UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS
School of Electronics & Computer Science

Exploiting Gene Expression and Protein Data for Predicting Remote Homology and Tissue Specificity

by
Daniela Wieser

A thesis submitted for the degree of
Doctor of Philosophy

June 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS
SCHOOL OF ELECTRONICS & COMPUTER SCIENCE

Doctor of Philosophy

EXPLOITING GENE EXPRESSION AND PROTEIN DATA FOR PREDICTING REMOTE HOMOLOGY AND TISSUE SPECIFICITY

by Daniela Christine Wieser

In this thesis I describe my investigations of applying machine learning methods to high throughput experimental and predicted biological data. The importance of such analysis as a means of making inferences about biological functions is widely acknowledged in the bioinformatics community. Specifically, this work makes three novel contributions based on the systematic analysis of publicly archived data of protein sequences, three dimensional structures, gene expression and functional annotations: (a) remote homology detection based on amino acid sequences and secondary structures; (b) the analysis of tissue-specific gene expression for predictive signals in the sequence and secondary structure of the resulting protein product; and (c) a study of ageing in the fruit fly, a commonly used model organism, in which tissue specific and whole-organism gene expression changes are contrasted.

In the problem of remote homology detection, a kernel-based method that combines pairwise alignment scores of amino acid sequences and secondary structures is shown to improve the prediction accuracies in a benchmark task defined using the Structural Classification of Proteins (SCOP) database. While the task of predicting SCOP superfamilies should be regarded as an easy one, with not much room for performance improvement, it is still widely accepted as the gold standard due to careful manual annotation by experts in the subject of protein evolution.

A similar method is introduced to investigate whether tissue specificity of gene expression is correlated with the sequence and secondary structure of the resulting protein product. An information theoretic approach is adopted for sorting fruit fly and mouse genes according to their tissue specificity based on gene expression data. A classifier is then trained to predict the degree of specificity for these genes. The study concludes that the tissue specificity of gene expression is correlated with the sequence, and to a certain extent, with the secondary structure of the gene's protein product.

The sorted list of genes introduced in the previous chapter is used to investigate the tissue specificity of transcript profiles obtained from a study of ageing in the fruit fly. The same list is utilised to investigate how filtering tissue-restricted genes affects gene set enrichment analysis in the ageing study, and to examine the specificity of age-associated genes identified in the literature. The conclusion drawn in this chapter is that categorisation of genes according to their tissue specificity using Shannon's information theory is useful for the interpretation of whole-fly gene expression data.

Contents

List of Figures

List of Tables

DECLARATION OF AUTHORSHIP

The work presented in this dissertation was started at the University of Sheffield and, due to relocation of the supervisor, continued at the University of Southampton. This thesis is the result of a part-time PhD study. It was partly carried out in collaboration with the European Bioinformatics Institute in Hinxton, Cambridge. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this thesis nor anything substantially the same has been or is being submitted for any qualification at any other university.

I, Daniela Wieser declare that the thesis entitled *"Exploiting gene expression and protein data for predicting remote homology and tissue specificity"* and the work presented in it are my own. I confirm that:

- this work was done mainly while in candidature for a research degree at the University of Southampton;
- 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;
- where I have consulted the published work of others, this is always clearly attributed;
- 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; I have acknowledged all main sources of help;
- 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;
- Publications during the time of this PhD
  - **D. Wieser** and M. Niranjan (2009). Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores. In Silico Biol, 9(3), 89–103.
    My contribution:
    I contributed to the planning of the project, performed the experiments and wrote the paper. This paper is discussed in Chapter 3 and Chapter 4.

- S. Patient, **D. Wieser**, M. Kleen, E. Kretschmann, M. J. Martin, and R. Apweiler (2008). UniProtJAPI: a remote API for accessing UniProt data. Bioinformatics, 24(10), 1321–1322.

  My contribution:

  I contributed to the planning, design, implementation and testing of the software and wrote large parts of the manuscript. The library was used in Chapter 5 to extract protein data from UniProtKB, but is not further discussed in this document because the creation of the library was largely a collaborative effort and is not directly research related.

- A. Freitas, **D. Wieser**, and R. Apweiler (2010). On the importance of comprehensible classification models for protein function prediction. IEEE/ACM Trans Comput Biol Bioinform, 7(1), 172–182.

  My contribution:

  I contributed to writing the manuscript. A part of the manuscript is described in Chapter 2.2.

- C. Selman, J. M. A. Tullet, **D. Wieser** E. Irvine, S. J. Lingard, A. I. Choudhury, M. Claret, H. Al-Qassab, D. Carmignac, F. Ramadani, A. Woods, I. C. A. Robinson, E. Schuster, R. L. Batterham, S. C. Kozma, G. Thomas, D. Carling, K. Okkenhaug, J. M. Thornton, L. Partridge, D. Gems, D. J. Withers (2009). Ribosomal protein S6 kinase 1 signaling regulates mammalian life span. Science, 326 (5949), 140–144.

  My contribution:

  I performed the data analysis of the microarray experiments of this study. This work is related to Chapter 6 where age-associated genes in the fruit fly are investigated. However, the paper was a collaborative project and is not directly related to this thesis. Therefore it is not further discussed in this manuscript.

- C. Slack, C. Werz, **D. Wieser**, N. Alic, A. Foley, H. Stocker, D. Withers, J. Thornton, E. Hafen and L. Partridge (2010). Regulation of lifespan, metabolism and stress responses by the Drosophila SH2B protein Lnk. PLoS Genetics. **Accepted.**

  My contribution:

  I performed the data analysis of the microarray experiments of this study. However, the paper was a collaborative project and is not directly related to this thesis. Therefore it is not further discussed in this manuscript.

- Part of Chapter 5 of this thesis was presented as a poster at the 12th Microarray and Gene Expression Data (MGED) meeting, 2009, Phoenix/USA.
- Part of Chapter 7 (WFSM) of this thesis was presented as a poster at the ISMB/ECCB conference, 2007, Vienna/Austria.

Daniela Wieser, June 2010

Acknowledgements

Many people have assisted me along the way of completing this work. Their help and guidance was greatly appreciated, and I would like to offer my regards to everyone who supported me in any respect during the last couple of years.

I especially want to thank my supervisor, Prof. Mahesan Niranjan, whose support, guidance and encouragement has enabled me to pursue a part-time PhD program. Despite me being an off-site part-time student he was always accessible and willing to help me with my research. He has sacrificed several weekends to meet me in Cambridge, and often came to see me for meetings at my workplace. His positive and progressive attitude helped me to stay on track and keep my enthusiasm throughout all stages of the program. I also like to thank all fellow PhD students I met at the University of Southampton and the administrative staff who always made me feel most welcome. I am most grateful to the University of Sheffield and the University of Southampton for their generous funding.

I am indebted to many of my colleagues at the European Bioinformatics Institute. This thesis would not have been possible without the generous support of the group members I worked with. In particular, I would like to thank Ernst Kretschmann and Dr. Rolf Apweiler for letting me work on automated protein annotation. The associated experience broadened my perspective on many aspects in the field of Bioinformatics, and considerably strengthened my programming skills. It was during the time working on this project that my research interest was sparked. Prof. Janet Thornton deserves special thanks for hiring me in her research group in the interesting field of ageing at the EBI. Her rigor and passion on research has helped me gain a better understanding on all the subjects covered in this thesis. I'd also like to thank all members of the Thornton group, and groups I previously worked with at the EBI, for all the advice and friendship. I'd also like to thank Prof. Linda Partridge and the members of the Partridge lab at the University College London in particular for their input on Chapter 6.

At the early stages of my PhD studies I was lucky enough to meet the best English teacher one could wish for, Denise Swallow. Denise has taught me so many things beyond English grammar and vocabulary and I will always be grateful for the inspiration and support she has given.

My deepest gratitude goes to my partner Markus Brosch for his tireless love and support throughout the entire course of this study. He was always there for me and supported me in any respect.

Finally, I'd like to thank PNH and Leon who taught me to always have a plan and never loose sight of my goals.

Nomenclature

AA	Amino Acid
AUC	Area Under the ROC Curve
BLOSUM	Blocks Substitution Matrix
CDS	Coding Sequence
DNA	Deoxyribonucleic Acid
cDNA	complementary or copy DNA
DSSP	Definition of Secondary Structure of Proteins
E value	Expectation value
FFS	Forward Feature Selection
FP	False Positive
FN	False Negative
FDR	False Discovery Rate
FPR	False Positive Rate
GFP	Green Fluorescent Protein
GNF	The Genomics Institute of the Novartis Research Foundation
IIS	Insulin/IGF (Insulin-like Growth Factor)-like Signalling
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
RBF	Radial Basis Function
RHD	Remote Homology Detection
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
tRNA	transfer RNA
SCOP	Structural Classification of Proteins
SS	Secondary Structure
SSEA	Secondary Structure Element Assignment
SVM	Support Vector Machine
TPR	True Positive Rate
TP	True Positive
TFBS	Transcription Factor Binding Site
TN	True Negative
TP	True Positive
UTR	Untranslated Region
S/MAR	Scaffold/Matrix Attachment Region
WFSM	Weighted Finite State Machine

Chapter 1

Motivation and Outline

1.1 Motivation for this work

The widespread use of high-throughput technologies in molecular biology has led to a wealth of publicly available data on genes and proteins (Flicek *et al.*, 2010; The UniProt Consortium, 2010). However, the interrelationships of these genes and proteins are as yet poorly understood and are further complicated by the immense amount of data available, making manual characterisation by experts unfeasible. The general aim of this PhD work was to use machine learning techniques to detect biological relationships hidden in both gene and in protein data, and to make inferences based on these relationships. Even though 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 will use 'we' throughout the document, since for most parts of the thesis other people were involved, as stated in the acknowledgements.

The first problem that we approached concerned remote homology detection of proteins. The motivation for this work was that distant evolutionary relationships between proteins with low amino acid sequence similarity are difficult to recognise by computational methods. Consequently, many sequences obtained from large-scale sequencing projects cannot be assigned to any known proteins or families despite being evolutionarily related. Various sequence-based methods have been developed to predict remote homology of proteins. Some of these methods have been modified to make use of the better conserved secondary structure to boost sensitivity. Our motivation was to develop a kernel-based remote homology detection method that allows for a combination of sequence and secondary structure similarity scores in a discriminative approach. This work is described in detail in the chapter *"Remote Homology Detection Using a Kernel Method that Combines Sequence and Secondary Structure Similarity Scores"*.

Building on this work we developed a similar method to investigate tissue specificity of gene expression. Tissue specificity of gene expression is important for a number of studies, but is often difficult to determine by experimental methods. This leads to many genes being uncharacterised in terms of the tissues in which they are expressed. Various gene properties have been shown to be different between tissue-restricted and housekeeping genes. The motivation for this part of the thesis was to investigate whether tissue specificity of gene expression is also correlated with the sequence and secondary structure of the resulting protein product , and whether this information can be used to predict gene tissue specificity. For this, we used an information theoretic approach to sort fruit fly and mouse genes according to their tissue specificity. The method is based on an adaptation of Shannon's information theory to the transcriptome framework. We then trained support vector machine (SVM) classifiers to predict classes of genes that display various degrees of tissue specificity. This work is detailed in the chapter *"Tissue Specificity of Gene Expression is Correlated with the Sequence and Secondary Structure of Resulting Protein Product"*. When we use the term *gene specificity* we refer to gene tissue specificity throughout this thesis.

Related to this work we investigated in the final part of this thesis the tissue-specific contribution to whole-body RNA transcript profiles in the fruit fly *Drosophila melanogaster*. The fruit fly is widely used to investigate mechanisms underlying diverse biological processes, including development, metabolism, neurobiology and ageing. Whole-body RNA microarray profiling has been applied to monitor gene expression changes in various conditions related to these mechanisms. However, there is little information concerning the capacity of microarrays to capture tissue-specific effects of these processes in whole-fly samples. Building on the work described in the previous chapter we used the sorted list of fly genes to investigate transcript profiles obtained from a study of ageing in the adult fly in terms of tissue specificity. The sorted list of genes was also used to investigate how filtering tissue-specific genes affects gene set enrichment analysis in the ageing study, and to study the tissue specificity of age-associated genes from the literature. This work is detailed in the chapter *"Analysis of the Tissue-Specific Contribution to Whole-Body RNA Transcript Profiles in Drosophila Melanogaster"*.

1.2 Outline of this thesis

This thesis is structured as follows:

- Chapter 2 provides background information on biological entities, microarray technology, machine learning techniques and statistical methods relevant or used in this work.

- Chapter 3 introduces the three topics investigated in this work, and summarises relevant literature.

- Chapter 4 presents the detailed methods used in the first project examined in this thesis, i.e. the prediction of protein remote homologoues. Results are also presented for this project and a discussion is included.

- Chapter 5 provides detailed methods for the work done on prediction of tissue specificity, as well as results and discussion.

- Chapter 6 details the methods used to examine whole-body transcript profiles in the fruit fly and presents the results and discussion of this chapter.
- Chapter 7 presents final conclusions by summarising the contributions of this thesis and discusses directions for future work.

Chapter 2

Background and Methods

This chapter gives a brief overview of biological terms that are relevant for this work and describes them in a general manner. Associated technologies to obtain data related to these terms are explained if they have direct relevance to this work. The statistical methods and associated evaluation metrics used are also introduced.

2.1 Biological background

2.1.1 From DNA to RNA to proteins

Deoxyribonucleic acid (DNA) stores the genetic information that enables cells to reproduce and perform their functions. It consists of two strands of nucleotide bases. There are four types of bases: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). It is the sequence of these four nucleotide bases that encodes information on how to build a protein. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins, whereas three nucleotides (codon) code for one amino acid. The code is read by copying stretches of DNA into the related ribonucleic acid (RNA), in a process called transcription (Figure 2.1a). The precursor messenger RNA (pre-mRNA) molecule contains two types of segments in eukaryotes, exons and introns, the latter of which is removed during splicing. This process enables the construction of alternate products. The combination of the coding regions of all these exons is called the coding sequence (CDS). The complementary or copy DNA, cDNA, is a DNA molecule usually obtained by a reverse transcription of an mRNA molecule. Spliced RNA sequences are referred to as messenger RNA (mRNA) that consists of an open reading frame (ORF) and untranslated regions (UTRs). UTR refers to either of two sections on each side of a coding sequence on a strand of mRNA. The ends of mRNA strands are called the 5' (five prime) and 3' (three prime) ends. If the UTR is found on the 5' side, it is called the 5' UTR, or if it is found on the 3' side, it is called the 3' UTR. The proteins are built based upon the ORF in the RNA in a process called translation. The mRNA is processed by a ribosome, which, with the aid of transfer RNA (tRNA), strings together the prescribed amino aids of the protein. An mRNA may or may not cover the complete coding sequence of a gene (alternative splicing). The resulting product of this process, proteins, are large polymers required for the structure, function, and regulation of the body's cells, tissues, and organs. They are made of 20 common amino acids (see Figure 2.2). Throughout this work the amino acid alphabet is used to refer to the amino acids. The amino acid alphabet is a twenty-character alphabet consisting of the characters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y, each representing one of the 20 amino acids coded for by DNA. Amino acids contain an carbon that is connected to an amino (NH_3) group, a carboxyl group (COOH), and a variable side group (R). Amino acids bind together via peptide bonds, which occur between the amino and carboxyl groups of adjacent amino acids.

Protein structures

Proteins have four levels of structure: primary, secondary, tertiary and quaternary. Figure 2.3a depicts a **primary protein sequence**, which is simply a string of letters each of which represents one of the 20 amino acids. The order in which the amino acids appear in this sequence is important because it largely determines the structure

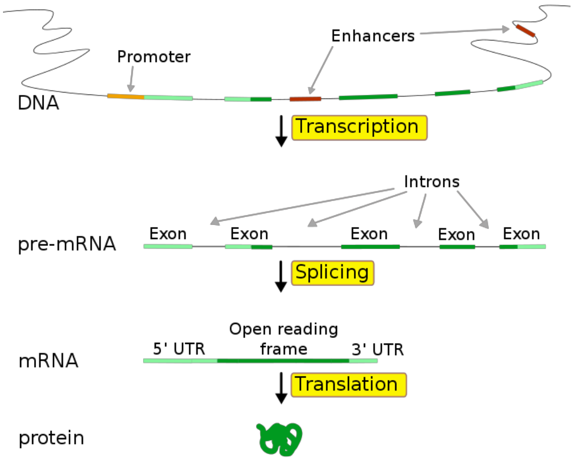


Figure 2.1: **From DNA to RNA to proteins**. Schematic view of gene transcription and translation (from `http://commons.wikimedia.org/wiki/File:Gene2-plain.svg`). DNA is transcribed into pre-mRNA. Introns, and sometimes exons, are removed from the pre-mRNA during splicing, resulting in a mRNA molecule. The mRNA molecule is translated into a protein product. Enhancers are regulatory regions that increase the level of expression of a gene. Promoters are another type of regulatory regions in DNA sequences.

of a protein that in turn determines its function. Alteration to this order may result in the loss of function of the protein. Normally proteins range from 10 to 10,000 amino acids. The largest protein is Titin which consists of 34,350 amino acids. Titin is a key component in the assembly and functioning of vertebrate muscles. The UniProtKB (The UniProt Consortium, 2010) is a public repository for the collection of primary amino acid sequences alongside functional information. UniProtKB/Swiss-Prot, the manually curated section of UniProtKB, contains 514,789 sequence entries comprising 181,163,771 amino acids. The remainder, UniProtKB/TrEMBL, contains 10,376,872 sequence entries comprising 3,344,735,583 amino acids (release 57.14, February 2010). The majority of these sequences were predicted from the DNA sequences, while $\approx$ 26% of sequences in Swiss-Prot have experimental evidence at either the transcript or protein level.

The surrounding chemical environment, which is composed of water and other solvents at different concentrations and temperatures, and the amino acid side chains, determine the way in which these are arranged in space relative to each other. The

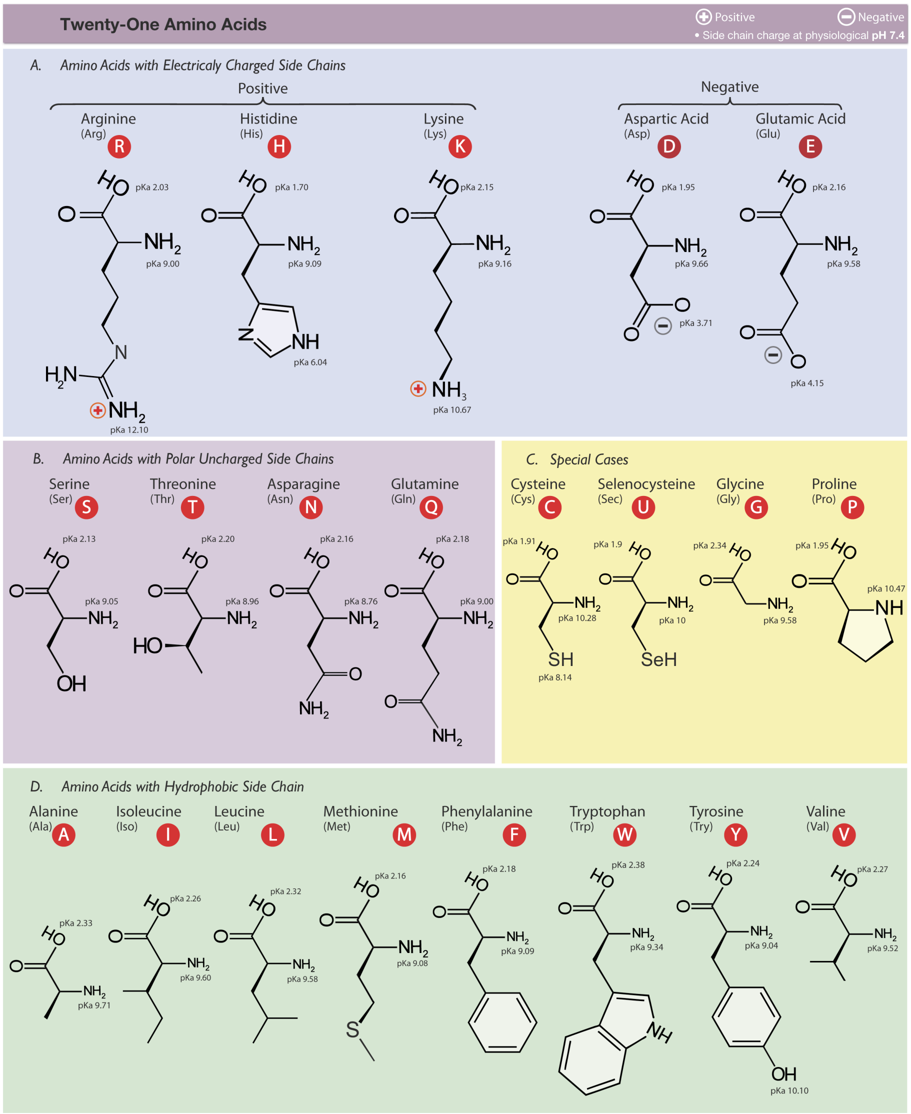


Figure 2.2: **Amino Acids**. Grouped table of the twenty common eukaryotic amino acids' structures, nomenclature, and their side groups' pKa's and charge at pH 7.4. The pH is a measure of the acidity or basicity of a solution while pKa indicates the acid or basic properties of an amino acid. A pKa < 2 means strong acid, pKa > 2 but < 7 means weak acid, pKa > 7 but < 10 means weak basic and pKa > 10 means strong basic. The figure also includes an additional amino acid, Selenocysteine. However, none of the protein sequences used in this work contained this amino acid. The figure is taken from `http://upload.wikimedia.org/wikipedia/commons/0/0f/Amino_acids.png`

secondary structure of proteins is defined by the conformation of the polypeptide and the hydrogen bonding between the carbonyl oxygen atoms and amide atoms in the peptide bonds. Hydrogen bonds are well established and their role in secondary structure architecture and protein folding has been studied extensively (Baker and Hubbard, 1984; Jeffrey and Saenger, 1991; McDonald and Thornton, 1994).

The basic secondary structures that form are known as α-helices, β-sheets and turns; these are also known as basic **secondary structures**. In addition, there are random coils. Random coils are highly flexible portions of a polypeptide chain that have no fixed three-dimensional structure. On average, 60% of a protein exists as α-helices and β-sheets. The remainder of the molecule is in coils and turns (Lodish *et al.*, 2007). The protein domains used in Chapter 4.1.1 consist of 31% α-helices and 25% β-sheets as assigned by the DSSP program (Kabsch and Sander, 1983a). The PSIPRED program (Jones, 1999a) predicted a similar average secondary structure content for these protein domains, namely 32% α-helices and 24 % β-sheets. Following explanations on α-helices and β-sheets have been adapted from Lodish *et al.* (2007).

An α-helix contains 3.6 amino acids per turn. The helix is stabilised by hydrogen bonding between the backbone carbonyl of one amino acid and the backbone NH of the amino acid four residues away. All main chain amino and carboxyl groups are hydrogen bonded, and the R groups stick out from the structure in a spiral arrangement. This structure is very stable but flexible, and it is often seen in parts of a protein that need to bend or move. There are both hydrophilic and hydrophobic helices, depending on the characteristics of the side chains of the amino acids. The former is often found on protein surfaces, whereas the latter tend to be buried within the core of the folded protein. The amino acid proline is usually not found in α-helices. A specialised form of an α-helix is called a coiled coil, a rodlike quaternary protein structure formed by two or three α-helices interacting with each other.

In a β-sheet, two or more strands of amino acids are involved. These line up to form a pleated like structure that tends to be rigid and less flexible than alpha helices. Each strand is made up of five to eight residues. Hydrogen bonding in the β-sheet occurs between backbone atoms in separate, but adjacent, β-strands. These distinct β-strands may be either within a single polypeptide chain, with short or long loops between the β-strand segments, or on different polypeptide chains. In some proteins, β-sheets form the floor of a binding pocket or a hydrophobic core; in other proteins embedded in membranes the β-sheets curve around and form a hydrophilic central pore through which ions and small molecules may flow.

Turns are usually related to proline and glycine, which are common and small amino acids and are often responsible for sharp bends and twists in α-helices and hairpins in β-sheets. They are composed of four residues and located on the surface of a protein.

By knowing which spatial geometry neighbouring amino acids adopt when they bind together it is possible to determine which secondary structure a protein may have. The DSSP (Define Secondary Structure of Proteins) algorithm is a standard method for assigning secondary structure to the amino acids of a protein where atomic-resolution coordinates are available (Kabsch and Sander, 1983a). The assignment is based on the detection of hydrogen-bonds defined by an electrostatic criterion. Secondary structure elements are then assigned according to characteristic hydrogen-bond patterns. DSSP defines eight types of secondary structure depending on the pattern of hydrogen bonds: *H = α-helix*, *B = residue in isolated β-bridge*, *E = extended strand, participates in β ladder*, *G = 3 helix (3_{10} helix)*, *I = 5 helix (pi helix)*, *T = hydrogen bonded turn*, *S = bend*, and *L = loop or other*. STRIDE (Structural identification) is another algorithm for the assignment of protein secondary structure elements given the atomic coordinates of the protein (Frishman and Argos, 1995). In addition to the hydrogen bond criteria used by the DSSP algorithm, the STRIDE assignment criteria also include dihedral angles. The assignment of STRIDE is close to the one done by DSSP (95% of identity). DSSP remains the most widely-used program for secondary structure assignment, and is used in this thesis.

For proteins where no crystal structure is available methods have been developed to predict secondary structure elements from the amino acid sequence. These methods typically define three states: α-helix, β-strand and others. The performance of predictions of secondary structure are measured via the 3-state accuracy, also termed the Q3 score. The Q3 score is the percent of residues for which a method's predicted secondary structure is correct.

The first secondary structure prediction methods were introduced around 30 years ago. Early secondary structure prediction methods have a Q3 score of 50-60% (Kabsch and Sander, 1983b). The most well-known methods include the Chou-Fasman (Chou and Fasman, 1978) and GOR methods (Garnier *et al.*, 1978). These considered single amino acid statistics and are based on the observation that different amino acids have different preferences in adopting secondary structure elements. Later approaches of secondary structure prediction incorporated local dependencies i.e. the neighbouring amino acids (Bowie *et al.*, 1991; Holley and Karplus, 1989; Levin *et al.*, 1986; Nishikawa and Ooi, 1986; Qian and Sejnowski, 1988; Yi and Lander, 1993). These methods achieved Q3 scores above 60%.

The performance of prediction programs was further boosted through the inclusion of evolutionary information into the methods (Hua and Sun, 2001; Kloczkowski *et al.*, 2002; Pollastri *et al.*, 2002; Salamov and Solovyev, 1995; Zvelebil *et al.*, 1987). Conservation evident in multiple sequence alignments of homologs can reveal which amino acids are functionally or structurally important. For instance, surface-exposed loop regions that are not important functionally tend to be part of non-conserved regions in sequence alignments. Including evolutionary conservation knowledge led to the first program to surpass 70% (Rost and Sander, 1993).

Further improvement came from better remote homology detection or HMMs and larger sequence databases. The PSIPRED program (Jones, 1999a) is an example of a secondary structure prediction program in this category. It consists of feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST and achieved a Q3 score of around 77%. Jpred is a secondary structure prediction server that provides α-helix, β-strand and coil predictions (Cole *et al.*, 2008) with a Q3 score of 81.5%. It is also based on multiple sequence alignments and neural networks. Another method whose Q3 score is given as above 80% is called PROTEUS (Montgomerie *et al.*, 2006). PROTEUS exploits the information that is available in the protein structure databases. The accuracy of current protein secondary structure prediction methods can be assessed for example in EVA (Eyrich *et al.*, 2001), which automatically analyses protein secondary structure prediction servers.

In Chapter 4 we develop a classification method that integrates primary and secondary structures (assigned by DSSP and predicted by PSIPRED) of proteins to predict membership of protein families. Primary and secondary protein structures are also used to predict tissue specificity of gene expression in Chapter 5. Tissue specificity of gene expression is clarified in the next subsection. However, first the terms tertiary and quaternary protein structure are clarified.

Once the process of protein synthesis is completed, the protein takes its final shape. This stable form of the protein is known as the **tertiary structure** (Figure 2.3c). Each protein ultimately folds into a three dimensional shape with a distinct inside and outside. The interior of a protein molecule contains mainly hydrophobic amino acids, which tend to cluster and exclude water. By contrast, the exterior of a protein molecule is largely composed of hydrophilic amino acids, which are charged or able to H-bond with water. The Protein Data Bank (PDB) is a public repository for the threedimensional structural data of proteins (Weissig and Bourne, 2002). It contained 63,559 structures in February 2010. For most proteins that have been identified to date, only the primary sequence is available in public databases. Since the function of a protein is determined through its three-dimensional structure, computational methods have been developed that aim to predict the three-dimensional structure for these proteins from sequence, and detect similarities to proteins with known structures.

An important community effort is CASP (Critical Assessment of Techniques for Protein Structure Prediction) that takes place every two years since 1994 to assess predictions to monitor progress in this direction (Moult *et al.*, 2009).

The **quaternary structure** is the arrangement of multiple folded protein molecules in a multi-subunit complex (Figure 2.3d).

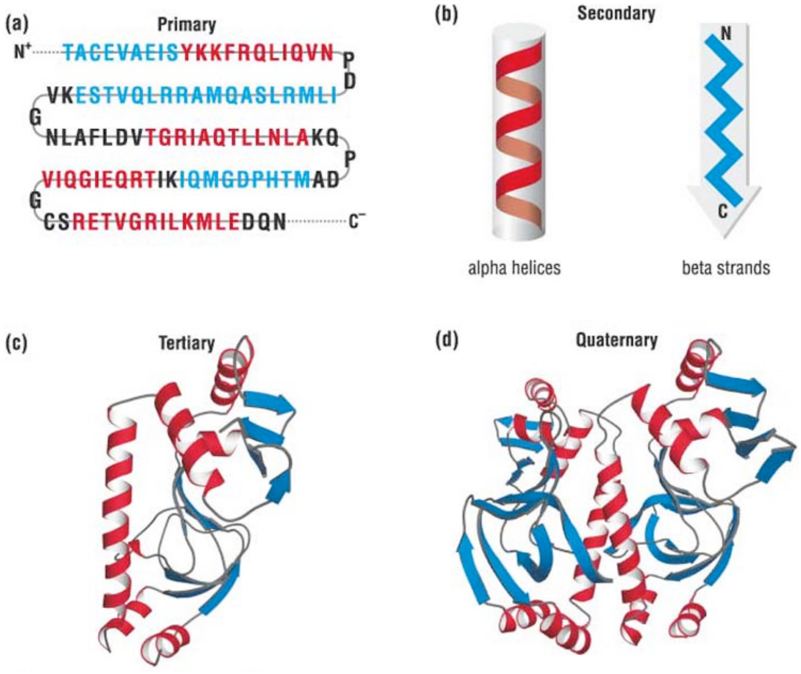


Figure 2.3: Protein conformations. The architecture of proteins at four levels of organsiation is shown: a) primary b) secondary c) tertiary d) quartenary

2.1.2 Tissue specificity of gene expression

In multicellular organisms some genes are expressed and translated to proteins in essentially all tissues, whereas others are expressed predominantly in only one or a few tissues. Housekeeping genes are constitutively expressed in all tissues to maintain cellular functions. It is often assumed that housekeeping genes are expressed at the same level in all cells and tissues, but there are some variances, in particular during cell growth and organism development. Human cells have several hundreds of housekeeping genes, but the exact number is unclear. An example for a housekeeping gene is GAPDH (glyceraldehyde 3-phosphate dehydrogenase) that codes for an enzyme that is vital to glycolysis. Another important housekeeping gene is albumin, which assists in transporting compounds throughout the body. Several housekeeping genes code for structural proteins that make up the cytoskeleton such as beta-actin and tubulin. Others code for subunits of the ribosome. Examples of genes that are expressed in a tissue specific manner include various transcription factors and germ line transcripts. A specific example is the the glycoprotein hormone alpha subunit that is produced only in certain cell types of the anterior pituitary and placenta, but not in lungs or skin.

RNA in situ hybridization (ISH) can be used to identify the spatial pattern of expression of a particular mRNA. The probe is labelled, either radioactively or by chemically attaching a fluorochrome. A tissue is soaked in a solution of single-stranded probes under conditions that allow the probe to hybridize to complementary RNA sequences in the cells. Unhybridized probes are then removed. Radioactive probes are detected by autoradiography. Fluorochrome is detected by fluorescence microscopy. Another technology frequently used to investigate tissue specific expression are microarrays. These are explained in the following paragraph.

2.1.3 Microarray technology

Although all of the cells in a living organism contain identical genetic material, every cell shows a different gene expression profile. Studying which genes are active and which are inactive in different cell types helps to understand both how these cells function normally and how they are affected when various genes do not perform properly. With the development of microarray technology, scientists can examine how active thousands of genes are at any given time. A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides each containing picomoles of a specific DNA sequence. This can be a short section of a gene or other DNA element that are used as probes to hybridize a cDNA. Probe-target hybridization is usually detected and quantified by fluorescence-based detection of fluorophore-labeled targets to determine relative abundance of nucleic acid sequences in the target. In standard microarrays, the probes are attached to a solid surface by a covalent bond to a chemical matrix. The solid surface can be glass or a silicon chip. Affymetrix GeneChip arrays are commonly used. In this technology each gene is typically represented by a set of 11-20 pairs of probes.

Next-generation sequencing (NGS) platforms, that are a relatively new development, also allow transcriptional profiling (RNA-Seq). Microarrays as described above are expected to be superseded by RNA-Seq in the next couple of years. They provide a more precise measurement of levels of transcripts and are often more cost-effective (Wang *et al.*, 2009c).

In this thesis, we used Affymetrix GeneChip array data from the fruit fly and the mouse to investigate tissue specificity of gene expression. In future, similar data based on RNA-Seq may become available that could further refine the methods and results of Chapter 5 and 6 due to advances in technologies.

2.1.4 Sites, regions and modifications in DNA sequences

Following definitions and explanations on sites, regions and modifications in DNA sequences are relevant for Chapter 5.

DNA methylation and CpG islands

After DNA replication, several modifications occur in the DNA, and methylation is one such post-synthesis modification. DNA methylation has been implicated with a number of biological processes including regulation of imprinted genes, X chromosome inactivation, and tumor suppressor gene silencing in cancerous cells. DNA methylation usually occurs in the CpG islands, a CG rich region (cytosine followed by guanine), upstream of the promoter region. The letter p signifies that the

C and G are connected by a phosphodiester bond. In humans, DNA methylation is carried out by a group of enzymes called DNA methyltransferases. DNA methylation systems are well-characterised in vertebrates, but methylation in the fruit fly and other invertebrates remains controversial (Wang *et al.*, 2006). CpG islands are often located around the promoters of housekeeping genes or other genes frequently expressed in a cell. At these locations, the CG sequence is not methylated. By contrast, the CG sequences in inactive genes are usually methylated to suppress their expression. The methylated cytosine may be converted to thymine by accidental deamination. The cytosine to thymine mutation can be corrected only by an inefficient repair mechanism. Hence, over evolutionary time scales, the methylated CG sequence will be converted to the TG (thymine followed by guanine) sequence. This explains the deficiency of the CG sequence in inactive genes.

The ratio of observed to expected CpGs can be used to predict methylated and unmethylated genomic regions (Equation 2.1).

$$CpG[o/e] = \frac{\text{frequency of observed CpG}}{\text{frequency of expected CpG}} = \frac{\text{frequency of observed CpG}}{\text{frequency of C} \times \text{ frequency of G}} \tag{2.1}$$

A number of relationships were found between the occurrence or location of CpG islands and the extent of tissue-specific expression of the associated genes (e.g. Elango *et al.* (2009); Gardiner-Garden and Frommer (1987); Schug *et al.* (2005)). These are reviewed in Chapter 3.2. In Chapter 5 we revisit and investigate the occurrence of CpG islands in the fruit fly and in the mouse.

Scaffold matrix attachment region

Scaffold matrix attachment regions (S/MARs) are genomic elements thought to delineate the structural and functional organisation of the eukaryotic genome. Originally, S/MARs were identified through their ability to bind to the nuclear matrix (i.e. the network of fibres found throughout the inside of a cell nucleus). Binding is dispersed over a region of several hundred base pairs. These elements are found flanking a gene or a small cluster of genes and are located often in the vicinity of cis-regulatory sequences. This has led to the suggestion that they contribute to higher order regulation of transcription by defining boundaries of independently controlled chromatin domains. S/MARs may act as boundary elements for enhancers, restricting their long range effect to only the promoters that are located in the same chromatin domain.

A study on chromatin compactness showed that putative S/MARs were more abundant upstream of tissue-specific genes than upstream of housekeeping genes. S/MARs attach themselves to the nuclear matrix and help the formation of chromatin loops. Tissue-specific genes appear to have less accessible and more compact DNA in their promoter regions, and hence more S/MAR sequences (Ganapathi *et al.*, 2005).

S/MAR sites are identified in Chapter 5 and used to help to discriminate between tissue-specific and broadly expressed fruit fly and mouse genes.

Simple sequence repeats

Simple sequence repeats (SSRs) in DNA sequences are composed of tandem iterations of short oligonucleotides. For example, CGG CGG CGG is a repeat because CGG is repeated three times. The number of repeated copies can also be fractional as in CCCCA CCCCA CCCC. SSRs have been shown to differ between housekeeping and tissue-specific genes in human and mouse in the 5'-UTR region and other genomic regions (Lawson and Zhang, 2008). In this thesis SSRs were computed for the 5'UTR regions for the fruit fly and mouse genes investigated in Chapter 5. Compared to the work above we not only determined the SSRs for the most broadly expressed and tissue-specific genes, but we also determined them for genes with mid-range pattern of expression. The `mreps` (Kolpakov *et al.*, 2003) program is a software for identifying serial repeats in DNA sequences used in this work. In particular we make use of the periods and exponents identified by `mreps` and use them as attributes in the classification models trained to predict tissue-specific genes. The period of the repeat describes the length of a repeated string while the exponent describes the number of repeated copies. For example, the period and exponent of CGG CGG are 3 and 2 respectively.

2.1.5 Polysomes

As mentioned earlier, ribosomes read the sequence of messenger RNAs and assemble proteins out of amino acids bound to tRNAs. The simultaneous translation of a single mRNA molecule by multiple ribosomes increases the overall rate at which cells an synthesise a protein. Complexes containing several ribosomes are referred to as polyribosomes or polysomes. Simultaneous translation of a single mRNA molecule is observable in electron micrographs and by sedimentation analysis. Two numbers of importance related to simultaneous translation are the ribosome occupancy and ribosome density. Ribosome occupancy refers to the fraction of a given gene's transcripts associated with ribosomes while ribosome density refers to the average number of ribosomes bound per unit length of coding sequence.

In Chapter 5 we use ribosomal occupancy data for *Drosophila melanogaster* to investigate the relationship between ribosomal occupancy and tissue specificity.

