Buried-Count_Length	465
			NetsurfP_RSA_Buried-Average_Length	466
			NetsurfP_RSA_Buried_MaxDiff	467
			NetsurfP_RSA_Buried_AverageDiff	468
			NetsurfP_RSA_AverageDiff_Length	469
			NetsurfP_RSA_Buried_or_Exposed-1ifmoreExposedsites1ifmore	470
Phobius	28	A combined transmembrane topology and signal peptide predictor.	Phobius-Numbe_of_TransMem	471
			Phobius-SignalPeptide_predicted	472

			Phobius-Numbe_of_TransMem_Length	473
SPAAN	29	Predicts if a proteins is an adhesion	Pad-value	474
PickPocket	30	PickPocket server predicts binding of	PickPocket-Max	475
		peptides to any known MHC molecule	PickPocket-Avg	476
		using positiion specific weight	PickPocket-Count_Score	477
			PickPocket-Average_score	478
			PickPocket_Count_Score_Length	479
			PickPocket-A-Max	480
			PickPocket-A-Avg	481
			PickPocket-A-Count_Score	482
			PickPocket-A-Average_score	483
			PickPocket-A_Count_Score_Length	484
			PickPocket-B-Max	485
			PickPocket-B-Avg	486
			PickPocket-B-Count_Score	487
			PickPocket-B-Average_score	488
			PickPocket-B_Count_Score_Length	489
NetMhcPan	31	NetMHCpan server predicts binding	NetMhcPan-MaxMHC	490
		of peptides to any known MHC	NetMhcPan-MaxRank	491
		molecule using artificial neural	NetMhcPan-AvgMHC	492
			NetMhcPan-AvgRank	493
			NetMhcPan-Count_Score	494
			NetMhcPan-Average_MHC_Length	495
			NetMhcPan_Count_Score_Length	496
			NetMhcPan_Average_Comb_Length	497
			NetMhcPan_WeakBinders	498
			NetMhcPan_StrongBinders	499
			NetMhcPan_WeakBinders_Length	500
			NetMhcPan_StrongBinders_Length	501
			NetMhcPan-A-MaxMHC	502
			NetMhcPan-A-MaxRank	503
			NetMhcPan-A-AvgMHC	504
			NetMhcPan-A-AvgRank	505
			NetMhcPan-A-Count_Score	506
			NetMhcPan-A-Average_MHC_Length	507
			NetMhcPan-A_Count_Score_Length	508
			NetMhcPan-A_Average_Comb_Length	509
			NetMhcPan-A_WeakBinders	510
			NetMhcPan-A-StrongBinderst	511
			NetMhcPan-A-WeakBinders_Length	512
			NetMhcPan_a_StrongBinders_Length	513
			NetMhcPan-B-MaxMHC	514
			NetMhcPan-B-MaxRank	515
			NetMhcPan-B-AvgMHC	516
			NetMhcPan-B-AvgRank	517
			NetMhcPan-B-Count_Score	518
			NetMhcPan-B-Average_MHC_Length	519
			NetMhcPan-B_Count_Score_Length	520
			NetMhcPan-B_Average_Comb_Length	521
			NetMhcPan-B_WeakBinders	522
			NetMhcPan-B_StrongBinders	523
			NetMhcPan-B_WeakBinders_Length	524
			NetMhcPan-B_StrongBinders_Length	525

AnnotationFeature name consisits of Tool first followed by setting or specific site, such as amino acid, followed by the variable name.

Annotation Tool was used in pad136

Feature Tag	Meaning
MaxScore	Maximum score of all predicted sites from a tool
AvgScore	Average score of all predicted sites from a tool
No_Sites_Score	Number of all predicted sites output from a tool
No_Sites	Number of predicted sites output from a tool
AvgDiff	Average of the Differnce between all predited sites and the cutoff/threshold score from a tool
Max_Threshold	Maximum threshold score from all predicted sites
MaxPosDiff	Maximum difference between all predicted site and threshold as absolute values.
_Length	The variable normalized by protein length (divided by the proteins length).
Largest_Diff	The maximum difference between all predicted sites and threshold
Isoelectric	Predicted isoelectric point
Instability	Predicted instability index
MolecWeight	Predicted Molecular Weight
Aromaticity	Predicted protein aromaticity
GRAVY	Predicted grand average of hydropathy
PercTurn	Percent of protein predicted to be a part of a turn
PercHelix	Percent of protein predicted to be a part of an alpha-helix
PercSheet	Percent of protein predicted to be a part of a beta-sheet
PercAlanine	Percentage of residues that are alanine
PercCysteine	Percentage of residues that are cysteine

PercAsparticAcid	Percentage of residues that are aspartic acid
PercGlutamicAcid	Percentage of residues that are glutamic acid
PercPhenylalanine	Percentage of residues that are phenylalanine
PercGlycine	Percentage of residues that are glycine
PercHistidine	Percentage of residues that are histidine
PercIsoleucine	Percentage of residues that are isoleucine
PercLysine	Percentage of residues that are lysine
PercLeucine	Percentage of residues that are leucine
PercMethionine	Percentage of residues that are methionine
PercAsparagine	Percentage of residues that are asparagine
PercProline	Percentage of residues that are proline
PercGlutamine	Percentage of residues that are glutamine
PercArginine	Percentage of residues that are arginine
PercSerine	Percentage of residues that are serine
PercThreonine	Percentage of residues that are threonine
PercValine	Percentage of residues that are valine
PercTryptophan	Percentage of residues that are tryptophan
PercTyrosine	Percentage of residues that are tyrosine

Appendix K: Listing the SBA positive protein dataset and the rank of recall in the *Neisseria meningitidis MC58* proteome as well as the classification prediction for each protein.

Id	Rank of Recall in *MC58* Proteome	Prediction
NP_274477.1	2	BPA
NP_275083.1	12	BPA
NP_274574.1	27	BPA
NP_274809.1	30	BPA
NP_274028.1	45	BPA
NP_273394.1	48	BPA
NP_273507.1	65	BPA
NP_273150.1	67	BPA
NP_274618.1	75	BPA
NP_274002.1	179	BPA
NP_274087.1	187	BPA
NP_273508.1	188	BPA
NP_274980.1	209	BPA
NP_274547.1	213	BPA
NP_274441.1	249	BPA
NP_273967.1	565	non-BPA
NP_274440.1	608	non-BPA
NP_273462.1	823	non-BPA

Shading represents that the protein was a known BPA included in the classifier's (BPAD200+N+B+AF) training data as well as a curated SBA positive protein.

Int Health 2015; **7**: 85–89
doi:10.1093/inthealth/ihv002

REVIEW

The promise of reverse vaccinology

Ashley I. Heinson, Christopher H. Woelk* and Marie-Louise Newell

Faculty of Medicine, University of Southampton, Southampton, UK

*Corresponding author: Tel: +44 203 8120 5928; E-mail: C.H.Woelk@soton.ac.uk

Received 22 October 2014; revised 6 January 2015; accepted 7 January 2015

Reverse vaccinology (RV) is a computational approach that aims to identify putative vaccine candidates in the protein coding genome (proteome) of pathogens. RV has primarily been applied to bacterial pathogens to identify proteins that can be formulated into subunit vaccines, which consist of one or more protein antigens. An RV approach based on a filtering method has already been used to construct a subunit vaccine against *Neisseria meningitidis* serogroup B that is now registered in several countries (Bexsero). Recently, machine learning methods have been used to improve the ability of RV approaches to identify vaccine candidates. Further improvements related to the incorporation of epitope-binding annotation and gene expression data are discussed. In the future, it is envisaged that RV approaches will facilitate rapid vaccine design with less reliance on conventional animal testing and clinical trials in order to curb the threat of antibiotic resistance or newly emerged outbreaks of bacterial origin.

Keywords: Bacterial pathogen, Epidemic, Reverse vaccinology, Subunit vaccine

Since the introduction of vaccination into western medicine in 1796 with the smallpox vaccine developed by Edward Jenner, vaccines have risen to be the most cost-effective way of controlling infectious diseases.[1,2] Conventional vaccinology remains an active area of research, where a pathogen is cultured in the laboratory and proteins are purified from the culture, to be used as potential vaccine candidates. However, the availability of sequence data for entire bacterial genomes, and by translation their proteomes, led to the advent of reverse vaccinology (RV). Using bioinformatics approaches, RV selects proteins that could be potential candidates for subunit vaccines. When RV first emerged the advantages of this approach were heralded to result in rapid vaccine formulation, the ability to identify vaccine candidates from bacteria that could not be cultured, and all at reduced costs compared to conventional approaches. One drawback of RV is that it cannot be used to predict polysaccharides or lipids, which are often included in vaccines as active compounds.[3] RV approaches have been primarily focused on bacterial pathogens.[1,4–9] Although the development of antibiotics led to a tremendous decrease in the mortality and morbidity associated with bacterial infections in the past, the emergence of antibiotic resistance represents a pressing medical problem resulting in a renewed interest in vaccine development. It is estimated that in the United States alone at least 2 million people become infected with antibiotic resistant bacteria per annum, with an associated cost in excess of US$20 billion dollars to the US healthcare system each year.[10] The growing worldwide antibiotic resistance problem was echoed by a European intergovernmental conference (Antibiotic resistance action to promote new technologies, Birmingham, UK, 12–13 December 2005), which called for increased funding and research into vaccines since they represent the best option to counteract the rise of antibiotic resistant bacteria.[11]

RV identifies proteins, primarily in bacterial pathogens, which may be used to formulate subunit vaccines. Such vaccines are composed of one or more purified components, commonly proteins (i.e., subunit), and not the whole microorganism. Therefore, subunit vaccines induce protective immunity without the risk of side effects or immune reactions caused by other parts of pathogenic bacteria.[12] Additional benefits of this type of vaccine are that they incorporate proteins in their most native form, facilitating correct protein folding and reconstitution of conformational (non-linear, discontinuous) epitopes. It has been estimated that up to 90% of B cell epitopes from native proteins are conformational, and host antibodies that bind to conformational epitopes may have a larger neutralizing effect than linear epitopes.[13] By incorporating more than one protein into a subunit vaccine, it is possible to derive immunity to more than one strain or serotype of a bacterial pathogen.[12] Potential drawbacks of subunit vaccines are their moderate immunogenicity, limited half-lives in vivo, and requirement for adjuvants to generate robust immune responses.[12,14] Some examples of prominent subunit vaccines currently in clinical use include Recombivax and Engerix[15] for hepatitis B virus, BOOSTRIX[16,17] for use in adolescents and adults as a booster immunization against diphtheria, tetanus and pertussis, and Bexsero[18] for use in infants and adolescents against *Neisseria meningitidis* serogroup B (MenB).