2.1.6 Ageing

Ageing is often described as the accumulation of damage to macromolecules, cells, tissues and organs over time. Understanding and characterising the genetic effects on ageing is an ongoing effort, but with the advent of high-throughput sequencing and microarray technologies this task is greatly facilitated. A central step in unraveling the mechanisms involved is to record differences during ageing by comparing gene expression profiles of old animals with young animals. Genetic alterations and environmental interventions in the laboratory to extend lifespan include reducing insulin/IGF-like signaling via mutations, dietary restriction, and reducing stress or temperature (Partridge, 2008). Bioinformatics methods to analyse these data and suggest novel biological hypotheses include profiling of changes in gene expression, evolutionary considerations, or finding orthologues via sequence similarity. In Chapter 6 we investigate how filtering tissue-specific genes from whole-fly gene expression data can help to find more subtle connections to ageing in the fruit fly.

2.2 Computational methods and tools

In the following section we will describe the machine learning methods and evaluation measures used throughout this thesis.

2.2.1 Support vector machines

Support vector machines (SVMs) are widely used to solve data classification problems (Vapnik, 1999). Their flexible structure allows the modeling of diverse sources of data. Further, they are able to deal with high-dimensional and large data-sets, making SVMs a popular choice for application in bioinformatics. In Chapter 4 and 5, SVMs are used for discriminating protein domains belonging to SCOP superfamilies and to predict classes of tissue specificity for fruit fly and mouse genes. In the following, some basic principles and background information on SVMs are given.

SVM models are trained from a set of positively and negatively labeled training vectors. The trained model can be used to classify new unlabeled test samples. SVM learns the model by mapping the input training samples x_1, ..., x_n into a feature space and seeking a hyperplane in this space which separates the two types of examples with the largest possible margin. If the training set is not linearly separable, SVM finds a hyperplane, which optimises a trade-off between good classification and large margin. Kernel functions can also be used to train non-linear classifiers. The two key concepts of SVMs, large margin separation and kernel functions, are explained below.

Large margin separation

Figure 2.4a shows a simple example of linear separable data. The stars and circles can be separated by drawing a straight line so that the circles (negative points) lie on one side of the line and the stars (positive points) on the other side. Large margin separation draws the line so that it is as far away as possible from the points in both data sets (Figure 2.4b). For large margin separation, not the exact location but only the similarity of the data points to each other is important. Similarity of two feature vectors can be computed by the dot-product also known as the scalar or inner product between the corresponding feature vectors (Equation 2.2, as explained below).

Let $\mathbf{x}$ denote a vector with M components x_j, j=1,...,M, i.e. a point in an M-dimensional vector space. The notation $x_{\boldsymbol{i}}$ will denote the i^{th} vector in a data set $\{(x_i, y_i)_{i=1}^{n}\}$ where y_i is the label associated with $\mathbf{x}_i$ and n is the number of examples.

$$\langle w, x\rangle = \sum_{j=1}^{M} w_j \times x_j \tag{2.2}$$

A linear classifier is based on a linear discriminant function of the form

$$y(x) = w \times x + b \tag{2.3}$$

The discriminant function y(x) assigns a score for the input x, and is used to decide how to classify it. The vector w is known as the weight vector, and the scalar b is called the bias. In two dimensions, the points satisfying the equation $\langle w, x\rangle$ = 0 correspond to a line through the origin in the two-dimensional classifier. In a three-dimensional classifier the line is substituted by a plane, and in an n-dimensional

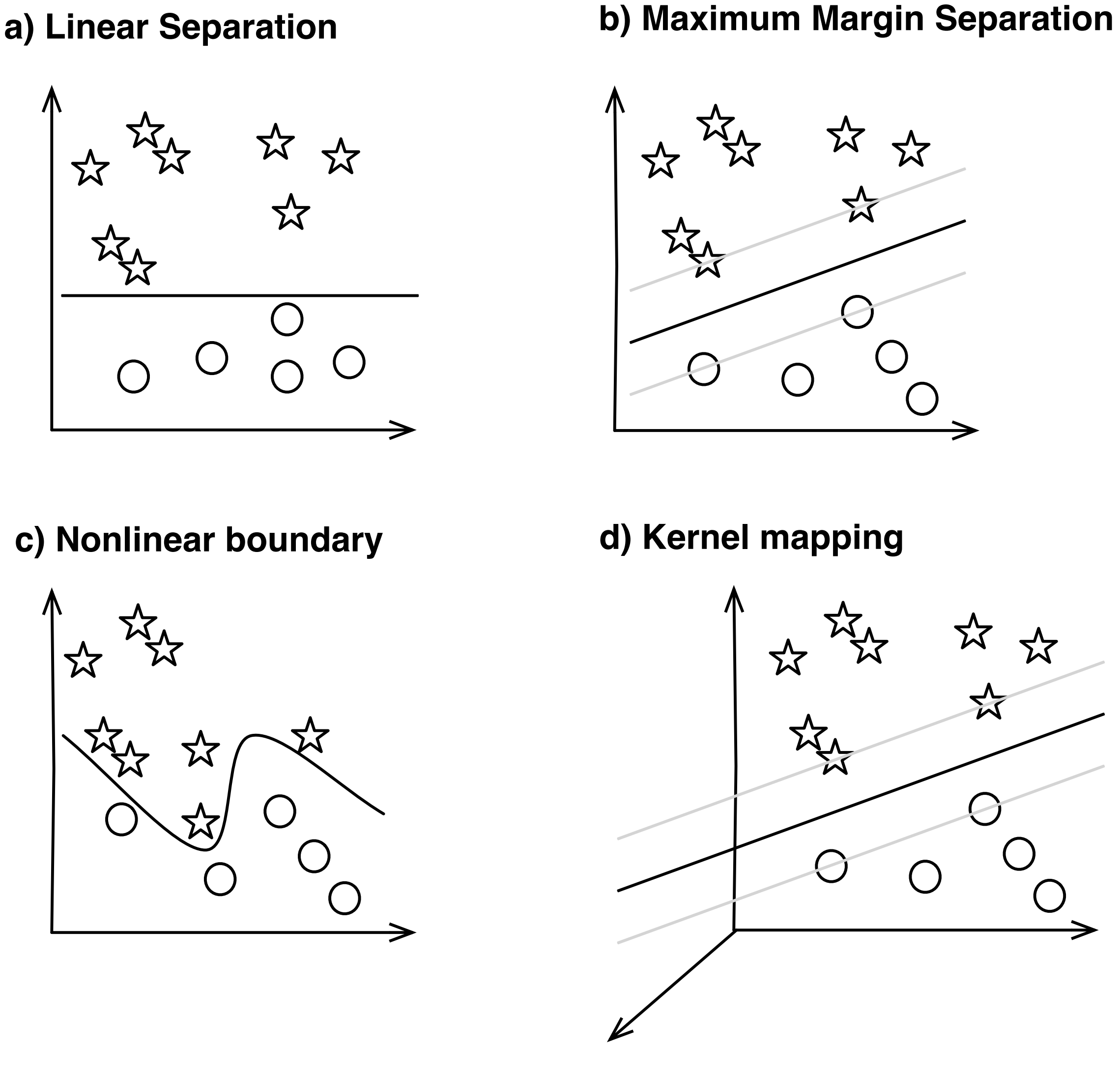


Figure 2.4: **Support vector machine classification - key concepts**. Panel **a** shows a possible hyperplane that separate the data points (positive and negative instances displayed as stars and circles) in two dimensions. Panel **b** shows the optimal hyperplane, the maximum margin boundary, which separates the positive from the negative instances. Panel **c** shows a non-linear boundary. Panel **d** maps the input data from Panel **c** to a separable problem using a kernel function in a higher dimensional feature space.

classifier by a hyperplane. The bias translates the hyperplane with respect to the origin. The hyperplane divides the space into two half spaces according to the sign of y(x), that indicates on which side of the hyperplane a point is located. If y(x)>0, then one decides for the positive class, otherwise for the negative.

Kernel methods

In machine learning, the kernel trick is a method for using a linear classifier algorithm to solve a non-linear problem (Figure 2.4c) by mapping the original non-linear observations into a higher-dimensional space, where the linear classifier is subsequently used; this makes a linear classification in the new space equivalent to non-linear classification in the original space (Figure 2.4d).

The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. The two commonly used families of kernels are polynomial kernels and radial basis functions (RBF). These are the two kernels of choice in Chapter 4 and 5.

Implementation

All computational results presented in this thesis were generated using one of two implementations of SVMs. First, we used the freely available `SVM-light` package (Joachims, 1999) implemented in Java for the work on protein remote homology detection. Second, we used R's implementation in the package `e1071` for the work on predicting tissue-specific classes. The reasons for using these different software packages was simply a switch from the Java to the R programming language due to the multitude of packages available for microarray data analysis for the latter.

2.2.2 Forward feature selection

SVMs generally aim at maximising predictive accuracy, ignoring the important issues of validation and interpretation of discovered knowledge which can lead to new insights and hypotheses which are biologically meaningful and advance the understanding of domain knowledge by biologists. Knowing which features led to a prediction increases the confidence of the biologist in the system's predictions, leading to new insights about the data and the formulation of new biological hypotheses, and detecting errors in the data (Freitas *et al.*, 2010). Decision trees and models are examples of systems that are immediately interpretable. In Chapter 5 we use forward feature selection (FFS) (Miller, 1990) to extract the feature with highest discriminative power in SVMs. Feature selection is the technique of selecting a subset of relevant features for building robust learning models. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models by alleviating the effect of the curse of dimensionality, enhancing generalisation capability, speeding up learning process, and improving model interpretability. In forward feature selection one adds the best feature at each round of classification.

2.2.3 Shannon entropy

The Shannon entropy was introduced by Claude E. Shannon in his seminal paper *A Mathematical Theory of Communication* (Shannon, 1948). It measures the degree of uncertainty for a given variable in a system. The entropy is calculated as a product of probability and the logarithm of probability for each possible state of the targeted variable. Suppose we have the discrete probability distribution p(x$_i$), for the probability of events x$_i$ for 'i' in [1..N], i.e., p(x$_i$) is a discrete probability distribution with N states. Then, the Shannon entropy is defined as:

$$H(X) = -\sum_{x} P(x)\log_2[P(x)] \tag{2.4}$$

bits, where P(x) is the probability that X is in the state x, and $Plog_2P$ is defined as 0 if P=0. H is maximal when all states occur at equal probability. The minimum is taken on if one state occurs at probability 1, the others being ”forbidden”; then H=0 holds. One of the original usages (Shannon, 1948) for Shannon entropy was the measure of information conveyed on average for symbols in a given language, but it has been generalised and applied to many fields in bioinformatics to quantify information content (Herman and Schneider, 1992; Loewenstern and Yianilos, 1999; Ritchie *et al.*, 2008; Schmitt and Herzel, 1997; Schneider, 2000; Strait and Dewey, 1996).

Schug *et al.* (2005) demonstrated the effectiveness of using Shannon information entropy for ranking genes according to their tissue specificity ranging from tissue-specific to ubiquitous expression. This approach was validated using gene expression data from human and mouse, demonstrating that most genes show statistically significant tissue-dependent variations in expression levels. An investigation of the promoter regions of tissue-specific and ubiquitously expressed genes revealed distinct DNA motifs for these classes. Kadota *et al.* (2006) extend this method to account for the fact that entropy alone can only measure the overall tissue-specificity of a gene, but it does not explain to which tissue a gene is specific to. Shannon's entropy was also used to define and estimate the diversity and specialisation of transcriptomes and gene specificity in human data (Martinez and Reyes-Valdes, 2008). We use a similar approach as in the latter paper to define gene specificities in fruit fly and mouse data, as detailed in Chapter 5.1.2.

A toy example is given in the following to further assist the reader with understanding the entropy measurement used in this work. Let us assume there is a mouse, and there are 64 different tissues or organs in the mouse. Let us further assume that there is exactly one gene expressed in one tissue, for instance in tissue number five. According to the probability distribution, the chance that the gene is expressed is the same in each of the tissues i.e., $\frac{1}{64}$=0.015625. The task is to guess in which tissue the gene is expressed by asking questions with yes and no answers. One strategy that always leads to the answer using a minimal number of questions is to divide the search space. For example, the first question one would ask is: "Is the tissue number we are looking for < 32?". If it is, one can go on to ask if the number is < 16, and then < 8, and so on until the search space is narrowed down and the right answer is reached. If the total number of tissues is 64, one will always find the right answer by asking exactly six questions. And that is the basic principle of Shannon entropy. It can be computed by taking the negative binary logarithm of the probability that the gene is expressed in a tissue (here: $-\log_2(\frac{1}{64})=6$). In cases were the gene is expressed in several tissues, the average of all cases is computed using Equation 2.4. In Chapter 5.1.2 the average frequency of a gene among tissues is also taken into account for normalisation purposes.

2.2.4 Performance measures

Receiver Operating Characteristic (ROC) curves

To evaluate classifier performance in Chapter 4 we use receiver operating characteristic (ROC) curves, which show the true positive rates (TPR) on the y-axis over the full range of false positive rates (FPR) on the x-axis. The distribution of the test results of a classifier often overlaps, as shown in Figure 2.5. For every possible threshold value that is selected to discriminate between the two classes, there will be some cases with the positive class correctly classified as positive (TP = true positive fraction), but some cases within the positive class will be classified negative (FN = false negative fraction). On the other hand, some cases in the negative class will be correctly classified as negative (TN = true negative fraction), but some cases will be classified as positive (FP = false positive fraction).

In a ROC curve the true positive rate (sensitivity (recall), Equation 2.5) is plotted in function of the false positive rate (1-specificity; see Equation 2.6 for specificity) for different threshold values. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination has a ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test (Zweig and Campbell, 1993).

$$sensitivity\,(recall) = \frac{TP}{TP + FN} \tag{2.5}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.6}$$

$$precision = \frac{TP}{TP + FP} \tag{2.7}$$

It is possible to average the curves from several runs. In Chapter 4 we use vertical averaging to combine the results of several benchmark sets. Vertical averaging takes vertical samples of the ROC curves for fixed FP rates and averages the corresponding TP rates.

Another measure of accuracy that is frequently used by biologists is precision. Precision is defined as the ratio of true positives over the sum of true positives plus false positives (Equation 2.7). If there are no false positives, the precision is 100%. In a typical analysis, there is a trade-off between recall and precision. A combined measure of these two numbers is accuracy, which is defined as the ratio of true positive plus true negative cases over the total number of cases. Precision/recall curves are presented in Chapter 4. We mainly use ROC curves to measure the performance of the classifiers, as used by other work to which we compared our results (De Ferrari and Aitken, 2006; Handstad *et al.*, 2007).

Area Under the ROC curve (AUC)

An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers it is often useful to reduce ROC performance to a single scalar value

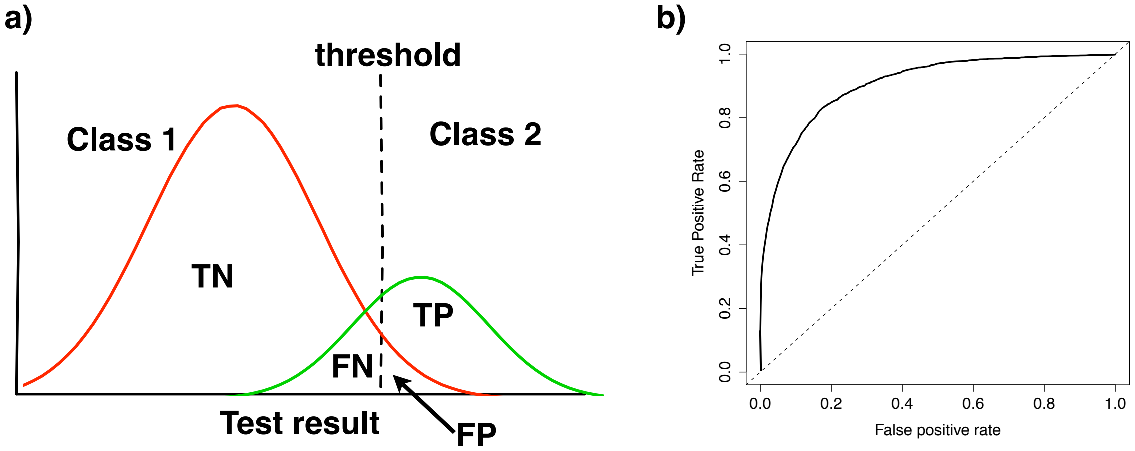


Figure 2.5: **Receiver Operating Characteristic (ROC)**. Panel **a** shows an overlap of the distribution of the test results of a classification problem. Panel **b** shows an example ROC curve. The dashed line in the latter panel represent the result of a random classifier.

representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC. AUCs are used in Chapter 4 and 5 to compare classifier performance. The AUC quantifies the quality of the classifier, and a larger value indicates better performance. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Values range from 0 to 1.

Fisher ratio

In Chapter 5 the Fisher ratio is used to select the most discriminant amino acid that separates tissue-specific from broadly expressed genes. The Fisher ratio is a measure of class distinction which reflects the difference between classes relative to the standard deviation within the classes. It is calculated as:

$$\text{Fisher's ratio} = \frac{(m_1 - m_2)^2}{v_1 + v_2} \tag{2.8}$$

where m_1, and m_2 are the means of class 1 and class 2, and v_1, and v_2 the variances. Tighter classes have smaller variances. The difference between the means should be higher and the standard deviation of each class should be lower for linearly separable cases. Fisher ratio provides an insight of how much two classes are separable. The higher the score the more separable are the two classes.

p value

P values are used in Chapter 4 and 6 to indicate the significance of a particular result, and are explained in a generic manner below. A p value describes the probability that a particular result, or a result more extreme than the result observed, could have occurred by chance, if the null hypothesis were true. The null hypothesis typically proposes a general or default position, such as that there is no relationship between two quantities, or that there is no difference between a treatment and the control. The lower the p value, the less likely the null hypothesis, so the more significant the result. Commonly used thresholds for rejecting a null hypothesis are p values < 0.05 or < 0.01, corresponding to a 5% or 1% chance respectively of an outcome at least that extreme. The threshold is often represented by the Greek letter α (alpha).

False discovery rate (FDR) control and q value

The false discovery rate (FDR) is often used in multiple hypotheses testing to correct for multiple comparisons. In a list of rejected hypotheses, the FDR controls the expected proportion of incorrectly rejected null hypotheses. The FDR can be considered as the expected false positive rate. For instance, if 1000 functional terms were over-represented in a comparison, and a maximum FDR for these observations was 0.10, then 100 of these observations would be expected to be false positives. The q value (Storey, 2003) of an hypothesis test measures the minimum FDR that is obtained when calling that test significant. In Chapter 6 we estimate q values for each functional term found to be over-represented in sets of genes.

Student's t-test

In Chapter 4, a paired t-test is used to test if the differences in classifier performances were significant when predicting remote homologues. The paired t-test provides an hypothesis test of the difference between population means for a pair of random samples whose differences are approximately normally distributed. The test statistic is calculated as:

$$t = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}} \tag{2.9}$$

$$variance = s^2 = \frac{\sum(y - \bar{y})^2}{n - 1} \tag{2.10}$$

$\bar{y}_A$ and $\bar{y}_B$ are the means of the two samples A and B, s_A^2 and s_B^2 are their variances, and n is the number of elements in the two samples.

Standard linear regression

Linear regression refers to any approach to modelling the relationship between two variables, such that the model depends linearly on the unknown parameter to be estimated from the data. Such a model is called a linear model. Linear models are used in Chapter 4 and 5 to describe the relationship between gene specificity and other variables (Equation 2.11). The R function `lm` which stands for 'Linear Model' was used for this purpose.

$$y = a + bx \tag{2.11}$$

Correlation coefficient

A widely-used type of correlation coefficient is Pearson *r*. The correlation coefficient determines the extent to which values of two variables are proportional to each other i.e. linearly related. The correlation is high if it can be approximated by a straight line sloped upwards or downwards. This line is called the regression line or least

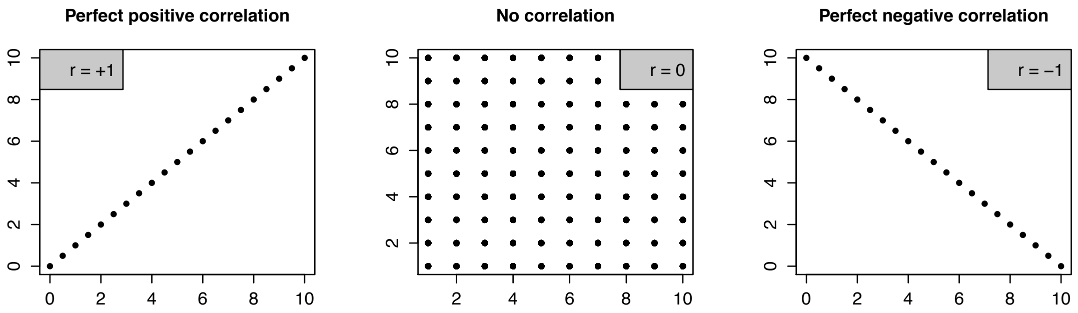


Figure 2.6: **Corrleation coefficient example**. The three extreme cases are shown were there is perfect positive correlation, no correlation and perfect negative correlation between two variables.

squares line, because it is determined such that the sum of the squared distances of all the data points from the line is the lowest possible. The Pearson product moment correlation coefficient for two variables x and y is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.12}$$

where $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively.

The statistic r ranges from -1, through 0, to 1 where -1 corresponds to perfect negative correlation, 0 to no correlation, and 1 to perfect positive correlation (Figure 2.6). The closer the coefficients are to +1.0 or –1.0, the greater is the strength of the linear relationship.

The correlation coefficient r are used in Chapter 5 and 6 of this work to quantify the strength as well as direction of relationships of several variables e.g. between tissue specificity of gene expression and sequence length or mean expression signal.

Boxplots

Boxplots are used throughout the document (e.g. Figure 4.4), and rather than to explain them each time individually, they are explained in a generic form here. In boxplots, the box itself contains the central 50% of the data. The upper edge of the box indicates the 75th percentile of the data set, and the lower edge indicates the

25th percentile. The range of the middle two quartiles is known as the inter-quartile range, with the lines within the boxes representing the median data values. The ends of the vertical lines signify the minimum and maximum data values, unless outliers are present in which case the vertical lines extend to a maximum of 1.5 times the inter-quartile range. The points outside the ends of the vertical lines are outliers.

Chapter 3

Introduction and Literature Review

3.1 Remote homology detection

The following paragraphs explain the motivation that led to the first main project pursued during the course of this PhD study entitled ”Remote Homology Detection Using a Kernel Method that Combines Sequence and Secondary Structure Similarity Scores”. Relevant literature related to this topic is reviewed, and the method is introduced. More detailed methods, results and discussion are given in Chapter 4.

The motivation for this project originated from the observation that protein sequences are being accumulated in public data repositories at an exponential rate. The number of proteins, however, for which a three-dimensional atomic structure has been determined and for which the biochemical function has been experimentally verified, is comparatively low (Berman *et al.*, 2010; The UniProt Consortium, 2010). To make matters worse, metagenomics projects swamp the scientific community with even more sequences: for example, 6.12 million proteins from the Global Ocean Sampling Expedition were published recently and await characterisation (Rusch *et al.*, 2007). Machine learning and data-driven statistical modelling techniques are being developed in order to assist characterisation; these techniques aim to predict structural, functional and evolutionary relationships between these proteins automatically (Friedberg, 2006). Generally, distinctions are made between instance-based learning, and generative and discriminative methods. Instance-based learning methods typically classify an unknown sequence based on the nearest training sequences in a database of known proteins. An example is the k-nearest neighbour algorithm (Shakhnarovich *et al.*, 2005), which is often used in conjunction with the pairwise alignment algorithms Smith-Waterman (Smith and Waterman, 1981) or BLAST (Altschul *et al.*, 1990).

A widely used generative approach is the hidden Markov model (Durbin, 1998) that characterises the likelihood of a given biological sequence being generated by a statistical model. Decision trees (Quinlan, 1990) and SVMs (Vapnik, 1999) are discriminative approaches. They train a classification model to distinguish a group of proteins, which has some property of interest (positive examples) from a set that is known not to have this property (negative examples) (Kretschmann *et al.*, 2001; Liao and Noble, 2003).

Of the above-mentioned methods, instance-based learning is the simplest. It is efficient in detecting homologues if sequence similarity is close. Much of the challenge in making predictions from amino acid sequences, however, arises from the fact that a higher degree of variability can accumulate at the sequence level than at the atomic structure level during evolution; i.e. multiple sequences can give rise to similar structures (Chothia and Lesk, 1986). An example is shown for the SCOP (Murzin *et al.*, 1995) domain pair *d1emy* (myoglobin) and *d1it2a* (hagfish hemoglobin) in Figure 3.1. The sequence identity between these domains is only 16.4 per cent, while their structural similarity is close which is indicated by a RMSD of structure alignments of only 1.667 Å(angstrom) with 688 atoms aligned. According to the protein data bank (PDB)(Berman, 2008) classification they are both concerned with *Oxygen transport* and according to the UniProtKB (Wu *et al.*,

a) Sequence Alignment

```
d1emy__             2 LSDGEWELVLKTWGKVEADIPGHGETVFVRLFTGHPETLEKFDKFKHLKT     51
                      |:||:.:.:.|.|.|:..:...:...:.:|.....|:....|.||...|:
d1it2a_            11 LTDGDKKAINKIWPKIYKEYEQYSLNILLRFLKCFPQAQASFPKFSTKKS     60

d1emy__            52 EGEMKASEDLKKQGVTVLTALGGILKKKGHHE---AEIQPLAQSHATKHK     98
                        .::...::|.|.|.:...:..|:....:.|   ..::.|:|.|.|..|
d1it2a_            61 --NLEQDPEVKHQAVVIFNKVNEIINSMDNQEEIIKSLKDLSQKHKTVFK    108

d1emy__            99 IPIKYLEFISDAIIHVLQSKHPAEF    123
                      :...:.:.:|...:..:..  .|||
d1it2a_           109 VDSIWFKELSSIFVSTIDG--GAEF    131
```

b) Secondary Structure Alignment (SS defined by DSSP, 3 States)

```
d1emy__             1 LLLHHHHHHHHHHHHHHHLLHHHHHHHHHHHHHHHLHHHHHHLLLLLLLL     50
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
d1it2a_            10 LLLHHHHHHHHHHHHHHHLLHHHHHHHHHHHHHHHLHHHHHHLLLLLLLL     59

d1emy__            51 LHHHHHLLH---HHHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHH--     95
                            |||   ||||||||||||||||||||||||||||||||||||
d1it2a_            60 ------LLHHHLHHHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHH    103

d1emy__            96 ----LLLLLHHHHHHHHHHHHHHHHHHLLLLLLHHHHHHHHHHHHHHH      139
                          |||||||||||||||||        |||||||||||||||||||
d1it2a_           104 HHLLLLLLLHHHHHHHHHHHH--------LLLLHHHHHHHHHHHHHHH      143
```

c) Secondary Structure Alignment (SS predicted by PSIPRED)

```
d1emy__            21 HHHHHHHHHHHHHHHHHHHHHHCHH--------CCCCCCHHHCCCCHHHH     62
                      |||||||||||||||||||||||||        ||||||   ||||||||
d1it2a_            23 HHHHHHHHHHHHHHHHHHHHHHCHHHHHHHHHCCCCCCC---CCCCHHHH     69

d1emy__            63 HHHHHHHHHHHHHHHHCCC---HHHHHHHHHHHHHHHCCCCHHHHHHHHH    109
                      |||||||||||||||||||   ||||||||||||||||||||||||||||
d1it2a_            70 HHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHH    119

d1emy__           110 HHHHHHHHHCCCCCCHHHHHHHHHHHHHHHH    140
                      |||||||||      ||||||||||||||||
d1it2a_           120 HHHHHHHHH------HHHHHHHHHHHHHHHH    144
```

d) Alignment by PyMol

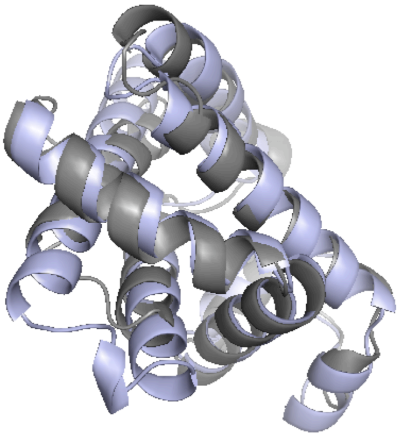


Figure 3.1: Alignments of (a) sequence, (b & c) secondary and (d) tertiary structures of proteins *d1emy* (myoglobin) and *d1it2a* (hagfish hemoglobin) from the SCOP family *a.1.1.2* (globins). Bold symbols and the pipe symbols indicate an alignment match between two columns. Colon and period symbols indicate conservative and semi-conservative substitutions, respectively. While there is a close structural similarity between these proteins, their sequence alignment is poor. The secondary structures, both true and predicted, show close similarities. Chapter 4 of this thesis exploits this observation by augmenting sequence similarities with predicted secondary structures. The alignment was created using the JAligner software.

2006) they share the Swiss-Prot keywords *Heme*, *Iron*, *Metal-binding* and *Oxygen transport*. Sequence-based methods are not very efficient in detecting this similarity. A BLAST query against UniProtKB Release 14.5 using the accession number *P02186* (*d1emy*) returns the protein *Q7SID0* (*d1it2a*) with an E value of 0.067. This is above the E values 10^{-4} to 10^{-30} that are often used as a threshold for confident function assignment (Cai *et al.*, 2006; Engelhardt *et al.*, 2005; Gopal *et al.*, 2001; Martin *et al.*, 2004; The Arabidopsis Initiative, 2000). Approaches to detecting such similarities are called remote homology detection methods. Generally, remote homology detection methods use a generative or discriminative model, because these models are able to detect subtle sequence similarities. One approach, which combines generative and discriminative models, is the work of Jaakkola and Haussler. They showed how score functions derived from a generative model of positive examples can be used in a discriminative setting (Jaakkola *et al.*, 2000); their method is known as the *Fisher kernels*. Recent research suggests that the best-performing methods are discriminative. In this category several techniques use protein sequences to train SVM classifiers. For instance, Liao and Noble introduced an SVM method, called *SVM-pairwise*, which uses Smith-Waterman similarity scores (Liao and Noble, 2003). Each sequence is represented as a vector of the pairwise sequence-similarities scores to all the sequences of the training set. *SVM-pairwise* was shown to perform better than the above-mentioned *Fisher kernel* method. Other alignment-based methods include the *LA-kernel* (Saigo *et al.*, 2004) and *SVM-SW* (Rangwala and Karypis, 2005). Instead of representing the sequences as a vector of features these methods directly calculate the kernels using an explicit protein similarity measure. Leslie and colleagues introduced several string kernels for use with SVMs (Leslie *et al.*, 2002, 2004): *spectrum*, *mismatch* and *profile kernels*. These kernels measure sequence similarity based on shared occurrences of fixed-length patterns in the data. More recently, the *GPkernel* (Handstad *et al.*, 2007) was introduced. The *GPkernel* is a motif kernel based on discrete sequence motifs where the motifs are evolved using genetic programming. The *GPkernel* method achieved better results than the *SVM-pairwise* method, the *mismatch kernel* and a PSI-BLAST (Altschul *et al.*, 1997) based approach. The methods mentioned in the last paragraph primarily use amino acid sequence information to train the classifiers. It is known that sequences which are distantly related but which have similar functions also tend to have highly conserved patterns of secondary structures (Russell and Barton, 1994). Several researchers have demonstrated that the prediction performance of remote homology detection (and fold recognition) methods can be improved by incorporating secondary structure information. Wallqvist *et al.* (2000), for instance, report an increase in sensitivity of their fold-recognition method after modifying the Smith-Waterman algorithm to consider an alignment of both amino acid and secondary structure elements. Ginalski *et al.* (2003) have shown that the addition of predicted secondary structures to conventional sequence profiles is able to boost the sensitivity of profile-profile comparison methods for sequence similarity searches. The technique described in the latter paper is known from fold recognition algorithms, for example (Jones, 1999b; Shi *et al.*, 2001). Chung and Yona present a method for classifying protein families into superfamilies (Chung and Yona, 2004). The authors use statistical models of protein families in the form of profiles and augment the profiles with structural information. The authors note that true structure performs significantly better. Secondary structure element alignments methods (SSEA) have been shown to provide a rapid prediction of the fold for given sequences and have also been applied to the related problem of novel fold detection (McGuffin and Jones, 2002; McGuffin *et al.*, 2001). *DomSSEA* is a modified form of this method that uses predicted secondary structure to predict continuous domains (Marsden *et al.*, 2002). HHsearch (Soding, 2005) is a program based on profile hidden Markov models that augments the sequence profiles with secondary structure.

These and similar studies have indicated that the incorporation of secondary structure information, even if predicted, can increase sensitivity of a protein comparison model. While these studies, which have been covering instance-based and generative learning systems, are clearly important, it is also important to investigate discriminative approaches since it has been pointed out that discriminative approaches generally outperform instance-based methods in remote homology detection. One approach that uses secondary structures in a discriminative setting was introduced in (Hou *et al.*, 2004). Their *SVM-I-sites* method encodes structure information into feature vectors after comparing sequence profiles to the I-sites library of local structural motifs (Bystroff and Baker, 1998); it achieves a comparable performance to the sequence-based *SVM-pairwise* method. One limitation of this method is that it uses secondary structures only, thus disregarding potentially useful information that is encoded in the amino acid sequence. A method that was tested for using both kinds of information in a discriminative setting is the previously mentioned *profile kernel*. The authors report that secondary structure profiles can help the *profile kernel* achieve better performance. The prediction performance of a classifier that uses the secondary structures alone, however, was not investigated. The results were calculated for the entire SCOP database but not for its respective classes. Our approach is also based on combining sequence and secondary structures in a discriminative setting. Instead of using string kernels based on shared occurrences of fixed length patterns, we exploit the performance improvement gained by using a kernel that measures similarity based on all-against-all Smith-Waterman similarity scores. We calculated Smith-Waterman similarity scores from sequences, from observed secondary structures and from predicted secondary structures; the sets of scores were fed into SVM classifiers separately and in combination. Further, we carried out score re-weighting experiments in which more influence was given to either the sequence or secondary structure similarity scores. We compare our method with the sequence-based *SVM-pairwise* method and with the mean achieved by the *GPkernel* method. We show that a complementary classifier is superior to these sequence-only based methods overall, for the different SCOP classes and for the majority of the families participating in the benchmark. We note that it has been reported before that remote homology detection is improved by integrating secondary structures. The difference to most other works lies in the use of SVMs and kernels thus investigating performance in a discriminative setting in contrast to instance-based and generative models. This work has been published in 2009 (Wieser and Niranjan, 2009). Since the publication of our paper another method was developed that combines sequence and secondary structure information (Wang *et al.*, 2009a). This method complements profile-profile comparison with various structure- and function-related patterns revealed by multiple sequence alignments. The resulting tool, PROCAIN (Wang *et al.*, 2009b), improves homology detection and alignment quality beyond the range of other tools, e.g. COMPASS, a tool for comparison of multiple protein alignments (Sadreyev and Grishin, 2003). In another recent paper, Dobson *et al.* (2009) showed that machine learning approaches to predicting SCOP categories can be improved by performing a sequence enrichment step that exploits unannotated sequences within genomic sequence databases. Margelevicius and Venclovas (2010) presented a new homology detection method based on sequence profile-profile comparison that integrated position-dependent gap penalties. Evaluation results showed that at the level of protein domains the method compared favorably to other tested methods including HHsearch, COMPASS and PSI-BLAST.

Methods and results of this study are detailed in Chapter 4. Next the remaining research problems addressed in this thesis are introduced.

3.2 Tissue specificity of gene expression

A similar method as used in the remote homology detection problem was developed to investigate whether tissue specificity of gene expression is correlated with the sequence and secondary structure of the resulting protein product. The motivation for this is described in the following paragraphs, and a corresponding literature review is also included.

Virtually all of the cells in multi-cellular organisms are genetically identical, i.e., they carry the same DNA. Different cells take on different roles by activating only parts of the DNA they carry. Cell fate changes are accompanied by gene expression changes and consequently, gene expression varies widely amongst cell types and tissues. Capturing these differences is important for a number of studies including developmental and disease or age-related studies. For instance, the success of gene manipulation experiments designed to extend lifespan often depends on the tissue in which the gene is deleted or over-expressed (Broughton and Partridge, 2009). Knowledge of tissue specificity of gene expression can help in targeting the correct tissue in which the pathway to be manipulated has the desired function. The tissue-specific genes can be used as a targeting agent in order to reach a particular tissue or organ. A complete knowledge of the tissue specificity of gene expression in model organisms is desirable for the study of ageing and other biological processes.

Various experimental techniques have been developed to identify tissue-specific gene expression signatures. Microarray analysis (Brown and Botstein, 1999) and in situ hybridization (Jones and Robertson, 1970; Tautz and Pfeifle, 1989) are the most commonly used techniques for the study of spatial patterns of mRNA expression. For instance, high-density microarrays have been used to interrogate the gene expression of the vast majority of protein-encoding genes in 79 human and 61 mouse tissues (Su *et al.*, 2004). A digital atlas containing the expression patterns of around 20,000 genes in the adult mouse brain also exists (Lein *et al.*, 2007). These data were generated using ISH (see Chapter 2.1.2). FlyAtlas is a microarray-based atlas of gene expression in multiple adult tissues in the fruit fly. It currently provides gene expression profiles for 17 adult tissues (Chintapalli *et al.*, 2007). Image-based data from hundreds of *Drosophila* blastoderm embryos has been used to build a model ”virtual embryo” that captures the average spatial expression patterns for 95 genes (Fowlkes *et al.*, 2008). An atlas of gene expression patterns during fruit fly embryogenesis has been assembled based on ISH (Tomancak *et al.*, 2007). About 40% of the genes with detectable expression showed tissue-restricted expression in this study. The Berkeley Drosophila Genome Project (BDGP; http://www.fruitfly.org) is a further resource for tissue-specific data in the fruit fly.

However, experimental data is not always available for all organisms, and our knowledge of tissue-specific expression is incomplete. Experimental techniques have their limitations. For example, microarray studies depend on the ability to isolate mRNA from a particular tissue, which requires dissection of the animal. Frequently this is a formidable task. In the case of worms, this is barely feasible and, in the case of flies, it is laborious and time-consuming to obtain sufficient tissue for analysis. Clean tissue separation is critical. In higher organisms, such as humans and mice, the sheer number of different cell types, organs and tissues make it difficult to obtain a complete picture of genes that are truly expressed in a tissue-specific fashion. In contrast to microarray or other similar assays, ISH does not require dissection. However, the drawback to ISH is that it is not quantitative (Wilcox, 1993). In addition, parameters such as genetic variability, nutritional state, pathogen exposure and effects of the isolation procedure add to the complexity of obtaining accurate and complete tissue-specific measurements. Another limitation is the cost, which limits the number of tissues that can be profiled. If, for example, a cost of £350 per array is assumed (data taken from `http://affy.arabidopsis.info/costs.html` in February 2010), then the sampling of 60 mouse tissues, with 4 replicates each, results in a total cost of £84,000.