The first clinically registered vaccine developed by an RV approach was for the bacterial pathogen MenB. Vaccines against

other meningococcal serogroups (A, C, Y and W135) utilized capsular polysaccharides to stimulate protective immunity. However, the capsular polysaccharide from MenB was a poor immunogen and may elicit autoantibodies in humans.[19,20] Pizza and colleagues,[19] in the laboratory of Rino Rappuoli, employed an early RV method based on a filtering approach to identify antigenic proteins for incorporation into a subunit vaccine against MenB. The protein-coding genes in the genome of MenB (strain MC58) were identified and then initially filtered based on predicted subcellular localization using bioinformatics programs (e.g., PSORTB, ProDom and Blocks databases, Table 1). A total of 570 proteins were identified as surface-exposed of which 350 could be cloned and expressed in *Escherichia coli*. These 350 proteins were filtered down to a subset of 7 with confirmed surface expression (ELISA and FACS analysis) and the ability to induce bactericidal antibodies in mice, as well as sufficient sequence conservation across other strains and serogroups of *N. meningitidis*. Subsequently, Novartis Vaccines formulated a subunit vaccine with the brand name Bexsero from three of these proteins: Factor H binding protein, Neisserial adhesion A and Neisseria heparin binding antigen, in conjunction with a detergent extracted outer membrane vesicle (DOMV) suspension from MenB strain NZ98/254.[21,22] The primary antigenic component of the DOMV suspension is the previously described immunogenic MenB protein Porin A (PorA). Bexsero has now been licensed in the EU, Canada and Australia.[18] Over 500 000 doses of Bexsero have been shipped to the UK and approved for use by the National Health Service (NHS), which currently aims to incorporate Bexsero into the childhood vaccination program.[23,24] In summary, the development of Bexsero represents truly pioneering work, but the promise of rapid vaccine development through an RV approach must be tempered since the time from inception to license was more than a decade.

Previous RV studies utilizing filtering approaches did not share standard methodologies and used different sets of bioinformatics tools and programs.[19,25–31] The New Enhanced Reverse Vaccinology Environment (NERVE) was the first attempt to provide an automated RV approach in a downloadable tool, and by way of novelty, included an autoimmunity filter, which was a protein blast (blastp) that compared the predicted antigens to human proteins. If they were significantly similar there would be a chance that the antigens may stimulate an autoimmune response or be weakly immunogenic and thus were filtered out.[7] Vaxign enhanced the functionality of NERVE by creating a user-friendly web interface and incorporated binding predictions to MHC class I and II epitopes.[9] Recently, the Jenner-Predict server has been shown to outperform these other methods, although it still relies on a filtering approach whereby all cytosolic proteins are removed in the first step.[6] The main limitations of all filtering approaches is that a large number of surface-orientated proteins are identified as potential vaccine candidates, which then require extensive laboratory characterization, and proteins from other subcellular localizations with the potential to be vaccine candidates may be overlooked.

Machine learning has been incorporated into RV approaches in order to reduce the number of potential vaccine candidates and evaluate the entire proteome for antigenic proteins.[5,32–34] For example, Flower and Doytchinova[32] constructed a training data set of 100 positive and 100 negative antigens curated from the literature. These training data were annotated with auto-cross covariance transformations, which annotated the proteins for basic characteristics associated with size, hydrophobicity and weight. A regression-based classifier was able to discriminate antigens from non-antigens with 82% accuracy, but these training data were later shown to be contaminated with non-bacterial sequence data. It should be noted that machine-learning methods are reliant on the quality of the training data and will give varying results depending on how a protective antigen is defined. To address this, our own work, Bowman et al.,[5] used a rigorous definition of a bacterial protective antigen (BPA) by defining a BPA as a bacterial protein that results in significant ($p<0.05$) protection (i.e., bacterial load reduction or survival assay) in an animal model following immunization and subsequent challenge with the bacterial pathogen. Using this definition a support vector machine (SVM) classifier was able to discriminate antigens from non-antigens with 92% accuracy. In addition, a larger training data set and annotation of protein data using bioinformatics tools with greater biological relevance contributed to this increased accuracy. In total, 122 features were derived from 19 bioinformatic tools for protein annotation (Table 1). Furthermore, Bowman et al,[5] were able to demonstrate that the trained SVM classifier could significantly recall known bacterial protective antigens curated from the literature when embedded in the background of entire bacterial proteomes.

Machine learning based RV approaches may be further improved by incorporating epitope-binding information and gene expression data for the annotation of proteins used in training data sets. The incorporation of epitope information into RV approaches is in its infancy.[5,9] Traditionally, epitope information has not been included into RV approaches due to the high rate of false positive predictions.[6] However, it is hoped that machine learning techniques with their tremendous capability to separate signal from noise will be able to utilize predictions from epitope-binding software such as Bepipred[5] and SVMHC[35] (Table 1) as the sensitivity and specificity of such programs improves. Another way in which RV can be improved is by taking into account gene expression data.[6] With the cost of microarray and next generation sequencing (RNA-Seq) technologies constantly dropping, the accumulation of bacterial gene expression data in public repositories is rapidly growing. This information could be used to annotate training data to enhance machine-learning based RV approaches. One would envisage that bacterial proteins that are not expressed during infection would not be presented to the host immune system and thus not be viable vaccine candidates.

The examination of a bacterial outbreak in the past can be used to illustrate the potential utility of an RV approach in an epidemiological setting. At the American Legion convention in Philadelphia, 1976, a bacterial outbreak of unknown origin resulted in a pneumonia-like illness in 182 individuals resulting in 29 deaths.[36] The illness was dubbed Legionnaire's disease, but it took 6 months to identify the pathogenic cause as the gram-negative bacterium *Legionella pneumophila*.[36] This previously uncharacterized species of bacteria was originally thought not to infect humans. If such an outbreak occurred today, high throughput sequencing of nucleic acids in biological specimens from infected individuals could be used to rapidly diagnose the causative pathogen and also to assemble the complete pathogen genome or transcriptome. The protein-coding genes from the sequenced genome could then be subjected to an RV approach in order to identify candidates for rapid subunit vaccine formulation. As it happens, *L. pneumophila* is not presently transmitted from humans to humans[37] and can be treated with antibiotics (e.g.,

Table 1. Bioinformatic tools utilized in reverse vaccinology approaches

Software/ program	Main feature	URL/PMID
BepiPred[a]	Predicts linear B-cell epitopes	http://www.cbs.dtu.dk/services/BepiPred/
Blast[b,d]	Basic Local Alignment Search Tool, searches sequences of DNA/ proteins and compares similarity as an alignment	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome
Blocks databases	Database containing multiple alignments of conserved regions in protein families	http://blocks.fhcrc.org/help/
DictyOGlyc[a]	Predicts glycosylation	http://www.cbs.dtu.dk/services/DictyOGlyc/
HMMTOP[a,b,c,d]	Predicts transmembrane helices	http://www.enzim.hu/hmmtop/
Jenner-Predict server	Most recent and accurate filtering pipeline published for reverse vaccinology	http://14.139.240.55/vaccine/validation.html
LipoP[a]	Predicts lipoproteins & signal peptides	http://www.cbs.dtu.dk/services/LipoP/
NERVE	New Enhanced Reverse Vaccinology Environment, was the first automated RV approach and also was the first to include an autoimmunity filter	16848907
NetAcet[a]	Acetylation	http://www.cbs.dtu.dk/services/NetAcet/
NetGlycate[a]	Predicts glycation	http://www.cbs.dtu.dk/services/NetGlycate/
NetNGlyc[a]	Predicts N-glycosylation	http://www.cbs.dtu.dk/services/NetNGlyc/
NetOGlyc[a]	Predicts O-glycosylation	http://www.cbs.dtu.dk/services/NetOGlyc/
NetPhosBac[a]	Predicts phosphorylation	http://www.cbs.dtu.dk/services/NetPhosBac-1.0/
NetPhosK[a]	Predicts phosphorylation	http://www.cbs.dtu.dk/services/NetPhosK/
NetPhosYeast[a]	Predicts phosphorylation	http://www.cbs.dtu.dk/services/NetPhosYeast/
OrthoMCL[c]	Predicts orthologs and conserved sequences in more than one genome	http://www.orthomcl.org/orthomcl/
Pfam[d]	Searchable database which stores protein family information	http://pfam.xfam.org/
ProDom	Database of protein domain families	http://prodom.prabi.fr/prodom/current/html/home.php
ProP[a]	Predicts cleavage sites	http://www.cbs.dtu.dk/services/ProP/
ProtParam[a]	Basic protein stats such as molecular weight amino acid composition as examples	http://ca.expasy.org/tools/protparam.html
PSORTB[a,b,c,d]	Predicts subcellular localization of proteins	http://www.psort.org/psortb/
SignalP[a]	Predicts the presence of signal peptides	http://www.cbs.dtu.dk/services/SignalP/
SPAAN[b,c]	Predicts adhesins and adhesin like proteins using neural networks	15374866
SVMHC	Machine learning techniques (Support Vector Machine) to predict MHC class I and class II binding peptides	http://www.sbc.su.se/~pierre/svmhc/
TargetP[a]	Subcellular localization	http://www.cbs.dtu.dk/services/TargetP/
TMHMM[a]	Transmembrane helices	http://www.cbs.dtu.dk/services/TMHMM/
Vaxign	First online interfaced RV tool. It was the first to include MHC type I and type II epitopes to make predictions	http://www.violinet.org/vaxign/
Vaxitope[c]	MHC class I and class II prediction	20671958
YinOYang[a]	O-linked beta-N-acetylglucosamine	http://www.cbs.dtu.dk/services/YinOYang/

MHC: major histocompatibility complex; PMID: PubMed identifier.
Superscripts identify bioinformatic tools that were implemented in RV approaches as follows:
[a] This tool was used in Bowman et al. (2010) Machine Learning approach to RV. Eighteen tools are listed in the table of which the 19th tool referenced in the paper but not presented in the table is protein length.
[b] This tool is implemented in the NERVE pipeline.
[c] This tool is implemented in the Vaxign pipeline.
[d] This tool is implemented in the Jenner-Predict pipeline.

fluoroquinolones and macrolides).[38] However, there is currently no vaccine available to protect against *L. pneumophila* infection and should this pathogen develop antibiotic resistance, RV could be used to identify candidates for subunit vaccines.

In summary, RV has already led to the formulation of a vaccine that is utilized in current clinical practice (i.e., Bexsero), and continued improvements will further enhance the utility of this approach. The speed with which an RV approach could result in

a clinically viable vaccine has not been fully realized due to the conventional requirement for extensive animal testing and clinical trials. However, one can envisage a future where increased understanding of pathogen and host immune responses leads to highly accurate in silico modeling. This will hopefully reduce the amount of unsuccessful candidates that proceed to clinical trials and also the amount of conventional testing required to prove the efficacy and safety of vaccine candidates. Furthermore, RV approaches may be complemented with other methods to construct subunit vaccines. Specifically, OMV suspensions (i.e., Bexsero) and structural vaccinology approaches can be used to enhance the immunogenicity of subunit vaccines developed through an RV approach. For Gram-negative bacterial pathogens, vaccines based on OMV preparations from circulating strains have been used to successfully protect susceptible individuals during geographically confined outbreaks.[39] Structural vaccinology has been used to improve the biochemical characteristics of vaccine candidates and could be used to increase the immunogenicity of protein antigens identified by RV.[40] The current Ebola outbreak, although due to a virus, has demonstrated the requirement for the rapid development of therapeutics in an epidemiological setting. Should a highly pathogenic outbreak of unknown bacterial origin emerge in the future, RV will hopefully be in a position to aid the rapid formulation of subunit vaccines that can be administered to infected individuals following appropriate clinical evaluation, as the epidemiological situation requires.