Thus, because of lack of suitable experimental techniques, difficulties with dissecting, high costs and time demands to dissect and investigate each and every tissue in many organisms, tissue-specific information is frequently lacking.

There have been efforts to use computational models to predict tissue specificity. For instance, De Ferrari and Aitken (2006) trained a Naive Bayes classifier, based on physical and functional characteristics of genes, to discriminate tissue-specific from housekeeping genes. The input features included cDNA length, CDS length, number of exons, measures of chromatin compactness, percent of GO terms for the gene that match with a housekeeping GO terms list, and percent of GO terms for the gene that match with a tissue-specific GO terms list. Their classifier achieved a 90%, 93% and 97% success rate in classification of human, mouse and fruit fly housekeeping genes. The performance of the classifiers was good. However, it should be noted that including classifier attributes that are based on functional characteristics (here GO terms) lead to at least two issues. First, it is known that there are functional differences between tissue-specific and housekeeping genes. One of the housekeeping GO terms used by the authors is "GO:0005840 ribosome". Ribosomes are the components of cells that make proteins from amino acids and thus they are present in every single cell and tissue. Labelling genes that are annotated with this GO term as housekeeping genes is therefore a simple task. Second, many of the GO terms have been transferred automatically via computational methods and are thus predicted themselves. Basing a prediction method on the outcome of another one is generally not practical because errors can be easily propagated. Features such as the cDNA length, CDS length and number of exons are better suited to be used with a classifier since they are directly observed from the genomic sequence.

In the first part of this thesis, we investigated how combining sequence and secondary structures in a discriminative setting affects the ability of support vector machines to discriminate protein domains. Here, we use a similar method to investigate if tissue specificity of gene expression is correlated with the sequence and secondary structure of the resulting protein product. We are also interested in knowing to what extent this information can be used to predict tissue specificity of genes computationally. Our interest in pursuing this question stems mainly from the observations discussed in the next paragraph.

First, housekeeping genes are known to evolve, on average, more slowly than tissue-specific genes (Hastings, 1996; Larracuente *et al.*, 2008; Zhang and Li, 2004; Zhu *et al.*, 2008). The hypothesis is that broadly expressed genes are subjected to greater negative selection pressure because they must function in a more diverse biochemical environment than do narrowly expressed genes. Substitution rates at nonsynonymous sites show strong negative correlation with tissue distribution breadth. Conversely, silent substitution rates do not vary with expression pattern, even in ubiquitously expressed genes (Duret and Mouchiroud, 2000). Genes selectively expressed in one human tissue can often be discriminated from genes expressed in another tissue purely on the basis of their synonymous codon usage (Plotkin *et al.*, 2004). Gene expression has also been shown to evolve faster in narrowly expressed, compared to broadly expressed, genes (Yang *et al.*, 2005).

Second, the types of proteins encoded by housekeeping and tissue specific genes differ. A mouse study established a relationship between the domains encoded by a gene and its degree of tissue specificity (Lehner and Fraser, 2004). Many protein domains in both tissue-specific or widely expressed genes were enriched. The authors found that genes that encoded domains associated with receptors, ligands and extracellular matrix proteins or in DNA- or nucleic acid- binding proteins were expressed in significantly fewer tissues than were other genes. In contrast, genes encoding protein domains that functioned in protein degradation, in the cytoskeleton or in RNA-binding, were expressed significantly more widely than were other genes. Another study showed that proteins with a universal distribution tend to be predominantly enzymes and transporters, while the tissue-specific forms are dominated by regulatory proteins such as transcription factors (Freilich *et al.*, 2005, 2006).

Third, the two classes showed several different genomic features. For example, human housekeeping genes were found to be more compact and shorter than other genes (Eisenberg and Levanon, 2003). The average length for introns, exons, 3'UTR and 5'UTR was shorter for housekeeping genes than for tissue-specific genes. In a human and mouse study (Schug *et al.*, 2005), most tissue-specific genes were found to typically contain a TATA box, but no CpG island, and they often coded for extracellular proteins. In contrast, the widely-distributed (i.e. least tissue-specific) genes frequently contained CpG islands and often coded for nuclear or mitochondrial proteins. The genes in the class that had no CpG island or TATA box were commonly mid-specificity genes that coded for membrane proteins. Sp1, a binding site for selected ubiquitous transcription factors, was found to be a weak indicator of less-specific expression. YY1 binding sites, also binding sites for ubiquitous transcription factors were strongly associated with the least-specific genes. Several other studies also observed a relationship between the occurrence of CpG content or CpG islands and tissue specificity of genes in various organisms (Elango *et al.*, 2009; Foret *et al.*, 2009; Gardiner-Garden and Frommer, 1987). However, a different study on human genes showed that the majority of tissue-specific genes possessed neither CpG islands nor TATA boxes in their core promoters (Zhu *et al.*, 2008).

A study on chromatin compactness showed that putative Scaffold/Matrix Attachment Regions (S/MAR) were more abundant upstream of tissue-specific genes than upstream of housekeeping genes. S/MARs attach themselves to the nuclear matrix and help the formation of chromatin loops. Tissue-specific genes appear to have less accessible and more compact DNA in their promoter regions, and hence more S/MAR sequences (Ganapathi *et al.*, 2005).

In human, the use of SVMs allowed the identification of DNA hexamers that discriminate tissue-specific gene promoters or regulatory regions from those that are not tissue-specific (Rao *et al.*, 2007). It was also shown that housekeeping and tissue-specific genes in human and mouse differ in simple sequence repeats (SSR) in the 5'-UTR region (Lawson and Zhang, 2008). SSR densities in 5'-UTRs in housekeeping genes were about 1.7 times higher than those in tissue-specific genes. Other regions, i.e., introns, coding exons, 3'-UTRs and upstream regions, also contained different, but less obvious, SSRs.

Recently, the correlation of tissue specificity with genomic structure, phyletic age, evolutionary rate and promoter architecture of human genes was re-evaluated. Again, housekeeping genes were found to be less compact and older than tissue-specific genes. It was confirmed that they evolved more slowly in terms of both coding and core promoter sequences. Housekeeping genes primarily use CpG-dependent core promoters, whereas the majority of tissue-specific genes possess neither CpG-islands nor TATA-boxes in their core promoters (Zhu *et al.*, 2008). A study in the nematode worm, *Caenorhabditis elegans*, showed that gene expression data from whole-animal microarray data can be used to predict tissue specificity of genes (Chikina *et al.*, 2009). The authors leveraged existing whole-animal *C. elegans* microarray data to generate predictions of tissue-specific gene expression and experimentally validated these predictions. SVMs were used to build a predictive model of tissue-specific microarray profiles. The SVMs automatically identified expression patterns that separated genes expressed in a particular tissue from other genes (e.g., neuronal and non-neuronal genes). Their predictions reached a precision of 90% for all of the major tissues of the worm (intestine, hypodermis, muscle, neurons, and pharynx) except germ line. This study concentrated on genes expressed in a particular tissue, but housekeeping genes were not considered.

Lastly, in *Saccharomyces cerevisiae*, a correlation was found between the expression level of a gene and the amino acid composition of its protein (Raghava and Han, 2005). The authors analysed 3,468 *S. cerevisiae* genes. Some amino acid residues were observed to have significant positive correlation, while other residues had negative correlation with the expression level of genes. A significant negative correlation was also found between length and gene expression.

Combining these observations, we speculate that the amino acid content may differ among those genes that show different tissue specificity. The ability to successfully predict whether a given gene is a housekeeping or tissue-specific gene would add to the understanding of tissue specificity and how it arises.

We investigate the use of sequence and secondary structure information with

SVMs to discriminate between tissue-specific and broadly expressed genes. We determined the amino acid sequences and secondary structures in proteins that varied in their tissue specificities and then used the sequences and structures to calculate Smith-Waterman similarity scores. The sets of scores were fed into SVM classifiers, both separately and in combination. We also investigate the effects of combining these attributes with genomic features such as lengths of various genomic regions on the performance of the classifier. We also examine whether a relationship exists between ribosomal occupancy and tissue specificity.

We apply our definition to microarray-based expression data for fruit fly genes and then validate and contrast these results with similar data derived from mouse genes. We concentrate on the fruit fly, *Drosophila melanogaster*, because it is the least studied organism (Table 3.1) for which good data have recently become available. For the fruit fly, we used a data set that contains the expressions of a large number of genes in 17 fruit fly tissues (Chintapalli *et al.*, 2007). For the mouse, we used a data set that contains the expression data of a large number of genes in 61 mouse tissues (Su *et al.*, 2004). These data sets allowed a detailed examination of the relationship between tissue expression of genes and their related protein products.

We found that tissue specificity of gene expression is correlated with the sequence and secondary structure of the resulting protein products, and that this specificity can be predicted to a certain extent in both the fruit fly and the mouse. Simple amino acid counts perform almost as well as using Smith-Waterman similarity scores. Integrating secondary structure does not improve upon prediction performance, but still results in performance that is better than random. Combining amino acid percentages with other attributes, such as cDNA length, CDS length, 5'UTR length, further improves the classifier performance.

The experiments performed and results related to this topic are detailed in Chapter 5.

Table 3.1: **Computational studies on tissue specificity**. The table summarises previously published work that investigated housekeeping (HK) or tissue-specific (TS) genes. TF= transcription factor.

Author,Year	**Species**	**Nr. of genes used**	**Short description**
Eisenberg and Levanon (2003)	human	575 HK genes 5404 non-HK genes	Lengths of genomic regions were studied. HK genes were found to be shorter + more compact compared to other genes.
Schug *et al.* (2005)	human mouse	ranked list of genes **human:** 12,626 probesets (8,571 genes) **mouse:** 12,655 probesets (7870 genes)	Shannon's entropy was used to rank genes according to their tissue specificity. Investigation of correlation between TS and base composition of promoters, CpG islands + TF motifs.
Ganapathi *et al.* (2005)	human	525 HK genes 532 TS genes	Investigation of the differences in the chromatin features (S/MAR regions) between the two groups of genes.
De Ferrari and Aitken (2006)	human mouse fruit fly	**human:** 76 HK genes (103 transcripts) 326 TS genes (580 transcripts) **mouse:** 93 HK genes (113 transcripts) 286 TS genes (564 transcripts) **fruit fly:** 40 HK genes (80 transcripts) 193 TS genes (412 transcripts)	A Naive Bayes classifier was trained to discriminate between the two classes based on: cDNA +CDS lengths, number of exons, S/MAR regions, Poly(dA-dT) + (CCGNN)2–5 motifs, and percentage of GO terms.
Rao *et al.* (2007)	human	2273 non-TS 1817 TS	SVMs were trained to discriminate TS genes from those that are not TS based on 6-mers in the promoter regions.
Lawson and Zhang (2008)	human mouse	**human:** 1914 HK genes 275 TS genes **mouse:** 1597 HK genes 890 TS genes	Investigation of differences of SSRs between HK and TS genes.
Zhu *et al.* (2008)	human	885 TS genes 3140 HK genes	Correlation of tissue specificity with genomic structure, phyletic age, evolutionary rate + promoter architecture was re-evaluated.
Chikina *et al.* (2009)	worm	2872 genes expressed in one or more tissues	Predictions of TS gene expression in various tissues using SVMs based on expression profiles (e.g. brain vs non-brain).
This study	fruit fly mouse	**fruit fly:** 11,804 genes (15,560 transcripts) **mouse:** 21,900 genes (11,355 transcripts) ranked lists of genes	Discriminating TS from HK genes (with various degrees) based on sequences, secondary structures + genomic features (cDNA, CDS, 5'UTR, 3'UTR and protein length, CpG islands and content, SSRs).

3.3 Whole-fly body gene expression data and the ageing fruit fly

The sorted list of genes introduced above was also used to investigate the tissue specificity of transcript profiles obtained from a study of ageing in the fruit fly. The motivation for this is detailed in the next paragraphs, alongside a literature review on relevant topics.

Small invertebrate model organisms such as the nematode worm and the fruit fly are widely used to investigate mechanisms underlying diverse biological processes, including development, metabolism, neurobiology and ageing. Whole genome transcript profiles have become an important tool in such investigations, to direct attention to candidate genes and processes for targeting in subsequent experimental analysis. Because of the small size of these organisms and the difficulties of dissecting specific tissues for analysis, molecular methods have been developed to isolate RNA from specific tissues (Chalfie *et al.*, 1994; Jin and Lloyd, 1997). However, these methods have their own technical limitations (Jin and Lloyd, 1997; Klebes *et al.*, 2002), and RNA expression-profiling is still often applied to RNA extracted from whole organisms or from body parts of heterogeneous tissue composition (Girardot *et al.*, 2006; Kim *et al.*, 2005; Landis *et al.*, 2004; Magwire, 2007; Pletcher *et al.*, 2005). It would therefore be helpful if bioinformatics methods could be developed to extract tissue-specific information from these whole body expression profiles. We have developed such a method for the fruit fly and we have applied the method to RNA transcript profiles obtained from studies of ageing in the adult fly.

The fruit fly has been much employed to unravel mechanisms of ageing and to identify genes that may have a functional role in it (McElwee *et al.*, 2007; Partridge, 2008; Piper and Bartke, 2008; Piper and Partridge, 2007; Pletcher *et al.*, 2005; Skorupa *et al.*, 2008). More than 80 genes whose manipulation increases or decreases lifespan, or alters the phenotypic ageing process, have been identified (de Magalhães and Toussaint, 2004; Kaeberlein *et al.*, 2002). For instance, null mutation of chico, a gene that codes for the single fly insulin receptor substrate, has been found to extend lifespan up to 48% in females (Clancy *et al.*, 2001), while fly mutants for methuselah displayed approximately a 35% increase in average lifespan (Lin *et al.*, 1998). It is well established that the insulin/insulin-like growth factor signaling (IIS) pathway is central to regulation of lifespan in various organisms (Bartke, 2008; Clancy *et al.*, 2001; Cowen, 2001; Holzenberger *et al.*, 2003; Tissenbaum and Ruvkun, 1998). Fat and neuronal tissues appear to be of particular importance for lifespan extension via lowered IIS (Broughton and Partridge, 2009). However, the exact signalling mechanisms and biochemical changes by which this, and other evolutionarily conserved pathways such as TOR (target of rapamycin) signalling (Greer and Brunet, 2008), promote longevity in various tissues are not yet fully understood.

Gene expression profiling has been a useful method for identifying candidate processes for lifespan-extension by reduced IIS and other interventions (Cao *et al.*, 2001; Fu *et al.*, 2006; Hong *et al.*, 2008; Kim *et al.*, 2005; Landis *et al.*, 2004; Magwire, 2007; McElwee *et al.*, 2006, 2007; Miller, 2009; Park and Prolla, 2005; Pletcher *et al.*, 2005; Thompson *et al.*, 2009; Zahn *et al.*, 2006). Identifying groups of genes that are differentially expressed in long-lived nematode worms (McElwee *et al.*, 2006; Miller, 2009; Thompson *et al.*, 2009) and in flies (Kim *et al.*, 2005; Landis *et al.*, 2004; Magwire, 2007; McElwee *et al.*, 2007; Pletcher *et al.*, 2005) has been implicated, for example, with enhanced stress resistance (Miller, 2009) and xenobiotic metabolism (McElwee *et al.*, 2007). However, accurate identification and interpretation of differentially expressed genes have been limited by several factors. Gene expression profiles vary among tissues, and expression of a significant fraction of the genome is highly tissue-specific in the adult fly (Chintapalli *et al.*, 2007; Whitehead and Crawford, 2006) and worm (Hunt-Newbury *et al.*, 2007). Many ageing-related changes can be associated with changes in tissue-specific gene expression, while there also may be a common set of genes that change equivalently in different tissues (Cao *et al.*, 2001; Girardot *et al.*, 2006; Kayo *et al.*, 2001; Lee *et al.*, 2000, 2002; Park and Prolla, 2005; Pattison *et al.*, 2003; Rodwell *et al.*, 2004; Welle *et al.*, 2003; Zhan *et al.*, 2007). For example, genes that make up the mitochondrial electron transport chain appear to decrease in expression with age in multiple tissues (Zahn *et al.*, 2006). In contrast, the effects of ageing are particularly pronounced in the brain, where a reduction in synaptic density has been observed in various organisms (Girardot *et al.*, 2006; Hong *et al.*, 2008). The age-related genes in a fruit fly study in seven tissues at six adult life stages showed tissue-specific patterns (Zhan *et al.*, 2007). However, the authors also identified overlaps of the age-related functional groups among tissues. The seven tissues for which genome-wide expression profiles were measured were the brain, thoracic muscle, gut, malpighian tubule, accessory gland, testis and abdominal adipose tissue. In each tissue hundreds of age-related differentially expressed genes were found. Less than 10% of them in each tissue were in common with any other tissue. Similarly, less than 20% of the biological processes enriched with the age-related genes were in common between any two tissues.

Ideally, to capture and interpret both specific and common transcriptional responses during the ageing process and as a result of interventions that extend lifespan, the animal should be dissected into several tissues before gene expression is profiled. The trade-off is that, in the case of worms, this is barely feasible and, in the case of flies, it is laborious and time-consuming to obtain sufficient tissue for analysis. Consequently, whole flies (Landis *et al.*, 2004; Magwire, 2007; Pletcher *et al.*, 2005), or tissues that are technically easy to separate from the fly body such as the head and thorax (Girardot *et al.*, 2006; Kim *et al.*, 2005) are frequently used. While it is known that a significant fraction of the genome will be missed or under-represented in whole-fly samples (Chintapalli *et al.*, 2007) and that tissue-specific expression may be inadequately captured, there is little information concerning the capacity of microarrays to capture tissue-specific effects of ageing in whole-fly samples. To date, knowledge of tissue specificity of differentially expressed genes is often insufficiently considered in the subsequent data analysis.

We describe a computational approach, which partitions genes according to their tissue specificity, that can be used to address some of the above shortcomings and clarify tissue-specific fly transcripts and gene expression in the ageing fly, and in general. Based on an information theoretical approach, we investigate how to utilise FlyAtlas (Chintapalli *et al.*, 2007), a microarray-based atlas of gene expression in multiple adult tissues, to delineate tissue-specific from ubiquitous expression in whole-fly experiments. We begin by taking the sorted list of fruit fly genes according to their degree of tissue specificity introduced in the previous chapter, obtained from the FlyAtlas gene expression profiles. We then use the defined tissue specificity to determine the capacity of Affymetrix high-density oligonucleotide whole-genome micro arrays to capture tissue-specific age-associated changes in whole-fly samples. Importantly, we find that genes with tissue-specific expression are associated with higher fold-changes amongst significantly differentially expressed genes and a lower mean expression signal indicating that changes in tissue-specific expression might be easier to detect using whole-fly arrays. We also describe how filtering genes with tissue-specific expression from data from a whole-fly ageing experiment affects data analysis and the derivation of meaningful information from the data. The significance of several age-related GO terms was increased after removing tissue-specific differentially expressed genes. This is due to a bias in GO annotation towards broadly expressed genes, and to differences in function of broadly and tissue-specifically expressed genes. This study is complemented by an analysis of the tissue specificity of age-associated genes in the fly. We found that most known age-associated genes are broadly expressed.

The investigations concerning this study are described more fully in Chapter 6.

Chapter 4

Remote Homology Detection Using a Kernel Method that Combines Sequence and Secondary Structure Similarity Scores

Below, we explain the experiments performed to investigate the ability of our method to predict protein remote homologues and present and the discuss the results.

4.1 Methods

4.1.1 Benchmark sets

All the experiments are based on protein domains retrieved from the manually curated SCOP database (Murzin *et al.*, 1995). SCOP classifies proteins with known structures hierarchically into classes, folds, superfamilies and families based on their evolutionary relationships and structural or functional similarities. Figure 4.1 shows a schematic representation of the SCOP hierarchy. The top level places domains with similar secondary structure elements in classes. Examples include

All α protein and *All β proteins.* All the available classes are summarised in Table 4.1. The second level assigns the domains belonging to a class to different folds, depending on the topology of their secondary-structure elements. A fold is divided up into superfamilies. Domains sharing the same superfamily have a probable common evolutionary origin, which is usually suggested by their structural and functional features. The superfamilies are further divided up into families of domains with sequence identity $> 30\%$, or domains with very similar functions and structures. Domain pairs belonging to the same superfamily, but to different families are considered to be remote homologues. The benchmarks used in this work were those defined previously to test the performance of the *GPkernel* (Handstad *et al.*, 2007). Figure 4.1 also shows a schematic representation of the benchmark set. Briefly, domains belonging to one family constitute the positive test set. Domains inside the same superfamily but of different families form the positive training set. Negative training and test domains are taken from outside the superfamily. The negative test set consists of one random family from each of the other superfamilies while the negative training set is composed of the rest of the families in these superfamilies. A total of 102 classification tasks were carried out, each positive training set holds at least ten domains and each positive test set holds at least ten domains (4,019 domains were used in total).

Table 4.1: Available SCOP classes. Classes a, b, c, d and g were used in the benchmarks.

Symbol	Description
a	α-helical domains
b	β-sheet domains
c	α/β domains which consist of β-α-β structural units or motifs that form mainly parallel β-sheets
d	$\alpha+\beta$ domains formed by independent α-helices and mainly antiparallel β-sheets
e	multi-domain proteins
f	membrane and cell surface proteins and peptides (not including those involved in the immune system)
g	small proteins
h	coiled-coil proteins
i	low-resolution protein structures
j	peptides and fragments
k	designed proteins of non-natural sequence

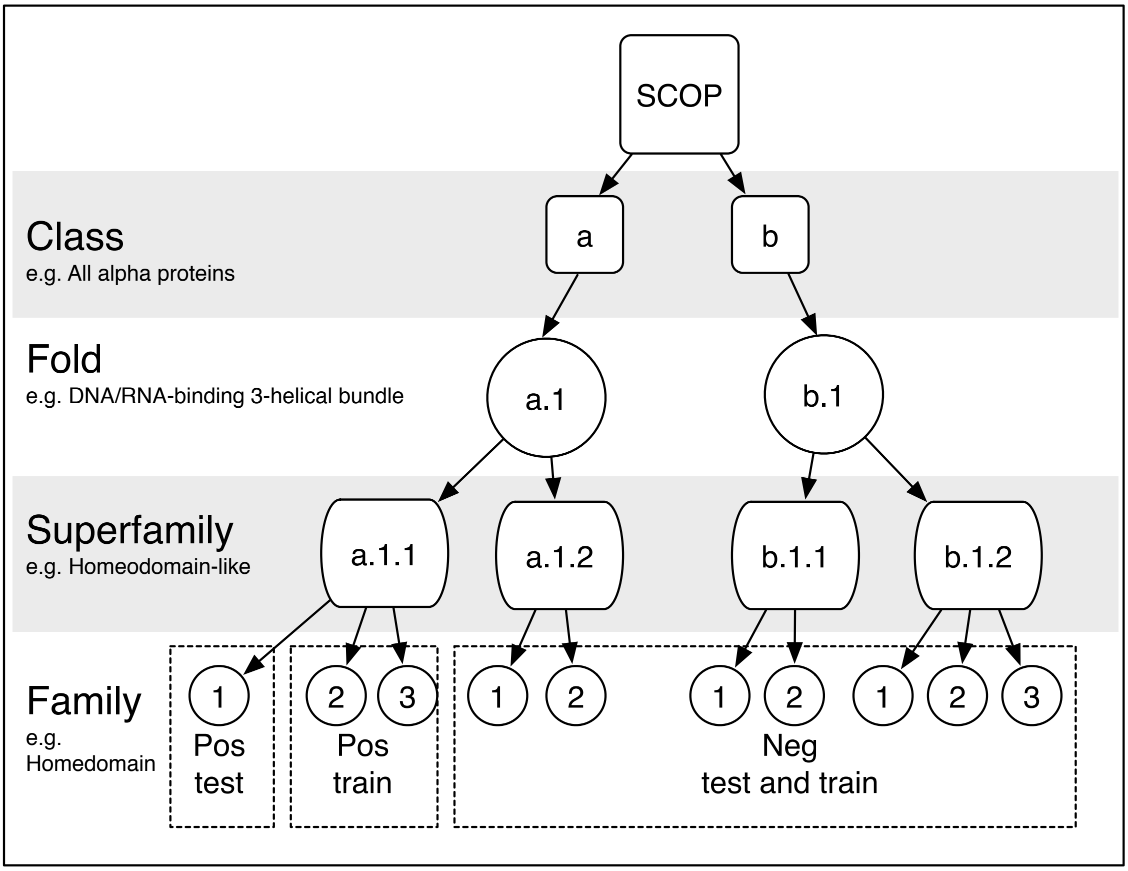


Figure 4.1: Schematic representation of the SCOP hierarchy and setup of the benchmark set. Domains belonging to one family constitute the positive test set. Domains inside the same superfamily but of different families form the positive training set. Negative training and test domains are taken from outside the superfamily.

4.1.2 Data preparation

The data preparation steps are illustrated in Figure 4.2, and further explained below. The SCOP domains were downloaded from the ASTRAL database (Chandonia *et al.*, 2004) (SCOP version 1.67; $<$ 95% identity) and supplemented with observed and predicted secondary structures. The definition of secondary structure of proteins (DSSP) programme (Kabsch and Sander, 1983a) was executed in order to assign an observed secondary structure to each domain included in the benchmark. The DSSP sequences were calculated from PDB entries; PDB-style files were also downloaded from the ASTRAL database. As mentioned earlier, the DSSP distinguishes between eight secondary structure states,namely *H = α-helix*, *B = residue in isolated β-bridge*, *E = extended strand, participates in β ladder*, *G = 3 helix (3_{10} helix)*, *I = 5 helix (pi helix)*, *T = hydrogen bonded turn*, *S = bend.* The PSIPRED programme (Jones, 1999a), version 2.5, was executed in order to retrieve predicted secondary structures for the domains participating in the benchmark. NCBI Toolkit Version 6.1 and blast-2.2.15 were installed and used with PSIPRED. PSIPRED also requires the installation of a sequence database for which we compiled UniProtKB/Swiss-Prot. The latter was filtered to remove low-complexity regions (repetative short fragments), transmembrane regions, and coiled-coil segments, using the `pfilt` programme that is included in PSIPRED. The E value threshold for the `blastpgp` programme used for PSIPRED was 0.001; otherwise default values were used throughout. The PSIPRED programme distinguishes between three secondary structure states, i.e. *H = Helix*, *E = Strand*, *C = Others.* All-against-all pairwise similarity scores were calculated using the JAligner software (Moustafa, 2007), with gap opening and gap extension penalties of 10.0 and 0.5, respectively. JAligner is an open source Java implementation of the Smith-Waterman algorithm for performing local sequence alignments. Here, it was employed to compute alignments between two proteins based on their amino acid or secondary structure symbols. The alignments were scored so that comparatively high scores were given to highly similar alignment

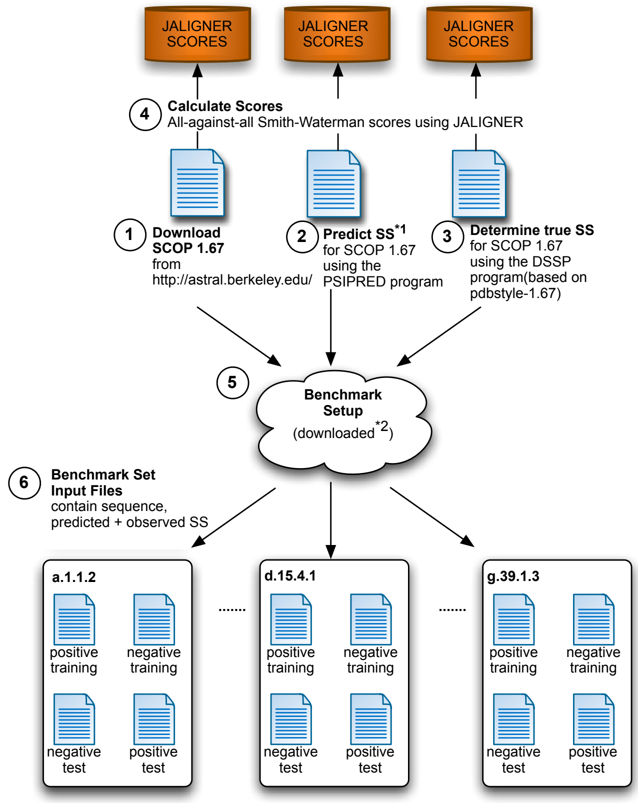
Table 4.2: Scores for different alignment situations. The sequences are toy examples. Scores have been computed using the JAligner software as explained in the methods.

Sequence 1	Sequence 1	score	comment
YYYY	AAAA	0	Unrelated sequences (short)
YYYYYYYYYYYY	AAAAAAAAAAAA	0	Unrelated sequences (long)
YYYYYY	AAA	0	Unrelated sequences (different length)
VDAA	VDAD	14	similar sequences (short)
VDAA	VDAA	18	identical sequences (short)
VDAAVAKVC	VDAAVA	26	identical sequences (different length)
VDAAVAKVC	VDADVAKVD	29	similar sequences (long)
VDAAVAKVC	VDAAVAKVC	44	identical sequences (long)

regions and low scores otherwise. The raw score for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. The similarity matrices used for the sequence alignments and for the secondary structure alignments were BLOSUM62 and IDENTITY, respectively. The JAligner software uses the affine gap penalty model, which charges the score -a for the existence of a gap, and the score -b for each residue in the gap. A gap of k residues thus receives a total score of -(a+bk) while a gap of length 1 receives the score -(a+b). Table 4.2 shows some examples that demonstrate how the scores vary for sequences with different similarities. The score increases with the length for alignments of identical sequences. The scores were not normalised for length, so if A is longer than B, then A aligned with A has a higher score than B aligned with B. The scores (base-ten logarithm) were then used to assemble the SVM input vectors, as described below.

4.1.3 SVM training

We trained six types of SVMs to identify proteins belonging to a superfamily in the SCOP database. We refer to them as *SVM-pairwise+*, *SVM-pairwise (AA)*, *SVM-pairwise (DSSP)*, *SVM-pairwise (PSIPRED)*, *SVM-pairwise (AA+DSSP)* and *SVM-pairwise (AA+PSIPRED)*. AA indicates that the method uses amino acid sequence information, while *DSSP* and PSIPRED indicate the use of secondary structure information and the type of secondary structure assignment program used.



*1 Secondary Structure
*2 http://www.biomedcentral.com/content/supplementary/1471-2105-8-23-S1.txt

Figure 4.2: Data preparation workflow. A total number of 102 benchmark sets were prepared. Sequence and secondary structure alignment scores were calculated for each of the 4019 protein domains participating in the benchmark sets.

SVM-pairwise+ is a simplified version of *SVM-pairwise (AA)* that uses a subset of the all-against-all pairwise similarity scores only (see below). The distance metrics used when combining sequence and secondary structure similarity scores are as given in Equation 4.1:

$$d(i, j) = \alpha d(s_i, s_j) + (1 - \alpha)d(ss_i, ss_j) \tag{4.1}$$

The similarity between protein i and protein j is measured by summing their sequence similarity score calculated from the sequences s_i and s_j and their secondary structure similarity score calculated from the secondary structures ss_i and ss_j. Setting α to its extreme values of one and zero give classifiers that are based purely on sequence alignment scores and secondary structure alignment scores, respectively. The range of i is over both positive and negative training-sequences. In *SVM-pairwise+* the range of j is over the positive training-sequences only, while for all other methods scores are computed by computing alignments over both positive and negative training-sequences. The LIBSVM library (Chih-Chung and Chih-Jen, 2001) was employed to train and test the classifiers using a Radial Basis Function (RBF) kernel for binary classification. Default settings were used for the kernel parameters C and gamma. These were 1 and 1/k respectively where k equals the number of attributes in the input data. The RBF kernel was used because it performed better than the other kernel types available in LIBSVM, all of which were tested using default settings. It is possible that sensitivity can be further boosted if the kernel parameters are systematically optimised. However scarcity of the training-data makes tuning the parameters difficult. The positive classes, for instance, contain only 30 instances on average; to tune parameters properly yet another validation set is needed, which would reduce the amount of data that is available for estimating the SVM parameters. We took the base-ten logarithm of all values of the input feature vector. The feature vectors were normalised before training and a linear scaling applied to range between -1 and 1. The output of the SVM is a discriminant score that was used to rank the members of the test set.

4.1.4 Other methods

The areas under the curves (AUCs) were computed using the R package ROCR (Sing *et al.*, 2005). The same package was used to calculate the average ROC curves in Figure 4.4 (vertical averaging). Following Handstad *et al.* (2007) we compared the significance of the performance differences between the methods by means of p values. We used two-sided paired t-tests, with a confidence interval of 0.95. Thus, $p \leq 0.025$ implies a significant difference between the two methods. Using the more conservative Wilcoxon test resulted in the same conclusions (results not reported).

4.2 Results and discussion

4.2.1 Overall performance

The results of the experiments are generally in agreement with the expectation that a method that combines sequence and structure-similarity scores into one kernel should significantly increase the classifier's performance: the *SVM-pairwise (AA+DSSP)*, *SVM-pairwise (AA+PSIPRED)* methods performed best over all (Figure 4.3, Figure 4.4). The latter methods gave the highest medians - 0.981 and 0.977 (Table 4.4), and show the smallest inter-quartile range after excluding extreme values, i.e. the dispersion of the AUCs is small. The low p values of 0.0007 and 0.0009 suggest that these methods performed significantly better than the sequence-based method (Table 4.3). The averaged ROC curves (Figure 4.4) demonstrate that these classifiers mostly achieve higher or equal true positive rates for arbitrary thresholds of the false positive rate than classifiers that are based on sequence or secondary structure similarity scores only. These two methods also achieve higher precision for arbitrary

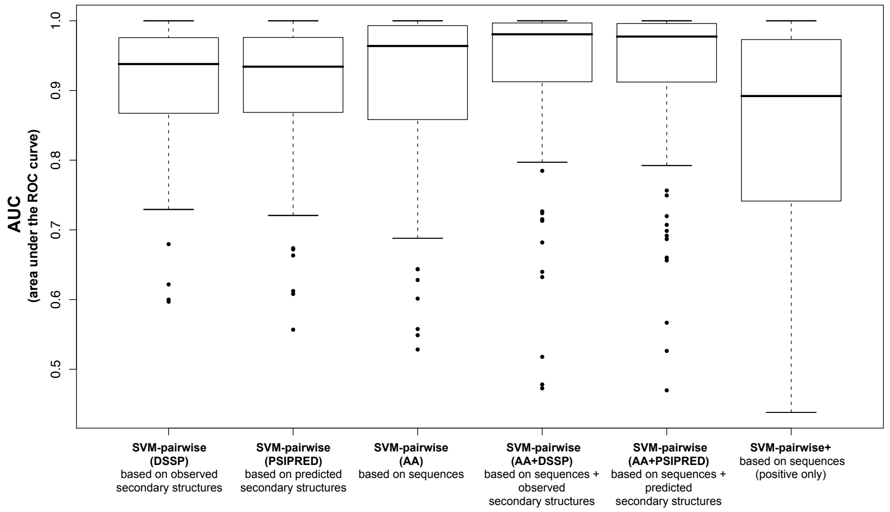


Figure 4.3: Boxplots comparing the performance of six methods. The methods were tested for their ability to predict SCOP superfamily memberships. Performance was measured as AUC (area under the ROC curve). The methods that combine sequence and secondary structure similarity scores - *SVM-pairwise (AA+DSSP)* and *SVM-pairwise (AA+PSIPRED)* - showed the highest median AUCs.

thresholds of the recall rate (black straight and green dotted line in Figure 4.5). The two methods which use secondary structures only - *SVM-pairwise (DSSP)* and *SVM-pairwise (PSIPRED)* - showed similar performances (p value: 0.838); their medians were 0.938 and 0.934 respectively. This indicates that the secondary structures predicted by the PSIPRED programme were useful for improving the detection of remote homologues. The sequence-based *SVM-pairwise (AA)* method achieved a higher mean than the method using secondary structures only, but again the difference is not significant (p values: 0.9416 and 0.9799). The comparison of *SVM-pairwise (AA)* with *SVM-pairwise+* accords with previous observations made by Liao and Noble (2003), where the former performs slightly better than the latter. Figure 4.4 shows the relative performance of the classifiers at a relatively low false positive rate of 0.05. The corresponding mean (averaged over 102 families for each method) true positive rates were 0.70, 0,70, 0.65, 0.60, 0,60 and 0.48 for *SVM-pairwise (AA+DSSP)*, *SVM-pairwise (AA+PSIPRED)*, *SVM-pairwise (AA)*, *SVM-pairwise (DSSP)*, *SVM-pairwise (PSIPRED)* and *SVM-pairwise+* respectively. The corresponding median true positive rates were 0.86, 0.84, 0.75, 0.62, 0.64 and 0.45 for *SVM-pairwise (AA+DSSP)*, *SVM-pairwise (AA+PSIPRED)*, *SVM-pairwise (AA)*, *SVM-pairwise (DSSP)*, *SVM-pairwise (PSIPRED)* and *SVM-pairwise+* respectively. We also calculated the mean and median true positive rates for a false positive rate of 0.01. The resulting means were 0.58, 0.58, 0.53, 0.39 and 0.35 and the medians were 0.66, 0.67, 0.61, 0.35 and 0.26 (reported in the method order as above). Our methods performed better than recently published results for this data set. The *GPkernel*, for instance, achieved a mean 0.902. The best performing method in their study was the LA-kernel that achieved a mean of 0.919 (Table 4.4).

4.2.2 Performance for SCOP classes

Table 4.5 shows the same results divided into the different SCOP classes. The classes participating in the benchmark were: *All α proteins*, *All β proteins*, *α and β proteins (a/b)*, *α and β proteins (a+b)* and *Small proteins*. The medians and means achieved by *SVM-pairwise (AA+DSSP)* and *SVM-pairwise (AA+PSIPRED)* were the highest for all classes. In particular, domains belonging to the classes *All β proteins* and *α and β proteins (a+b)* are generally classified more easily by these methods. For example, the median and mean achieved by *SVM-pairwise (AA+DSSP)* for the class *All β proteins* were 0.963 and 0.900, respectively; the medians calculated by the other methods ranged from 0.817 to 0.920 and their means ranged from 0.782 to 0.891. The fact that the mean is only slightly higher than the means calculated by some of the other methods is largely due to two outlier families (Figure 4.6) which cause the overall mean to drop: *b.40.2.1 Bacterial AB5 toxins, B-subunits* and *b.40.2.2 Superantigen toxins, N-terminal domain*; this is further discussed in the next section. *SVM-pairwise (AA+DSSP)* also calculated the highest median and mean for all the other SCOP classes. Note that domains belonging to the class *e: Small proteins* are assigned easily by using the sequence-based method; *SVM-pairwise (AA)* achieved a median of 0.995 and mean of 0.991. Therefore the improvement of

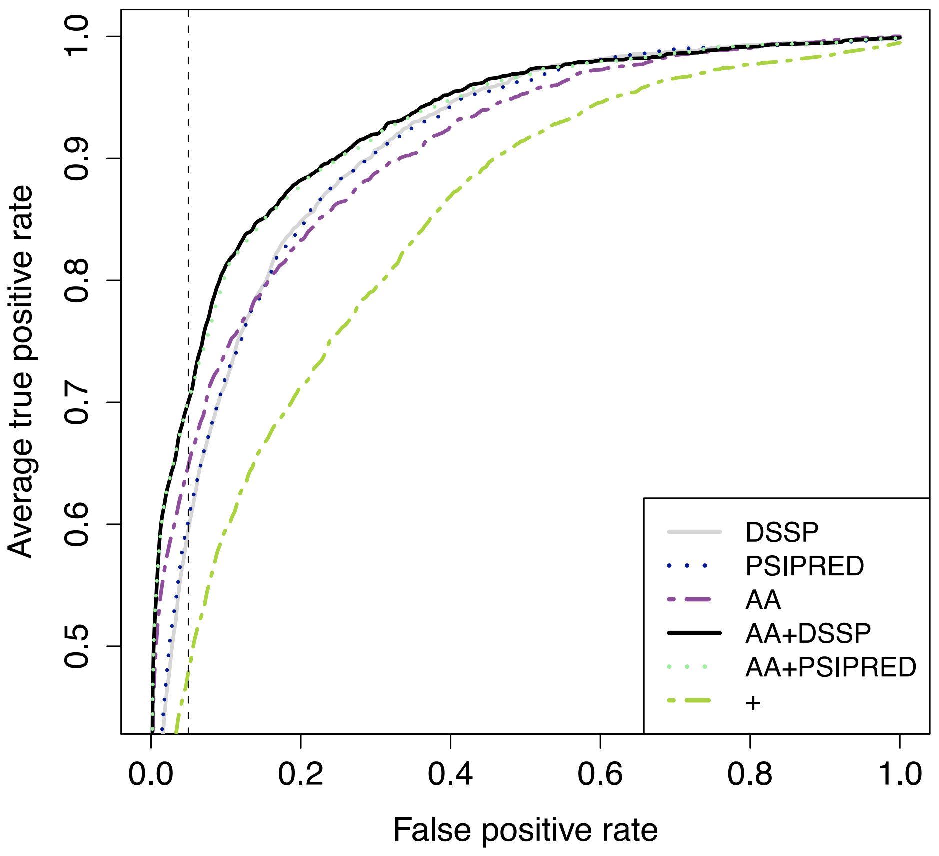


Figure 4.4: **Averaged ROC curves**. For each method the 102 ROC curves – one for each superfamily benchmark set - were averaged using vertical averaging. We trained six types of SVMs to identify proteins belonging to a superfamily in the SCOP database. We refer to them as *SVM-pairwise+*, *SVM-pairwise (AA)*, *SVM-pairwise (DSSP)*, *SVM-pairwise (PSIPRED)*, *SVM-pairwise (AA+DSSP)* and *SVM-pairwise (AA+PSIPRED)*. The vertical dashed line indicates the performance of the method at a false positive rate of 0.05.

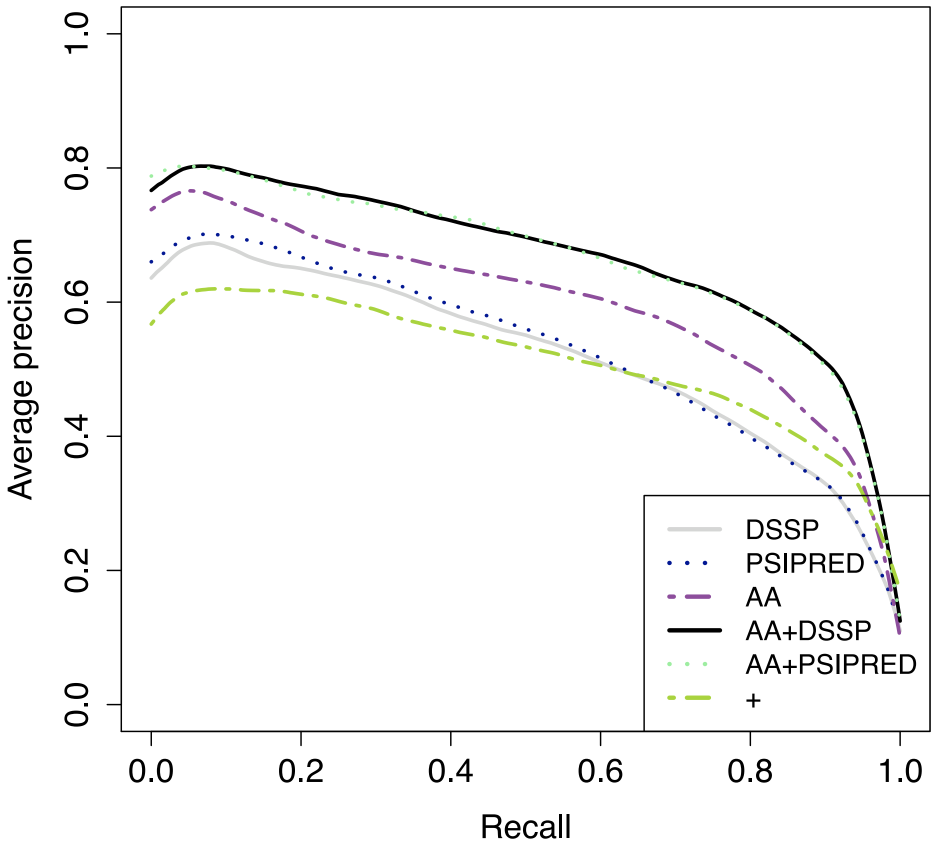


Figure 4.5: **Averaged precision/recall curves**. For each method the 102 precision/recall curves – one for each superfamily benchmark set - were averaged using vertical averaging. We trained six types of SVMs to identify proteins belonging to a superfamily in the SCOP database. We refer to them as *SVM-pairwise+*, *SVM-pairwise (AA)*, *SVM-pairwise (DSSP)*, *SVM-pairwise (PSIPRED)*, *SVM-pairwise (AA+DSSP)* and *SVM-pairwise (AA+PSIPRED)*.

Table 4.3: P values indicating the significance of different AUCs. A p value < 0.025 suggests that the difference is significant.

SVM-pairwise method 1	SVM-pairwise method 2	p-value
AA	*PSIPRED*	0.9799
AA	*DSSP*	0.9416
AA	*AA+DSSP*	0.0007
AA	*AA+PSIPRED*	0.0009
AA+DSSP	*AA+PSIPRED*	0.5214
DSSP	*PSIPRED*	0.8381

SVM-pairwise (AA+DSSP) over this method is only slight. Another reason could be that the secondary structure element *L (Loops and irregular elements)*, as defined by the *DSSP*, is the most frequent element in this class. Proteins belonging to this class might be less structured, and therefore, the method might benefit less from the structural annotations in this case. It can be concluded that the complementary methods showed better performances over all and in each individual SCOP class. The two classes *All* β *proteins* and α *and* β *proteins (a+b)* benefit the most by using sequence and secondary structures.

4.2.3 Performance for constituent families

To establish whether the hypothesis that *SVM-pairwise (AA+DSSP)* performs better than *SVM-pairwise (AA)* is universal across all 102 families, a family-centric version of the data was plotted (Figure 4.6). The majority of the families - the dots in the plot that have a positive value on the x-axis - display a classification improvement, or show equal performance. A performance improvement of ≥ 0.1 was observed for the families: Hemoglobin I (*a.1.1.2*); Phycocyanin-like phycobilisome proteins (*a.1.1.3*); Ferritin (*a.25.1.1*); Galectin (animal S-lectin) (*b.29.1.3*); Pepsin-like (*b.50.1.2*); Pleckstrin-homology domain (*b.55.1.1*); DnaQ-like 3'-5' exonuclease (*c.55.3.5*); Bacterial dinuclear zinc exopeptidases (*c.56.5.4*); and Transferrin (*c.94.1.2*). Discordant families with a performance drop of ≥ 0.1 are: Bacterial AB5 toxins, B-subunits

Table 4.4: Median (Mean) AUCs for SCOP classes. For comparison to other methods tested on the same data set: the GPkernel achieved a mean of 0.902, the LA-kernel a mean of 0.919 (Handstad *et al.*, 2007).

Class: Descriptions	Families	*SVM-pairwise (DSSP)*	*SVM-pairwise (PSIPRED)*	*SVM-pairwise (AA)*	*SVM-pairwise (AA+PSIPRED)*	*SVM-pairwise (AA+DSSP)*	*SVM-pairwise+*
a: All α proteins	16	0.964 (0.938)	0.966 (0.936)	0.993 (0.930)	0.996 **(0.976)**	**0.997 (0.976)**	0.912 (0.841)
b: All β proteins	26	0.918 (0.888)	0.920 (0.891)	0.917 (0.874)	0.959 **(0.904)**	**0.963** (0.900)	0.817 (0.782)
c: α and β proteins (a/b)	39	0.911 (0.906)	0.920 (0.902)	0.958 (0.915)	0.958 (0.924)	**0.966 (0.930)**	0.929 (0.887)
d: α and β proteins (a+b)	12	0.905 (0.863)	0.906 (0.867)	0.898 **(0.899)**	**0.942** (0.870)	0.940 (0.875)	0.680 (0.772)
g: Small proteins	9	0.967 (0.966)	0.968 (0.967)	0.995**(0.991)**	**0.996** (0.990)	**0.996 (0.991)**	0.957 (0.937)
Overall:	102	0.938 (0.907)	0.934 (0.906)	0.964 (0.906)	0.977 (0.927)	**0.981 (0.928)**	0.892 (0.844)

(*b.40.2.1*); and Superantigen toxins, N-terminal domain (*b.40.2.2*). We account for the lower performance of *SVM-pairwise (AA+DSSP)* with two explanations. First, it is possible that there is similarity in the secondary structures, but that the Smith-Waterman algorithm cannot capture this similarity using the IDENTITY matrix. In order to check this hypothesis, we applied a spectrum kernel which measures secondary structure similarity based on shared occurrences of fixed-length patterns in the data. Surprisingly, domains belonging to Bacterial AB5 toxins, B-subunits (*b.40.2.1*) can be predicted easily by using such a spectrum kernel. A *spectrum kernel* with pattern length of 9, 11 and 13 achieved an AUC of 0.983. The AUC calculated by *SVM-pairwise (AA+DSSP)* was 0.473 only. However, overall this method is not very efficient; it achieved a median of 0.827 and a mean of 0.797 only, and was therefore not longer included in the results and discussions. The second explanation is that there is no structural similarity, but that the proteins are grouped for functional reasons. An investigation of Swiss-Prot keywords in the respective families shows that the keyword Toxin is over-represented. Toxins vary greatly in purpose and mechanism, and can be highly complex; presumably this makes them difficult to predict. Intuitively, it is expected that families for which the sequence similarity scores are relatively low and the secondary structure similarity scores are relatively high, benefit most from the usage of *SVM-pairwise (AA+DSSP)*. This hypothesis can be confirmed in several cases; for instance, Hemoglobin I (*a.1.1.2*), Phycocyanin-like phycobilisome proteins (*a.1.1.3*), and Ferritin (*a.25.1.1*). However, it is not universally true; for example, Pepsin-like (*b.50.1.2*) also benefits despite a high level of sequence similarity.

4.2.4 SCOP benchmarks - weighting kernels

We use a weighted sum of two alignment scores to integrate information: the first representing sequence similarity and the second representing secondary structure similarity (Equation 4.1). The results discussed in the previous sections are based

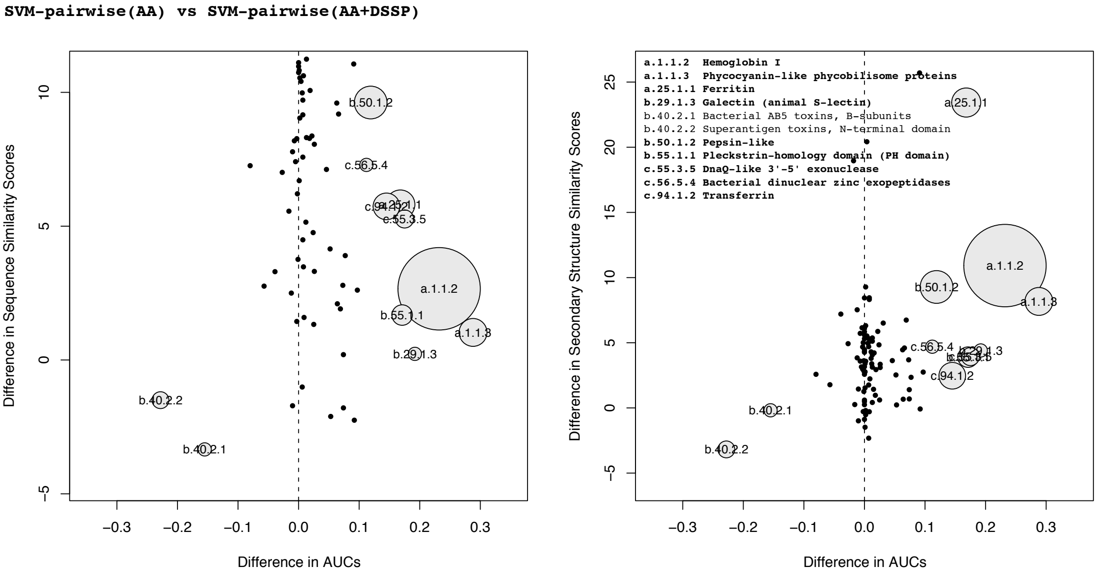


Figure 4.6: Scatterplots showing the relationship between improvement in homology detection and similarity scores. Each of the 102 points in the scatter diagram corresponds to a domain family. A performance increase was observed for all families that are located to the right of the dashed line. Those families for which the performance difference between amino acid sequence based methods and a complementary method (x-axes) is greater than 0.1 are shown as circles, the areas of which are proportional to the numbers of protein domains in them. The y-axes show the discrepancy of average similarity scores between the domains in the positive and the negative test sets.

on experiments that used equal weight on both sets of scores, i.e. $\alpha = 0.5$. Table 4.5 shows the variation in prediction performance of *SVM-pairwise (AA+DSSP)* with α, the differential weighting. We carried out evaluations for only for the highlighted families in Figure 4.6, i.e. those families that showed a performance difference of ≥ 0.1 compared to *SVM-pairwise (AA)*. It was found that Hemoglobin I (*a.1.1.2*) domains can be predicted more easily by using the complementary classifier. The AUCs gradually increase with the weight given to the secondary structure similarity scores; they range from 0.688 to 0.967. The effect of setting different α values for this family is further illustrated in Figure 4.7. Similar to *a.1.1.2*, the AUCs for the families Phycocyanin-like phycobilisome proteins (*a.1.1.3*), DnaQ-like 3'-5' exonuclease (*c.55.3.5*) and Transferrin (*c.94.1.2*) also increase with the weight given on the secondary structure similarity scores; they range from 0.644 to 0.988, from 0.549 to 0.947 and from 0.691 to 0.948 respectively. The performance does not always increase if the highest weight is given to the secondary structure score. Pepsin-like (*b.50.1.2*), for instance, reaches the maximum AUC with an α value of 0.5, i.e. if equal weight is set to sequence and to secondary structure. We observed previously that proteins belonging to Bacterial AB5 toxins, B-subunits (*b.40.2.1*) and Superantigen toxins, N-terminal domain (*b.40.2.2*) are less clearly attributed using the complementary classifier. Weighting the kernels did not result in a clear improvement of performance, indicating that this family is generally difficult to predict.

From this study it is difficult to conclude which weight should be given to the sequence and to the secondary structure similarity scores. However, it is clear intuitively that there is overlapping information in the two scores used in this work - both sequences and secondary structures describe the same protein. Investigating how sequences and their corresponding secondary structures may be combined more effectively in ways other than summing their independently derived alignment scores

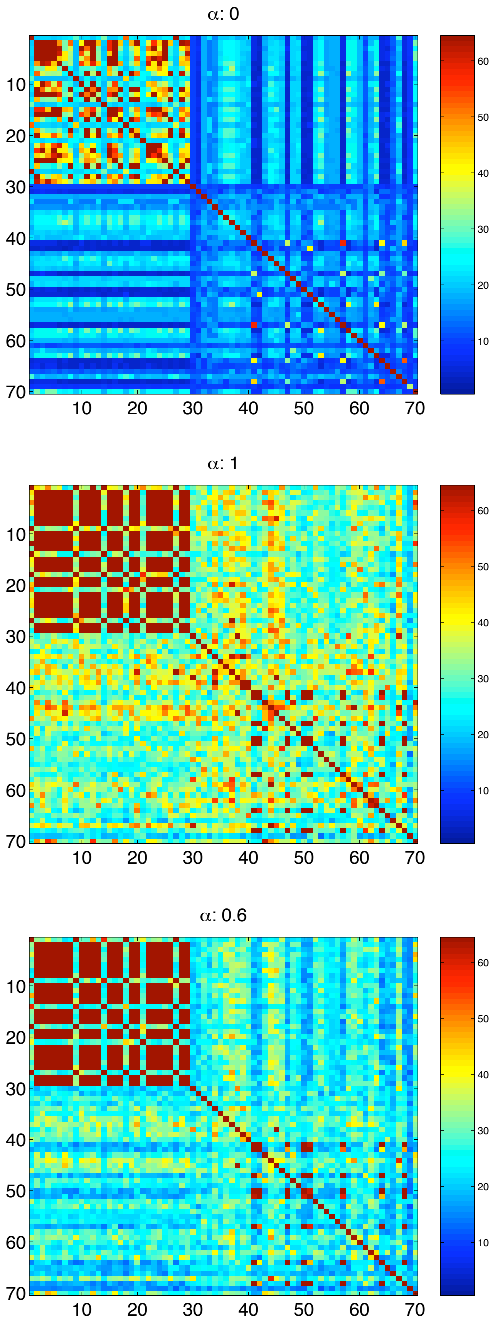


Figure 4.7: Variation in pairwise similarity scores - computed from sequence and predicted secondary structure alignments- plotted as intensities for three different values of α, on a subset of the data i.e. from the training set domains corresponding to the family *a.1.1.2* (globins). All 31 positive examples and a subset (40) of negative examples are shown. The top left corners correspond to the target (positive) class of examples, where sequence (and structure) similarities are high because these are proteins from the same SCOP family. With predicted secondary structures only ($\alpha = 0$), structural correlations within a family decrease, but similarities across proteins from the negative class almost disappear (high blue). When combining these sources of information ($\alpha = 0.6$), we see that the scores of alignment to the negative class can be suppressed, while maintaining high similarities for proteins within the target family.

Table 4.5: AUCs calculated by weighting. Setting α to its extreme values of one and zero give classifiers that are based purely on sequence alignment scores and secondary structure alignment scores respectively.

α	a.1.1.2	a.1.1.3	a.25.1.1	b.29.1.3	b.40.2.1	b.40.2.2	b.50.1.2	b.55.1.1	c.55.3.5	c.56.5.4	c.94.1.2
1.0	0.688	0.644	0.831	0.723	0.628	**0.706**	0.793	0.811	0.549	0.601	0.691
0.9	0.764	0.737	0.916	0.782	0.595	0.629	0.809	0.846	0.581	0.627	0.733
0.8	0.822	0.809	0.962	0.829	0.569	0.582	0.853	0.890	0.586	0.632	0.757
0.7	0.865	0.865	0.980	0.864	0.532	0.435	0.882	0.925	0.604	0.656	0.761
0.6	0.893	0.902	0.994	0.895	0.510	0.435	0.900	0.957	0.667	0.684	0.780
0.5	0.915	0.935	0.999	0.918	0.475	0.483	**0.907**	0.977	0.732	0.703	0.812
0.4	0.925	0.962	**1.000**	0.932	0.574	0.536	0.899	0.987	0.758	0.736	0.858
0.3	0.929	0.973	**1.000**	0.944	0.606	0.594	0.889	0.989	0.768	0.781	0.883
0.2	0.933	0.982	**1.000**	**0.950**	0.646	0.602	0.893	**0.990**	0.750	0.845	0.902
0.1	0.940	0.986	**1.000**	**0.950**	**0.706**	0.606	0.841	0.985	0.813	**0.892**	0.924
0.0	**0.967**	**0.988**	0.999	0.938	0.597	0.600	0.898	0.966	**0.947**	0.888	**0.948**

could extend this work. For example, secondary structures can be used to guide a sequence alignment algorithm in order to get a more biologically relevant alignment, in particular when sequence similarity is low. It has been demonstrated, for instance, that sequence alignments can be improved by limiting the number of gaps in the regions of secondary structures (Barton and Sternberg, 1987; Gerstein and Levitt, 1996; Lesk *et al.*, 1986). Yet the Smith-Waterman alignment algorithm used in this work disregards this knowledge. The algorithm also disregards the fact that mismatches between residues become, in principle, more likely if they correspond to the same secondary structure. Chapter 7 expands on this issue and gives directions on how secondary structures could help to guide sequence alignments using weighted finite state machines (WFSMs)(Cortes and Mohri, 2005). Other ideas for improving this work includes the use cross validation to set an optimal value for α, monitoring performance on a hold-out subset of the training data. In the results reported here on SCOP domains, we did not pursue this due to scarcity of data. Chapter 7 adds some further critical evaluation of this work.

4.2.5 Comparison to PSI-BLAST

Handstad *et al.* (2007) used the same benchmark set as we did in this study. The authors also tested a PSI-BLAST based approach, which is briefly reviewed in the following. The authors first created a multiple sequence alignment of the positive training set. This alignment was given as input to PSI-BLAST, which was then run for 1 iteration against the test set using standard parameter values. The E values of the resulting alignments were used to rank the test set. An average ROC score of 0.575 was reported for the superfamily benchmark, as compared to scores of 0.797-0.919 for the SVM based methods tested in this study. An important issue with the PSI-BLAST based approach described above is the relatively small amount of labelled data used. Through iterative heuristic alignment, PSI-BLAST can leverage unlabeled data from a large sequence database to obtain a much richer profile representation of each sequence, but in the method used by Handstad *et al.* (2007) only the 4,019 protein domains from SCOP were utilised. It is conceivable that had they performed an iterative search against all the sequences in UniProtKB, for example, the search might have found matches outside SCOP via which remote homologous in the test set could have been linked. Alternatively one could also start with the test sequences as starting point for the search rather than with the training sequences, and check if their superfamilies can be successfully detected.

We implemented an improved version of PSI-BLAST that is more sensitive in detecting remote homologous. For this we downloaded and installed PSI-BLAST version ncbi-blast-2.2.23. We also downloaded UniProtKB release 2010_06 including the SwissProt and TrEMBL databases comprising 517,100 and 10,867,798 sequence entries, respectively. The UniProt sequences were combined with the 4,019 SCOP benchmark domains, resulting in 11,388,917 sequences. We generated a combined BLAST database using the `makeblastdb` command. This database was then searched with each positive test sequence available in the benchmark set using an iterative PSI-BLAST search. There were 3,128 positive test sequences in total. Default values were used for running PSI-PLAST (Matrix: BLOSUM62; gap existence penalty: 11 gap extension penalty: 1). The results were recorded and after the 10th iteration (or at the point of convergence which happened earlier in few cases) we checked which SCOP domains were detected with E values < 10. Appendix A shows the resulting hits for one of the families. i.e. family 'a.1.1.2'. All hits are listed, except hits to members of the same family as the query family.

We report the recall rate of such a PSI-BLAST based method for various E value thresholds based on the following method. For each query sequence, we recorded the superfamily of the best hit. We counted a true positive if it matched the superfamily of the query sequence, and a false negative otherwise. Figure 4.8 shows the recall rate that was achieved for various E value cut-offs. If all hits are taken into account a recall rate of 0.57 is achieved. If a more restrictive E value cut-off of 0.01 or 0.0001 are used, the recall rate drops to 0.54 and 0.51, respectively. The effect of further reducing the E value cut-off on the recall rate is illustrated in Figure 4.8. Additional experiments would have to be performed to determine which E value cut-off should be used to detect similarities to remote homologues. Even though it is unclear how to count the false positives in such a method, it is obvious that restrictive cut-offs yield the fewest false positives, improving precision, but reducing the recall rate. A direct comparison to the family-based discrimination methods above is difficult since the method described is not a classification approach in which positive and negative classes were predicted in a rank-based manner. In Chapter 4.2.1 we reported a true positive rate (recall) of 0.58 at a false positive rate of 0.01 for the two best performing methods of SVM-pairwise(AA+DSSP) and SVM-pairwise(AA+PSIPRED). If we assume an equally low false positive rate for the PSI-BLAST result, taking into account all hits up to an E value threshold of 10, then this method achieves a similar recall rate of 0.57.

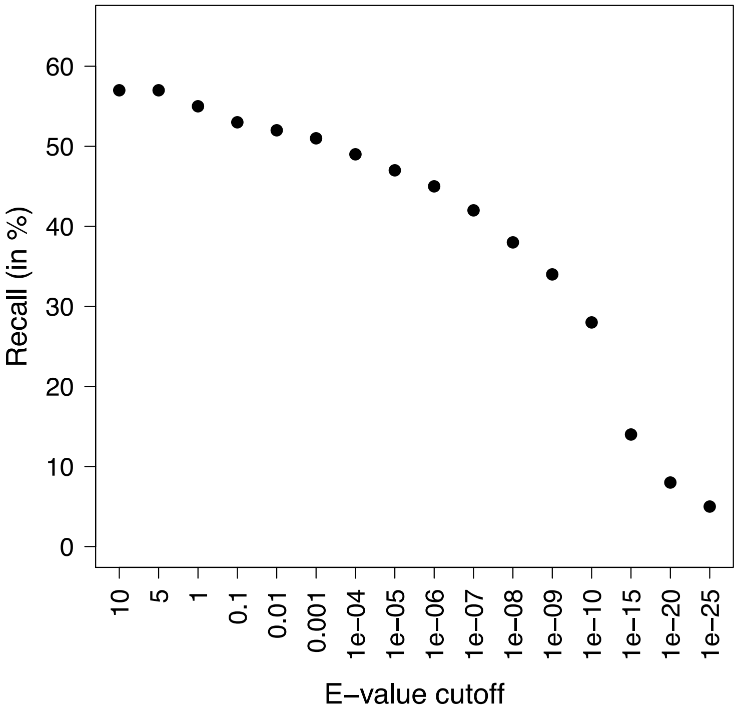


Figure 4.8: The recall (sensitivity) of the PSI-BLAST based method is shown for various E value cut-offs. For each of the 3,128 query sequences, we recorded the superfamily of the best BLAST hit outside the query family. We counted a true positive if this superfamily matched the superfamily of the query sequence, and a false negative otherwise.

Chapter 5

Tissue Specificity of Gene Expression is Correlated with the Sequence and Secondary Structure of Resulting Protein Product

The successful prediction of protein remote homologues through combining sequence and secondary structure similarity scores in a discriminative setting prompted us to investigate a similar method to predict tissue specificity of gene expression, as introduced in Chapter 3.2. Two data sets were prepared for this purpose. These are explained in the following paragraphs, followed by other methods and the results and discussion of this topic.

5.1 Methods

5.1.1 Data collection

5.1.1.1 Fruit fly

Pre-processed data files were downloaded from the FlyAtlas website (Chintapalli *et al.*, 2007). The data consist of probe expression levels using Drosophila Genome 2.0 Arrays. This comprises whole-fly data, and data from 17 tissues dissected from the adult fruit fly. The tissues investigated are the brain, head, eye, thoracicoabdominal ganglion, crop, midgut, hindgut, ovary, testis, accessory gland, carcass, heart, salivary gland, tubule, fat body, spermatheca mated and spermatheca virgin. Genes that were not expressed in any of the tissues were removed from the data set, leaving 14,171 present probe sets. A gene was considered expressed if it was called present in at least 3 of the 4 replicates. The 14,171 Affymetrix IDs were mapped to 11,804 FlyBase gene IDs. For this, all individual probes were mapped against all known and predicted transcripts of the *Drosophila melanogaster* genome release version 5.4. Probes that mapped to more than one gene in the genome and probes that did not map to any known or predicted gene in the genome were excluded from further analysis.

Protein sequences for the 11,804 present genes were downloaded from the Ensembl database (Ensembl 55, BDGP5.4). These are uniquely identified via the FlyBase transcript IDs. In total, 8,598 FlyBase gene IDs could be mapped to exactly one FlyBase transcript ID, while 3,206 IDs were mapped to more than one transcript, resulting in a total of 18,133 protein sequences. If several transcripts IDs mapped to the same sequence, one of the transcripts was removed, resulting in 15,560 unique sequences for the final data set.

The PSIPRED programme (Jones, 1999a), version 2.61, was executed in order to retrieve predicted secondary structures for the proteins participating in the experiment. NCBI Toolkit and Blast (blast-2.2.18) (Altschul *et al.*, 1990) were installed and used with PSIPRED. The PSIPRED programme also requires the installation of a sequence database for which we compiled UniProtKB/Swiss-Prot (version 14.9) (The UniProt Consortium, 2010). As before, the latter was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments, using the `pfilt` programme that is included in PSIPRED. The E value threshold for the blastpgp programme used for PSIPRED was 0.001; otherwise default values were used throughout. The PSIPRED programme distinguishes between three secondary structure states, i.e., H = Helix, E = Strand and C = Others.

The number of exons, the cDNA, CDS, 5'UTR length and 3'UTR length were downloaded from the Ensembl BioMart for each transcript.

5.1.1.2 Mouse

The GNF mouse expression atlas (Su *et al.*, 2004) was downloaded from its website (http://biogps.gnf.org/downloads/). The GEO code for this data set is GSE1133. We also downloaded the chip annotation file from the same site (GEO platform accession: GPL1073). Affy IDs were mapped to Entrez Gene GeneIDs. Rows where an affy ID corresponded to none or several GeneIDs were disregarded. Rows that corresponded to the same GeneIDs were merged and the mean value taken. This resulted in a data set with 21,900 genes. Protein sequences corresponding to these IDs were downloaded from UniProtKB using a Java library named UniProtJAPI (Patient *et al.*, 2008), whereas fragments were excluded. The final data set consisted of 11,356 genes. If several Entrez Gene GeneIDs mapped to the same sequence, one of the genes was removed, resulting in 11,352 unique sequences for the final data set used for the SVM training.

As above, PSIPRED was executed to predict secondary structures for each protein sequence.

This downloaded set comprised data from 61 mouse tissues. The tissues included B220 B cells, CD4 T cells, CD8 T cells, adipose tissue, adrenalgland, amygdala, bladder, blastocysts, bone, bonemarrow, brown fat, cerebellum, cerebral cortex, digits, dorsal root ganglia, dorsal striatum, embryo day 10.5, embryo day 6.5, embryo day 7.5, embryo day 8.5, embryo day 9.5, epidermis, fertilized egg, frontal cortex, heart, hippocampus, hypothalamus, kidney, large intestine, liver, lung, lymph node, mammary gland, medial olfactory epithelium, olfactory bulb, oocyte, ovary, pancreas, pituitary, placenta, preoptic, prostate, retina, salivary gland, skeletal muscle, small intestine, snout epidermis, spinal cord lower, spinal cord upper, spleen, stomach, substantia nigra, testis, thymus, thyroid, tongue, trachea, trigeminal, umbilical cord, uterus and the vomeralnasal organ.

The number of exons, the cDNA, CDS, 5'UTR length and 3'UTR length were downloaded from the Ensembl BioMart for each Entrez Gene GeneID.

5.1.2 Ranking genes according to their tissue specificity

We calculated the degree of tissue specificity for each of the 11,804 fruit fly genes and 21,900 mouse genes by measuring the degree by which a given gene's expression profile differed from a ubiquitous expression profile. We used the gene specificity index S_i to measure the tissue specificity of each gene. The formula used to calculate the gene specificity is based on an adaptation of Shannon's information theory to the transcriptome framework, as described in Martinez et al. (Martinez and Reyes-Valdes, 2008). First, the gene expression profiles were converted into frequency matrices by calculating the relative frequency p_{ij} for the ith gene (i = 1,2, ..., g) in the j^{th} tissue (j = 1, 2, ..., t). The average frequency p_i of the i^{th} gene among tissues of an organism is defined by

$$p_i = \frac{1}{t} \sum_{j=1}^{t} p_{ij} \tag{5.1}$$

$$S_i = \frac{1}{t} \left(\sum_{j=1}^{t} \frac{p_{ij}}{p_i} log_2 \frac{p_{ij}}{p_i} \right) \tag{5.2}$$

$$C_i = log_2 \frac{t}{i} \tag{5.3}$$

where t is the number of tissues examined in each organism (Equation 5.1). The gene specificity then is defined as the information that its expression provides about the identity of the source tissue (Equation 5.2).

The S_i values range from 0 to $\log_2$(t), with higher values indicating higher tissue specificities. A minimum of 0 is achieved if the gene is expressed at the same level in all tissues, and the maximum of $\log_2(17) = 4.09$ in the fruit fly, and $\log_2(61) = 5.93$ in the mouse, is achieved if the gene is expressed in only one tissue. Examples of S_i values for genes with various expression patterns are shown in Figure 5.1. We used Equation 5.3 to define 17 *bins* for the fruit fly and 61 *bins* for the mouse, with each *bin* defining a group of genes with a certain degree of tissue specificity. Bin C_i (i = 1,2, ... , t) defines the central S_i value associated with a group of genes. The differences between two C_i values of two neighbouring bins were divided in half so that the upper 50% of S_i values was attributed to the first *bin* and the lower 50% to the second bin. Figure 5.2 summarises how many genes were attributed to the respective bins in both organisms. Note that bin *1* contains genes that are expressed in one tissue only, whereas *bin 17* in the fruit fly, and *bin 61* in the mouse, contain genes that are expressed in all tissues. Note also that the distribution to the bins could have been done differently. For example, we could have chosen to distribute an equal number of genes to each bin. We used the formula above because it separates the data into the bins observed after visual inspection of the gene-specificity against the mean expression signal. Using this formula results in a class imbalance for some of the bins. For some of the experiments, we chose the n-top an n-bottom genes from the sorted list of gene specificities, in order to reduce dependency on this imbalance.

For evaluation purposes we compared and contrasted the ranked gene list calculated using Shannon's information theory with the gene list published in Suppl.

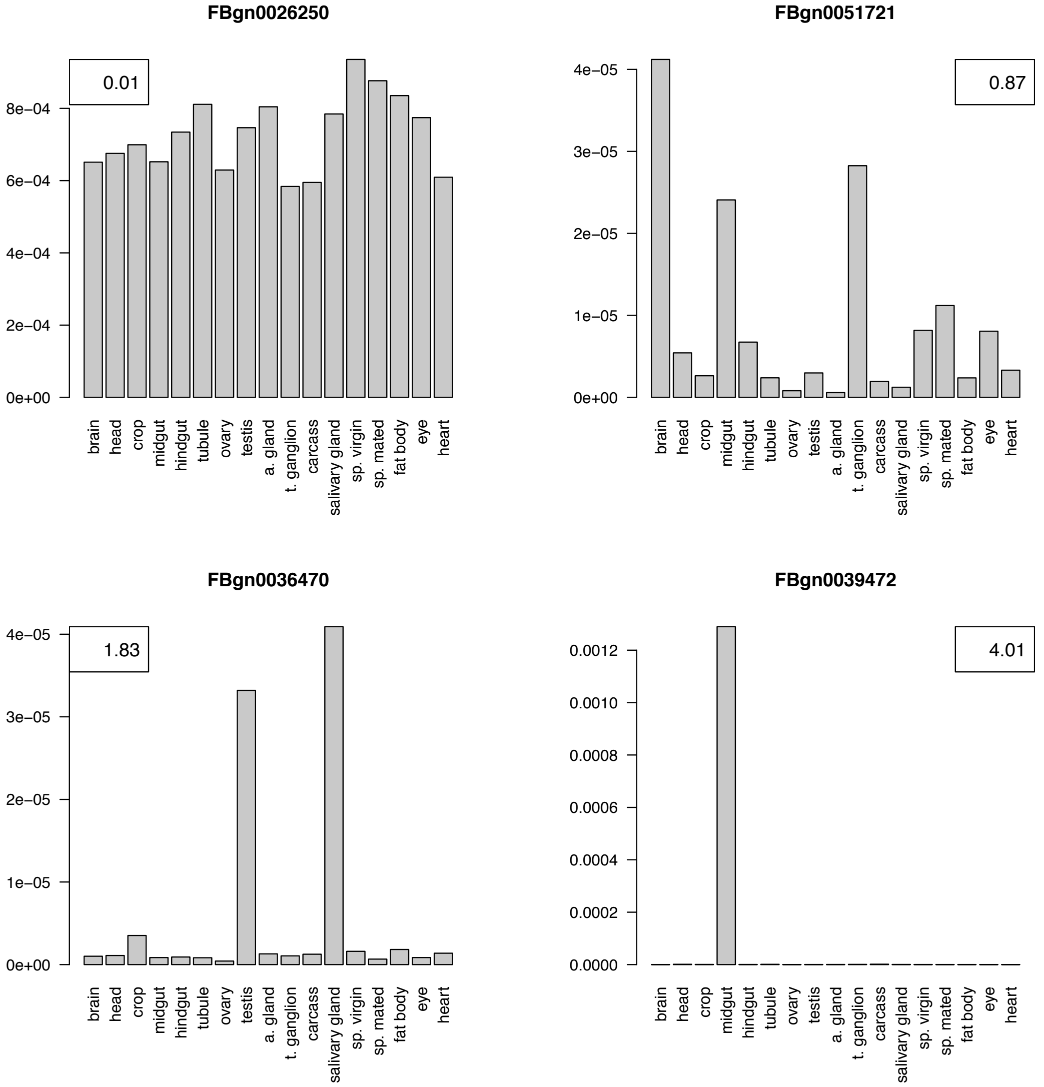


Figure 5.1: **Example gene specificities.** The gene specificity values (legend boxes) are shown for 4 fruit fly genes with various expression profiles in 17 tissues. Genes with a ubiquitous expression profile have gene specificity values close to 0, while genes with specific profiles have gene specificity values close to 4. (a. = accessory , t. = thoracicoabdominal , sp.= spermatheca)

tables S1 and S2 of FlyAtlas to address the differences and similarities of these gene lists. S1 lists 45 highly specific genes, with the five top ranking genes included for each tissue. These represent genes present in all four replicates in that tissue and absent in all four replicates everywhere else. The genes also have the highest signals possible compared with other tissues. Similarly, a list of 50 genes was published in S2 containing genes that are ubiquitously expressed. For this, the mean expression value for each tissue and the standard deviation of all means were calculated for each probe set. The resulting standard deviations were divided by the mean of all means. The resulting list was sorted and the top 50 genes were selected. The Shannon's entropy method is largely in agreement with the selection procedure explained above (median S_i for S1= 3.825, median S_i for S2= 0.049), except for some genes that are classified highly tissue-specific by FlyAtlas but show a midrange pattern of expression by the method used in this work ($S_i < 2$ for: *FBgn0029090*, *FBgn0033419*, *FBgn0033702*, *FBgn0038526*, *FBgn0052815*). These correspond to lowly expressed genes that are detected in only one tissue using present/absent counts. An important advantage of using Shannon's entropy formula is that it also considers potential biases in the expression levels across the tissues that express the gene (Schug *et al.*, 2005). The S_i value for the disparate genes is relatively low because the expression levels measured for these genes are similar in most tissues. Another important advantage of using Shannon's entropy formula is that it not only clearly separates the tissue-specific from ubiquitously expressed genes, but also allows us to distinguish between genes that show mid-range expression. Furthermore, using a single measurement for defining gene specificity makes it easier to globally analyse and inspect the properties of genes assigned to a specific category. Thus, the ranked gene list presented here is more detailed than previously presented, and a clearly defined statistical framework is used that has previously been proven valuable. The bins containing genes with various degrees of tissue specificity overlap with previously identified categories for tissue-specific (De Ferrari and Aitken, 2006) and housekeeping genes (Farre *et al.*, 2007). Many significantly upregulated terms in the ubiquitous bins relate to genes whose products are involved in RNA processing. Overall, the tissue-specific genes have a more varied set of functions.