Authors' contributions: AIH and CHW contributed equally to this manuscript; AIH, CHW and MLN discussed and planned the content of the manuscript; AIH and CHW wrote the preliminary draft of the manuscript, which was then edited by all co-authors. All authors read and approved the final manuscript. CHW is the guarantor of the paper.

Funding: This work was supported by the European Commission who have funded AIH's graduate studies through a Marie Curie Career Integration Grant with acronym "RevVac" and grant agreement number [PCIG13-GA2013-618334].

Competing interests: None declared.

Ethical approval: Not required.

References

1 Delany I, Rappuoli R, Seib KL. Vaccines, reverse vaccinology, and bacterial pathogenesis. Cold Spring Harb Perspect Med 2013;3: a012476.

2 Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. Immunity 2010;33:530–41.

3 Kanampalliwar A, Rajkumar S, Girdhar A, Archana T. Reverse vaccinology: basics and applications. J Vaccines Vaccin 2013;4:194.

4 Talukdar S, Zutshi S, Prashanth KS et al. Identification of potential vaccine candidates against *Streptococcus pneumoniae* by reverse vaccinology approach. Appl Biochem Biotechnol 2014;172:3026–41.

5 Bowman BN, McAdam PR, Vivona S et al. Improving reverse vaccinology with a machine learning approach. Vaccine 2011;29:8156–64.

6 Jaiswal V, Chanumolu SK, Gupta A et al. Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. BMC Bioinformatics 2013;14:211.

7 Vivona S, Bernante F, Filippini F. NERVE: new enhanced reverse vaccinology environment. BMC Biotechnology 2006;6:35.

8 Esposito S, Castellazzi L, Bosco A et al. Use of a multicomponent, recombinant, meningococcal serogroup B vaccine (4CMenB) for bacterial meningitis prevention. Immunotherapy 2014;6:395–408.

9 He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010;2010:297505.

10 U.S. Department of Health and Human Services CfD, Control and Prevention. Antibiotic resistance threats in the United States, 2013. Atlanta: U.S. Department of Health and Human Services, Centers for Disease, Control and Prevention; 2013.

11 Finch R, Hunter PA. Antibiotic resistance--action to promote new technologies: report of an EU Intergovernmental Conference held in Birmingham, UK, 12–13 December 2005. J Antimicrob Chemother 2006;58(Suppl 1):i3–22.

12 Hansson M, Nygren PA, Stahl S. Design and production of recombinant subunit vaccines. Biotechnol Appl Biochem 2000;32:95–107.

13 Huang J, Honda W. CED: a conformational epitope database. BMC Immunol 2006;7:7.

14 Coffman RL, Sher A, Seder RA. Vaccine adjuvants: putting innate immunity to work. Immunity 2010;33:492–503.

15 Michel ML, Tiollais P. Hepatitis B vaccines: protective efficacy and therapeutic potential. Pathol Biol (Paris) 2010;58:288–95.

16 Jones T. Boostrix (GlaxoSmithKline). IDrugs 2005;8:656–61.

17 Scott LJ, McCormack PL. Reduced-antigen, combined diphtheria, tetanus, and acellular pertussis vaccine, adsorbed (boostrix (R)): a guide to its use as a single-dose booster immunization against pertussis. BioDrugs 2013;27:75–81.

18 Friend R, Staton T. Novartis Bexsero® meningitis B vaccine receives clinical recommendation for use in infants and adolescents in Australia. FierceVaccines, 17 March 2014. http://www.fiercepharma.com/press-releases/novartis-bexsero-meningitis-b-vaccine-receives-clinical-recommendation-use [accessed 10 September 2014].

19 Pizza M, Scarlato V, Masignani V et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000;287:1816–20.

20 Giuliani MM, Adu-Bobie J, Comanducci M et al. A universal vaccine for serogroup B meningococcus. Proc Natl Acad Sci USA 2006;103:10834–9.

21 Bai X, Findlow J, Borrow R. Recombinant protein meningococcal serogroup B vaccine combined with outer membrane vesicles. Expert Opin Biol Ther 2011;11:969–85.

22 Tani C, Stella M, Donnarumma D et al. Quantification by LC-MS(E) of outer membrane vesicle proteins of the Bexsero(R) vaccine. Vaccine 2014;32:1273–9.

23 NHS. Meningitis B Vaccine. http://www.nhs.uk/Conditions/vaccinations/Pages/meningitis-B-vaccine.aspx [accessed 10 May 2014].

24 Wise J. Meningitis B vaccine to be introduced in UK after U turn on its cost effectiveness. BMJ 2014;348:2327.

25 Finco O, Bonci A, Agnusdei M et al. Identification of new potential vaccine candidates against *Chlamydia pneumoniae* by multiple screenings. Vaccine 2005;23:1178–88.

26 Gamberini M, Gomez RM, Atzingen MV et al. Whole-genome analysis of *Leptospira interrogans* to identify potential vaccine candidates against leptospirosis. FEMS Microbiol Lett 2005;244:305–13.

27 Ariel N, Zvi A, Grosfeld H et al. Search for potential vaccine candidate open reading frames in the Bacillus anthracis virulence plasmid pXO1: in silico and in vitro screening. Infect Immun 2002;70:6817–27.

28 Gat O, Grosfeld H, Ariel N et al. Search for *Bacillus anthracis* potential vaccine candidates by a functional genomic-serologic screen. Infect Immun 2006;74:3987–4001.

29 Wizemann TM, Heinrichs JH, Adamou JE et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect Immun 2001;69: 1593–8.

30 Lei B, Liu M, Chesney GL, Musser JM. Identification of new candidate vaccine antigens made by *Streptococcus pyogenes*: purification and characterization of 16 putative extracellular lipoproteins. J Infect Dis 2004;189:79–89.

31 Bhatia V, Sinha M, Luxon B, Garg N. Utility of the *Trypanosoma cruzi* sequence database for identification of potential vaccine candidates by in silico and in vitro screening. Infect Immun 2004;72:6245–54.

32 Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 2007;8:4.

33 Goodswen SJ, Kennedy PJ, Ellis JT. Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. Bioinformatics 2014;30:2381–3.

34 Goodswen SJ, Kennedy PJ, Ellis JT. A novel strategy for classifying the output from an in silico vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. BMC Bioinformatics 2013;14:315.

35 Donnes P, Kohlbacher O. SVMHC: a server for prediction of MHC-binding peptides. Nucleic Acids Res 2006;34(Web Server issue):W194–7.

36 Altman LK. In Philadelphia 30 years ago, an eruption of illness and fear. http://www.nytimes.com/2006/08/01/health/01docs.html?pagewanted=all&_r=0 [accessed 10 November 2014].

37 United States Environmental Protection Agency. Legionella: Drinking water fact sheet. Washington DC: United States Environmental Protection Agency Office of Wate; 2000.

38 Yu LL, Hu BJ, Huang SL et al. Activity of macrolides and fluoroquinolones against intracellular *Legionella pneumophila* [in Chinese]. Zhonghua Jie He He Hu Xi Za Zhi 2011;34:409–12.

39 Acevedo R, Fernandez S, Zayas C et al. Bacterial outer membrane vesicles and vaccine applications. Front Immunol 2014;5:121.

40 Dormitzer PR, Grandi G, Rappuoli R. Structural vaccinology starts to deliver. Nat Rev Microbiol 2012;10:807–13.

International Journal of
Molecular Sciences

MDPI

Article

Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology

Ashley I. Heinson [1], Yawwani Gunawardana [1], Bastiaan Moesker [1,†], Carmen C. Denman Hume [2,‡], Elena Vataga [3], Yper Hall [4], Elena Stylianou [5], Helen McShane [5], Ann Williams [4], Mahesan Niranjan [6] and Christopher H. Woelk [1,*]

1 Faculty of Medicine, University of Southampton, Southampton SO17 1BJ, UK; a.heinson@soton.ac.uk (A.I.H.); y.p.gunawardana@soton.ac.uk (Y.G.); bastiaanmoesker@gmail.com (B.M.)
2 London School of Hygiene and Tropical Medicine (LSHTM), Department of Pathogen Molecular Biology, London WC1E 7HT, UK; carmen.denman@gmail.com
3 iSolutions, University of Southampton, Southampton SO17 1BJ, UK; e.vataga@soton.ac.uk
4 Public Health England, National Infection Service, Porton Down Salisbury, SP4 0JG, UK; yper.hall@phe.gov.uk (Y.H.); ann.rawkins@phe.gov.uk (A.W.)
5 The Jenner Institute, University of Oxford, Oxford OX3 7DQ, UK; elena.stylianou@ndm.ox.ac.uk (E.S.); helen.mcshane@ndm.ox.ac.uk (H.M.)
6 Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK; mn@ecs.soton.ac.uk
* Correspondance: c.h.woelk@soton.ac.uk; Tel.: +23-8120-5928
† Since completing this work, has moved to Thermo Fisher Scientific, Inchinnan Business Park, 3 Fountain Drive, Paisley PA4 9RF, UK.
‡ Since completing this work, has moved to King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Kingdom of Saudi Arabia.

Academic Editor: Susanna Esposito
Received: 15 November 2016; Accepted: 17 January 2017; Published: 1 February 2017

Abstract: Reverse vaccinology (RV) is a bioinformatics approach that can predict antigens with protective potential from the protein coding genomes of bacterial pathogens for subunit vaccine design. RV has become firmly established following the development of the BEXSERO® vaccine against *Neisseria meningitidis* serogroup B. RV studies have begun to incorporate machine learning (ML) techniques to distinguish bacterial protective antigens (BPAs) from non-BPAs. This research contributes significantly to the RV field by using permutation analysis to demonstrate that a signal for protective antigens can be curated from published data. Furthermore, the effects of the following on an ML approach to RV were also assessed: nested cross-validation, balancing selection of non-BPAs for subcellular localization, increasing the training data, and incorporating greater numbers of protein annotation tools for feature generation. These enhancements yielded a support vector machine (SVM) classifier that could discriminate BPAs (n = 200) from non-BPAs (n = 200) with an area under the curve (AUC) of 0.787. In addition, hierarchical clustering of BPAs revealed that intracellular BPAs clustered separately from extracellular BPAs. However, no immediate benefit was derived when training SVM classifiers on data sets exclusively containing intra- or extracellular BPAs. In conclusion, this work demonstrates that ML classifiers have great utility in RV approaches and will lead to new subunit vaccines in the future.