5.1.3 Sequence alignments and similarity scores

All-against-all pairwise similarity scores were calculated using the R package `Biostrings` (Pages *et al.*, 2009), with default gap opening and gap extension penalties of 10.0 and 4, respectively. The Smith-Waterman algorithm for performing local sequence alignments was chosen in the settings. The software was employed to compute alignments between two proteins, based on their amino acid or secondary structure symbols. The alignments were scored so that comparatively high scores were given to highly similar alignment regions and low scores were given otherwise. The similarity matrices used for the sequence alignments and for the secondary structure alignments were BLOSUM62 and IDENTITY, respectively. The scores were then used to assemble the SVM input vectors. Note that computing all-against-all sequence alignments is computationally expensive. For the mouse, we computed $2 * (\frac{11{,}356 \times 11{,}356}{2} + \frac{11{,}356}{2}) = 128,970,092$ alignments while we computed $2 * (\frac{18{,}133 * 18{,}133}{2} + \frac{18{,}133}{2}) = 328,823,820$ alignments alignments for the fruit fly.

5.1.4 SVM training

The R package `e1071` (Dimitriadou *et al.*, 2009) was employed to train and test the classifiers, using a polynomial kernel for binary classification. Model parameters were chosen by searching possible values and identifying those that minimised prediction errors on the training data. The polynomial kernel was used because it performed slightly better than did the other kernel types available in `e1071` (linear, sigmoid, rbf) when tested on a small subset of the data.

The models were trained 100 times on a randomly selected 90% of the data and

A) Fruit fly

	a) *Bin* Name	b) Gene specificity value (S_i)		c) Number of genes	d) Number of Proteins (all)	e) Number of Proteins (unique)
		Min	Max			
Tissue-specificity increases ↑	1	3.59	4.09	624	703	676
	2	2.79	3.59	1008	1166	1129
	3	2.29	2.79	529	645	612
	4	1.93	2.29	498	714	657
	5	1.63	1.93	430	635	556
	6	1.39	1.63	416	620	542
	7	1.18	1.39	418	683	603
	8	1.00	1.18	495	739	639
	9	0.84	1.00	482	830	687
	10	0.70	0.84	486	843	716
	11	0.57	0.70	589	984	840
	12	0.44	0.57	685	1251	1020
	13	0.33	0.44	928	1558	1309
	14	0.23	0.33	1247	2192	1801
	15	0.13	0.23	1560	2501	2058
	16	0.04	0.13	1359	1994	1658
	17	0.00	0.04	50	75	57
all	1-17	0.00	4.09	11804	18133	15560

B) Mouse

	a) *Bin* Name	b) Gene specificity value (S_i)		c) Number Genes	d) Number Proteins	e) Number Proteins (unique)
		Min	Max			
Tissue-specificity increases ↑	1	5.43	5.93	72	39	39
	2	4.64	5.43	212	121	121
	3	4.14	4.64	184	109	109
	4	3.77	4.14	160	97	97
	5	3.48	3.77	159	95	95
	6	3.23	3.48	159	98	97
	7	3.03	3.23	129	79	79
	8	2.85	3.03	136	83	83
	9	2.68	2.85	127	71	71
	10	2.54	2.68	141	92	92
	11	2.41	2.54	134	79	79
	12	2.29	2.41	137	74	74
	13	2.18	2.29	144	83	83
	14	2.07	2.18	118	66	66
	15	1.98	2.07	154	85	85
	16	1.89	1.98	148	88	88
	17	1.80	1.89	136	82	82
	18	1.72	1.80	148	86	86
	19	1.65	1.72	137	73	73
	20	1.57	1.65	141	79	79
	21	1.50	1.57	138	85	85
	22	1.44	1.50	150	83	83
	23	1.38	1.44	128	70	70
	24	1.32	1.38	147	96	96
	25	1.26	1.32	152	90	90
	26	1.20	1.26	132	77	77
	27	1.15	1.20	141	76	76
	28	1.10	1.15	151	80	80
	29	1.05	1.10	170	96	96
	30	1.00	1.05	184	113	113
	31	0.95	1.00	148	92	92
	32	0.91	0.95	172	94	94
	33	0.86	0.91	173	100	100
	34	0.82	0.86	189	115	115
	35	0.78	0.82	184	95	95
	36	0.74	0.78	176	92	92
	37	0.70	0.74	204	134	134
	38	0.66	0.70	177	105	105
	39	0.63	0.66	204	113	113
	40	0.59	0.63	218	132	132
	41	0.56	0.59	234	143	143
	42	0.52	0.56	263	140	140
	43	0.49	0.52	262	151	151
	44	0.46	0.49	308	181	181
	45	0.42	0.46	297	168	168
	46	0.39	0.42	318	180	180
	47	0.36	0.39	343	197	197
	48	0.33	0.36	340	204	204
	49	0.30	0.33	412	220	220
	50	0.27	0.30	454	268	268
	51	0.24	0.27	501	289	289
	52	0.22	0.24	516	277	277
	53	0.19	0.22	592	347	347
	54	0.16	0.19	742	394	394
	55	0.14	0.16	824	423	423
	56	0.11	0.14	1042	537	537
	57	0.09	0.11	1358	635	635
	58	0.06	0.09	2131	971	971
	59	0.04	0.06	3514	1458	1458
	60	0.01	0.04	1235	556	556
	61	0.00	0.01	0	0	0
all	1-61	0.00	5.93	21900	11356	11355

C) **Example bin calculation:**
upper/lower threshold for bin 10

$$\left.\begin{aligned} C_9 &= \log_2 \frac{17}{9} = 0.92 \\ C_{10} &= \log_2 \frac{17}{10} = 0.77 \\ C_{11} &= \log_2 \frac{17}{11} = 0.63 \end{aligned}\right\} \text{Equation 5.3}$$

$$\mathrm{Bin}_{10\mathrm{lower}} = 0.77 + \frac{0.77 - 0.92}{2} = \mathbf{0.70}$$

$$\mathrm{Bin}_{10\mathrm{upper}} = 0.77 + \frac{0.77 - 0.63}{2} = \mathbf{0.84}$$

D)

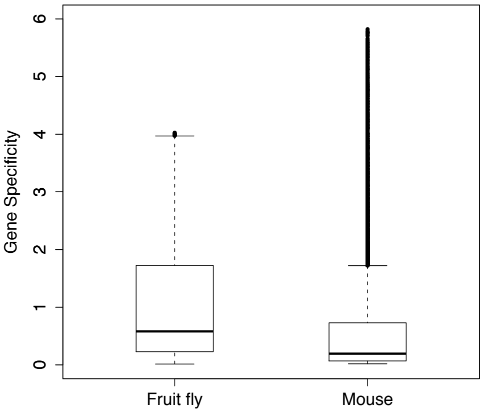


Figure 5.2: **Splitting fruit fly and mouse genes into bins according to their tissue specificity values (S_i).** The genes are grouped into 17 (fruit fly) and 61 (mouse) different bins as described in the methods section using Equation 5.3. Tissue specificity is highest for *bin 1* and lowest for *bin 17* (fruit fly) and *bin 61* (mouse). a) Bin names used in the manuscript b) The S_i cutoffs used to assemble the respective bins c) The number of FlyAtlas genes assigned to the bins d) The number of proteins assigned to the bins e) The unique number of proteins assigned to the bins (for the mouse these numbers are almost identical with the numbers in d). A) Fruit fly B) Mouse C) Example on how the upper S_i threshold for bin 10 was calculated for the fruit fly D) Boxplots showing the gene specificity distribution of all genes in both organisms.

performance was tested on the remaining 10% of the data. The resulting ROC curves for each run were combined using vertical averaging.

5.1.5 SVM input vectors

The SVM input vectors assembled are described in the following subsections. Their predictive performances were compared using mean AUCs which are given a a later section.

5.1.5.1 Composition of amino acid residues

Each protein in the training data set of proteins was characterised by a vector (i = 1, ..., 21) representing the amino acid composition, together with a positive or negative label for discriminating the two different groups (e.g. tissue-specific and broadly expressed genes). The vector had 20 elements for the amino acid composition since there are 20 possible amino acids. Amino acid composition is defined as the ratio between the number of occurrences of a specific amino acid residue and the total number of residues in a protein.

5.1.5.2 Composition of secondary structure symbols

Each protein in the training data set of proteins was characterised by a vector (i = 1, ..., 4) representing the secondary structure element composition, together with a positive or negative discrimination label. The vector had 3 elements for the secondary structure composition since there are 3 states predicted by PSIPRED. Secondary structure composition is defined as the ratio between the number of occurrences of a specific secondary structure state and the total number of residues in a protein.

5.1.5.3 Smith-Waterman similarity scores

Each protein in the training data set of n proteins was characterised by a vector (i = 1, ..., n+1) representing the Smith-Waterman similarity scores computed against all other proteins in the training data set, together with a positive or negative discrimination label.

The sequence based classifier was trained on sequence similarity scores, while the secondary structure based classifier was trained on secondary structure sequence similarity scores. For the combined classifier, each protein in the training data set of n proteins was characterised by a vector (i = 1, ..., 2× n+1) representing the Smith-Waterman similarity scores computed against all other proteins in the training data set, using both the amino acid sequences and secondary structure sequences, together with a positive or negative discrimination label.

5.1.5.4 Genomic features

Each protein in the training data set of proteins was characterised by a vector representing the genomic features, together with a positive or negative discrimination label.

The attributes used for each transcript were: 1. protein length (log), 2. number of exons, 3. presence of S/MAR in the 5' region (binary), 4. presence of S/MAR in the 3' region (binary), 5. cDNA length (log), 6. CDS length (log), 7. 5'UTR length (log), 8. 3'UTR length (log), 9. the number of CpG islands, 10. the CpG content and 11. the mreps period and the mreps exponent.

If a value was not available for a transcript, it was set to 0.

5.1.5.5 Combined input vector

The attributes used for each transcript were as described above for the genomic features. In addition, the percentage of each of the 20 amino acids was used.

5.1.6 Other methods

The areas under the curves (AUCs) were computed using the R package `ROCR` (Sing *et al.*, 2005). The same package was used to calculate the average ROC curves based on vertical averaging.

SSRs were computed using the `mreps` program (Kolpakov *et al.*, 2003) for the 5'UTRs. A minimum length of 10 was required for the sequence repeats, otherwise default values were used. A Java program was written to parse the `mreps` output. The `mreps` period and exponent were used in the SVMs.

The program `newcpgreport` was used to detect CpG islands in the sequences 1,000 bp upstream of the transcription start sites. Default settings were used, i.e., the window size was 100, the shift increment was 1, the minimum length 200, the minimum ratio between observed and expected CpG content was 0.6, and the minimum percentage 50%. A Java program was written to parse the number of CpG islands in each sequence.

The CpG bias of a sequence is defined as the ratio of the observed frequency of CpG dinucleotides divided by their expected frequency (Equation 2.1). The expected number of CpG dinucleotides is the product of the frequency of C and G nucleotides in a given sequence. A Java program was written to calculate the number of CpG dinucleotides.

Genomic features, such as cDNA length, etc., were downloaded from Ensembl using BioMart.

The EMBOSS program `marscan` was executed to find MAR/SAR sites in nucleic sequences (here 1,500 upstream). A Java program was written to extract the number of MAR sites identified.

All Affymetrix arrays taken from Zid *et al.* (2009) were normalised using gcrma background correction to correct for non-specific binding, followed by a quantiles and loess normalisation using the corresponding Bioconductor packages.

5.2 Results

5.2.1 Prediction performances

Prediction performances for all classifiers tested are summarised in Tables 5.4-5.7. Median AUCs are presented in Tables 5.4 and 5.6 while mean AUCs are presented in Tables 5.5 and 5.7. The mean AUCs are presented with a confidence interval to give an indication of unreliability.

5.2.1.1 Protein sequence-based classifier

First, we studied the ability of SVMs to discriminate the various tissue specificities of 17 groups of fruit fly genes and 61 groups of mouse genes, based on the amino acid content of their protein products. Columns 3 in tables 5.4 and 5.6 summarise the AUCs obtained for discriminating the tissue-specific *bin 1* from *bins 1-17* in the fruit fly, and *bins 1-61* in the mouse, based on frequency counts of amino acids in the transcripts. All groups of genes with various tissue specificity showed better than random predictions (AUCs fruit fly = 0.541-0.795, AUCs mouse = 0.596-0.942). The discriminating power increased with the bin number, i.e., the negative training set number, in most cases (black dots in Figure 5.3). In the fruit fly, the performance was best for discriminating *bin 1* and *bin 14* (AUC= 0.795), and worst for genes belonging to the same *bin* i.e., *bin 1* (AUC= 0.541). In the mouse, the performance was best for discriminating *bin 1* and *bin 41* (AUC= 0.942), and worst for genes belonging to the same bin i.e. *bin 1* (AUC= 0.596).

Next, we investigated whether the use of Smith-Waterman sequence similarity scores, rather than simple amino acid counts, improves the ability of the SVMs to discriminate between tissue-specific and broadly expressed genes. Pairwise mean homologies within and between tissue specificity bins are presented for both organisms in Figures B.1, B.2 and B.3 (Appendix B).

Overall, the classifiers using Smith-Waterman similarity scores had more discrimi-

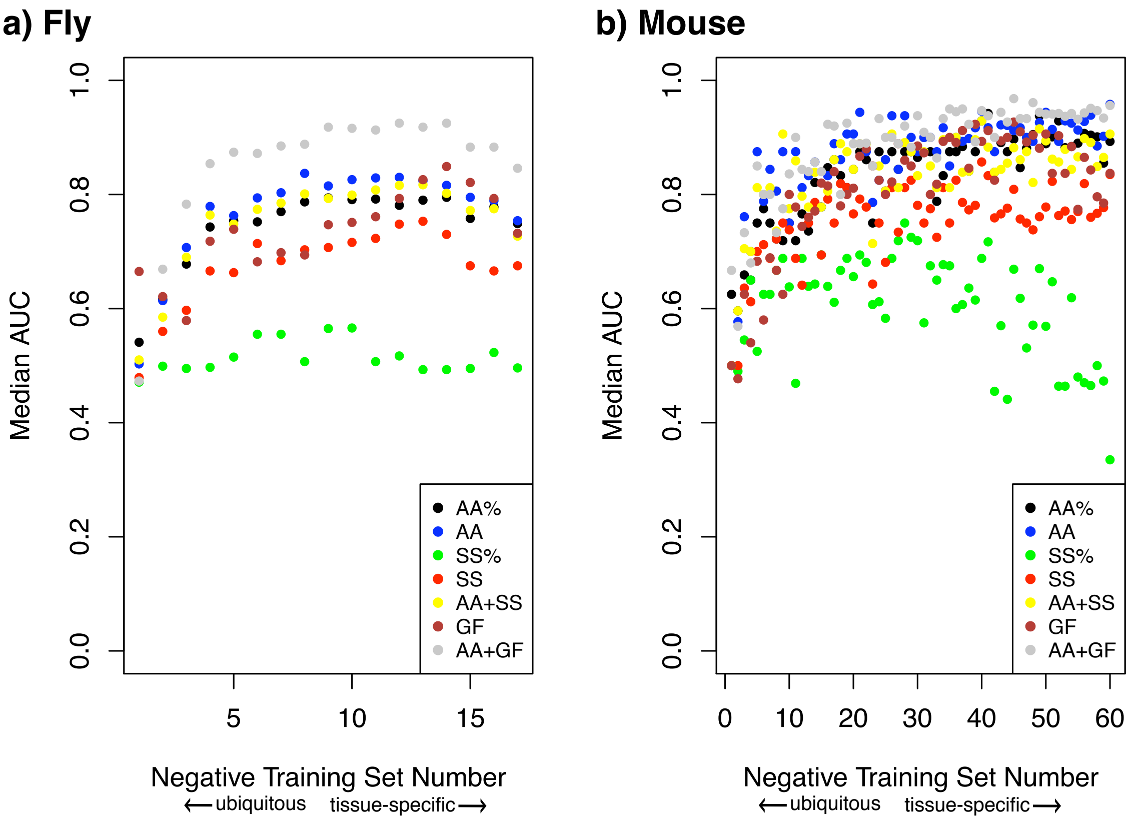


Figure 5.3: **Discriminating tissue-specific genes from broadly expressed genes in the fruit fly and in the mouse**. Prediction performances of SVMs for discriminating tissue-specific genes (*bin 1*) from tissue-specific and broadly expressed genes (*bins 1-17* in the fruit fly and *bins 1-61* in the mouse) based on sequences, predicted secondary structures and genomic features. The prediction performances were measured as median AUC averaged over 100 runs. *AA%* and *SS%* indicate that the input vector for the SVM contained amino acid percentages and the secondary structure symbol percentages for each gene. *AA* is based on Smith-Waterman similarity scores of the protein sequences, *SS* is based on Smith-Waterman similarity scores of the secondary structures and *AA+SS* is a combination of the latter two. GF indicates that *genomic features* where used to assemble the input vector. **a)** results for the fruit fly **b)** results for the mouse.

native power than did the classifiers based on the amino acid content (blue and black dots in Figure 5.3). We observed an increase of 3.2% and 4.1% in the mean AUCs for the fruit fly and mouse, respectively. A total of 15 out of the 17 benchmarks showed higher AUCs using this method in the fruit fly (Table 5.4) while we observed an improvement in 46 out of 60 benchmarks in the mouse (Table 5.6). However, the additional benefits of these became negligible when compared to the computational expense added by calculating all-versus-all Smith-Waterman similarity scores. For this reason, in many of the classifiers below, we used only the amino acid counts.

Because the distribution of genes to groups using Equation 5.3 results in an imbalance in the number of genes attributed to classes, we looked for a different way to build the test and training sets. To do this, we used the ordered list of genes and took the most tissue-specific and most broadly expressed genes, starting with at least 30 in the positive and negative training sets, up to 5,000 transcripts (i.e., 10,000 in total if considering both groups). We ran classifiers for different numbers of genes, and determined the best classifier using amino acid counts.

Figures 5.4a and c summarise the results of these experiments for both organisms. It shows that the prediction performance decreases with the number of genes added to the training sets. This was not surprising since intuitively it should be more difficult to discriminate genes that have more similar gene specificity values. The best discriminative power was found if the 60 most specific genes were compared to the 60 most broadly expressed genes in the fruit fly (AUC = 0.833), while 30 was the best number of genes in the mouse (AUC=1). Figure B.4 lists the gene names and descriptions for the 30 most tissue-specific and 30 most broadly expressed genes responsible for the AUC=1 result in the mouse. There is an increased occurrence of 'olfactory receptor genes' in the list of broadly expressed genes (5 genes), while several instances of the 'kallikrein 1-related peptidase' (5 genes), 'crystallin' (2 genes) and 'carcinoembryonic antigen-related cell adhesion molecule' (2 genes) might bias the list of tissue-specific genes. Figure B.5 presents the amino acid contents of these genes. These were used as input for the classifier that achieved the AUC=1 result. It is not immediately obvious which amino acids separated clearly between the classes, even though some tendencies are apparent. For example, the amino acid asparagine (N) appears to be more frequent in the list of tissue-specific genes. A more detailed analysis of amino acid content in the different classes is presented below. Figures B.6 and B.7 present the all-against-all sequence similarity scores calculated for these 60 genes.

We were also interested in the prediction performance of a model that is trained on genes from one of the two organisms and tested on the other. Figure 5.5a shows the results of a model that was built using mouse genes, and tested on fly genes while Figure 5.5b shows the results after swapping the organisms. The classifier that was trained on mouse genes and tested on fly genes performed best when 65 genes were used to train and test the model (AUC of 0.711). In contrast, the classifier that was trained on fly genes and tested on mouse genes performed best when 15 genes were used to train and test the model (AUC of 0.751). The models trained on mouse genes overall performed better (median AUC = 0.596) than the models trained on fly genes (median AUC = 0.537) when averaged over all benchmark sets tested (99 benchmark sets in which the number of genes ranged from 15 to 500).

5.2.1.2 Secondary structure-based classifier

The classifier based on secondary structure symbols was generally poor. The average AUCs were 0.515 and 0.604 for the fruit fly and mouse, respectively. Again, the full results are given in tables 5.4, 5.6 and Figure 5.3 (green dots). The use of secondary structure similarity scores as input vectors for the SVMs resulted in an increase in AUCs compared to those obtained using the secondary structure symbols alone (see red dots in Figure 5.3). The average AUCs were 0.674 and 0.751 for the fruit fly and mouse, respectively. Compared to the classifiers using amino acid similarity scores, no performance improvement could by gained for any of the benchmarks in either

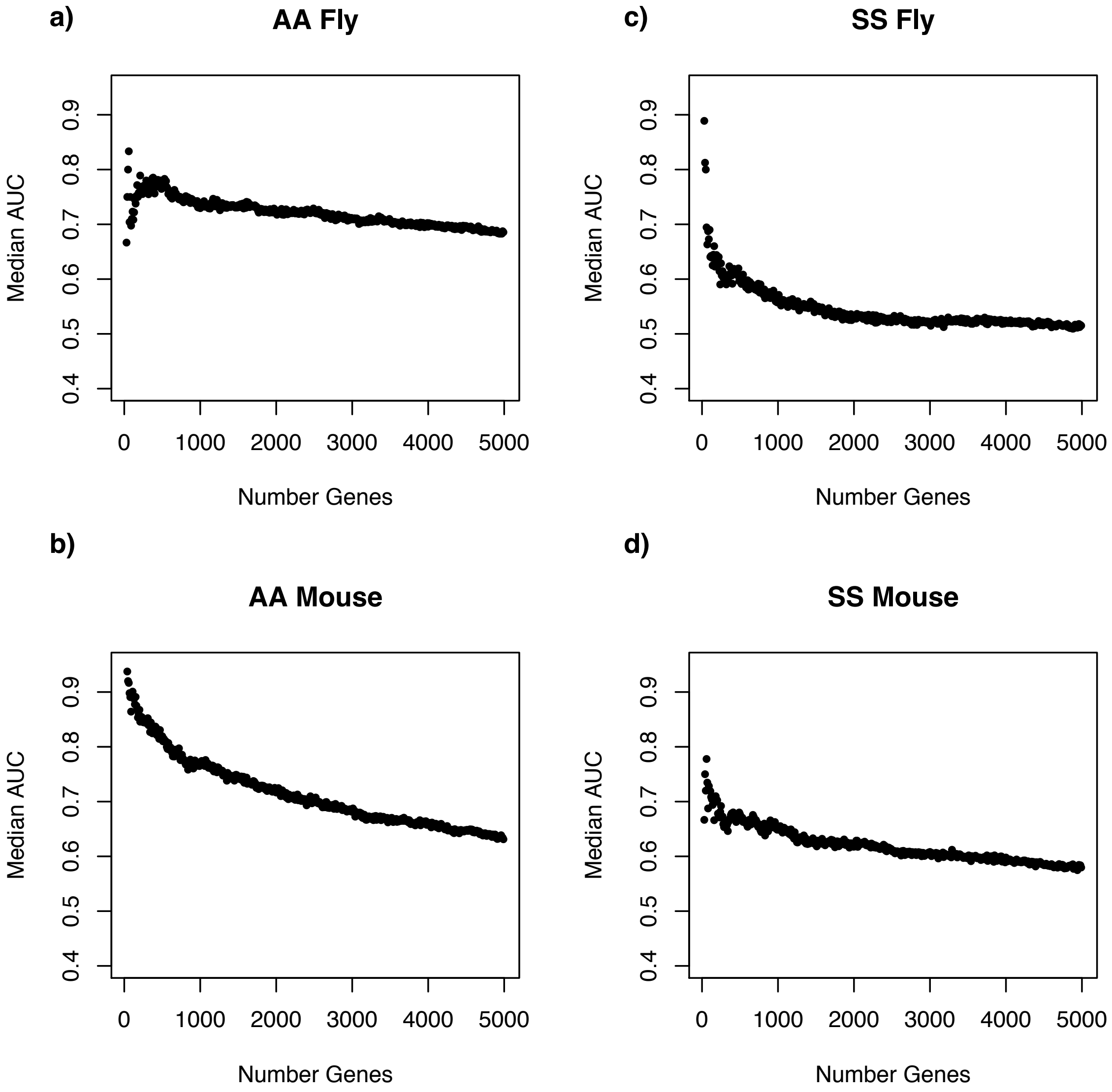


Figure 5.4: **Relationship between the number of genes in the training set and the prediction performance**. The n-top and n-bottom genes from the ordered list of genes, according to their gene specificities, were used as positive and negative training sets. The x-axes indicate how many genes have been used in the positive and the negative training set. The prediction performance is plotted in the y-axis and is measured as median AUC averaged over 100 runs. The classifiers were based on **a+b)** Amino acid percentage. **c+d)** secondary structure symbol percentage.

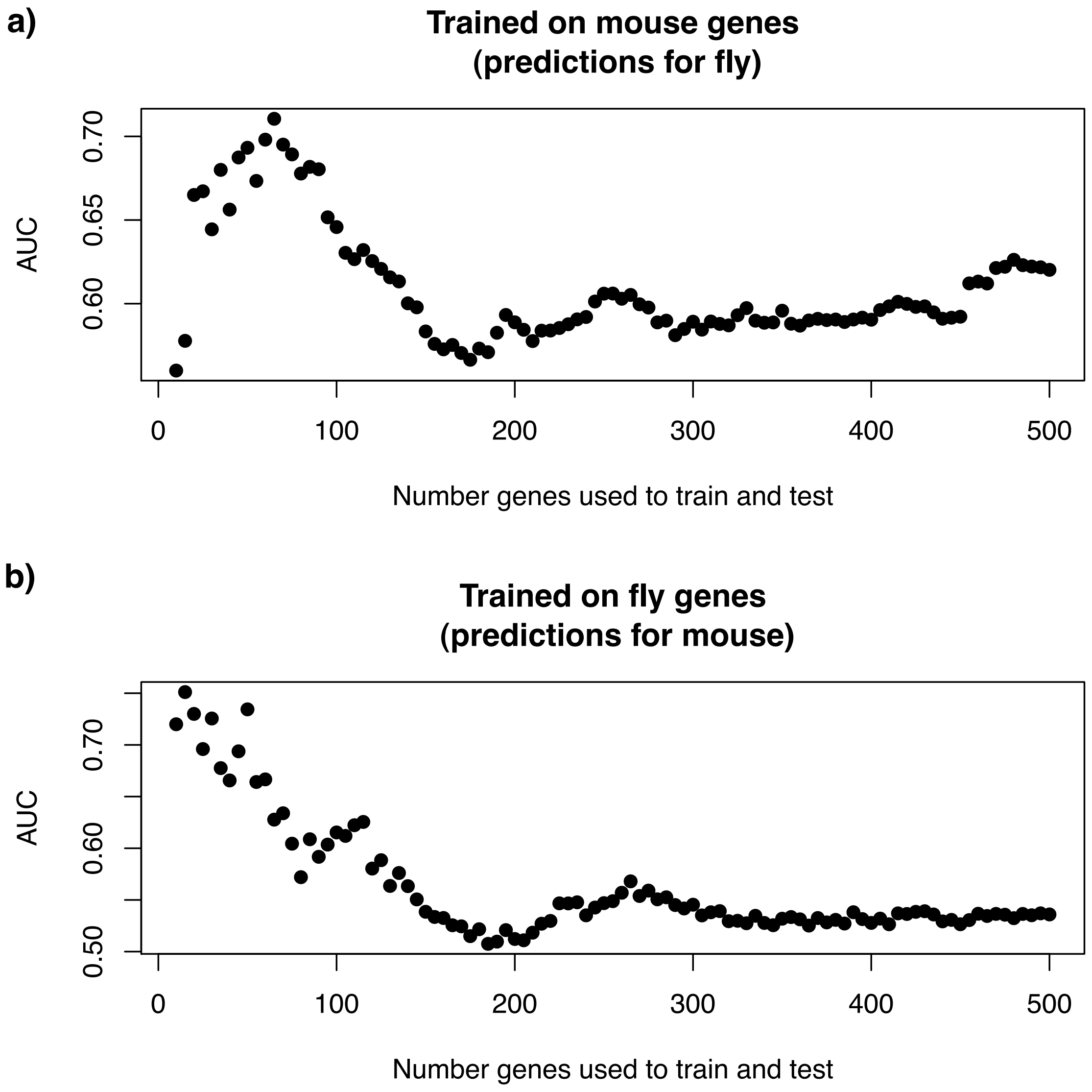


Figure 5.5: **Relationship between the number of genes in the training set and the prediction performance**. The n-top and n-bottom genes from the ordered list of genes, according to their gene specificities, were used as positive and negative training sets. The x-axes indicate how many genes have been used in the positive and the negative training set. The prediction performance is plotted in the y-axis and is measured as AUC. The classifiers were trained on **a)** mouse genes (tested on fly genes) **b)** fly genes (tested on mouse genes).

organism.

We investigated how the prediction performance changes with the number of genes attributed to the positive and negative training sets taken from the bottom and top of the ordered list of gene specificities. Figures 5.4b and d summarise the results of these experiments for both organisms. Prediction performance decreases with the number of genes added to the training sets. For the secondary structure similarity scores, the best number of genes was 30 for the fruit fly (AUC = 0.889) and 60 for the mouse (AUC = 0.778).

5.2.1.3 Combined sequence and secondary structure-based classifier

Combining sequence and structure similarity scores did not increase the performance of the sequence based classifier (yellow dots in Figure 5.3). In the fruit fly, none of the benchmark showed any performance improvement, except if genes belonging to *bin 1* were tested against themselves, which was likely to represent a random effect. In the mouse, only 7 out of the 60 benchmark could be improved (*bin 1* against *bins 2, 6, 9, 10, 18, 22* and *24*).

5.2.1.4 Classifier based on genomic features

Next, we investigated the influence on classifier performance of other features that had previously been found to discriminate between tissue-specific and housekeeping genes in human, mouse and, to some extent, fruit fly samples. The prediction performances for the *bin* experiments ranged from 0.579-0.849 in the fruit fly and 0.477-0.927 in the mouse. The performance of this classifier was roughly in the range of the classifier that used amino acid percentages (brown dots in Figure 5.3).

5.2.1.5 Combining sequence-based classifier with genomic features

We also investigated if combining amino acid percentages and the genomic features results in an improvement in the classifier performance. The prediction performances for the *bin* experiments ranged from 0.473-0.925 in the fruit fly and 0.569-0.968 in the mouse. In the fruit fly, this was the best classifier tested for all benchmarks, while in the mouse, it was the best classifier for 41 groups of the 61 groups tested (brown dots in Figure 5.3). This amounts to a 13.86% and 6.32% increase in average AUCs for the fruit fly and mouse, respectively, if compared to the classifier based on amino acid percentages only, and to a 5.69% and 9.25% increase compared to the classifiers based on genomic features only.

5.2.2 Relation to other work

De Ferrari and Aitken (2006) trained a Naive Bayes classifier to discriminate between housekeeping and tissue-specific genes in human, mouse and fruit fly data. We downloaded the supplementary material for the fruit fly and mouse experiments from this study to compare the data set with ours and test our SVM method on these data. For the fruit fly, data for 20,016 transcripts were downloaded. Of these, 80 were labelled housekeeping while 412 were labelled tissue-specific. A Naive Bayes classifier was trained by the authors of the study to classify transcripts in the benchmark sets and also the remaining transcripts. The resulting probabilities, for a transcript to be housekeeping, were included in the supplementary data files. First, we examined the gene specificity of the training examples. The median gene specificity values for the 80 housekeeping and 412 tissue-specific transcripts, using the S_i values calculated from the FlyAtlas data, were 0.082 and 0.400, respectively (Figure 5.6a, first two boxes). While the housekeeping genes have low gene specificity values indicating broad expression, the genes labelled tissue-specific in this study, show a mid-range pattern of expression rather than a clear tissue-specific expression. The tissue-specific genes were originally identified by mapping homologues to human tissue-specific genes that in turn have been identified by various sources. Our data shows, that these appear not to be truly tissue-specifically expressed in the fruit fly. For instance, the gene FBgn0003071 (Phosphofructokinase; mapped to the transcripts CG4001-RA, CG4001-RB, CG4001-RC) that is involved in glycolysis, is expressed in all tissues examined by FlyAtlas, yet the transcripts were used as tissue-specific examples in the work above. It should be noted that our gene specificity values are based on genes, while the authors worked with gene transcripts. However, the gene has three annotated transcripts and it is unlikely that all three of them are tissue-specific transcripts.

Second, we investigated the gene specificity values for transcripts that were predicted to be housekeeping or tissue-specific (Figure 5.6a, boxes 2-10). Several thresholds were used to identify the class for a transcript (ranging from 50% to 95% probability). Housekeeping genes could be reliably identified using a probability threshold $> 90\%$. However, below that threshold many genes with restricted expression were predicted to be housekeeping. Tissue-specific genes could not be identified reliably at all thresholds used according to the gene specificity values S_i calculated from FlyAtlas. However, it should be noted that the downloaded supplementary data was partly inconsistent with the published paper. For instance, 3,410 transcripts were predicted to be housekeeping using a probability threshold >50% in the main paper, while in the supplementary data only 1,081 transcript were labelled housekeeping when using the same threshold. It was unclear which data was the correct one.

Similar results were observed for the mouse (Figure 5.6b). The housekeeping transcripts had low S_i values, while the tissue-specific genes showed midrange patterns of expression. Many outliers were present when investigating the gene specificities of the predicted class labels. Again, there was an inconsistency with the downloaded data. The lines in the downloaded data set for the mouse did not add up to the numbers presented in the paper, and is incomplete. Therefore only the fruit fly data set was further investigated in the following for which all the training examples were available for download.

We mapped the 80 housekeeping and 412 tissue-specific transcripts from the fruit fly data set to the sequence data and genomic features defined in previous paragraphs. Out of the 412 tissue-specific transcripts only 392 transcripts could be mapped to these data, hence a few training data examples were lost during the mapping. Next, we trained several SVMs to discriminate between the two classes of genes. Figure 5.7 shows the prediction performance of several SVM classifiers trained on three different sets of features: (1) on the amino acid counts in the protein sequences as explained in chapter 5.1.5.1 (2) on the amino acid counts in the protein sequences and various genomic features as explained in chapter 5.1.5.4 (3) on the features used by De Ferrari and Aitken (2006) (reviewed in chapter 3.2), excluding the percentage of GO terms. As explained before, we do not think that functional characteristics should be used to train the classifier and therefore were re-trained the classifier without that information and used it for comparison reasons. The figure also indicates the performance of the Naive Bayes classifier on the same data-set as reported by the authors using the full set of features (4). We compared the TPR at a constant FPR of 20% of all four classifiers. The TPRs were 85%, 93%, 62% and 77% for classifier 1-4, respectively. The SVM classifier based on amino acid percentages and genomic features performed best.

5.2.3 Additional information on features that discriminate between the classes

5.2.3.1 Amino acids with best discriminative power

Some amino acids are not independent and do not provide any additional advantage when evaluated together. A forward feature selection (Miller, 1990) was used to select the amino acid combinations which gave most discriminative power between the test set and the training set for the 250 most tissue-specific and 250 most ubiquitously expressed genes. The number 250 was chosen because we felt it was a good compromise between the number of genes in the training and test sets and the computational expense, as well as the degree of tissue specificity, which

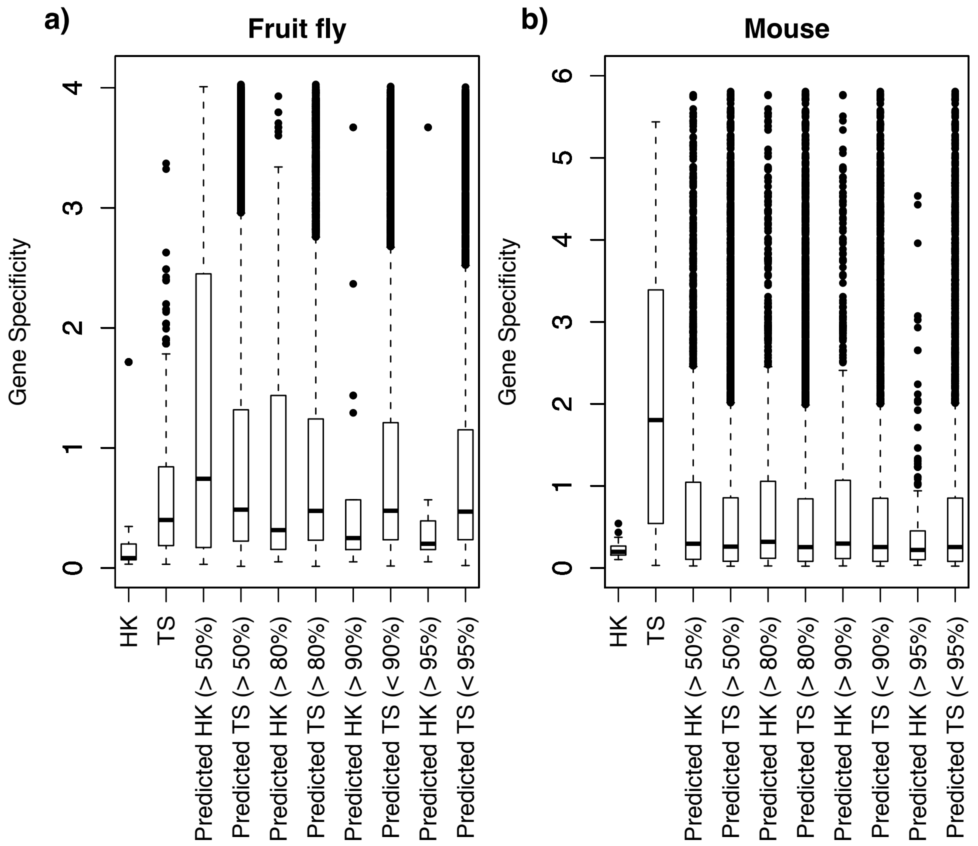


Figure 5.6: **Gene specificity of housekeeping (HK) and tissue-specific (TS) genes (transcripts) defined by De Ferrari and Aitken (2006)**. The first two boxes in each figure correspond to the genes (transcripts) in the training-set. All other boxes show the gene specificities of transcripts predicted to be HK or TS by the Naive Bayes classifier at various probability thresholds.