Keywords: reverse vaccinology; machine learning; support vector machine; bacterial protective antigen; bacterial pathogen

1. Introduction

Reverse vaccinology (RV) is a form of vaccine research that uses bioinformatics approaches to identify putative vaccine candidates in the protein coding genomes of bacteria (i.e., proteomes). The most successful RV study to date was by Pizza and colleagues [1] and led to a subunit vaccine against *Neisseria meningitidis* (*N. meningitidis*) serogroup B [2]. Initially, open reading frames in the genome of *N. meningitidis* serogroup B were identified and bioinformatics programs (psortB [3], ProDom [4], and Blocks databases [5]) used to predict the subcellular localization for every protein in the proteome. Extracellular predicted proteins were then cloned and expressed as recombinant proteins, purified and had their predicted surface expression confirmed through techniques such as enzyme-linked imunosorbent assay (ELISA) and fluorescence activated cell-sorting (FACS). Proteins with confirmed surface expression were then used to generate antibodies in the serum of immunized mice and the bactericidal activity of this serum was assessed. Examples that met these criteria were then screened for conservation across multiple MenB strains and their suitability for manufacturing in bulk was assessed [6]. In the final subunit vaccine, BEXSERO®, three proteins were selected for incorporation, (i.e., Factor H binding protein, Neisserial adhesion A, and Niesseria heparin binding antigen) along with a detergent extracted outer membrane vesicle (DOMV). The BEXSERO® vaccine is now licensed in over 35 countries and has already had an impact on the mortality and morbidity associated with *N meningitidis* serogroup B [7].

The RV approach of Pizza et al. [1] may be classified as a "filtering" approach, i.e., the organism's proteome is passed through a series of filters until a subset of proteins are identified that represent potential vaccine candidates. Several utilities have been developed to implement filtering approaches to RV, for example Violin [8], Jenner Predict [9], and Ivax [10]. Drawbacks of filtering methods include the necessity of assessing large numbers of candidates in the laboratory and potential candidates with predicted subcellular localization other than extracellular (e.g., cytoplasmic) are discarded [11]. The latter is a significant limitation since proteins predicted to be cytoplasmic or of unknown localizations have been shown to confer significant levels of protection in animal models [12–19]. Machine learning (ML) approaches to RV circumvent these problems since they do not discard such proteins but are able to successfully model the entire proteome of a bacterial species and rank predicted antigens for their likelihood of being a vaccine candidate [20,21].

The first ML study in the field of RV was published by Doytchinova and Flower [21], in which a training dataset was generated of 100 known antigens through a literature curation that defined a known antigen as a protein (or part of a protein) that, "has been shown to induce a protective response in an appropriate animal model after immunization". A negative training dataset was constructed by randomly sampling 100 proteins or non-antigens from the same bacterial species that corresponded to each known antigen in the positive training dataset. The proteins in this training dataset were annotated with auto cross-covariance (ACC) transformations, which reflect hydrophobicity, molecular size, and polarity. The annotated proteins were used to train a classifier based on discriminant analysis by partial least squares (DA-PLS), which was able to achieve an accuracy of 82% when distinguishing non-antigens from known antigens. In an extension to Doytchinova and Flower's [21] work, our initial RV study [20] focused exclusively on bacterial protective antigens (BPAs) defined as, "a whole protein that led to significant protection ($p < 0.05$) in an animal model (i.e., bacterial load reduction or survival assay) following immunization and subsequent challenge with the bacterial pathogen". Focusing on bacterial proteins the size of the training data (136 BPAs and 136 non-BPAs) was increased and then annotated with biologically-relevant protein annotation tools (e.g., PSORTb [3], LipoP [22], and Bepipred [23]) for the training of support vector machine (SVM) classifiers. This work showed that higher accuracies were obtained when using SVMs (i.e., 92%) when separating BPAs and non-BPAs in the training data and when recalling known antigens in the background of entire bacterial proteomes [20].

Building on our previous work in the field of ML applied to RV, this current study implemented a nested approach to cross-validation, removed an artificial bias associated with the selection of

non-BPAs for the negative training data, increased the size of the training data by approximately a third, and incorporated new protein annotation tools to model different aspects of immunogenicity (e.g., T-cell epitope prediction and Adhesin prediction [24]). The resulting SVM classifier was used to demonstrate that a significant signal for protection could be captured through the literature curation of BPAs as assessed through comparisons to randomly-permutated data.

2. Results

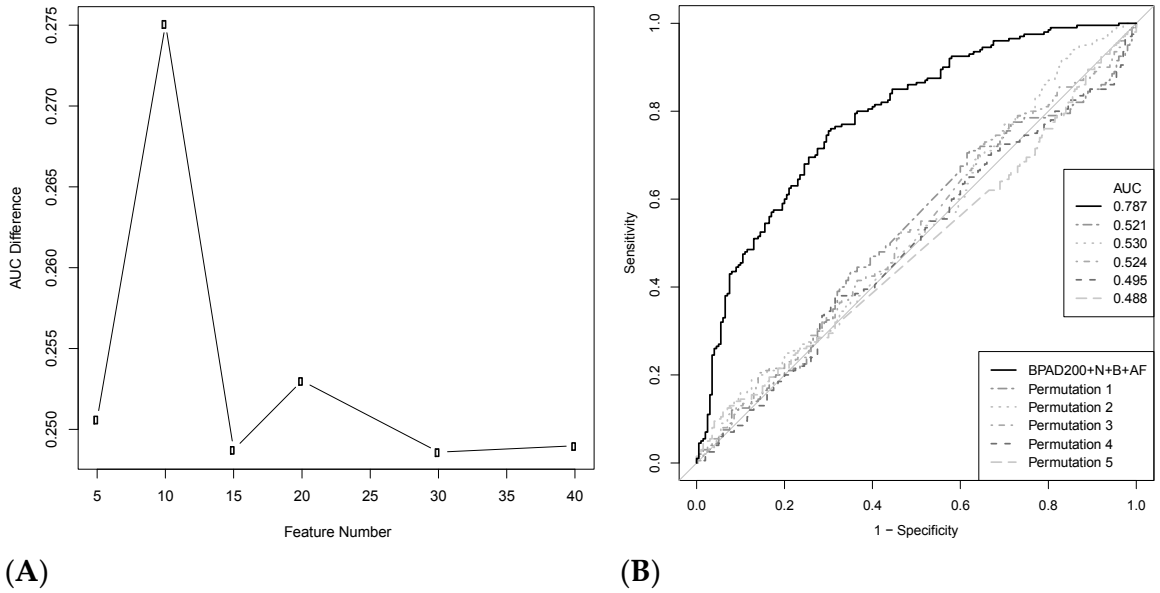
2.1. Permutation Analysis Reveals a Strong Protective Signal for BPAs Curated from the Literature

To enhance the biological relevance of ML classifiers for RV, several modifications were made to our previous approach [20]. These included adopting a nested (N) cross-validation approach, balancing (B) the negative training data for subcellular localization, increasing the size of the positive training dataset to 200 BPAs, and increasing the number of protein annotation tools by 15 to derive a total of 525 annotation features (AF) (full list Table S1). This resulted in a training data set with the designation BPAD200+N+B+AF consisting of 200 BPAs and 200 non-BPAs annotated with 525 features. Non-BPAs in the negative training data were selected by randomly sampling proteins from the same bacterial proteome for the relevant BPA pair while ensuring the same subcellular localization.

Initially, the BPAD200+N+B+AF dataset was used to determine if a signal for protective efficacy had been captured when curating BPAs from the literature record. SVM classifiers were trained in a nested leave tenth out cross-validation (LTOCV) approach to discriminate BPAs from non-BPAs in BPAD200+N+B+AF and in five additional data sets where the labels (i.e., BPA or non-BPA) were randomly permutated (permutation testing) and, thus, contained no biological information. Greedy backward feature elimination was used to select different numbers of features for SVM classifiers. For each feature set, the difference in the AUC calculated from receiver operator characteristic (ROC) curves between BPAD200+N+B+AF and the randomly permutated data was calculated (Figure 1A). An SVM classifier with the 10 most informative features led to a significant separation in AUC (average p-value 1.13×10^{-12}, DeLong test [25]) between BPAD200+N+B+AF and the randomly permutated data (Figure 1B). These results clearly demonstrated that the literature curation of BPAs captured a strong protective signal and that 10 features was the optimal number for discriminating BPAs from non-BPAs. Therefore, SVM classifiers containing 10 features were used to evaluate the changes in classification accuracies of each of the modifications to our RV approach.

2.2. A Nested Approach Has a Significant Impact on the Ability of SVMs to Classify BPAs

Having demonstrated that the SVM classifier BPAD200+N+B+AF trained on datasets curated from the literature had captured a biological signal reflective of protective antigens, the multiple modifications made to our RV approach were assessed in a stepwise manner. The starting point for this assessment was the BPAD136 classifier from our previous work [20], consisting of 136 BPAs and 136 non-BPAs. The following modifications to BPAD136 were then assessed in turn: nested cross-validation (BPAD136+N), balancing non-antigen selection for subcellular localization (BPAD136+N+B), increasing the size of training data, (BPAD200+N+B) and, finally, incorporating additional features (BPAD200+N+B+AF). Our previous work [20] had not implemented a truly nested cross-validated approach and it was hypothesized that previous SVM classifiers may have overfit the data. This was indeed the case, as reflected by a significant reduction in AUC (p-value = 9.69×10^{-5}, DeLong test [25]) when migrating from an overfit (BPAD136, AUC = 0.962) to a nested (BPAD136+N, AUC = 0.861) approach (Figure 2). Therefore, the implementation of a nested approach to cross-validation is recommended for RV studies and enables a better estimation of the performance of SVM classifiers.



(**A**) (**B**)

Figure 1. (**A**) Plot of the difference in area under the curve (AUC) between the support vector machine (SVM) classifier BPAD200+N+B+AF versus randomly permutated data with increasing feature numbers. SVM classifiers were trained to discriminate bacterial protective antigens (BPAs) and non-BPAs in BPAD200+N+B+AF and receiver operator characteristic (ROC) curves generated from a nested leave tenth out cross-validation approach for different numbers of features selected by greedy backward feature elimination. Five iterations were performed to assess the random breakage of ties during greedy backward feature elimination and AUC was averaged across iterations for each feature set. This analysis was then repeated for five datasets where the BPA and non-BPA labels were randomly permutated and average AUC calculated across randomly permutated data sets for each feature set; (**B**) ROC curves for the average of the five iterations of the 10 feature SVM classifier derived from BPAD200+N+B+AF (black solid line) and from each of the five randomly-permutated datasets (dotted grey lines).