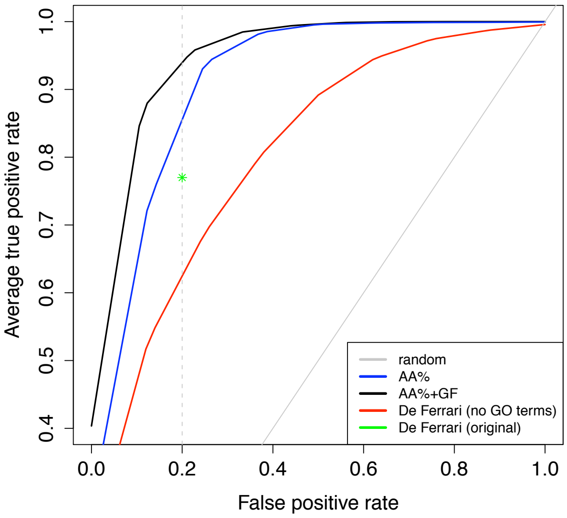


Figure 5.7: **Prediction performance compared with published work**. Our classifier based on amino acid counts and/or genomic features was applied on the fruit fly data used by De Ferrari and Aitken (2006). In this, 80 housekeeping transcripts were discriminated from 412 tissue-specific transcripts. The green star indicates the performance of the Naive Bayes classifier on the data set. The other curves represent the performances of SVMs trained in this work. The SVMs were trained on (1) amino acid percentages in the sequences (AA%), (2) amino acid percentages combined with genomic features (AA%+GF) and (3) the genomic features used by DeFerrari and Aitken, excluding the percentage of GO terms.

becomes less clean as more genes are added to the training set. Forward selection was started from the single amino acid that discriminated best between the classes according to the Fisher's ratio test (Equation 2.8). This test is based on the ratio of between-class variance to within-class variance. It evaluates how well a single amino acid is correlated with the separation between classes. We then built all of the two-dimensional feature subsets that include the amino acid already selected from the first step and finds the best one. This process was continued, building n-dimensional feature subsets until the subset reached a size of 20. We used AUCs for the selection criteria in this work. This procedure of attribute selection has been termed a greedy approach. A disadvantage of using this selection approach is that if two features have similar discriminative power, only one of them will be selected and appear important in the results.

Using this method, the amino acid N (asparagine) discriminated best between the two groups in both organisms in a 1-dimensional classifier, with asparagine being more frequent in tissue-specific genes than in broadly expressed genes (Figure 5.8). The medians for the broadly expressed genes were 4.12 and 2.88 and for the tissue-specific genes were 5.0 and 4.22 for the fruit fly and mouse, respectively. The robustness was investigated by evaluating which amino acids discriminated between the groups for n number of tissue-specific and broadly expressed genes, where n ranged from 30 to 1,000. In the fruit fly, asparagine was the best discriminator in 518 cases, followed by cysteine in 350 cases, leucine and proline in 51 cases and glutamine in one case. In the mouse, asparagine was the best discriminator in 549 cases, and leucine in 422 cases (5.9).

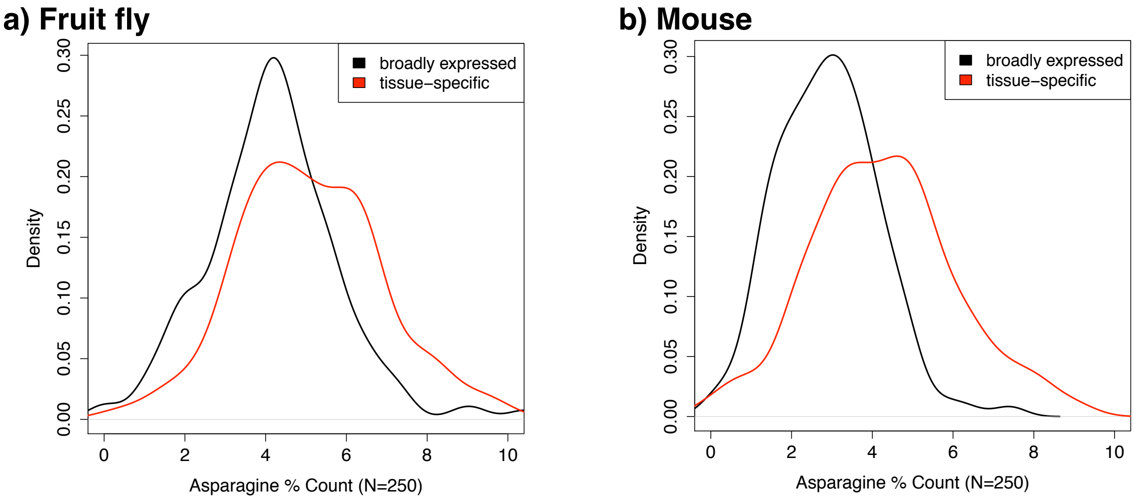


Figure 5.8: **Density distribution of asparagine** in the 250 most tissue-specific and the 250 most broadly expressed protein transcripts for the a) fruit fly and b) mouse.

Next, 19 two-dimensional classifiers were trained for finding the best combination of two amino acids, i.e., between asparagine and the other 19 amino acids. The pair of amino acids that achieved the highest AUC was recorded. This process was repeated 100 times to prevent situations where a well-scoring feature set might be found by chance. For the fly, the amino acid pair N+A (asparagine and alanine) performed best, while for the mouse the amino acid pair N+E (asparagine and glutamic acid) achieved the highest AUC (Figure 5.10). Similarly, the best combinations of 3-20 amino acids were determined. Table 5.1 lists the best combinations of features. In

Table 5.1: AUCs calculated for n-dimensional classifiers. The 250 genes with the lowest gene specificity value (broadly expressed) were compared with the 250 genes with the highest gene specificity value (tissue-specific expression). The column in grey indicates the amino acid combination that results in best performance. Classifiers were trained 100 times.

Amino Acid Combination	**AUCs**
Fly (Drosophila melanogaster)	
N	0.640
N+A	0.732
N+A+C	0.705
N+A+C+P	0.730
N+A+C+P+S	0.756
N+A+C+P+S+R	0.755
N+A+C+P+S+R+Q	0.754
N+A+C+P+S+R+Q+I	0.769
N+A+C+P+S+R+Q+I+F	0.776
N+A+C+P+S+R+Q+I+F+Y	0.773
N+A+C+P+S+R+Q+I+F+Y+K	0.771
N+A+C+P+S+R+Q+I+F+Y+K+V	0.770
N+A+C+P+S+R+Q+I+F+Y+K+V+M	0.768
N+A+C+P+S+R+Q+I+F+Y+K+V+M+W	0.769
N+A+C+P+S+R+Q+I+F+Y+K+V+M+W+D	0.773
N+A+C+P+S+R+Q+I+F+Y+K+V+M+W+D+A	0.763
N+A+C+P+S+R+Q+I+F+Y+K+V+M+W+D+A+L	0.774
N+A+C+P+S+R+Q+I+F+Y+K+V+M+W+D+A+L+T	0.774
N+A+C+P+S+R+Q+I+F+Y+K+V+M+W+D+A+L+T+G	0.772
all 20 amino acids	0.764
Mouse (Mus Musculus)	
N	0.732
N+E	0.781
N+E+K	0.795
N+E+K+A	0.810
N+E+K+A+H	0.823
N+E+K+A+H+C	0.834
N+E+K+A+H+C+W	0.833
N+E+K+A+H+C+W+L	0.827
N+E+K+A+H+C+W+L+V	0.836
N+E+K+A+H+C+W+L+V+Q	0.831
N+E+K+A+H+C+W+L+V+Q +R	0.835
N+E+K+A+H+C+W+L+V+Q +R+G	0.855
N+E+K+A+H+C+W+L+V+Q +R+G+T	0.851
N+E+K+A+H+C+W+L+V+Q +R+G+T+M	0.858
N+E+K+A+H+C+W+L+V+Q +R+G+T+M +P	0.851
N+E+K+A+H+C+W+L+V+Q +R+G+T+M +P+F	0.858
N+E+K+A+H+C+W+L+V+Q +R+G+T+M +P+F+ S	0.849
N+E+K+A+H+C+W+L+V+Q +R+G+T+M +P+F+ S +I	0.848
N+E+K+A+H+C+W+L+V+Q +R+G+T+M +P+F+ S +I +Y	0.843
all 20 amino acids	0.837

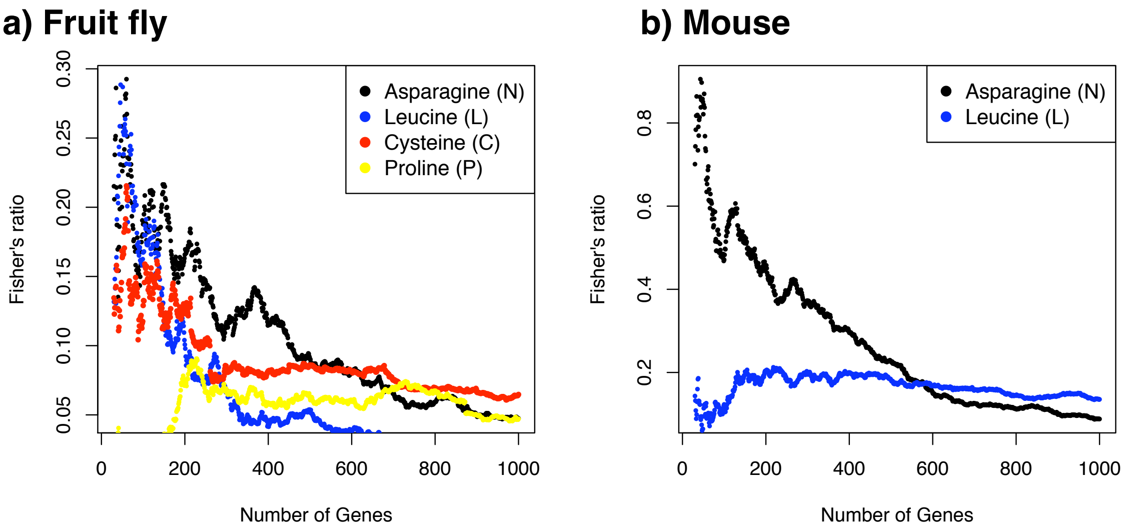


Figure 5.9: **Best discriminating amino acids in 1-dimensional classifiers**. Fisher's ratio versus the number of genes used for positive and negative training-sets. Fisher's ratios are shown for only those amino acids that had the highest discrimination power for at least one of the benchmark sets. a) fruit fly and b) mouse.

the fruit fly, the residue combination N+A+C+P+S+R+Q+I+F led to the best prediction performance (AUC= 0.776), while in the mouse, the residue combination N+E+K+A+H+C+W+L+V+Q +R+G+T+M (+P+F) led to the best prediction performance (AUC= 0.858).

5.2.3.2 Secondary structure elements with the best discriminative power

Forward feature selection was also used to select the secondary structure element combinations which gave most discrimination between the test set and the training set for the 250 most specific and ubiquitously expressed genes. C was the best secondary structure character for a 1-dimensional classifier, with an AUC=0.620 in the fruit fly, while E results in the best performance in the mouse, with an AUC =0.904 (Table 5.2). The overall C content was higher for the tissue-specific genes than for the broadly expressed genes (48.15 and 42.71) in the fruit fly and in the mouse (49.33 and 44.11). To investigate the robustness, we also checked which secondary structure elements best discriminated between the groups for n number of tissue-specific and broadly expressed genes, where n ranged from 30 to 1,000. In the

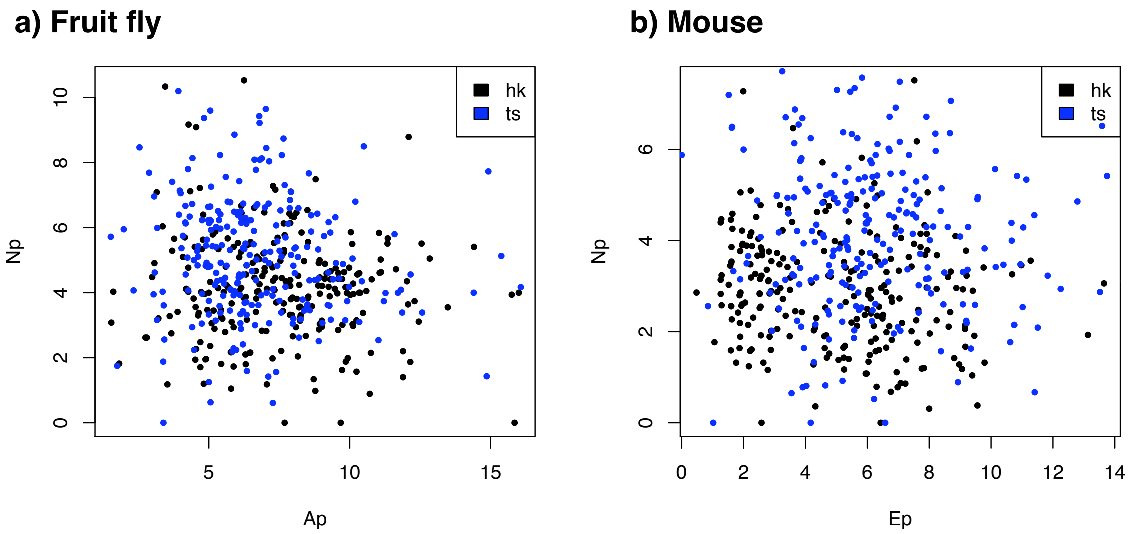


Figure 5.10: **Distribution of asparagine (N) and alanine (A)/glutamic acid (E)** in the 250 most tissue-specific (ts) and the 250 most broadly expressed (hk) protein transcripts for the a) fruit fly and b) mouse.

Table 5.2: AUCs calculated for n-dimensional classifiers. The 250 genes with the lowest gene specificity value (broadly expressed) were compared with the 250 genes with the highest gene specificity value (tissue-specific expression).

Secondary Structure Elements Fly	AUCs Fly		Secondary Structure Elements Mouse	AUCs Mouse
C	0.620		E	0.904
C+E	0.621		E+H	0.683
C+E+H	0.612		E+H+C	0.686

fruit fly, C was the best discriminator in 967 cases, followed by H in 4 cases. In the mouse, E was the best discriminator in 292 cases, and H in 679 cases (5.11).

Next, we analysed some features individually that have been shown to differ between tissue specific and housekeeping genes:

5.2.3.3 Sequence lengths

Housekeeping genes were observed to be shorter in human data (Eisenberg and Levanon, 2003). We did not find any correlation between sequence length and tissue specificity in the fruit fly and mouse data used in this work. Figure 5.12 shows the protein sequence lengths for each gene in the two data sets versus tissue specificity. There is no immediate obvious correlation between protein sequence length and gene

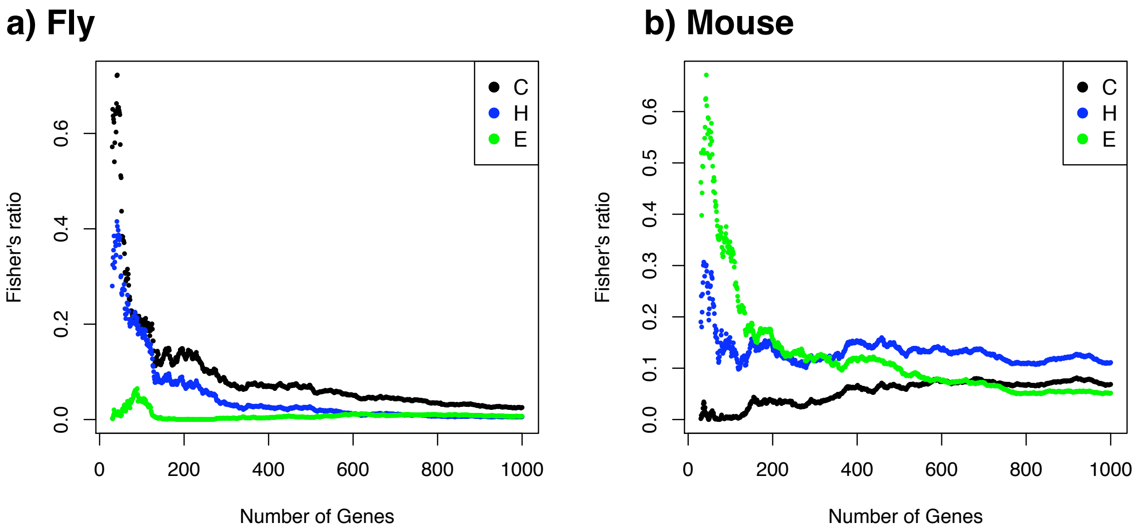


Figure 5.11: **Best discriminating secondary structure elements in 1-dimensional classifiers**. Fisher's ratio versus the number of genes used for positive and negative training-sets. Fisher's ratios are shown for the three secondary structure elements predicted by PSIPRED. a) fruit fly and b) mouse.

specificity. The correlation coefficients r were -0.108 and -0.065 for the fruit fly and mouse, respectively.

We also tested if a linear correlation existed between tissue specificity and cDNA, 3'UTR length, 5'UTR length and CDS length. Again, no clear linear relationship was found. For the mouse, the highest correlation was found for the 3'UTRs that had a correlation coefficient of -0.137. The other correlation coefficients were -0.07, -0.107 and -0.078 for CDS, cDNA and 5' UTR lengths respectively. Similarly, the highest correlation was found for the 3'UTR length for the fly genes (r=-0.126). The other correlation coefficients were -0.111, -0.150 and -0.116 for CDS, cDNA and 5' UTR lengths, respectively.

5.2.3.4 CpG islands and predicting the methylation status of transcription units

A correlation between the occurrence of CpG islands, and thus DNA methylation, in housekeeping and tissue-specific genes has been investigated in a comprehensive mouse data set (Schug *et al.*, 2005). Limited DNA methylation was observed in the

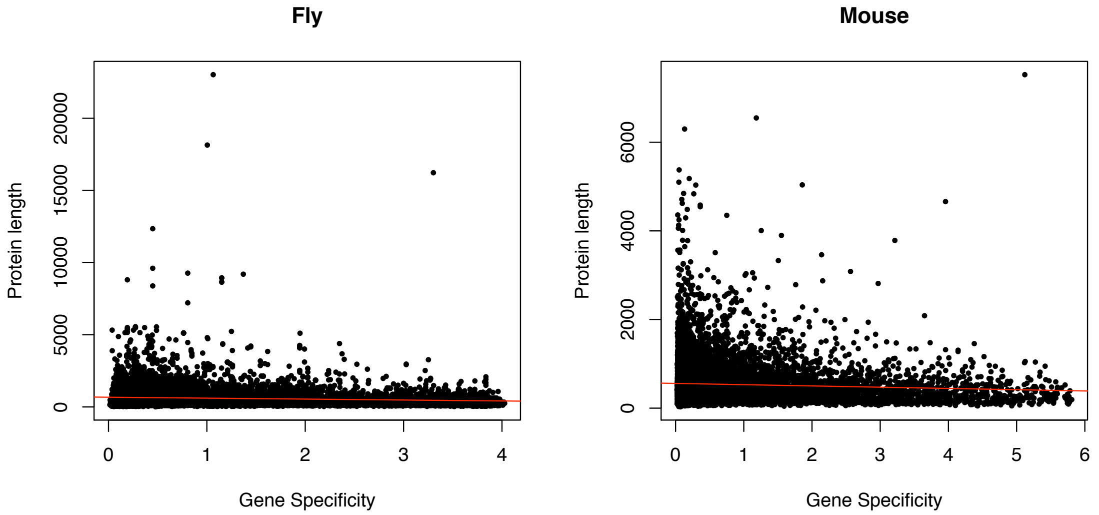


Figure 5.12: **Protein sequence lengths versus tissue specificity**. There was little evidence for any relationship between sequence length and tissue specificity in the fruit fly (15,560 sequences) and mouse data set (11,805 sequences). The correlation coefficients were -0.108 for the fruit fly and -0.065 for the mouse.

fruit fly (Lyko *et al.*, 2000; Wang *et al.*, 2006), but to the best of our knowledge there was no study that investigated a possible correlation between the expression breadth and DNA methylation in the fruit fly. We investigated the FlyAtlas data set for this purpose. We also investigated the mouse data set for such a correlation to confirm or negate previous observations.

We plotted the fraction of promoters with CpG islands against the gene specificity value for both organisms (Figure 5.13). In the fruit fly, the CpG islands appeared to be randomly distributed for genes with various tissue specificity (r = 0.212), while in the mouse there was a clear negative correlation between the fraction of promoters with CpG islands and tissue specificity (r = 0.922). This negative correlation has been shown previously in human and mouse genes (Schug *et al.*, 2005).

Next, we investigated the ratio of observed to expected CpGs (Equation 2.1) in the fruit fly and mouse sequences to predict methylated and unmethylated genomic regions. Figure 5.14 shows the frequency of all annotated protein coding genes in the fruit fly for which tissue specificity information is available, with CpG [o/e]

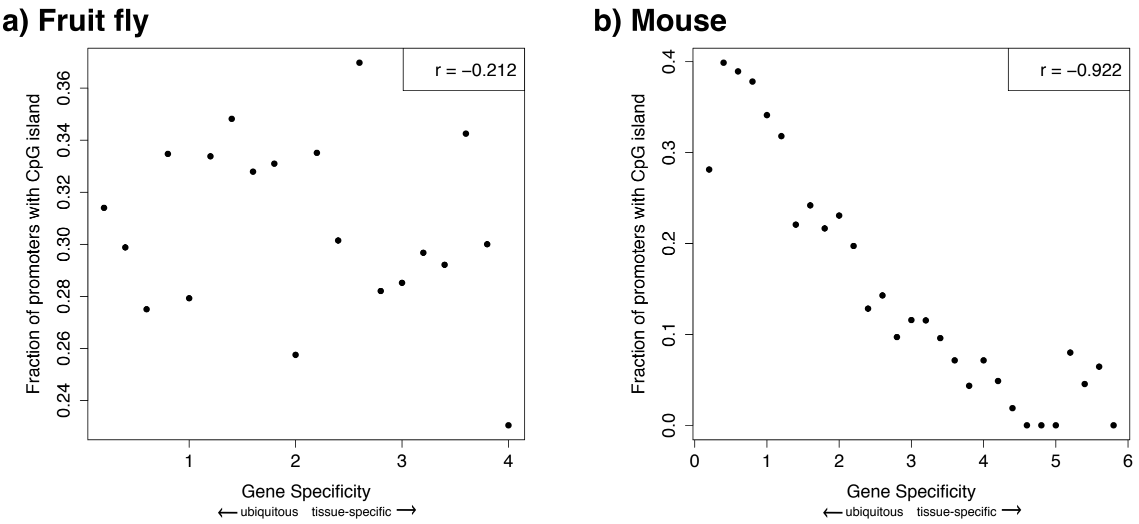


Figure 5.13: **Frequency of CpG islands in fruit fly and mouse promoter regions**. The fraction of promoters with CpG islands, as identified by the *newcpgreport* program, is plotted against the gene specificity value. The variable r in the legends indicates the correlation coefficient between the two variables.

frequencies between 0 and 2. For comparison, the contrasting distribution of all protein coding genes in the mouse is shown in panel b. In the fruit fly, most genes have a value around 1.0. Genes with a value < 0.7 represent CpG deficient genes. These are the genes that are expected to be methylated; these are very few compared to the number in the mouse. Again, this is in agreement with previous observations that there is limited DNA methylation in the fruit fly (Lyko *et al.*, 2000; Wang *et al.*, 2006). For both organisms, the density curves for broadly expressed genes are shifted towards the right when compared with the density curves for the tissue-specific genes. Although this signal was weak in the fruit fly data set, this indicated that the broadly expressed genes are less frequently methylated and agrees with previous observations that housekeeping genes are not methylated (Foret *et al.*, 2009).

To further assess the difference in CpG [o/e] frequencies between ubiquitous and tissue-specific transcripts, we compared these two categories. We observed an even distribution of the CpG-deficient genes (CpG [o/e] < 0.7) across varying gene specificity values. For both organisms, CpG-deficient genes cover all levels of tissue specificity (Figure 5.15). CpG over-represented genes are more commonly broadly

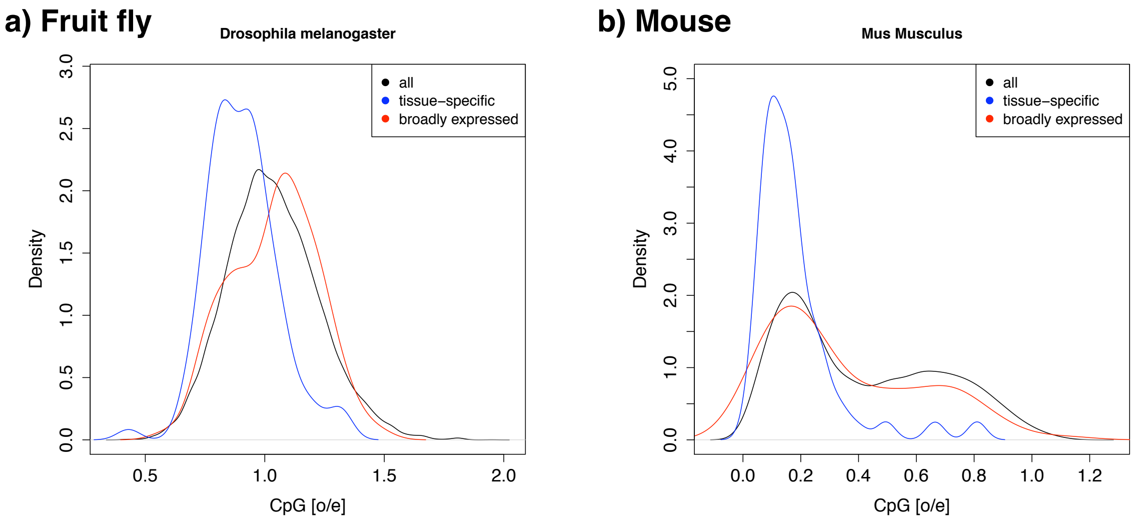


Figure 5.14: **Distribution of CpG bias** in the fruit fly and mouse. The 100 most tissue-specific and 100 most broadly expressed genes are plotted as well as the distribution for all genes.

expressed genes.

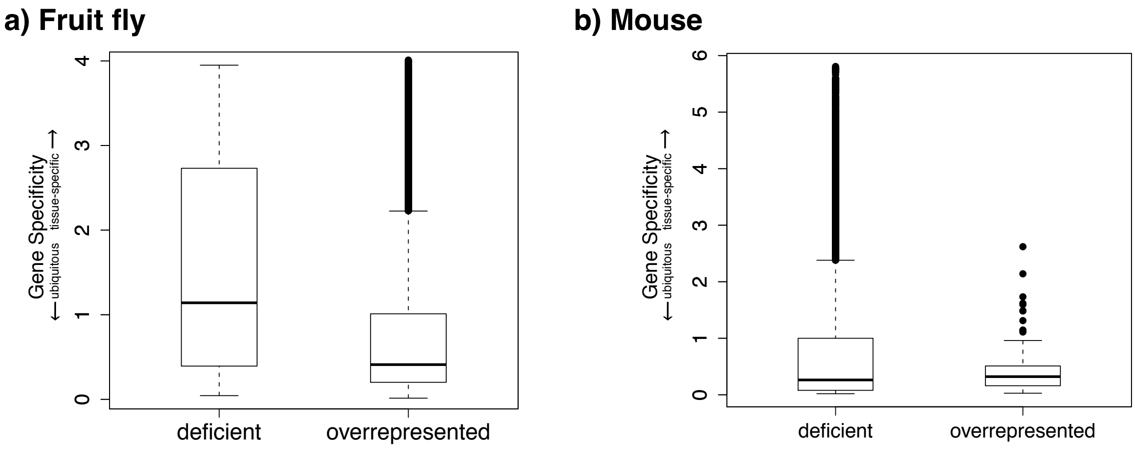


Figure 5.15: **Gene specificities** for CpG deficient (CpG [o/e] $<$ 0.7) genes and CpG overrepresented genes (CpG [o/e] $>$ 1.0).

5.2.3.5 Simple sequence repeats (SSR)

We contrasted the occurrence of SSRs in the 5'UTR region of housekeeping and tissue-specific genes to see if there are any distinguishable differences between the two classes of genes. The percentage of genes with SSR was higher for broadly expressed genes in both organisms (Figure 5.16). This is in agreement with previous observations in human and mouse studies (Lawson and Zhang, 2008) where SSR densities in 5'-UTRs in housekeeping genes have been shown to be higher than in tissue-specific genes. Our results confirmed this and also showed that this trend was repeated in the fruit fly.

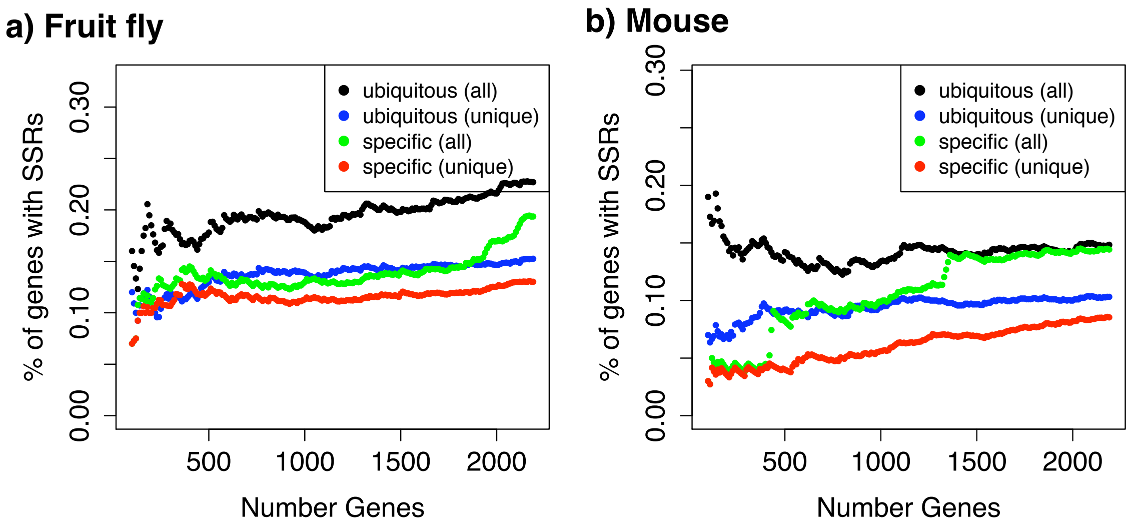


Figure 5.16: **Percentage of tissue-specific and broadly expressed genes with SSRs**. The x-axis indicates the number of tissue specific genes or broadly expressed genes investigated, and the y-axis indicates the percentage of the genes in which SSRs were identified. *Unique* indicates that transcripts with > 1 SSRs were counted once only, while *all* indicates that these transcripts were counted several times.

5.2.3.6 S/MAR

The promoter regions (1,500 upstream and downstream of TSS) of the transcripts were scanned for S/MAR regions using the EMBOSS `marscan` program. We noted the presence or absence of S/MAR regions in a transcript; however, when two regions were identified in a transcript, this information was disregarded. This allowed us to work with binary data. Start and end position of the regions were also discarded in this study.

For the fly, 4,260 S/MAR regions were identified in the upstream regions pertaining to 3,846 unique transcript IDs. These were mapped back to 2,644 FlyBase gene IDs. A total of 3,311 S/MAR regions were identified in the downstream regions pertaining to 2,962 unique transcript IDs. These were mapped back to 2,073 FlyBase gene IDs. For the mouse, 1,119 S/MAR regions were identified in the upstream regions pertaining to 1,058 unique Ensembl transcript IDs. These were mapped back to 1,053 EntrezGene gene IDs. In downstream regions 1,319 S/MAR regions were identified, pertaining to 1,250 unique transcript IDs. These were mapped back to 1,211 EntrezGene gene IDs.

No differences were immediately obvious between tissue-specific and broadly expressed genes. The tissue specificity of the genes with S/MAR region was similar to all genes (Figure 5.17). In human, it was shown that S/MARs are more abundant in the 5' regions of tissue specific genes as compared to the housekeeping genes (Ganapathi *et al.*, 2005) but we did not confirmed this here.

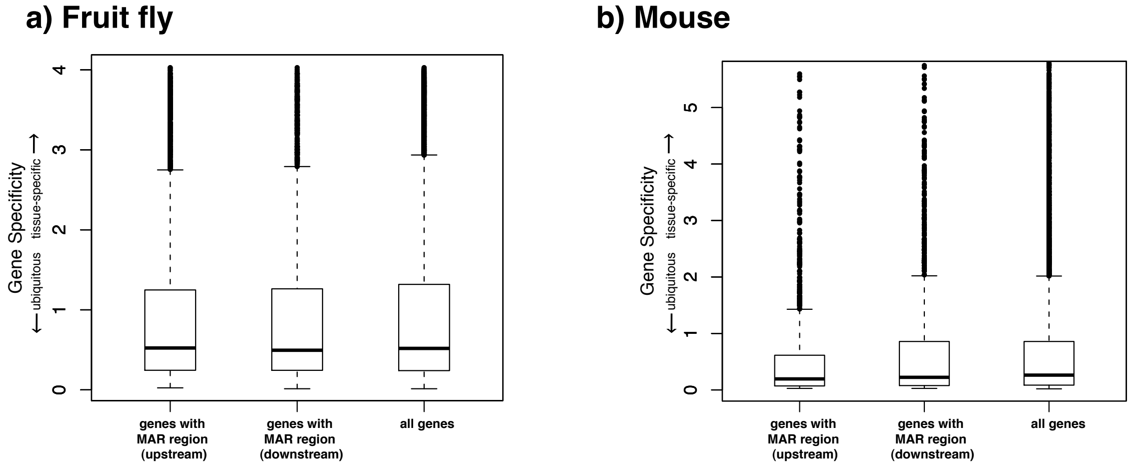


Figure 5.17: **Tissue specificity of genes with S/MAR regions in the a) fruit fly and b) mouse**. Boxplots are shown for all genes in which a S/MAR region was identified by the marscan program in the upstream regions, for all genes in which a S/MAR regions was identified by the marscan program in the downstream regions, and for all genes.

5.2.3.7 Ribosomal occupancy

Lastly, we investigated whether ribosomal occupancy is also correlated with tissue specificity in the fruit fly.

We looked at two data sets to determine any possible association between ribosomal occupancy and tissue specificity. The first data set was taken from a study on mRNA translational control during early *Drosophila* embryogenesis (Qin *et al.*, 2007) and concerns embryo data. In this study, sucrose polysomal gradient analyses and GeneChip analysis were used to investigate post-transcriptional regulation during *Drosophila* early embryonic development during the first 10 hours after egg-laying. The conclusion of the study was that genes involved in some biological processes are co-regulated at the translational level at certain developmental stages. Figure 5.18 shows the ribosomal loading for genes that were found to be expressed in the adult fruit fly. The plot shows tissue specificity (adult fruit fly data) versus occupancy (embryo data) at different embryonic stages. The occupancy spread ranged from 20-100% for all levels of tissue specificity, but was centred towards 60% for tissue-specific genes. The figures are somewhat biased because there are more broadly expressed genes than tissue-specific genes. The same trend was seen, but was somewhat clearer, when the genes were restricted to the 100 most tissue-specific and 100 most broadly expressed genes (Figure 5.18, Panel b+c).

The second data set was taken from a study of lifespan extension in adult flies upon a yeast restricted diet (Zid *et al.*, 2009). In this study, sucrose polysomal gradient analyses and GeneChip analysis were used to investigate the effects of dietary restriction on *Drosophila* lifespan. The experiment consisted of 12 arrays; these included 6 arrays under normal conditions and 6 under dietary restriction. Under each condition, 3 arrays with high ribosomal loading and 3 arrays with low ribosomal loading ($<$ 5 ribosomes) were prepared. Originally, these data were used to see translational differences between two conditions. In the present study, we only used the 6 arrays under normal conditions.

After normalisation and background correction, we looked at the differences in mean expression values between high and low ribosomal loading for the genes that showed a significant difference between the high and low fractions under normal conditions (1,011 probe sets, p<0.001, ANOVA analysis). We plotted the 100 most tissue-specific versus the 100 most broadly expressed genes (Figure 5.19). Overall,

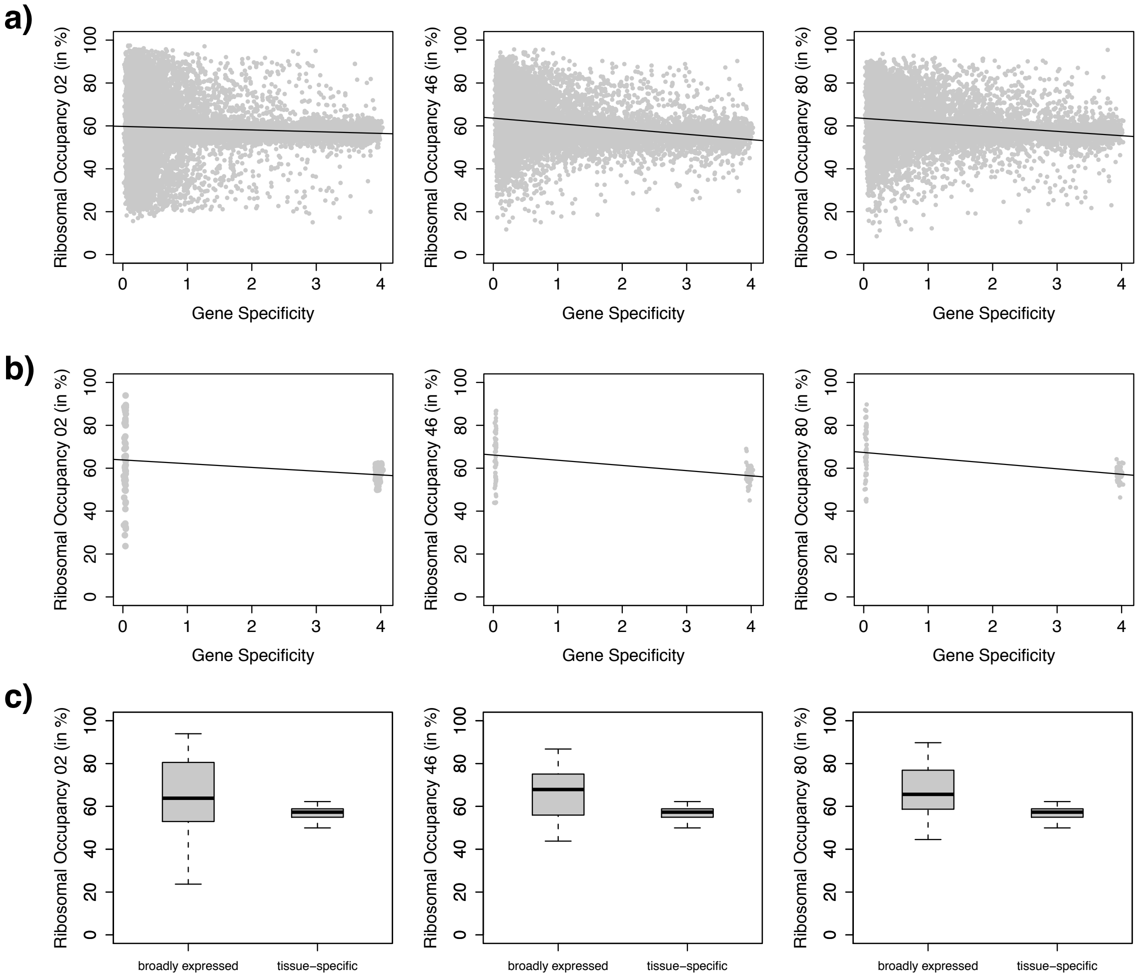


Figure 5.18: **Ribosomal occupancy in the fruit fly using embryo data at different embryonic stages.** Ribosomal occupancy, which is defined as the percentage of polysomal associated mRNAs for individual transcripts, is shown on the y-axes. Three major developmental stages of fruit fly embryogenesis are investigated: 0-2 hours, 4-6 hours and 8-10 hours after egg laying**a)** Ribosomal occupancy is plotted versus gene specificity. **b+c)** Ribosomal occupancy is shown for the 100 most broadly expressed genes and 100 most tissue-specific genes.

the difference in mean expression signals between high and low ribosomal loading was higher for the tissue-specific genes. This indicates that there was a greater variation in translational regulation for tissue-specific genes. This was expected since most housekeeping genes are required at a relatively constant level, while it is assumed that this is less critical for many tissue-specific genes.