2.3. Correcting a Bias in the Selection of Negative Training Data Impacts SVM Classification of BPAs

The subcellular localizations predicted by PSORTb [3] were compared between the positive (BPAs) and negative (non-BPAs) training data for BPAD136. This demonstrated that the negative training data had a larger proportion of proteins predicted to be located in the cytoplasm (Figure 3A,B). This resulted from the random selection of non-BPAs for the negative training data. Specifically, for each BPA, a non-BPA was randomly selected from the proteome of the same bacterial species. Since the majority of proteins in any given bacterial proteome are predicted to be cytoplasmic, random sampling led to a disproportionate number of non-BPAs with this subcellular location in the negative training data. To correct this bias, a new negative training dataset was generated, where each non-BPA was selected to match not only the bacterial species of the corresponding BPA, but also its subcellular localization (Figure 3C). Removing this bias in subcellular localization decreased the ability of the SVM classifier to discriminate between BPAs and non-BPAs as reflected in a significant reduction in AUC (p-value of 3.44×10^{-5}, DeLong test [25]) when comparing BPAD136+N to BPAD136+N+B (Figure 2). The performance of the SVM classifier is reduced because it can no longer utilize differences in subcellular localization to discriminate BPAs from non-BPAs. This is reflected by the removal of features related to bacterial subcellular localization (e.g., PSORTb and SignalP) from the top 10 features utilized by the classifier.

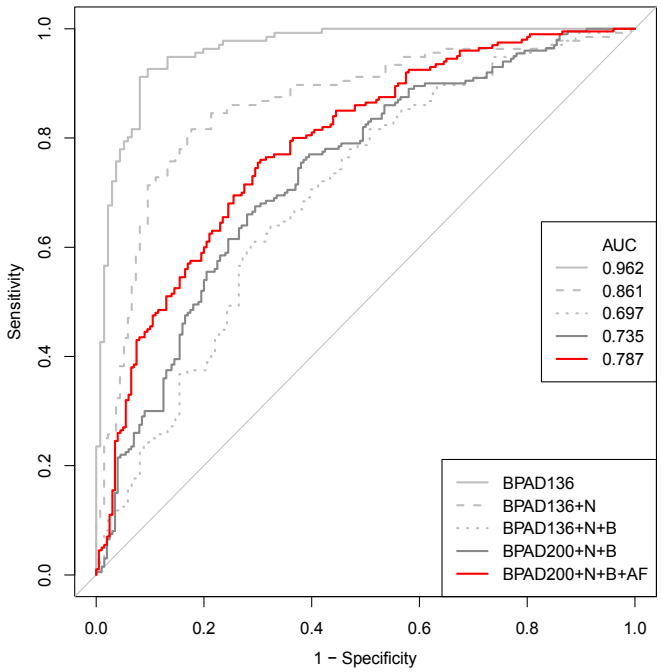


Figure 2. ROC curves were generated from SVM classifiers utilizing 10 features selected by greedy backward feature elimination in a LTOCV approach. Averages were plotted across five iterations of SVM classifiers implemented to randomly break ties resulting from the greedy backward feature elimination procedure. The benchmark to assess these modifications was a non-nested, non-balanced training data set of 136 BPAs and 136 non-BPAs annotated with 122 features from 19 protein annotation tools (BPAD136) [20]. Subsequent modifications were added in a stepwise fashion and included: a nested cross-validation approach (BPAD136+N), balanced selection of non-BPAs for predicted subcellular localization (BPAD136+N+B), increased size of training data (BPAD200+N+B), and additional features (525 total) derived from an increased number of protein annotation tools (BPAD200+N+B+AF).

2.4. Increasing the Size of the Training Data Has a Positive Impact on the Ability of SVMs to Classify BPAs

A literature curation yielded 64 new BPAs that were significantly protective ($p < 0.05$) in an animal model following immunization and subsequent challenge with the bacterial pathogen. These BPAs were used to expand the positive training data set from 136 to 200 BPAs and paired with non-BPAs balanced for subcellular localization to form the training data set designated BPAD200+N+B. Increasing the size of the training data in this manner led to improvements in AUC (Figure 2, 0.697 to 0.735), although this change was not statistically significant.

2.5. Increasing the Number of Protein Annotation Tools Enhances the Ability of SVMs to Classify BPAs

The effect of incorporating new annotation features derived from protein annotation tools (Table S1) with biological relevance for immunogenicity (e.g., Spaan [26], MHCpan [27], GPS-MBA [28]) on the ability of SVM classifiers to distinguish BPAs from non-BPAs was examined. Compared to our original approach [20], an additional 15 protein annotation tools were assessed resulting in a training data set (BPAD200+N+B+AF) annotated with a total of 525 features. The incorporation of these additional features resulted in an SVM classifier with an increased AUC (0.787) when compared to BPAD200+N+B. Although this increase did not attain significance, it should be stressed that when the incorporation of additional BPAs and features were considered together (i.e., BPAD136+N+B to BPAD200+N+B+AF) there is a significant increase in AUC ($p = 2.11 \times 10^{-2}$, DeLong test [25]). Finally, the top 10 features selected by greedy backward feature elimination for the discrimination of BPAs and non-BPAs in the BPAD200+N+B+AF dataset contained new features primarily related to T-cell epitopes (Table 1).

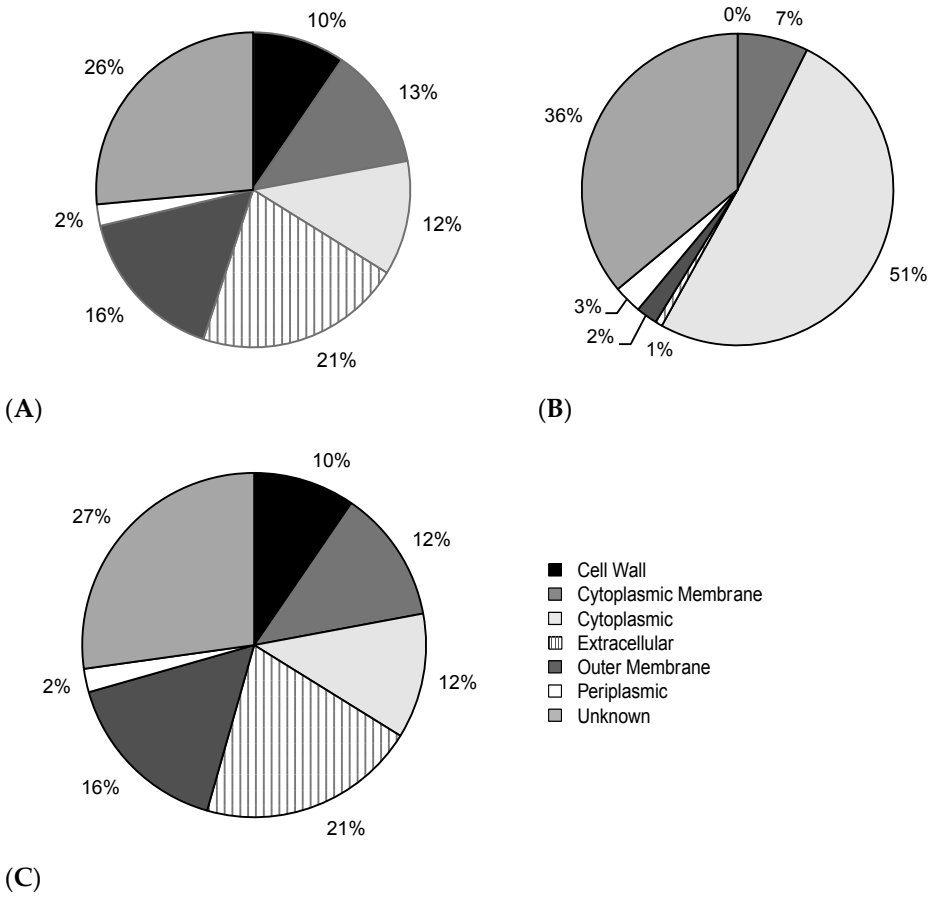


Figure 3. Pie charts showing subcellular localization as predicted by PSORTb [3] for the numbers of BPAs and non-BPAs in the following subsets of the BPAD136 dataset. (**A**) positive training data (i.e., 136 BPAs); (**B**) negative training data (i.e., 136 non-BPAs); and (**C**) negative training data balanced for subcellular localization (i.e., 136 non-BPAs).

Table 1. The top 10 annotation features selected by greedy backward feature elimination for discrimination of BPAs from non-BPAs by the SVM classifier trained on the BPAD200+N+B+AF data set.

Rank	Feature	Name of Bioinformatics Tool	Protein Annotation Tool Type	Correlated with BPA or Non-BPA
1	LipoP_Signal_Avr_Length	LipoP	Lipoprotein	BPA
2	YinOYang-T-Count	YinOYang	Glycosylation	BPA
3	NetPhosK-S-Count	NetPhosK	Phosphorylation	BPA
4	LipoP_SPI_Avr_Length	LipoP	Lipoprotein	BPA
5	**NetMhcPan-B-AvgRank**	**NetMhc**	T-Cell Epitope predictor (MHC Class II)	BPA
6	TargetP-SecretFlag	TargetP	Subcellular Compartmentalisation—In Eukaryotic Cells	BPA
7	YinOYang-Average-Difference1_Length	YinOYang	Glycosylation	Non-BPA
8	**MBAAgl7_CorCount**	**GPS-MBA**	T-Cell Epitope predictor	BPA
9	**PickPocket-Average_score**	**PickPocket**	MHC Peptide Binding	Non-BPA
10	PropFurin-Count_Score	ProP	Cleavage Sites—In Eukaryotic Cells	BPA

Features in bold represent those derived from protein annotation tools that were added in this study compared to our previous approach [20]. For a full list of bioinformatics tools utilized in this study and the annotation features derived from them please see Table S1.

2.6. Intracellular and Extracellular BPAs Utilize Different Features for Classification

An unsupervised approach was used to further explore the biological signals in the features derived from protein annotation tools used to annotate the 200 BPAs curated from the literature record. Hierarchical clustering of the 142 BPAs not predicted to have an unknown subcellular localization by psortB [3] using all 525 annotation features revealed two main groups that primarily corresponded to predicted subcellular localization, i.e., intra- or extracellular (Figure 4). This suggests that intracellular and extracellular BPAs may have fundamentally different biological properties. Therefore, it was hypothesized that greater accuracies would be achieved when SVM classifiers were trained separately on intra- and extracellular BPAs. To test this hypothesis, two training datasets were constructed from BPAD200: intracellular BPAD51 (iBPAD51) consisting of 51 BPAs and 51 non-BPAs with subcellular localization predicted as intracellular, and extracellular BPAD91 (eBPAD91) with 91 BPAs and 90 non-BPAs with subcellular localization predicted as extracellular. A matching non-BPA balanced for subcellular localization could not be found that met the inclusion criteria for ACF35754.1 (a BPA from *Salmonella enterica subsp. enterica serovar Paratyphi A*), detailed in the Methods (Section 4.1) and, thus, the negative training data was reduced by one.