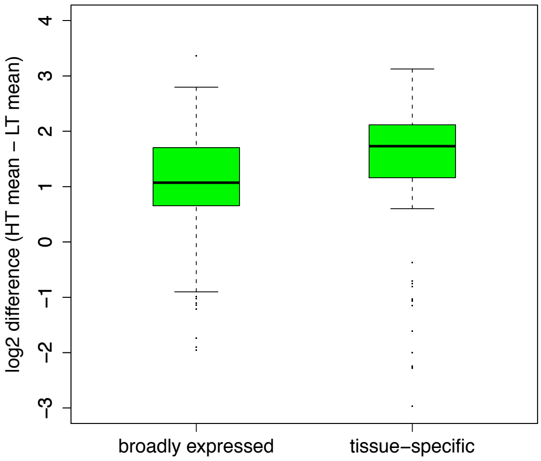


Figure 5.19: **Ribosomal occupancy.** Difference between low ribosomal loading and high ribosomal loading in adult flies for the 60 most broadly expressed and 60 most tissue-specific fruit fly genes.

5.3 Discussion

Several genomic features are known to differ between tissue-specific and housekeeping genes. In this work, we have investigated the tissue specificity of gene expression and its possible relationship with the sequence and secondary structure of the resulting protein product. The approach used here was to adopt an information theoretic approach to sort genes in FlyAtlas, and in the mouse GNF data set, according to their tissue specificity. Based on their gene specificity values, the genes were distributed to several groups that were then used to build positive and negative training sets for SVMs used to predict gene specificity. Our primary interest was in the sequence and predicted secondary structure, but we also considered other genomic features that have been previously implicated in discrimination of these two gene classes. The results obtained were clearly not random, and the classification performance was good for the sequence based classifier. Therefore, these computational studies confirmed the expectation, in both the fruit fly and the mouse, that amino acid sequences are different for genes with various degrees of tissue specificity. In most benchmark sets, the classifiers using Smith-Waterman similarity scores had more discriminative power than did the classifiers based on the amino acid contents. This suggests that there is additional discriminative power in the order of the amino acids. The performance of the classifier was not further improved by integrating secondary structure similarity information indicating that, in the investigated data set, there was not additional discriminative power in the predicted secondary structure.

The fruit fly was chosen as a primary organism because its tissues have been less well studied than human and mouse tissues in terms of their gene specificities (Table 3.1). The recent availability of FlyAtlas made it possible to revisit some of the features that have been associated with tissue specificity and housekeeping properties of human and mouse genes in much more detail in the fruit fly than was previously possible. We also tested our method on mouse data to confirm the results we saw in the fruit fly data and to investigate the range of applicability of the method based on sequence and/or secondary structures. Due to difficulties with mapping the GNF probe set IDs to protein sequences for the human data set, we did not further investigate the method on human genes. The human data set contained 44,760 probe sets, but annotation was available for only 22,558 of these probe sets. Only 8,282 probe sets (5,850 unique ones) out of the 22,558 are mapped to an EntrezGene ID in the annotation file provided by GNF. Only 25% of the remaining probe sets could be mapped to a protein sequence in UniProtKB, the majority of which corresponded to ubiquitously expressed genes. Hence, there were no good training data available for tissue-specific genes in the human data set. The mouse GNF probe set data used in this work was also mapped to protein sequences via their EntrezGene IDs. A more complete set might be obtained by mapping the probe sets to FlyBase (FBpp) protein IDs and then to extract the corresponding protein sequences from Ensembl.

One of the problems encountered in this study was the question of how to divide the genes into groups of certain tissue specificity. Depending on how one defines tissue specificity, genes may be divided differently. For instance, a gene may be considered tissue-specific if it is expressed in exactly one tissue only. However, for other cases, one might consider genes to be tissue-specific if they are expressed in, for example, 10% of the tissues. We had to decide how to distribute the genes to bins and the method chosen resulted in an imbalance in training sets for positive and negative training examples. To circumvent the problem of imbalance for the positive and negative training examples, we performed further experiments that included the same number of positive and negative examples. However, the question of how to best assemble the training sets remains. Another problem in this work was that microarray data are noisy and that the low resolution of microarrays can be a problem. Genes with low expression may in fact be broadly expressed, but they may be below the detection threshold limits of the microarray technologies used in this work.

Another difficulty was the presence of several transcripts for one gene, which occurred in especially in the fruit fly data set. Even though transcripts with the same amino acid sequences were removed from the data set, a bias might be introduced by those transcripts that are highly similar. One way of circumventing this problem is to remove genes with several transcripts altogether. However, this would reduce the fruit fly data set by almost 30%. Another possibility is to take the mean of the amino acid percentages and similarity scores for two transcripts. The problem with this strategy is that some of the transcripts were highly diverse, and thus it would be better to include them both, but separately. Similarly, paralogous genes that may have high sequence or structural similarity might introduce another bias. Additional study is

Table 5.3: **Paralogous genes**. A total number of 18 genes were extracted from the FlyMine (Lyne *et al.*, 2007) database, that were found to be paralogous genes to the gene FBgn0000024. The table lists the genes alongside their gene specificity values. In the gene name column, NA indicates that no gene name was available for that gene.

FlyBase gene	**gene symbol**	**gene name**	S_i
FBgn0000326	clt	cricklet	0.17
FBgn0001114	Glt	Glutactin	1.19
FBgn0001987	Gli	Gliotactin	0.48
FBgn0015568	alpha-Est1	alpha-Esterase-1	0.69
FBgn0015569	alpha-Est10	alpha-Esterase-10	0.70
FBgn0015570	alpha-Est2	alpha-Esterase-2	1.49
FBgn0015571	alpha-Est3	alpha-Esterase-3	0.64
FBgn0015572	alpha-Est4	alpha-Esterase-4	0.51
FBgn0015574	alpha-Est6	alpha-Esterase-6	2.14
FBgn0015575	alpha-Est7	alpha-Esterase-7	0.75
FBgn0015576	alpha-Est8	alpha-Esterase-8	0.45
FBgn0015577	alpha-Est9	alpha-Esterase-9	0.49
FBgn0027584	CG4757	NA	1.35
FBgn0032131	CG3841	NA	3.17
FBgn0033943	CG12869	NA	0.24
FBgn0034736	CG6018	NA	1.03
FBgn0037090	CG7529	NA	0.70
FBgn0039084	CG10175	NA	1.32

required to determine if paralogous genes typically display the same tissue specificity. If so, they should be removed from the data set. As a preliminary test we investigated the paralogous genes of the broadly expressed gene 'FBgn0000024 (Acetylcholine esterase)' regarding their ranges of tissue specificity. The gene specificity values for the 18 paralogous genes range from 0.17 to 3.17 indicating varying degrees of tissue specificity for these genes (Table 5.3).

The amino acid residue, asparagine, was found to discriminate best between the 250 most tissue-specific genes and broadly expressed genes in both the fruit fly and the mouse data, in a 1-dimensional classifier. Asparagine was, on average, more frequently found in tissue-specific genes than in broadly expressed genes. Deamidation of asparagine residues is one of the most common post-translational modifications and results in protein degradation. During the process of deamidation, the asparagine residue is converted to aspartate. The biological properties of the mutated proteins differ from those of the original material, due to this conversion. One hypothesis is that broadly expressed genes must be more stable because they often have housekeeping function and therefore must be expressed at a relatively constant level across many or all known conditions. Deamidation of glutamine residues can also occur but does so at a much lower rate. Glutamine did not appear to have additional discriminative power in the feature selection process. However, as mentioned earlier, a disadvantage of greedy forward selection is that if two attributes are equally good discriminators then only one of them will appear on top of the list.

The feature selection process showed that not all amino acids are different between the two classes of transcripts. For the fly a combination of 9 amino acids led to the best performance while 14 amino acids were required for the mouse to get best performance. All amino acids additionally added resulted in a decrease of performance, indicating that they have no discriminative power. Adding more features can increase the noise, and hence the decrease of performance may be observed. There is a certain risk of overfitting when selecting the best discriminating features using a greedy forward selection approach, and the results are likely to be less reliable for higher dimesions.

In general, the predicted performance, based on secondary structure elements was poor. The secondary structure element C (Others) was found to discriminate best between the 250 most tissue-specific genes and broadly expressed genes in the fruit fly, in a 1-dimensional classifier. In the mouse, the element E (extended strand) was the better discriminator for these two groups. However, care should be taken when drawing conclusions from this, since many of the benchmark sets showed classifier performance that was not much better than random, when based only on the secondary structure elements.

We applied and compared the performance of the SVM classifier based on amino acid percentages in the sequence combined with various genomic features to a previously used data set for the fruit fly (De Ferrari and Aitken, 2006). The performance of the SVM classifier compared favourably with that of De Ferrari and Aitken (2006). One difficulty with comparing the two works was the discrepancy between results reported in the paper, and provided with the supplementary data files. Another difficulty arose due to the fact that our gene specificity values were base on genes, whereas the authors worked with transcripts. It is possible, that some of the genes have transcript with distinct tissue specificity, but our measurement based on the FlyAtlas data does not capture this.

When compared to previous studies, we observed a discrepancy in the correlation between expression breadth and sequence length in this fruit fly and mouse study. According to our results, there is no obvious correlation between sequence length and tissue specificity. We did not find that protein sequences and various genomic regions are shorter for housekeeping genes. This may be due to the definition of housekeeping genes. The broadly expressed genes investigated here might not necessarily all have housekeeping functions.

We observed a good correlation between the frequency of CpG islands in the mouse promoter regions and gene specificity. However, in the fruit fly, no correlation was found. This is in agreement with our expectation, since *Drosophila* is not known to have DNA methylation.

Because of the flexible structure of the SVM classifier, additional attributes can be easily added: either attributes already studied or newly discovered ones. For instance, in addition to the amino acid composition, the amino acid pair compositions could be integrated. In addition, transcription factor binding sites that are discovered de-novo or via database searches could also be incorporated. The ribosomal occupancy data investigated here could also be integrated in the SVMs.

Considering the success in discriminating tissue-specific and broadly expressed genes within an organism, future work might include the prediction of tissue specificity of genes in other model organisms. For instance, the worm, *C. elegans* is a popular model organism, but tissue-specific information is only available for part of its genome. The fruit fly and mouse models trained in this work could potentially be used to infer tissue specificity for these worm data. However, we did not follow this up during the course of this PhD study due to time limitation. We showed, however, that a classifier trained on mouse genes and tested on fly genes, or vice versa, performed better than random in experiments using between 15 and 500 genes to train and test the model. Systematically optimising the number of genes in the training and test sets is expected to further increase the prediction performance since the 100 most tissue-specific genes in the mouse are not directly comparable to the 100 most tissue-specific genes in the fly.

Table 5.4: Median AUCs calculated for different classifiers discriminating genes with various tissue specificity in the fruit fly. Genes assigned to *bin 1* are tissue-specific and constitute the positive training set, while genes belonging to one of the other *bins* constitute the negative training set. Tissue specificity decreases with the *bin* number. *AA%* and *SS%* indicate that the input vector for the SVM contained amino acid percentages and the secondary structure symbol percentages for each gene. *AA scores* is based on Smith-Waterman similarity scores of the protein sequences, *SS scores* is based on Smith-Waterman similarity scores of the secondary structures and *AA+SS scores* is a combination of the latter two. The attributes of the SVM based on genomic features were protein sequence length, cds length, cDNA length, 5'UTR length, 3'UTR length, upstream marscan results, downstream marscan results, number of exons, number of CpG islands and the CpG content.

	Neg.	**Median AUCs**						
	Class	**AA**	**AA**	**SS**	**SS**	**AA+SS**	**genomic**	**AA % +**
	Bin	**%**	**scores**	**%**	**scores**	**scores**	**features**	**genomic features**
tissue specificity increases →	1	0.541	0.503	0.471	0.479	0.510	0.665	0.473
	2	0.616	0.614	0.499	0.560	0.585	0.621	0.669
	3	0.678	0.707	0.495	0.597	0.690	0.579	0.783
	4	0.743	0.779	0.497	0.666	0.764	0.718	0.854
	5	0.754	0.763	0.515	0.663	0.748	0.739	0.874
	6	0.752	0.794	0.555	0.714	0.774	0.682	0.872
	7	0.770	0.803	0.555	0.684	0.785	0.698	0.885
	8	0.787	0.837	0.507	0.703	0.801	0.694	0.888
	9	0.794	0.815	0.565	0.707	0.793	0.747	0.918
	10	0.791	0.826	0.566	0.716	0.799	0.751	0.916
	11	0.792	0.829	0.507	0.723	0.808	0.761	0.913
	12	0.781	0.830	0.517	0.748	0.816	0.793	0.925
	13	0.790	0.824	0.493	0.753	0.817	0.826	0.918
	14	0.795	0.816	0.493	0.730	0.802	0.849	0.925
	15	0.758	0.795	0.495	0.675	0.772	0.821	0.883
	16	0.778	0.789	0.523	0.666	0.775	0.793	0.883
	17	0.749	0.754	0.496	0.675	0.727	0.732	0.846

Table 5.5: Mean AUCs calculated for different classifiers discriminating genes with various tissue specificity in the fruit fly. The mean AUCs are averaged over 100 runs and are given with a confidence interval (±1.96×standard error). Genes assigned to *bin 1* are tissue-specific and constitute the positive training set, while genes belonging to one of the other *bins* constitute the negative training set. Tissue specificity decreases with the *bin* number. *AA%* and *SS%* indicate that the input vector for the SVM contained amino acid percentages and the secondary structure symbol percentages for each gene. *AA scores* is based on Smith-Waterman similarity scores of the protein sequences, *SS scores* is based on Smith-Waterman similarity scores of the secondary structures and *AA+SS scores* is a combination of the latter two. The attributes of the SVM based on genomic features were protein sequence length, cds length, cDNA length, 5'UTR length, 3'UTR length, upstream marscan results, downstream marscan results, number of exons, number of CpG islands and the CpG content.

	Neg. Class Bin	**Median AUCs**						
		AA %	**AA scores**	**SS %**	**SS scores**	**AA+SS scores**	**genomic features**	**AA % + genomic features**
tissue specificity increases →	1	0.546±0.012	0.502±0.014	0.466±0.012	0.481±0.014	0.503±0.014	0.664±0.006	0.472±0.012
	2	0.615±0.008	0.611±0.010	0.502±0.008	0.561±0.008	0.587±0.008	0.623±0.006	0.668±0.008
	3	0.680±0.010	0.701±0.010	0.500±0.010	0.595±0.010	0.692±0.010	0.578±0.008	0.78±0.008
	4	0.742±0.006	0.773±0.008	0.499±0.008	0.662±0.010	0.761±0.010	0.715±0.006	0.851±0.006
	5	0.751±0.008	0.763±0.008	0.513±0.008	0.659±0.012	0.750±0.008	0.740±0.006	0.876±0.006
	6	0.744±0.008	0.792±0.008	0.548±0.010	0.715±0.010	0.774±0.010	0.680±0.006	0.87±0.006
	7	0.766±0.008	0.802±0.010	0.553±0.010	0.687±0.010	0.777±0.010	0.700±0.006	0.885±0.006
	8	0.784±0.006	0.830±0.006	0.508±0.008	0.702±0.008	0.798±0.008	0.692±0.006	0.887±0.004
	9	0.800±0.008	0.809±0.008	0.566±0.010	0.706±0.010	0.789±0.010	0.746±0.006	0.916±0.004
	10	0.794±0.008	0.820±0.008	0.561±0.008	0.711±0.008	0.794±0.010	0.751±0.006	0.916±0.004
	11	0.787±0.006	0.828±0.008	0.510±0.010	0.721±0.010	0.809±0.008	0.761±0.004	0.912±0.004
	12	0.782±0.008	0.831±0.006	0.514±0.008	0.747±0.008	0.815±0.008	0.791±0.004	0.923±0.004
	13	0.790±0.006	0.825±0.006	0.496±0.010	0.752±0.008	0.815±0.006	0.780±0.004	0.916±0.004
	14	0.792±0.006	0.820±0.006	0.493±0.010	0.729±0.008	0.802±0.006	0.782±0.004	0.925±0.002
	15	0.752±0.008	0.794±0.006	0.497±0.008	0.674±0.008	0.768±0.006	0.795±0.004	0.881±0.004
	16	0.773±0.008	0.790±0.006	0.520±0.008	0.668±0.008	0.772±0.006	0.754±0.004	0.878±0.004
	17	0.740±0.018	0.751±0.018	0.503±0.020	0.655±0.024	0.719±0.022	0.743±0.014	0.845±0.016

Table 5.6: Median AUCs calculated for different classifiers discriminating genes with various tissue specificity in the mouse. Genes assigned to *bin 1* are tissue-specific and constitute the positive training set, while genes belonging to one of the other *bins* constitute the negative training set. Tissue specificity decreases with the *bin* number. *AA%* and *SS%* indicate that the input vector for the SVM contained amino acid percentages and the secondary structure symbol percentages for each gene. *AA scores* is based on Smith-Waterman similarity scores of the protein sequences, *SS scores* is based on Smith-Waterman similarity scores of the secondary structures and *AA+SS scores* is a combination of the latter two. The attributes of the SVM based on genomic features were protein sequence length, cds length, cDNA length, 5'UTR length, 3'UTR length, upstream marscan results, downstream marscan results, number of exons, number of CpG islands and the CpG content.

	Neg.	Median AUCs						
	Class	AA	AA	SS	SS	AA+SS	genomic	AA % + gen–
	Bin	%	scores	%	scores	scores	features	omic features
tissue specificity increases ↑	1	0.625	0.500	0.500	0.500	0.500	0.500	0.667
	2	0.596	0.577	0.490	0.500	0.596	0.477	0.569
	3	0.659	0.761	0.545	0.636	0.705	0.625	0.733
	4	0.650	0.700	0.650	0.612	0.700	0.540	0.680
	5	0.750	0.875	0.525	0.700	0.812	0.683	0.850
	6	0.775	0.788	0.625	0.712	0.800	0.580	0.800
	7	0.750	0.844	0.625	0.688	0.812	0.689	0.800
	8	0.722	0.806	0.667	0.722	0.736	0.667	0.733
	9	0.719	0.875	0.688	0.750	0.906	0.625	0.775
	10	0.750	0.750	0.638	0.738	0.775	0.800	0.836
	11	0.719	0.875	0.469	0.688	0.859	0.778	0.900
	12	0.766	0.812	0.688	0.641	0.797	0.744	0.844
	13	0.736	0.833	0.639	0.750	0.778	0.760	0.840
	14	0.821	0.857	0.643	0.786	0.839	0.771	0.857
	15	0.778	0.778	0.694	0.694	0.778	0.820	0.840
	16	0.847	0.833	0.639	0.792	0.806	0.815	0.923
	17	0.861	0.889	0.611	0.750	0.861	0.840	0.920
	18	0.833	0.861	0.667	0.819	0.889	0.780	0.800
	19	0.812	0.906	0.688	0.812	0.875	0.800	0.925
	20	0.844	0.906	0.656	0.766	0.844	0.811	0.889
	21	0.875	0.944	0.694	0.792	0.889	0.867	0.889
	22	0.861	0.875	0.681	0.778	0.889	0.880	0.890
	23	0.750	0.786	0.607	0.643	0.714	0.850	0.850
	24	0.875	0.825	0.612	0.750	0.850	0.825	0.933
	25	0.806	0.861	0.583	0.681	0.806	0.800	0.900
	26	0.875	0.938	0.688	0.812	0.906	0.822	0.900
	27	0.844	0.844	0.719	0.734	0.812	0.889	0.889
	28	0.875	0.938	0.750	0.812	0.891	0.860	0.920
	29	0.875	0.900	0.725	0.825	0.875	0.850	0.883
	30	0.865	0.875	0.719	0.781	0.833	0.885	0.938
	31	0.900	0.900	0.575	0.750	0.875	0.873	0.909
	32	0.875	0.900	0.675	0.775	0.850	0.800	0.900
	33	0.788	0.875	0.650	0.725	0.812	0.818	0.864
	34	0.833	0.917	0.677	0.812	0.896	0.893	0.933

Continued on next page

Table 5.6 – continued from previous page

	Neg.	Median AUCs						
	Class	AA	AA	SS	SS	AA+SS	genomic	AA % + gen–
	bin	%	scores	%	scores	scores	features	omic features
tissue specificity increases →	35	0.875	0.850	0.675	0.750	0.812	0.900	0.950
	36	0.875	0.900	0.600	0.825	0.900	0.891	0.945
	37	0.884	0.893	0.607	0.786	0.839	0.912	0.950
	38	0.841	0.898	0.636	0.773	0.841	0.846	0.923
	39	0.875	0.917	0.615	0.781	0.896	0.923	0.892
	40	0.929	0.946	0.688	0.857	0.929	0.893	0.947
	41	0.942	0.917	0.717	0.833	0.883	0.912	0.938
	42	0.839	0.875	0.455	0.759	0.839	0.825	0.900
	43	0.891	0.922	0.570	0.766	0.844	0.894	0.944
	44	0.875	0.921	0.441	0.776	0.855	0.900	0.927
	45	0.897	0.912	0.669	0.809	0.882	0.927	0.968
	46	0.847	0.917	0.618	0.757	0.861	0.910	0.933
	47	0.881	0.900	0.531	0.750	0.875	0.892	0.933
	48	0.905	0.893	0.571	0.738	0.821	0.904	0.961
	49	0.938	0.926	0.670	0.761	0.915	0.881	0.944
	50	0.889	0.944	0.569	0.778	0.894	0.906	0.941
	51	0.901	0.914	0.647	0.823	0.862	0.837	0.941
	52	0.929	0.893	0.464	0.763	0.857	0.903	0.942
	53	0.882	0.929	0.464	0.771	0.879	0.837	0.937
	54	0.884	0.925	0.619	0.756	0.844	0.884	0.942
	55	0.890	0.913	0.480	0.770	0.866	0.774	0.932
	56	0.907	0.928	0.470	0.819	0.898	0.865	0.943
	57	0.904	0.939	0.465	0.760	0.891	0.843	0.951
	58	0.901	0.885	0.500	0.767	0.849	0.796	0.947
	59	0.855	0.902	0.473	0.777	0.865	0.785	0.934
	60	0.893	0.958	0.335	0.835	0.906	0.837	0.956

Table 5.7: Mean AUCs calculated for different classifiers discriminating genes with various tissue specificity in the mouse. The mean AUCs are averaged over 100 runs and are given with a confidence interval (±1.96×standard error). Genes assigned to *bin 1* are tissue-specific and constitute the positive training set, while genes belonging to one of the other *bins* constitute the negative training set. Tissue specificity decreases with the *bin* number. *AA%* and *SS%* indicate that the input vector for the SVM contained amino acid percentages and the secondary structure symbol percentages for each gene. *AA scores* is based on Smith-Waterman similarity scores of the protein sequences, *SS scores* is based on Smith-Waterman similarity scores of the secondary structures and *AA+SS scores* is a combination of the latter two. The attributes of the SVM based on genomic features were protein sequence length, cds length, cDNA length, 5'UTR length, 3'UTR length, upstream marscan results, downstream marscan results, number of exons, number of CpG islands and the CpG content.

	Neg.	Median AUCs						
	Class	AA	AA	SS	SS	AA+SS	genomic	AA % + gen−
	Bin	%	scores	%	scores	scores	features	omic features
tissue specificity increases →	1	0.638±0.055	0.502±0.057	0.382±0.059	0.552±0.057	0.530±0.061	0.427±0.053	0.608±0.049
	2	0.597±0.027	0.585±0.029	0.490±0.035	0.508±0.037	0.605±0.033	0.478±0.029	0.574±0.027
	3	0.667±0.031	0.755±0.027	0.546±0.031	0.627±0.031	0.701±0.031	0.605±0.029	0.727±0.025
	4	0.656±0.031	0.700±0.031	0.644±0.025	0.618±0.033	0.690±0.031	0.541±0.027	0.659±0.027
	5	0.744±0.029	0.851±0.022	0.526±0.033	0.687±0.031	0.781±0.027	0.693±0.024	0.826±0.022
	6	0.745±0.029	0.793±0.024	0.614±0.033	0.709±0.031	0.798±0.025	0.578±0.025	0.776±0.024
	7	0.733±0.027	0.812±0.025	0.610±0.041	0.696±0.029	0.795±0.027	0.681±0.024	0.783±0.024
	8	0.698±0.029	0.751±0.033	0.672±0.039	0.704±0.031	0.723±0.031	0.656±0.031	0.735±0.022
	9	0.710±0.029	0.870±0.022	0.694±0.037	0.743±0.033	0.870±0.024	0.632±0.029	0.765±0.022
	10	0.728±0.027	0.738±0.027	0.622±0.041	0.739±0.027	0.755±0.027	0.788±0.024	0.828±0.022
	11	0.711±0.031	0.857±0.022	0.502±0.037	0.659±0.031	0.827±0.027	0.770±0.024	0.867±0.024
	12	0.769±0.027	0.783±0.029	0.650±0.039	0.664±0.035	0.797±0.027	0.725±0.027	0.810±0.024
	13	0.716±0.029	0.816±0.029	0.615±0.033	0.730±0.033	0.769±0.033	0.750±0.025	0.829±0.020
	14	0.795±0.031	0.850±0.024	0.638±0.039	0.797±0.031	0.820±0.027	0.769±0.025	0.827±0.025
	15	0.763±0.027	0.754±0.031	0.668±0.037	0.675±0.033	0.744±0.035	0.796±0.022	0.831±0.022
	16	0.829±0.027	0.817±0.024	0.641±0.035	0.758±0.033	0.786±0.027	0.820±0.020	0.909±0.014
	17	0.850±0.024	0.855±0.022	0.600±0.033	0.734±0.031	0.828±0.027	0.823±0.020	0.892±0.020
	18	0.815±0.024	0.837±0.024	0.674±0.033	0.802±0.027	0.878±0.020	0.778±0.022	0.795±0.024
	19	0.808±0.025	0.894±0.020	0.689±0.035	0.792±0.031	0.840±0.027	0.798±0.024	0.886±0.022
	20	0.829±0.024	0.861±0.024	0.658±0.035	0.758±0.029	0.826±0.025	0.785±0.024	0.881±0.020
	21	0.848±0.024	0.910±0.022	0.658±0.035	0.780±0.029	0.854±0.024	0.853±0.022	0.870±0.020
	22	0.844±0.024	0.859±0.022	0.667±0.037	0.770±0.031	0.866±0.024	0.871±0.018	0.871±0.018
	23	0.724±0.029	0.795±0.027	0.597±0.031	0.652±0.037	0.686±0.031	0.83±0.0220	0.832±0.024
	24	0.862±0.022	0.814±0.024	0.618±0.031	0.742±0.031	0.820±0.025	0.812±0.020	0.910±0.016
	25	0.809±0.025	0.843±0.025	0.609±0.041	0.681±0.033	0.778±0.025	0.777±0.024	0.886±0.020
	26	0.858±0.022	0.929±0.014	0.671±0.035	0.812±0.025	0.892±0.018	0.819±0.022	0.888±0.018
	27	0.847±0.022	0.845±0.022	0.688±0.033	0.730±0.029	0.814±0.025	0.864±0.020	0.870±0.018
	28	0.852±0.022	0.906±0.020	0.713±0.033	0.795±0.029	0.870±0.022	0.839±0.018	0.903±0.016
	29	0.866±0.022	0.884±0.018	0.698±0.031	0.808±0.027	0.844±0.024	0.831±0.020	0.876±0.016
	30	0.845±0.024	0.872±0.018	0.687±0.037	0.750±0.033	0.825±0.022	0.883±0.014	0.927±0.014
	31	0.867±0.020	0.875±0.024	0.588±0.035	0.743±0.031	0.846±0.024	0.848±0.020	0.897±0.016
	32	0.842±0.025	0.875±0.020	0.674±0.039	0.782±0.027	0.851±0.024	0.800±0.022	0.891±0.018

Continued on next page

Table 5.7 – continued from previous page

	Neg.	Median AUCs						
	Class	AA	AA	SS	SS	AA+SS	genomic	AA % + gen−
	bin	%	scores	%	scores	scores	features	omic features
tissue specificity increases →	33	0.785±0.024	0.867±0.018	0.674±0.033	0.718±0.029	0.804±0.027	0.798±0.022	0.863±0.018
	34	0.822±0.024	0.899±0.020	0.661±0.035	0.795±0.027	0.876±0.020	0.885±0.016	0.911±0.014
	35	0.871±0.020	0.838±0.022	0.670±0.033	0.730±0.031	0.797±0.025	0.879±0.018	0.930±0.012
	36	0.866±0.020	0.873±0.022	0.612±0.033	0.814±0.025	0.876±0.024	0.894±0.014	0.927±0.014
	37	0.875±0.020	0.877±0.020	0.605±0.033	0.771±0.025	0.829±0.025	0.907±0.012	0.929±0.014
	38	0.821±0.024	0.885±0.018	0.630±0.039	0.768±0.029	0.843±0.024	0.824±0.022	0.911±0.016
	39	0.871±0.020	0.901±0.020	0.616±0.035	0.769±0.027	0.862±0.024	0.913±0.014	0.888±0.016
	40	0.894±0.020	0.921±0.018	0.676±0.035	0.840±0.025	0.887±0.022	0.876±0.016	0.932±0.012
	41	0.905±0.020	0.900±0.020	0.676±0.037	0.808±0.025	0.859±0.020	0.894±0.014	0.917±0.014
	42	0.819±0.022	0.857±0.024	0.461±0.035	0.738±0.029	0.807±0.029	0.813±0.018	0.890±0.014
	43	0.875±0.020	0.894±0.016	0.555±0.039	0.756±0.027	0.825±0.025	0.879±0.018	0.929±0.012
	44	0.854±0.022	0.896±0.020	0.447±0.035	0.766±0.033	0.857±0.020	0.881±0.018	0.918±0.014
	45	0.865±0.022	0.906±0.016	0.656±0.035	0.798±0.024	0.859±0.020	0.920±0.012	0.950±0.012
	46	0.829±0.024	0.889±0.020	0.618±0.043	0.754±0.029	0.828±0.024	0.902±0.012	0.922±0.012
	47	0.855±0.024	0.879±0.018	0.534±0.039	0.751±0.027	0.847±0.024	0.858±0.020	0.918±0.014
	48	0.890±0.018	0.879±0.022	0.557±0.039	0.745±0.027	0.809±0.024	0.881±0.016	0.954±0.008
	48	0.890±0.018	0.879±0.022	0.557±0.039	0.745±0.027	0.809±0.024	0.881±0.016	0.954±0.008
	49	0.891±0.022	0.909±0.014	0.647±0.035	0.756±0.029	0.879±0.022	0.853±0.020	0.935±0.010
	50	0.855±0.022	0.924±0.016	0.570±0.037	0.787±0.025	0.856±0.025	0.883±0.018	0.920±0.016
	51	0.871±0.018	0.892±0.016	0.633±0.033	0.798±0.024	0.836±0.024	0.814±0.020	0.923±0.012
	52	0.887±0.020	0.882±0.016	0.449±0.035	0.769±0.025	0.848±0.022	0.890±0.014	0.928±0.012
	53	0.852±0.024	0.909±0.016	0.445±0.035	0.765±0.029	0.848±0.024	0.822±0.020	0.921±0.012
	54	0.855±0.024	0.890±0.022	0.584±0.035	0.758±0.027	0.821±0.025	0.871±0.014	0.918±0.016
	55	0.873±0.020	0.896±0.018	0.496±0.031	0.763±0.029	0.854±0.020	0.759±0.022	0.909±0.016
	56	0.881±0.020	0.902±0.016	0.494±0.039	0.771±0.031	0.868±0.022	0.837±0.020	0.924±0.012
	57	0.878±0.018	0.926±0.012	0.467±0.041	0.768±0.027	0.858±0.022	0.837±0.016	0.922±0.016
	58	0.883±0.020	0.874±0.018	0.480±0.039	0.750±0.029	0.823±0.024	0.790±0.020	0.919±0.014
	59	0.825±0.027	0.900±0.014	0.482±0.039	0.759±0.029	0.845±0.020	0.776±0.024	0.925±0.014
	60	0.873±0.020	0.945±0.010	0.369±0.039	0.826±0.024	0.885±0.018	0.819±0.018	0.940±0.012

Chapter 6

Analysis of the Tissue-Specific Contribution to Whole-Body RNA Transcript Profiles in *Drosophila Melanogaster*

The ordered list of genes for fruit fly genes introduced in the previous chapter is used in the following to analyse the capacity of whole-body microarrays to detect tissue-specific expression in the ageing fly and in general. The tissue specificity of age-associated genes is also investigated. The chapter starts by describing the methods used, and presents the results and a discussion of the results.

6.1 Methods

6.1.1 Ranking genes according to their tissue specificity

The same method was applied to rank genes according to their tissue specificity as in Chapter 5.1.

6.1.2 Whole-fruit fly gene expression data

An ageing experiment was used to investigate the tissue-specific contribution to whole-body RNA transcript profiles, and to investigate the tissue specificity of age-associated genes. Data on whole-genome *Drosophila melanogaster* gene expression for the aged fly were previously described (McElwee *et al.*, 2007). In this study, wild-type (Dahomey) and long-lived chico^1/+ heterozygotes were compared. Chico^1/+ is a null mutation in the fly insulin receptor substrate in the insulin/insulin-like growth factor-1 signalling (Insulin and Insulin-like growth factor signaling (IIS)) pathway, a pathway central to ageing. We used *Supplementary data file 9* of this study, which contains the results from the statistical analysis including probeset IDs, gene IDs, mean expression signals, fold changes and the results of the statistical analysis. The list was used to determine differentially expressed genes ($q < 0.1$). This identified 1,169 differentially expressed genes (893 upregulated gene, 276 downregulated gene in chico^1/+). FlyBase gene IDs were mapped to Gene Ontology (GO) IDs version 1.107 (Ashburner *et al.*, 2000).

Data for other whole-fly experiments, again based on Drosophila Genome 2.0 Arrays, were downloaded from ArrayExpress (Parkinson *et al.*, 2007). For all six datasets (E-GEOD-7763, E-GEOD-5404, E-GEOD-8775, E-MEXP-1594 (208), E-MEXP-1594 (301), E-GEOD-7614), raw data (cel files) were normalised, using eight different normalisation routines following the method used in McElwee *et al.* (2007).

6.1.3 Age-associated genes

A list of 46 fly genes that have been shown to extend lifespan was downloaded from GenAge (de Magalhães and Toussaint, 2004) to investigate their degree of tissue specificity.

Lists of age-related genes in seven tissues have been collected from a microarray study on 15-60 days old flies (Zhan *et al.*, 2007).

6.1.4 Others

For functional overrepresentation analysis, we used a modified Fisher's exact test. We re-implemented the EASE (Expression Analysis Systematic Explorer) software (Hosack *et al.*, 2003; Huang da *et al.*, 2007) in R. EASE calculates over-representation with respect to the total number of genes assayed and annotated within each system (here Gene Ontology annotations). We used all genes available in Supplementary data file 9 from (McElwee *et al.*, 2007) as background sequences. For the filtering experiments, tissue-specific genes were removed from both the set of differentially expressed genes and the full set of background sequences. We used the EASE score to determine the significance of categories. The EASE score is a conservative adjustment to the Fisher exact probability that favours more frequent categories (Hosack *et al.*, 2003) over less frequent categories. The EASE score is calculated by removing one gene within the given category from the list and calculating the Fisher exact probability for that category. This process is exemplified in the following. Assume a list of 200 genes is differentially expressed from a population of 12,000 genes. If there is only one gene in the population in a rare category, e.g. "Rare function", and that gene happens to appear on the list of 200 genes, the Fisher exact test would deem that category significant (p = 0.016). Similarly, the Fisher exact test would consider a more common category, "More common function", with 765 members in the population and 20 members on the list, as slightly less significant (p = 0.017). From the biological perspective, a category based on the presence of a single gene is rarely interesting. If the single gene is a false positive, then the significance of the corresponding category is false. The EASE scores for these combinations are p = 1 and p = 0.030 for categories "Rare function" and for category "More common function", respectively. Thus, the EASE score eliminates the significance of the infrequent category while only slightly penalising the significance of the more global theme. The EASE score penalises the significance of categories supported by fewer genes and favours more robust categories compared to the Fisher exact probability.

Linear regression to measure the correlation between mean expression signal and tissue specificity was performed using the R programming language (R, 2009).

Clover (Frith *et al.*, 2004) was used for motif detection. We used experimentally verified transcription factor binding sites (TFBS) from TRANSFAC (2007) (Matys *et al.*, 2003) to scan the fruit fly promoter sequences (846 motifs in total). The promoter sequences (1000 bp from the transcription start sites) were extracted from Ensembl using BioMart (Durinck *et al.*, 2005), a data integration system for large scale querying of biological data. Clover compares each motif to the given sequence set, and calculates a raw score that quantifies the degree of the motif's presence in the test sequences. The present genes on the microarray served as background DNA sequences. The Clover algorithm repeatedly extracts random fragments of the background sequences (here we used 1000 randomisations), matched by length to the target sequences, calculates a raw score for each set of fragments and uses these to estimate a p value. The proportion of times that the raw score of a fragment set exceeds or equals the raw score of the target set, e.g. 0.01, is taken as the p value. Thus the p value indicates the probability of obtaining a raw score of this size or greater merely by chance, computed using background sequence sets. For each motif, a separate p value was calculated. In this work we considered motifs with a score $>$ 15 and a p value < 0.01 to be over-represented in the given gene lists. We did not consider under-represented motifs.