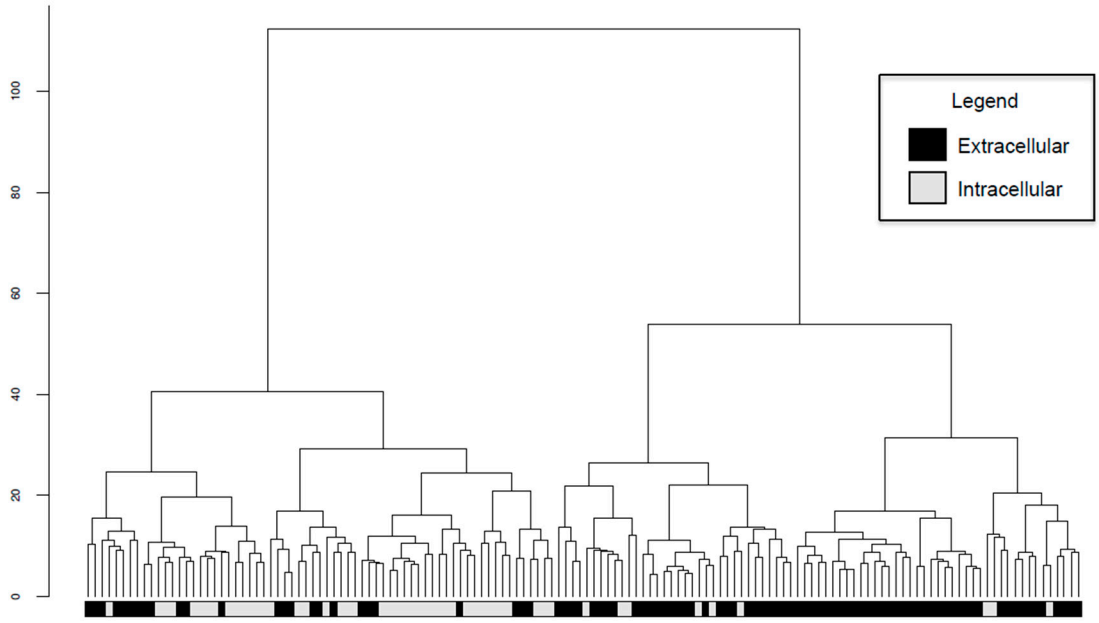
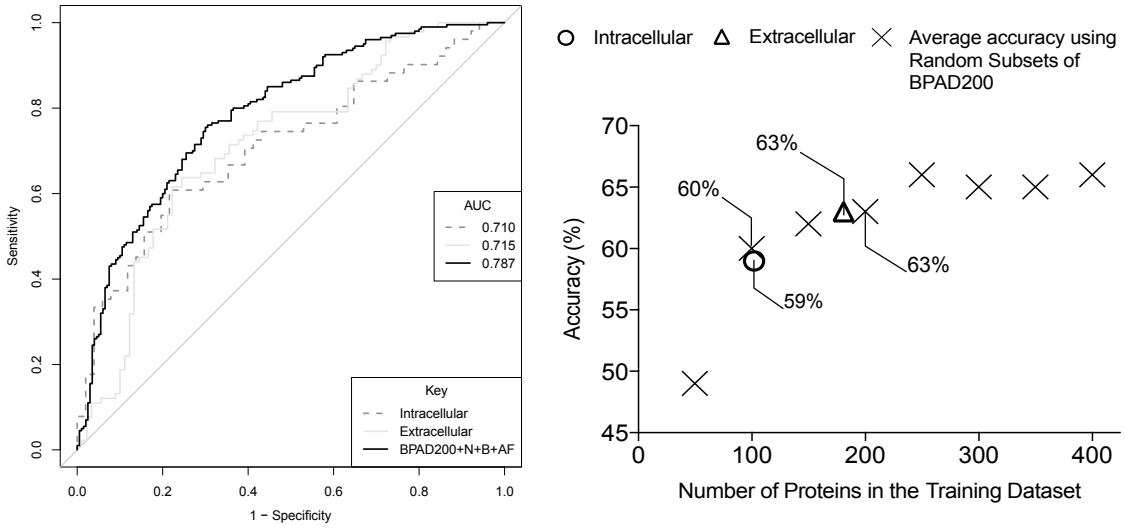


Figure 4. Hierarchical clustering of 142 BPAs from BPAD200+N+B+AF using all 525 annotation features, distances between BPAs were calculated using Euclidean metrics and then clustered using the Ward algorithm. White labels at the branch tips refer to BPAs with subcellular localization predicted by PSORTb [3] as intracellular (i.e., cytoplasm or cytoplasmic membrane) and black labels as extracellular BPAs (i.e., extracellular, periplasmic, outer membrane, cell wall).

Contrary to expected findings, SVM classifiers trained on iBPAD51 and eBPAD91 had lower AUC values compared to those trained on BPAD200 (Figure 5A). However, SVM classifiers had been trained on datasets of different sizes and this might explain the superior performance of BPAD200, which was the largest dataset. Therefore, SVM classifiers were trained on randomly selected subsets of BPAD200 of decreasing size (Figure 5B). This facilitated a better comparison of intra- (trained on iBPAD51) or extracellular (trained on eBPAD91) SVM classifiers, versus those trained using BPAs and non-BPAs from all subcellular localizations (BPAD200). However, the accuracies derived from iBPAD51 and eBPAD91 were similar to data sets of similar size consisting of BPAs from all subcellular localizations (Figure 5B). This suggests there is no immediate benefit to training separate SVM classifiers to recognize intra- or extracellular BPAs.



(**A**) (**B**)

Figure 5. (**A**) ROC curves obtained from SVM classifiers trained to distinguish BPAs from non-BPAs in the following data sets: iBPAD51 (dotted line), eBPAD91 (solid grey line) and BPAD200+N+B+AF (black line). Curves were drawn by averaging results from five iterations of SVM classifiers consisting of 10 features selected by greedy backward feature elimination assessed in a LTOCV approach; (**B**) Plot showing the average percentage accuracy (five iterations) of SVM classifiers of 10 features trained on different sized subsets of BPAD200+N+B+AF for comparison to SVM classifiers derived from iBPAD51 and eBPAD91.

Finally, to fully explore differences between SVM classifiers trained on intra- (iBPAD51) and extracellular (eBPAD91) BPAs, the top 10 features selected by greedy backward feature elimination for each classifier were interrogated (Table 2). The SVM classifier trained on eBPAD91 (extracellular classifier) utilized features derived from protein annotation tools that were not utilized by the classifier trained on iBPAD51 (intracellular classifier) and were related to the following diverse array of biological phenomena: an adhesin predictor SPAAN [26] (which describes if the protein adheres to the surface of cells), surface accessibility, and secondary structure predictor NetSurfP [29] (which predicts the relative likelihood of sections of each protein being exposed on the surface of a protein structure), lipoprotein prediction LipoP [22] (which predicts if a protein will interact with lipids), as well as a cleavage site predictor NetChop [30] (which predicts if a protein will be chopped by the human proteasome) (Table 2A). Features derived from protein annotation tools that were unique to the SVM classifier trained on iBPAD51 were largely related to immunogenicity and included: a B-cell epitope predictor [23] (predicts possible binding sites for B-cells), a T-cell epitope predictor [28] (predicts possible binding sites for T-cells) and a calpain cleavage predictor [31] (predicts a specific type of protein cleavage dependent on the presence of Ca^{2+}) (Table 2B). In summary, SVM classifiers appear to utilize features from different protein annotation tools to discriminate intra- and extracellular BPAs from their respective non-BPAs. There may be benefit to deriving separate classifiers for both types of BPA in the future as the literature record expands allowing larger training data sets to be interrogated.

Table 2. The top 10 annotation features selected by greedy backward feature elimination utilized by SVM classifiers trained on (**A**) eBPAD 91 and (**B**) iBPAD51.

(A)			
Rank	**Feature**	**Protein Annotation Tool Type**	**Rank in Intracellular Classifier**
1	Pad-value	**Adhesin**	42
2	DictOGlyc_Ser_Average_Threshold_Length	Glycosylation	189
3	LipoP_SPI_AvrScore	**Lipoprotein**	NF
4	Netsurfp_RSA_Exposed_AverageDiff	**Surface accesibility and secondary structure**	NF
5	PoloPhosphorylation_CorAvg	Phosphorylation	NF
6	Net_Chop_CorCount	**Predicts cleavage sites**	NF
7	DictOGlyc-No_Score_Sites_Length	Glycosylation	NF
8	GPS_SUMO_Sumoylation_Average_Score	Small ubiquitin like modifiers (SUMOs) binding site prediction	9
9	ProtParam-PercIsoleucine	General Annotation	144
10	ProtParam-PercGlutamicAcid	General Annotation	NF
(B)			
Rank	**Feature**	**Protein Annotation Tool Type**	**Rank in Extracellular Classifier**
1	Bepipred-Count_Length	**B-Cell Epitope**	149
2	CCD_av_diff	**Calpain Cleavage**	NF
3	YinOYang-T-Average-Difference1_Length	Glycosylation	NF
4	ProtParam-GRAVY	General Annotation	35
5	NetOGlyc-T-Max-I	Glycosylation	196
6	YinOYang-T-Average_Length	Glycosylation	NF
7	ProtParam-PercAlanine	General Annotation	97
8	NetPhosK-Y-MaxScore	Phosphorylation	NF
9	GPS_SUMO_Sumoylation_Average_Score	Small Ubiquitin like modifiers (SUMOs) binding site predictor	8
10	MBAAgl7_CorAvg	**T-Cell Epitope predictor**	NF

Protein annotation tools listed in bold represent those not present in the other classifier type. NF: not found in the top 200 features of the other classifier type that were submitted to the greedy backward feature elimination algorithm following non-specific *F*-score filtering.

3. Discussion

An SVM classifier capable of discriminating BPAs from non-BPAs was evaluated in a fully-nested approach while examining the impact of the addition of new BPAs curated from the literature record and additional annotation features derived from new protein annotation tools. The major finding was that a signal of protective efficacy can be curated from published data in the form of BPAs defined in this study. This was evident from the significant drop in accuracy of SVM classifiers trained on randomly permutated data (Figure 1). To reiterate, a BPA was defined as a whole protein that led to significant protection ($p < 0.05$) in an animal model (i.e., bacterial load reduction or survival assay) following immunization and subsequent challenge with the bacterial pathogen.

The most biologically relevant classifier generated in this study was built upon the dataset designated BPAD200+N+B+AF (Figure 2). The top 10 features used by this classifier may be examined to determine which features are reflective of protective efficacy (Table 1). The main signal related to protective efficacy in these top 10 features was the prediction of T-cell epitopes (bioinformatics tools NetMHC [27] and MBAAgl7 [28]) with a greater number of epitopes associated with protection as expected. In addition, annotation related to general biological processes, such as threonine glycosylation, phosphorylation, and lipoproteins, are positively linked with protection. These processes have previously been implicated in controlling both the humoral and cellular immune responses [32–36]. For example, lipoproteins have been shown to increase the influence of major histocompatibility complex-II (MHC-II) activation on T-helper cells (Th cells) [33]. This is achieved by lipid rich microdomains co-localizing and increasing the MHC-II molecules concentration on the cell surface, resulting in more efficient Th cell activation while requiring less antigen [33]. In summary, although epitope prediction clearly has value in an RV approach, other tools predicting general protein annotation features should also be considered and continue to be incorporated in future enhancements.

Clustering BPAD200+N+B+AF revealed that BPAs grouped based on extracellular and intracellular predicted subcellular localization (Figure 4). Unexpectedly, separate Intracellular (trained on iBPAD51), Extracellular (trained on eBPAD91), and Combined (BPAD200+N+B+AF) classifiers achieved similar accuracies (Figure 5). However, it was theorized that intracellular and extracellular classifiers would be selecting different features in order to capture biological differences whilst making predictions of BPA or non-BPA. A logical hypothesis would be that extracellular classifiers utilize features related to B-cell epitopes since this antigen type is surface exposed and that intracellular classifiers require features related to T-cell epitopes since this antigen type is internalized. However, this was not exactly the case, whereby the intracellular classifier utilizes features from both B-cell and T-cell epitope predictors but the extracellular classifier does not utilize features from any epitope prediction tools (Table 2). This could be due to the difficulty in predicting conformational epitopes from amino acid sequences [37], it is estimated that 90% of B-cell epitope binding is conformational [38]. To model this information CBTOPE [39] (a conformational B-cell epitope predictor) was included in this study, however annotation features derived from CBTOPE were not present in the top ten annotation features used in classification for any of the classifiers trained in this study. It is possible that with more advanced techniques these conformational epitopes will be more accurately predicted from amino acid sequences and may become an important part of the classification for extracellular BPAs from non-BPAs. Instead the extracellular classifier uses annotation features derived from more general protein annotation tools (e.g., adhesin prediction, surface accessibility, and general cleavage site prediction). Although the utility of separate intra- and extracellular classifiers has not been demonstrated in this study it is clear that these classifier types are modelling different aspects of antigen biology and future studies will explore this as more data becomes available.

There are a number of limitations that may be affecting the ability of SVM classifiers to discriminate BPAs from non-BPAs. Firstly, it is possible that the random selection of non-BPAs for the negative training data may result in the mistaken addition of protective antigens (i.e., BPAs) that simply have not been tested and documented in the literature record. To negate this limitation, non-BPAs with homology to BPAs were discarded. Furthermore, permutation analysis demonstrated that the non-BPAs represent useful negative training data since there was clearly a discriminatory signal between BPAs and non-BPAs (Figure 1B). Another limitation is that the features on which the classifiers are trained are all generated from protein annotation tools. These tools are not 100% accurate and therefore some tools will introduce noise into the annotation used to train SVM classifiers. Future studies should continue to add new features from protein annotation tools as they are developed to asses if improvements to classifier accuracy can be achieved. Regardless, the annotation features derived from protein annotation tools currently available produce sufficient signal to significantly distinguish classifiers from randomly permutated data (Figure 1B). Finally, most antigens previously tested and confirmed in the literature are predicted to be of unknown subcellular localization (29%) or are extracellular (45.5%, eBPAD of BPAD200). This led to a small training dataset (iBPAD51) when building the intracellular classifier described in this study (51 BPAs and 51 non-BPAs). Incorporating more proteins in the training datasets may enable better description of differences in protection derived through intracellular or extracellular proteins.

It is envisaged that the application of ML approaches to RV will build upon the success of filtering approaches that lead to the BEXSERO® vaccine. The SVM classifiers constructed in this study discriminate BPAs from non-BPAs and through comparisons to randomly permutated data clearly demonstrate that a signal for protective efficacy can be curated from the literature record. Future studies should concentrate on the expansion of the training data through the addition of further BPAs curated from the literature record and the incorporation of new protein annotation tools since these enhancements significantly increased the accuracy of SVM classifiers. This will increase the accuracy with which BPAs are predicted and reduce the number of laboratory assays that need to be performed in order to identify novel antigens with the required protective efficacy.

4. Methods

4.1. Training Data

A literature curation identified 64 new BPAs which were combined with 136 previously characterized BPAs [20] for a positive training dataset totaling 200 BPAs. A BPA is a bacterial protein that has led to significant protection ($p < 0.05$) in an animal model (i.e., bacterial load reduction or survival assay) following immunization and subsequent challenge with the bacterial pathogen. BPAs were only selected for inclusion if a consensus for this definition was met by two or more curators. Negative training data (non-BPA) was generated by randomly selecting a protein from the same bacterial species for each BPA. BLASTP was used to discard any non-BPAs that matched to known BPAs (i.e., >98% similarity) or non-BPAs already selected (E-value < 10×10^{-3}) [20]. Similarity is a measure of the extent to which sequences are related and E-values are a representation of the same sequence occurring by chance, both are described fully in the NCBI BLAST help manual [40]. In addition, unless otherwise stated, non-BPAs were selected from the same predicted subcellular localization as their BPA partner. One BPA (ACF35754.1) did not have a predicted extracellular protein (same subcellular localization) that matched the BLASTP inclusion criteria and a protein with unknown subcellular localization was sampled instead for inclusion in the negative training dataset. In summary, a dataset consisting of 200 BPAs and 200 non-BPAs was constructed and referred to throughout this study as BPAD200.

4.2. Permutation Analysis

Permutation analysis was used to determine the optimal feature number for SVM classifiers. Labels (BPA and non-BPA) were randomly permutated five times, generating five datasets. Permutation analysis should destroy the antigenic signal and is fully described by Good [41]. The AUCs achieved when training SVM classifiers using greedy backward feature elimination on these datasets were averaged and compared to five iterations of greedy backward feature elimination of BPAD200+N+B+AF. This process was repeated with SVM classifiers trained with different numbers of features.

4.3. Data Annotation

A second literature curation identified new protein annotation tools to generate novel annotation features for training SVM classifiers. This study increased the number of protein annotation tools from 19 in our previous work [20] to 34, and the output from these tools was parsed to generate 525 annotation features. Annotation tools included in the previous study [20] had their outputs parsed in a standardized manner for all annotation tools. Protein annotation tools for eukaryotic proteins were initially included to maximize the data on which SVM classifiers were trained. Classifiers trained on datasets including additional annotation features not previously utilised in ML approaches to RV have the AF tag, and a full list of protein annotation tools and annotations features can be found in Table S1.

4.4. Machine Learning Classification

Annotation features were scaled individually between −1 and 1 before training SVM classifiers on BPAs and non-BPAs [42]. A table of scaled annotation features as submitted to the classification pipeline can be found as Table S2. All implementations of the SVM in this manuscript used a non-specific filtering step based on F-score [43,44] to reduce the number of annotation features to 200. These features then underwent greedy backward feature elimination [45] to remove the least informative feature each round until the desired number of features remained. Backward selection was used to enable the interaction of all features to be considered. When features have equal information content the algorithm randomly selects which feature to remove, SVM classifiers were trained in five separate

iterations to assess the impact of randomly breaking such ties. A full description of SVM classifiers has been published by Noble [46].

To enable comparison to previous RV studies [20] SVMs with a radial bias function (RBF) kernel were used to construct classifiers, as implemented in the *libsvm* package [42] in Python following the *libsvm* user manual guidelines [47]. Unless otherwise indicated, classifiers were evaluated using a nested LTOCV model [48] to obtain SVM predicted probabilities for each protein within the training data of BPAs and non-BPAs. The first step in a fully nested approach was to split the data into 10 parts and isolate one of these tenths as the test dataset. Feature selection and parameter optimization were applied to the remaining training dataset only (i.e., the remaining 9/10 of the data). An SVM classifier was then built on only the training dataset and used to predict the class (BPA or non-BPA) of the test data in the one-tenth left out. This process was repeated a further nine times leaving the remaining tenths of the data out one at a time.

4.5. Statistics

ROC curves were used to evaluate the performance of SVM classifiers in this study, and generated by plotting the true positive rate (TPR, i.e., sensitivity) over the false positive rate (FPR, i.e., 1−specificity) [49]. Area under the curve (AUC) values were calculated from ROC curves and differences between AUC values was assessed with the DeLong [25] statistical test and $p < 0.05$ considered significant.

4.6. Hierarchical Clustering

The subcellular localization of BPAs from BPAD200+N+B+AF was predicted using the protein annotation tool PSORTb [3]. A BPA was labeled as extracellular if it was predicted to be localized close to the surface of the cell (i.e., cell wall, extracellular, outer membrane, or periplasmic). BPAs with a predicted subcellular localization of periplasmic were included in the extracellular group as these proteins clustered predominantly with the extracellular opposed to intracellular group. If a BPA was predicted to be localized to the cytoplasm or cytoplasmic membrane it was defined as intracellular. Intracellular and extracellular BPAs from BPAD200 were clustered on all annotation features using a Euclidean distance calculation and a Ward clustering metric using the *ClassDiscovery* [50], and *Dendextend* [51] packages in R [47].

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/18/2/312/s1.

Acknowledgments: This work was performed with the support of the IRIDIS High Performance Computing Facility and the Bioinformatics Core at the University of Southampton and was funded by a Marie Curie Career Integration Grant (CIG, PCIG13-GA2013-618334).

Author Contributions: Ashley I. Heinson, Bastiaan Moesker, Carmen C. Denman Hume and Christopher H. Woelk were responsible for the literature curation to identify novel bacterial protective proteins; Ashley I. Heinson, completed the computational work in this paper; Elena Vataga and Yawwani Gunawardana contributed analysis tools; Ashley I. Heinson, Christopher H. Woelk, Yawwani Gunawardana, Yper Hall, Elena Stylianou, Helen McShane, Ann Rawkins and Mahesan Niranjan designed analyses performed; Christopher H. Woelk Conceived and designed this project; Ashley I. Heinson and Christopher H. Woelk wrote this manuscript and all authors provided comment.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

RV	Reverse Vaccinology
ML	Machine Learning
BPA	Bacterial Protective Antigen
BPAD	Bacterial Protective Antigen Dataset
SVM	Support Vector Machine
LTOCV	Leave Tenth Out Cross Validation
iBPAD	Intracellular Bacterial Protective Antigen Dataset

eBPAD	Extracellular Bacterial Protective Antigen Dataset
ELISA	Enzyme-linked Imunosorbent Assay
FACS	Fluorescence activated cell-sorting
DOMV	Detergent Extracted Outer Membrane Vesicle
ACC	Auto Cross Covariance
DA-PLS	Discriminant Analysis by Partial Least Squares
RBF	Radial Bias Function
TPR	True Positive Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
N	Nested
B	Balanced
AF	Additional Features
HDL	High Density Lipoprotein
Th cells	T-helper cells

References

1. Pizza, M.; Scarlato, V.; Masignani, M.M.; Giuliani, B.; Arico, M.; Comanducci, G.T.; Jennings, L.; Baldi, E.; Bartolini, B.; Capecchi, B.; et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **2000**, *287*, 1816–1820. [CrossRef] [PubMed]
2. Crum-Cianflone, N.; Sullivan, E. Meningococcal Vaccinations. *Infect. Dis. Ther.* **2016**, *5*, 89–112. [CrossRef] [PubMed]
3. Yu, N.Y.; Wagner, J.R.; Laird, M.R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S.C.; Ester, M.; Foster, L.J. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **2010**, *26*, 1608–1615. [PubMed]
4. Corpet, F.; Servant, F.; Gouzy, J.; Kahn, D. ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **2000**, *28*, 267–269. [CrossRef] [PubMed]
5. Henikoff, S.; Henikoff, J.G.; Pietrokovski, S. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **1999**, *15*, 471–479. [CrossRef] [PubMed]
6. Giuliani, M.M.; Adu-Bobie, J.; Comanducci, M.; Arico, B.; Savino, S.; Santini, L.; Brunelli, B.; Bambini, S.; Biolchi, A.; Capecchi, B.; et al. A universal vaccine for serogroup B. meningococcus. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10834–10839. [CrossRef] [PubMed]
7. Watson, P.S.; Turner, D.P. Clinical experience with the meningococcal B vaccine, Bexsero®: Prospects for reducing the burden of meningococcal serogroup B disease. *Vaccine* **2016**, *34*, 875–880. [CrossRef] [PubMed]
8. He, Y.; Racz, R.; Sayers, S.; Lin, Y.; Todd, T.; Hur, J.; Li, X.; Patel, M.; Zhao, B.; Chung, M.; et al. Updates on the web-based VIOLIN vaccine database and analysis system. *Nucleic Acids Res.* **2013**, *14*, 1124–1132. [CrossRef] [PubMed]
9. Jaiswal, V.; Chanumolu, S.K.; Gupta, A.; Chauhan, R.S.; Rout, C. Jenner-predict server: Prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinform.* **2013**, *14*, 211. [CrossRef] [PubMed]
10. Moise, L.; Gutierrez, A.; Kibria, F.; Martin, R.; Tassone, R.; Liu, R.; Terry, F.; Martin, B.; de Groot, A.S. iVAX: An integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Hum. Vaccin. Immunother.* **2015**, *11*, 2312–2321. [CrossRef] [PubMed]
11. Heinson, A.I.; Woelk, C.H.; Newell, M.L. The promise of reverse vaccinology. *Int. Health* **2015**, *7*, 85–89. [CrossRef] [PubMed]
12. Sinha, K.; Bhatnagar, R. GroEL provides protection against *Bacillus anthracis* infection in BALB/c mice. *Mol. Immunol.* **2010**, *48*, 264–271. [CrossRef] [PubMed]
13. Velikovsky, C.A.; Cassataro, J.; Giambartolomei, G.H.; Goldbaum, F.A.; Estein, S.; Bowden, R.A.; Bruno, L.; Fossati, C.A.; Spitz, M. A DNA vaccine encoding lumazine synthase from Brucella abortus induces protective immunity in BALB/c mice. *Infect. Immun.* **2002**, *70*, 2507–2511. [CrossRef] [PubMed]

14. Fu, S.; Xu, J.; Li, X.; Xie, Y.; Qiu, Y.; Du, X.; Yu, S.; Bai, Y.; Chen, Y.; Wang, T.; et al. Immunization of mice with recombinant protein CobB or AsnC confers protection against *Brucella abortus* infection. *PLoS ONE* **2012**, *7*, 29552. [CrossRef] [PubMed]
15. Jain, S.; Kumar, S.; Dohre, S.; Afley, P.; Sengupta, N.; Alam, S.I. Identification of a protective protein from stationary-phase exoproteome of *Brucella abortus*. *Pathog. Dis.* **2013**, *70*, 75–83. [CrossRef] [PubMed]
16. Chang, Y.F.; Chen, C.S.; Palaniappan, R.U.; He, H.; McDonough, S.P.; Barr, S.C.; Yan, W.; Faisal, S.M.; Pan, M.J.; Chang, C.F. Immunogenicity of the recombinant leptospiral putative outer membrane proteins as vaccine candidates. *Vaccine* **2007**, *25*, 8190–8197. [CrossRef] [PubMed]
17. Mizrachi Nebenzahl, Y.; Bernstein, A.; Portnoi, M.; Shagan, M.; Rom, S.; Porgador, A.; Dagan, R. Streptococcus pneumoniae surface-exposed glutamyl tRNA synthetase, a putative adhesin, is able to induce a partially protective immune response in mice. *J. Infect. Dis.* **2007**, *196*, 945–953. [CrossRef] [PubMed]
18. Fritzer, A.; Senn, B.M.; Minh, D.B.; Hanner, M.; Gelbmann, D.; Noiges, B.; Henics, T.; Schulze, K.; Guzman, C.A.; Goodacre, J.; et al. Novel conserved group A streptococcal proteins identified by the antigenome technology as vaccine candidates for a non-M protein-based vaccine. *Infect. Immun.* **2010**, *78*, 4051–4067. [CrossRef] [PubMed]
19. Henningham, A.; Chiarot, E.; Gillen, C.M.; Cole, J.N.; Rohde, M.; Fulde, M.; Ramachandran, V.; Cork, A.J.; Hartas, J.; Magor, G.; et al. Conserved anchorless surface proteins as group A streptococcal vaccine candidates. *J. Mol. Med.* **2012**, *90*, 1197–1207. [CrossRef] [PubMed]
20. Bowman, B.N.; McAdam, P.R.; Vivona, S.; Zhang, J.X.; Luong, T.; Belew, R.K.; Sahota, H.; Guiney, D.; Valafar, F.; Fierer, J.; et al. Improving reverse vaccinology with a machine learning approach. *Vaccine* **2011**, *29*, 8156–8164. [CrossRef] [PubMed]
21. Doytchinova, I.A.; Flower, D.R. VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinforma.* **2007**, *8*, 4. [CrossRef] [PubMed]
22. Juncker, A.S.; Willenbrock, H.; von Heijne, G.; Brunak, S.; Nielsen, H.; Krogh, A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **2003**, *12*, 1652–1662. [CrossRef] [PubMed]
23. Larsen, J.E.; Lund, O.; Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2006**, *2*, 2. [CrossRef] [PubMed]
24. Kline, K.A.; Fälker, S.; Dahlberg, S.; Normark, S.; Henriques-Normark, B. Bacterial adhesins in host-microbe interactions. *Cell Host Microbe* **2009**, *5*, 580–592. [CrossRef] [PubMed]
25. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef] [PubMed]
26. Sachdeva, G.; Kumar, K.; Jain, P.; Ramachandran, S. SPAAN: A software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* **2005**, *21*, 483–491. [CrossRef] [PubMed]
27. Nielsen, M.; Lundegaard, C.; Blicher, T.; Lamberth, K.; Harndahl, M.; Justesen, S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* **2007**, *2*, 796. [CrossRef] [PubMed]
28. Cai, R.; Liu, Z.; Ren, J.; Ma, C.; Gao, T.; Zhou, Y.; Yang, Q.; Xue, Y. GPS-MBA: Computational analysis of MHC class II epitopes in type 1 diabetes. *PLoS ONE* **2012**, *7*, 33884. [CrossRef] [PubMed]
29. Petersen, B.; Petersen, T.N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009**, *9*, 51. [CrossRef] [PubMed]
30. Nielsen, M.; Lundegaard, C.; Lund, O.; Kesmir, C. The role of the proteasome in generating cytotoxic T-cell epitopes: Insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **2005**, *57*, 33–41. [CrossRef] [PubMed]
31. Liu, Z.; Cao, J.; Gao, X.; Ma, Q.; Ren, J.; Xue, Y. GPS-CCD: A novel computational program for the prediction of calpain cleavage sites. *PLoS ONE* **2011**, *6*, 19001. [CrossRef] [PubMed]
32. Norata, G.D.; Pirillo, A.; Ammirati, E.; Catapano, A.L. Emerging role of high density lipoproteins as a player in the immune system. *Atherosclerosis* **2012**, *220*, 11–21. [CrossRef] [PubMed]
33. Norata, G.D.; Catapano, A.L. HDL and adaptive immunity: A tale of lipid rafts. *Atherosclerosis* **2012**, *225*, 34–35. [CrossRef] [PubMed]
34. Rudd, P.M.; Elliott, T.; Cresswell, P.; Wilson, I.A.; Dwek, R.A. Glycosylation and the immune system. *Science* **2001**, *291*, 2370–2376. [CrossRef] [PubMed]

35. Liu, S.; Cai, X.; Wu, J.; Cong, Q.; Chen, X.; Li, T.; Du, F.; Ren, J.; Wu, Y.T.; Grishin, N.V.; et al. Phosphorylation of innate immune adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science* **2015**, *347*, 6227. [CrossRef] [PubMed]
36. Snapper, C.M.; Rosas, F.R.; Jin, L.; Wortham, C.; Kehry, M.R.; Mond, J.J. Bacterial lipoproteins may substitute for cytokines in the humoral immune response to T cell-independent type II antigens. *J. Immunol.* **1995**, *155*, 5582–5589. [PubMed]
37. Haste Andersen, P.; Nielsen, M.; Lund, O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* **2006**, *15*, 2558–2567. [CrossRef] [PubMed]
38. Huang Jian, H.; Honda, W. CED: A conformational epitope database. *BMC Immunol.* **2006**, *7*, 7.
39. Ansari, H.R.; Raghava, G.P. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Res.* **2010**, *6*, 6. [CrossRef] [PubMed]
40. Fassler, J.C.P. *BLAST Glossary, BLAST®Help*; National Center for Biotechnology Information: Bethesda, MD, USA, 2011.
41. Good, P. *Permutation Tests: A Practical Guide To Resampling Methods For Testing Hypotheses*; Springer Science & Business Media: Berlin, Germany, 2013.
42. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM TIST* **2011**, *2*, 27. [CrossRef]
43. Chen, Y.-W.; Lin, C.-J. Combining SVMs with Various Feature Selection Strategies. In *Feature Extraction*; Springer Science & Business Media: Berlin, Germany, 2006; pp. 315–324.
44. Polat, K.; Güneş, S. A new feature selection method on classification of medical datasets: Kernel *F*-score feature selection. *Expert Syst. Appl.* **2009**, *36*, 10367–10373. [CrossRef]
45. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
46. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
47. Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comp. Graph. Stat.* **1996**, *5*, 299–314. [CrossRef]
48. Simon, R.; Radmacher, M.D.; Dobbin, K.; McShane, L.M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **2003**, *95*, 14–18. [CrossRef] [PubMed]
49. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874. [CrossRef]
50. Coombes, K. ClassDiscovery: Classes and Methods for "Class Discovery" with Microarrays or Proteomics, R Package Version 2.1. Available online: http://bioinformatics.mdanderson.org/Software/OOMPA (accessed on 1 December 2016).
51. Galili, T. Dendextend: An R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* **2015**, *31*, 3718–3720. [CrossRef] [PubMed]