We tested whether each motif from a library of 846 is significantly overrepresented in a given sequence set. That means that it is likely that a few motifs will have p values more significant than 0.01 merely by chance. However, all the p values in this study were obtained by performing 1,000 randomisations, and motifs with p values < 0.01 and score > 15 are listed. Amongst the 90 overrepresented motifs we found 45 motifs with a p value of zero, i.e. the raw scores were never equalled in 1,000 randomisations, which is highly unlikely to occur by chance. We also find more motifs with p values < 0.01 than expected by chance. On average we would expect to get a false positive result about once every 100 times the test is used (1/0.01). This translates to 622 false positives (0.01× 72 groups × 864 motifs). A total of 2,355 motifs has been found at a p value < 0.01, whereas a total of 330 of these were unique motifs. These were further reduced to a final number of 90 motifs by taking into account only those motifs with a score > 15. Thus, we are confident that the majority of motif predictions made here are not merely due to chance.

6.2 Results and discussion

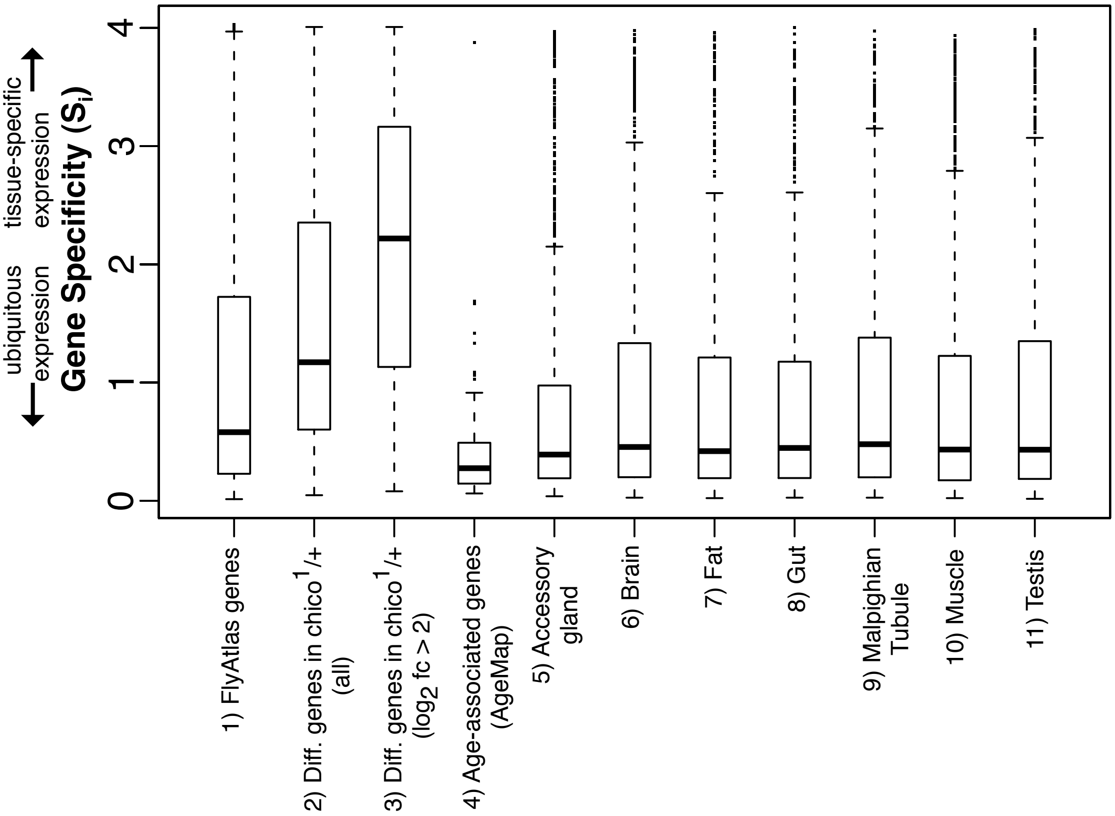


Figure 6.1: **Tissue specificity of genes in the fruit fly.** The boxplots show the average gene specificity of genes expressed in at least one of the FlyAtlas tissues (box 1), of all differentially expressed genes identified in a longevity experiment (box 2) and genes with a $\log_2$ fold change > 2 identified in this longevity experiment (box 3), of genes that have been shown to extend lifespan in the adult fruit fly (box 4), of genes that have been associated with ageing in various tissues (boxes 5-11).

We downloaded gene expression profiles for 17 tissues from FlyAtlas. For each gene that could be detected in at least one of these tissues (11,804 fruit fly genes) we calculated a gene specificity index S_i by measuring the degree by which its gene expression profile differs from a ubiquitous expression profile in which the same expression levels is observed for all tissues (Chapter 5). In this data set a minimum S_i value of 0.014 was achieved, and a maximum of 4.09. The median S_i value of 0.580 indicates that the majority of genes in this data set are broadly expressed (Figure 6.1, Box 1). We grouped the genes with variable tissue specificity into bins using Formula 5.3 as detailed in the previous chapter. Each bin defines a group of genes with a certain degree of tissue specificity (Figure 6.6).

6.2.1 Applying tissue-specific information to whole body expression profiles

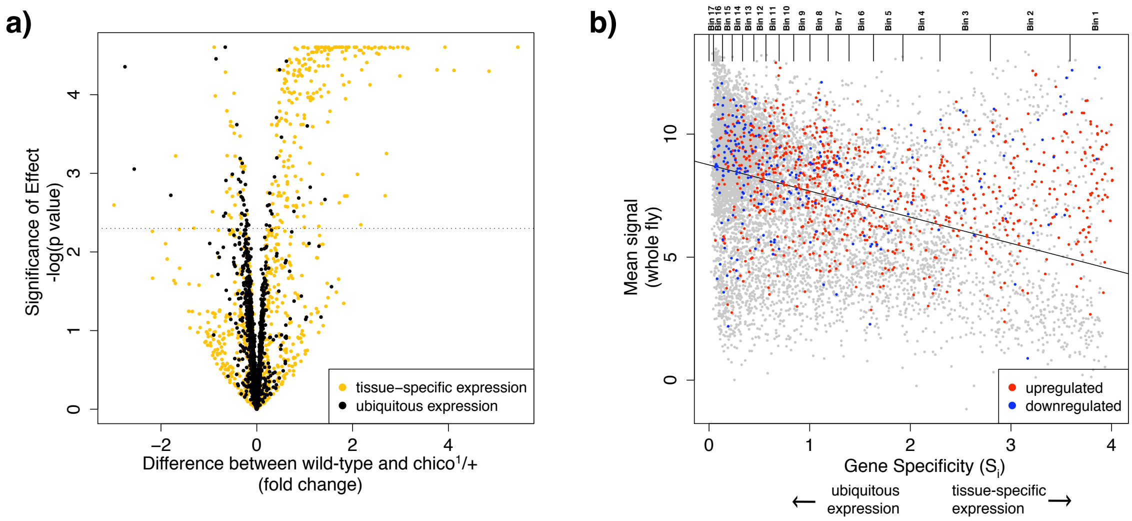


Figure 6.2: **Tissue specificity of differentially expressed genes in *chico*^{1}/+. a)** Volcano plot of $\log_2$ fold changes versus significance of differential expression. The 2,000 genes with the lowest (black dots) and highest (yellow dots) gene specificity values are plotted. The dotted horizontal line marks the threshold of p value = log (0.1) = 2.3 above which genes were considered differentially expressed in *chico*^{1}/+. **b)** The relationship between gene specificity and mean expression signal measured for these genes in the whole-fly longevity experiment. Each gene is plotted with the specificity value calculated from the FlyAtlas tissue data (X-axis) versus the mean expression value for that particular gene in the wild type and *chico*^{1}/+ whole-fly samples (Y-axis). Note the decrease of the average expression amplitude with increasing gene specificity. Blue dots indicate downregulation while red dots indicate upregulation in the long-lived animals ($\log_2$ fold change > 0).

To investigate the capacity of whole body expression profiles to capture the tissue-specific contributions regarding age-associated changes, we used data from a previously published longevity study (McElwee *et al.*, 2007), which included data in which wild-type (Dahomey) and long-lived *chico*^{1}/+ heterozygotes were compared. In this study, evolutionary conservation of regulated longevity assurance mechanisms was investigated using microarray data from long-lived mutant worms, mice with lowered IIS and a long-lived IIS mutant in flies (*chico*^{1}/+ heterozygotes). The whole fly gene expression data set was downloaded together with a set of 1,169 differentially expressed genes identified in the *chico*^{1}/+ flies.

We determined the average tissue specificity of these differentially expressed genes using the gene specificity values calculated from the FlyAtlas data (Figure 6.1, Box 2). The tissue specificity of differentially expressed age-associated genes covers the possible range. We found, however, that, in this experiment, tissue-specific genes are associated with higher fold changes between wild type and long-lived flies (Figure 6.1 Box 3, Figure 6.2). The median S_i value of all differentially expressed genes was 1.17 while it increased to 2.22 if only significantly differentially expressed genes with a $\log_2$ fold change > 1 were considered.

At least three explanations could account for these higher fold changes. First, we expect the absolute quantities of mRNA to be lower for tissue-specific genes than for broadly expressed genes. Lowly expressed genes are often associated with high variances and will tend to exceed higher fold change cut-offs (Mutch *et al.*, 2002). The `Cyber-T` software (Baldi and Long, 2001) used to detect the differentially expressed genes in this experiment penalises lowly expressed genes and, thus, many changes in tissue-specific expression with low fold changes might not be reported by this method. A test for association and a simple linear regression were performed on the data to determine if there was a significant relationship between the gene specificity value (S_i) and the mean expression signal (S) in this data (Figure 6.2b). There was evidence that S_i negatively correlated with S, with a Pearson's product moment correlation coefficient r of -0.39. The t-statistic for the slope was significant at the 0.05 critical alpha level, $p < 2.2\times10^{-16}$; 15.2% of the variability in mean expression signals could be explained by the gene specificity value ($r^2 = 0.152$). Thus, we rejected the null hypothesis and concluded that there was a negative significant relationship between the gene specificity values and the mean expression signals. This indicates that tissue-specific genes overall display a lower mean expression signal than broadly expressed genes in this sample. To investigate if this trend was specific for this experiment, or if this is generally observed, we tested if we could find a similar correlation in other whole-fly experiments (Ayroles *et al.*, 2009; Chintapalli *et al.*, 2007; Edwards *et al.*, 2006; Magwire, 2007; Morozova *et al.*, 2007). All data sets examined showed a negative correlation between tissue specificity and mean expression signal (Figure 6.3: $-0.323 < r < -0.143$).

The observation that gene expression signals correlate with the tissue specificity of genes is compatible with the previous findings that gene expression level and breadth are positively correlated in human data (Eisenberg and Levanon, 2003; Lercher *et al.*, 2002; Reverter *et al.*, 2008; Vinogradov, 2004; Zhu *et al.*, 2008). However, it is not compatible with a later study on human and mouse data where virtually no correlation was found between expression level and tissue specificity (Liao and Zhang, 2006). The lower mean expression signal for tissue-specific genes partially explains the higher fold changes for tissue-specific differentially expressed genes. However, the many points deviating from the regression line ($r^2 = 0.152$) indicate that there might be further reasons for the bias in fold changes, since not all tissue-specific genes display low expression signals and vice versa.

A second possible explanation for the bias in fold changes is that changes in tissue-specific expression might be easier to detect in whole-fly samples by microarray technologies: the resulting data for each gene on an array represents the sum of signals from every tissue and cell-type present in the sample. Thus, a change of the expression level of a constitutively expressed gene in one tissue can be compensated

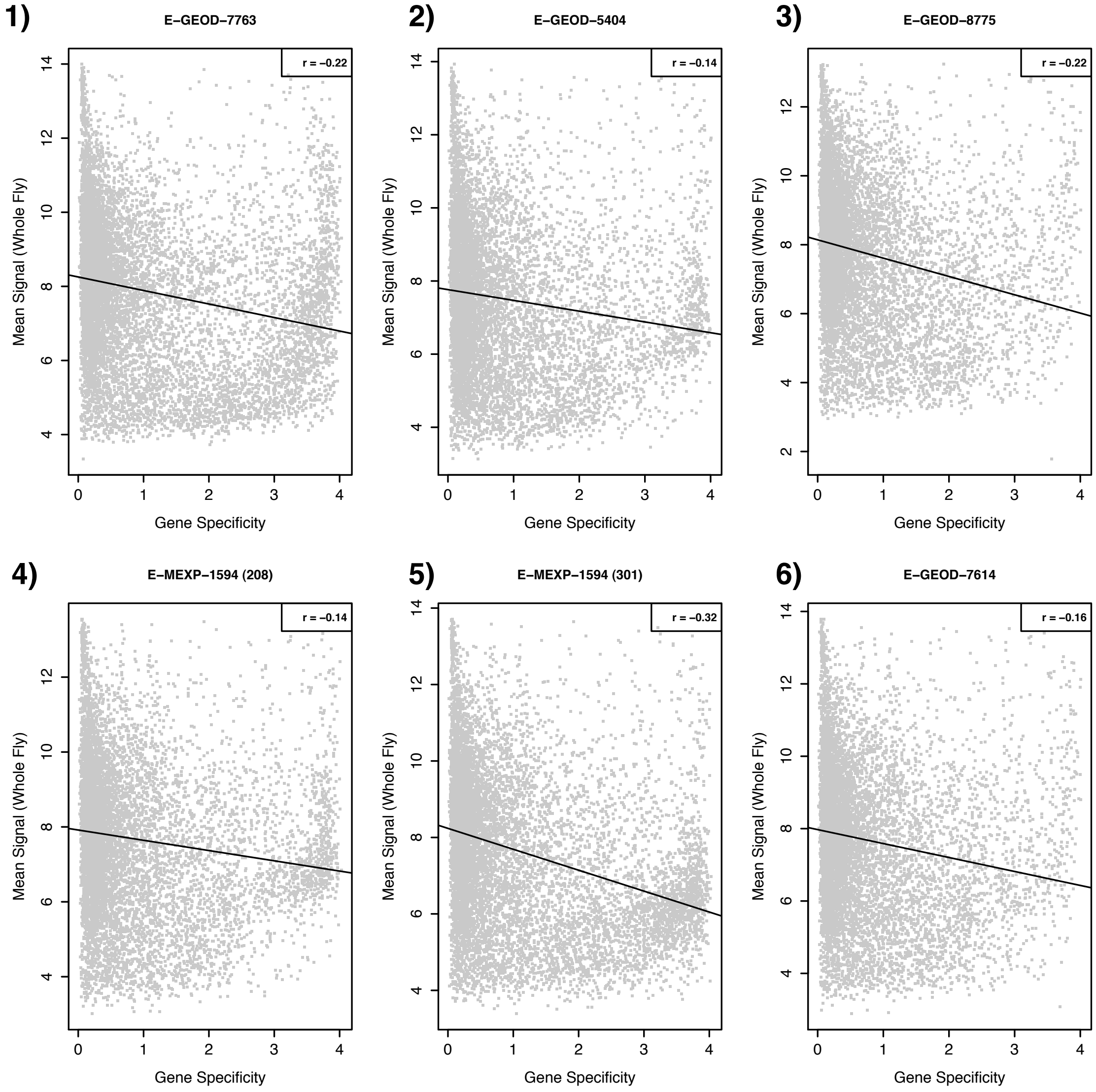


Figure 6.3: **The relationship between gene specificity and mean expression signal.** The relationship between gene specificity and mean expression signal measured in various whole-fly experiments. E-GEOD is the accession number for the respective experiment in ArrayExpress. The variables r in the legends indicate the values of the correlation coefficient. The data were extracted from studies of (1) whole-fly wildtype data, (2) aggressive behavior in fruit flies (3) lifespan extension (4+5) different Drosophila lines and (6) alcohol sensitivity in the fruit fly.

by changes in the opposite direction in one of the other tissues resulting in overall lower fold changes. In order to validate this a data set would be required that provides similar information to the FlyAtlas data set but with the additional dimension of ageing added. A comprehensive data set was not available at the time of writing, but we investigated another ageing data set that was performed on whole-body, thorax and head to see if the direction of up-/downregulation of genes is the same in different tissues (Girardot *et al.*, 2006). The expression levels of about two-third of the genes (3,034 out of 4,503 genes) changed in the same direction in the head, thorax and whole-body data while the remaining genes change in opposite directions (Figure 6.4). Note that this data set is based on Affymetrix Drosgenome 1 and therefore not fully comparable to the other data sets used in this study. Another study (Zhan *et al.*, 2007) found 16 genes that were consistently differentially expressed with age amongst several tissues. The expression levels of all but one of these genes was upregulated in some tissues, and downregulated in other tissues (Figure 6.4).

The third explanation is that tissue-specific genes change their expression more with age. Indeed it was recently suggested that it is possible that genes with lower maximum expression levels might be changing to a larger degree with age (Hong *et al.*, 2008). Again, a tissue-specific gene expression atlas with the additional dimension of ageing added may help to validate this hypothesis.

6.2.2 Filtering tissue-specific age-associated transcripts before enrichment analysis increases the significance of age-associated gene ontology terms

The usefulness of enrichment-based analysis greatly depends on the quality of the functional annotation associated with the input genes. False positive and false negative annotation errors in the GO database can adversely affect performance. In this data set, only 60% of the differentially expressed genes could be associated with

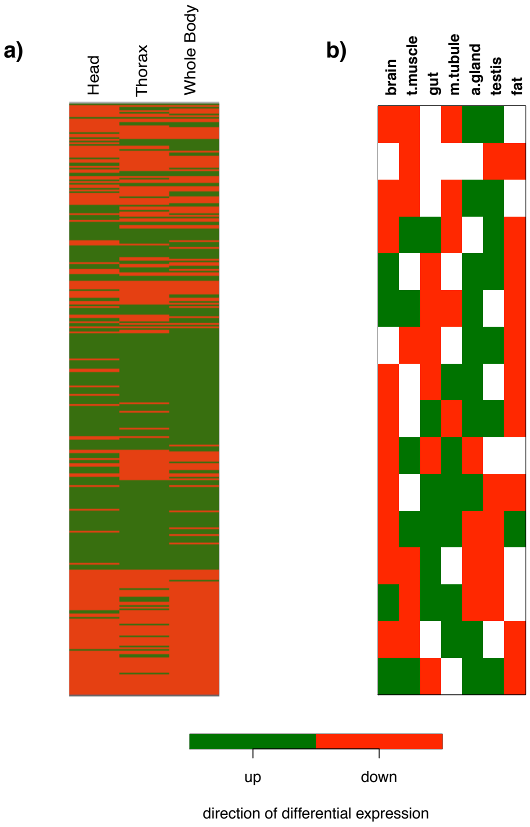


Figure 6.4: **Up-/downregulation of age-responsive genes in fruit fly body parts.** Green and red indicate up- and downregulation respectively in old flies (15-60 days old flies) compared to young flies (3 days old flies). **a)** 4,503 genes which have been identified as responsive in various ageing experiments. The figure shows the up-/downregulation of these genes in three body parts: whole body, head or thorax. **b)** 16 genes that were differentially expressed with age in at least 3 tissues according to spatial transcriptional profiles of aging in 7 tissues.

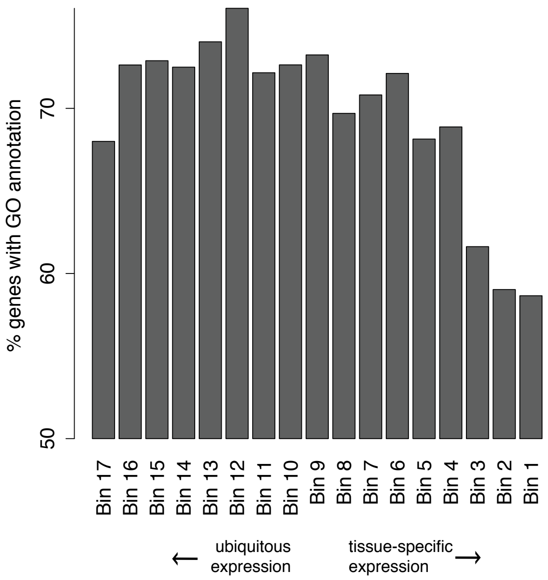


Figure 6.5: **GO annotation distribution.** Percentage of fruit fly genes in the data set that are annotated with GO annotation. The coverage is comparatively low for tissue-specific genes in bin 1.

one or more GO annotations. Tissue-specific genes are generally less well annotated than ubiquitously expressed genes (Figure 6.5).

The bias in functional annotation towards richer annotation for broadly expressed genes may have important repercussions on the conclusions drawn from an enrichment analysis study. Important connections to processes of interest, here ageing, may be missed just because of missing annotations associated with the data set. For instance, most of the 55 genes that are annotated with the term *determination of adult lifespan* in the full data set are broadly expressed (median $S_i = 0.29$) but it is unknown how many false negative annotations are associated with the data set. Removing tissue-specific genes from the data set before enrichment analysis may, thus, alter the significance of age-related terms associated with the differentially expressed genes. The results of such an analysis are reported in the following. Functional differences in tissue-specific and broadly expressed differentially expressed genes are also established.

The entire set of differentially expressed genes in *chico*^{1}/+ (819 up, 237 down) was divided into 17 groups based on their degree of tissue specificity (Figure 6.6). We then searched for over-represented GO terms (Ashburner *et al.*, 2000) in each list of genes, and combinations thereof (48 bins: *s1-s15*, *1-17*, *u1-u15*, *all*), using gene-annotation enrichment analysis. Note that only 1,070 of the 1,169 differentially expressed genes (92%) could be associated with a gene specificity value. As a result 99 genes could not be assigned to any of the filtered bins, but they are part of the *all* bin. Thus it is not surprising that some GO terms could be detected by investigating the full set of genes only. For instance, the full set of differentially expressed genes was required to detect the terms *insulin receptor binding* and *transforming growth factor beta receptor signaling pathway*. The gene *chico* that was mutated in this ageing study encodes an insulin receptor substrate that functions in an insulin/insulin-like growth factor (IGF) signaling pathway, and this accords with the over-representation of these terms.

We found 111 over-represented GO (81 up, 30 down) terms associated with one or more of the 48 groups of genes. Seven of these categories (2 up, 5 down) could only be identified using the full set of differentially expressed genes (bin *all*), but in none of the other groups (bins *s1-s15*, *1-17*, *u1-u15*). Conversely, 43 categories (11 down, 32 up) were identified in one or more of the filtered groups, but not in the full set of differentially expressed genes. Most of these categories were found after filtering tissue-specific genes (32 terms), some of them were found after removing broadly expressed genes (6 terms), and a few have been found in both of these groups (3 terms). Six GO terms could be detected if any of the 17 groups were tested alone (bins *1-17*), the other GO terms required a combination of groups of genes (bins *u1-u15*, *s1-s15*).

The upregulated terms that were significant only after removing tissue-specific genes from the set of differentially expressed genes include various terms related to metabolism (e.g. *hormone metabolic process* and *galactose metabolic process*) and oxidoreductase activity (e.g. *antioxidant activity* and *oxidoreductase activity*).

	a) Name	b) Clusters	c) Gene specificity value (S_i)		d) Number of genes				
			Min	Max	All	Down	Up	Down (FC 2)	Up (FC 2)
① Tissue-specificity increases ↑	1	1	3.59	4.09	624	3	70	0	43
	2	2	2.79	3.59	1008	19	100	2	49
	3	3	2.29	2.79	529	10	80	0	35
	4	4	1.93	2.29	498	7	48	2	14
	5	5	1.63	1.93	430	14	44	1	17
	6	6	1.39	1.63	416	9	54	3	15
	7	7	1.18	1.39	418	15	58	1	13
	8	8	1	1.18	495	15	86	1	13
	9	9	0.84	1	482	11	50	0	11
	10	10	0.7	0.84	486	10	47	0	6
	11	11	0.57	0.7	589	17	50	3	7
	12	12	0.44	0.57	685	16	41	2	6
	13	13	0.33	0.44	928	20	41	0	6
	14	14	0.23	0.33	1247	29	29	2	3
	15	15	0.13	0.23	1560	26	23	3	8
	16	16	0.04	0.13	1359	19	9	2	0
	17	17	0	0.04	50	0	0	0	0
② Tissue-specificity increases ↑	s1	1-2	2.79	4.09	1632	22	170	2	92
	s2	1-3	2.29	4.09	2161	32	250	2	127
	s3	1-4	1.93	4.09	2659	39	298	4	141
	s4	1-5	1.63	4.09	3089	53	342	5	158
	s5	1-6	1.39	4.09	3505	62	396	8	173
	s6	1-7	1.18	4.09	3923	77	454	9	186
	s7	1-8	1	4.09	4418	92	540	10	199
	s8	1-9	0.84	4.09	4900	103	590	10	210
	s9	1-10	0.7	4.09	5386	113	637	10	216
	s10	1-11	0.57	4.09	5975	130	687	13	223
	s11	1-12	0.44	4.09	6660	146	728	15	229
	s12	1-13	0.33	4.09	7588	166	769	15	235
	s13	1-14	0.23	4.09	8835	195	798	17	238
	s14	1-15	0.13	4.09	10395	221	821	20	246
	s15	1-16	0.04	4.09	11754	240	830	22	246
③ Tissue-specificity increases ↓	u1	16-17	0	0.13	1409	19	9	2	0
	u2	15-17	0	0.23	2969	45	32	5	8
	u3	14-17	0	0.33	4216	74	61	7	11
	u4	13-17	0	0.44	5144	94	102	7	17
	u5	12-17	0	0.57	5829	110	143	9	23
	u6	11-17	0	0.7	6418	127	193	12	30
	u7	10-17	0	0.84	6904	137	240	12	36
	u8	9-17	0	1	7386	148	290	12	47
	u9	8-17	0	1.18	7881	163	376	13	60
	u10	7-17	0	1.39	8299	178	434	14	73
	u11	6-17	0	1.63	8715	187	488	17	88
	u12	5-17	0	1.93	9145	201	532	18	105
	u13	4-17	0	2.29	9643	208	580	20	119
	u14	3-17	0	2.79	10172	218	660	20	154
	u15	2-17	0	3.59	11180	237	760	22	203
	all	1-17	0	4.09	11804	240	830	22	246

Figure 6.6: **Splitting 11,804 fruit fly genes into bins according to their tissue specificity values (S_i).** The genes are bined in 3 ways. **(1)** Into 17 different bins as described in the methods section using Equation 5.3. tissue specificity is highest for *bin 1* and lowest for *bin 17*. **(2)** Combining bins defined in (1) starting from the most tissue-specific genes (*s1-s15*). **(3)** Combining bins defined in (1) starting from the least tissue-specific genes (*u1-u15*). a) bin names used in the manuscript b) Indicates which bins are combined for the analysis c) The S_i cut-offs used to assemble the respective bins d) The total number of FlyAtlas genes assigned to the bins and the numbers of up- and downregulated genes in the ageing experiment studied. FC 2 indicates that a fold change cut-off of $log_2 > 1$ was used.

Ageing is known to influence certain key metabolic processes (Curtis *et al.*, 2005) and, thus, the over-representation of these terms is not surprising. The over-representation of the terms related to oxidoreductase activity (e.g. *oxidoreductase activity, acting on the CH-NH group of donors, NAD or NADP as acceptor and steroid dehyodrogenase activity*) and the term *antioxidant activity* is in agreement with previous observations that possible determinants of the enhanced life maintenance include increased resistance to oxidative stress provided by a shift to a highly reducing redox status (Houthoofd *et al.*, 2002). Some of the tissue-specific groups of genes were also associated with terms related to oxidative activity and metabolism. One mechanism involved in the defense of oxidative products is the family of glutathione transferases (Martinez-Lara *et al.*, 2003). This is reflected by an overrepresentation of the term *gluthatione transferase activity* in the upregulated set of genes, for primarily ubiquitously expressed ones. The latter term is significant for the all group and several of the bins (*all*, *u6-u13* and *u15*: $p < 0.05$).

The downregulated terms that were significant only after removing tissue-specific genes from the set of differentially expressed genes include the terms *determination of adult lifespan* (bins *u4* and *u5*: $p < 0.05$, *all*: $p = 0.148$) and *protein kinase activity.* Conversely, the terms *eggshell chorion formation*, *cell proliferation*, *skeletal muscle fiber development*, *response to other organism* were significant after removing broadly expressed genes.

In addition to the detection of age-associated GO terms after filtering tissue-specific and/or broadly expressed genes, the significance of some terms was increased after filtering. We were able to associate the term *response to starvation* with the downregulated set of primarily broadly expressed genes (*all*, *s13-s15* and *u4-u15*: $p < 0.007$; *u4-u15* < *all*). It is known that nutritional factors can exert major effects on ageing and interact with experimentally induced mutations that induce longevity. There is evidence that the *chico* mutation, analysed in this work, is involved in the same mechanisms as dietary restriction (Piper and Bartke, 2008), and this is in agreement with the overrepresentation of this term. In the fruit fly, starvation promotes the mobilisation of glycogen and lipid stores in response to increases in circulating adipokinetic hormone (Djawdan *et al.*, 1998; Kim and Rulifson, 2004). In our data set, this is reflected by an upregulation of monosaccharide, polysaccharide, lipid and carbohydrate biosyntetic processes (e.g. *trehalose biosynthetic process*, *lipid transporter activity* and *carbohydrate phosphatase activity*). Again, the significance of some of these terms was increased after filtering tissue-specific genes, although they were significant in the full set of differentially expressed genes too.

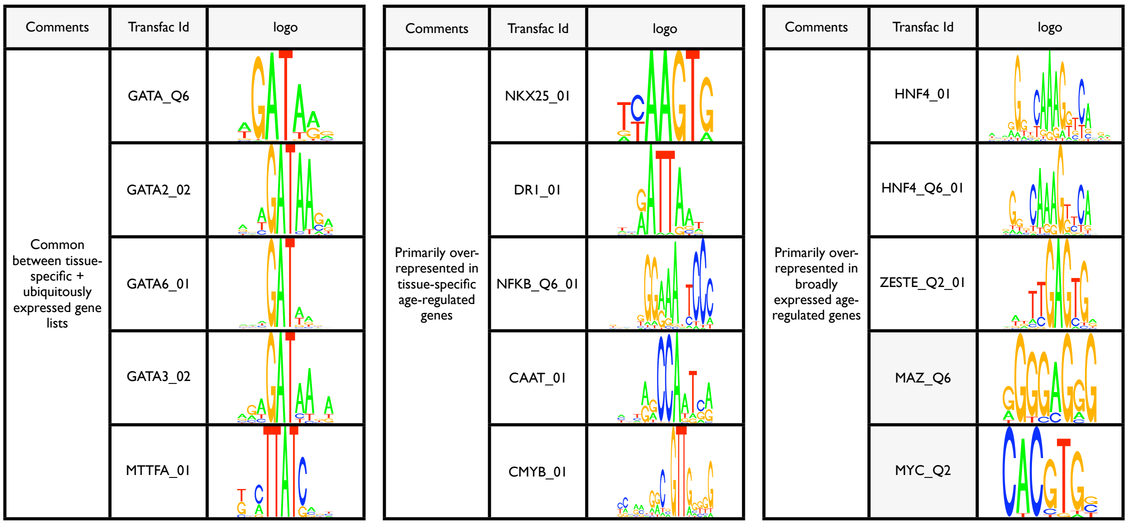
Thus our specificity analysis reveals functional bins which have previously been implicated in ageing encouraging us to believe that this approach is useful and powerful. We conclude that filtering tissue-specific genes prior to annotation enrichment analysis helps to find more subtle connections to ageing. The tissue-specific and ubiquitous bins represent different and complementary sets of genes, which can and should be studied for their expression during the ageing process taken together and separately. Note that the tissue-specific differentially expressed genes are predominantly midgut specific genes (68% for genes with an S_i value > 2.80). Thus, the GO terms over-represented in the tissue-specific groups are dominated by midgut specific terms.

It should be noted that in the approach used to detect significant GO terms each term was statistically tested independently. An important issue which arises is the effect of multiple testing on power. Each time we statistically test a term with a statistical test, we incur the risk of a false positive. It is standard practice in bioinformatics to use a p value threshold of 0.05 for the decision as to whether a term is significant or not. This p value is the probability of getting a false positive result, so on average we would expect to get a false positive result about once every 20 times the test is used (1/0.05). In the above experiment we tested 2,439 GO terms. This translates to 122 false positives (0.05 * 2,439 tests). The total number of terms that were significant was 333 (before parent nodes were removed from the result list).

It has to be assumed that approximately one third of the results are false positives, even though this calculation is conservative, as it ignores correlations between genes. Choosing a stricter EASE score cut-off to select significant GO terms might yield a lower false positive result. We used a relatively tolerant EASE score because we did not want to miss terms that have potential association to ageing. To further investigate multiple testing issues we used Storey's q value calculation to formally assess the false discovery rate (FDR) using the corresponding R package (Dabney *et al.*, 2010). The software, which we used with default values, takes a list of p values resulting from the simultaneous testing of many hypotheses and estimates their q values by using characteristics of the p value distribution. The q values, for the 111 GO terms that were deemed to be significant earlier, are reported in Tables C.1 and C.2 (Appendix C). They were first estimated from the EASE scores, and then also from the original Fisher p values. Choosing a q value of 0.05 means that we should expect 5% of all the terms with q value less than this to be false positives. Of the list presented in Appendix C we find 33 such GO terms if the q values are estimated from the EASE scores, while we find 77 if the q values are estimated from the Fisher p values. These are reported in Tables D.1 and D.2 (Appendix D). We expect 33*0.05= 1.65 and 77*0.05 = 3.85 false positives, respectively. When deciding on a cut-off or threshold value, we can now do this from the point of view of how many false positives will this result have rather than choosing an arbitrary cut-off.

Another possibility to address the multiple testing issues here would be to reduce the number of GO categories tested. A slim version of GO could be created that is specific to ageing and contains a reduced number of categories. Another possibility would be to exclude highly over- or underrepresented terms in the background gene set.

6.2.3 DNA motifs associated with tissue-specific and broadly expressed age-associated genes



Comments	Transfac Id	logo
Common between tissue-specific + ubiquitously expressed gene lists	GATA_Q6	
	GATA2_02	
	GATA6_01	
	GATA3_02	
	MTTFA_01	

Comments	Transfac Id	logo
Primarily over-represented in tissue-specific age-regulated genes	NKX25_01	
	DR1_01	
	NFKB_Q6_01	
	CAAT_01	
	CMYB_01	

Comments	Transfac Id	logo
Primarily over-represented in broadly expressed age-regulated genes	HNF4_01	
	HNF4_Q6_01	
	ZESTE_Q2_01	
	MAZ_Q6	
	MYC_Q2	

Figure 6.7: **Overrepresented TFBS.** Transcription factor binding sites, which are significantly over-represented in genes showing differential expression between wild type and long-lived fruit fly mutant. Several GATA regulatory motifs are enriched in the promoter regions of age-regulated genes for both tissue-specific and broadly expressed genes. Other motifs are primarily found in promoter regions of age-regulated genes that are tissue-specific (e.g. NFKB_Q6_01) or broadly expressed (e.g. MYC_Q2).

Differentially expressed genes often share a number of TFBS located upstream of the transcriptional start site, implying that they might be regulated by similar transcription factors that regulate a group of genes involved in a similar cellular function. In the previous section, it was shown that differentially expressed genes identified on whole-fly experiments capture both common and tissue-specific responses to ageing. These may or may not involve the same transcription regulatory machinery. We applied `Clover` (Frith *et al.*, 2004), a program for identifying functional sites in DNA sequences, to the promoter sequences of genes associated with extension of lifespan. As before the entire set of differentially expressed genes was divided into 17 groups based on their degree of tissue specificity. We then searched for over-represented functional sites in each list of genes, and combinations thereof. A total of 90 TFBS were identified. Of these, one TFBS could only be detected using the full set of differentially expressed genes (up: HNF) while 56 additional potential TFBS were detected using the filtered sets of genes. For instance FOXO1_01 was identified in ubiquitously expressed down-regulated genes (*u2*, *u3*) but not in the full set of down-regulated genes. This is interesting because the FOXO transcription factor is implicated in animal ageing (Greer and Brunet, 2008).

The data further suggests that tissue-specific and broadly expressed age-modulated genes share some TFBS (Figure 6.7). For example processes involving the GATA family of transcription factors appear to be shared across genes including all levels of tissue specificity. In flies GATA factors have a central role in heart specification (Qian and Bodmer, 2009), and they are also implicated in insect innate immune response (Bettencourt and Ip, 2004). In the worm several GATA transcription factors were found to be responsible for age regulation of several genes (Budovskaya *et al.*, 2008). Our data shows that the GATA transcription factors can be linked to age regulation of tissue-specific and broadly expressed genes in the fruit fly. An example for a TFBS that was primarily associated with the tissue-specific age-regulated genes is NFKB_Q6_01. The NF-kB system is the master regulator of the innate immunity, an ancient signaling pathway found in both insects and vertebrates. Recent studies have revealed that several key regulators of aging in *budding yeast* and *C. elegans* models, regulate the efficiency of NF-kB signaling and the level of inflammatory responses (Salminen *et al.*, 2008). Our data suggests a tissue-specific involvement of the NF-kB regulator in ageing for the adult fruit fly. A TFBS for MYC, which has been linked to ageing-related genes (Grandori *et al.*, 2003; Wu *et al.*, 1999) in human, was primarily found in the set of broadly expressed genes (down-regulation).

6.2.4 Age-associated genes

A total of 46 lifespan extending mutations are known for the fruit fly so far (de Magalhães and Toussaint, 2004). Little is known about the overall tissue specificity of these genes, i.e., whether these genes primarily perform housekeeping or tissue-specific functions. To address this question we scrutinised the ordered set of fruit fly genes for 46 fruit fly genes previously highlighted to be lifespan extending. Figure 6.1, Box 4, shows that most of these genes are ubiquitously expressed. The gene with the highest gene specificity value was *FBgn0029752* (Thioredoxin T, S_i=3.88), an evolutionarily conserved antioxidant and molecular chaperone whose over-expression in neurons was associated with lifespan extension of 15% in the fruit fly (Umeda-Kameyama *et al.*, 2007). This gene is highly expressed in the testis, but in none of the other adult tissues provided with the current FlyAtlas data set. An example of a gene with a midrange pattern of expression is *FBgn0037324* (Odorant receptor 83b, S_i=1.33). Loss-of-function mutation of *FBgn0037324* that is expressed in the brain, head and eye resulted in olfactory defects, altered adult metabolism, enhanced stress resistance and life-extension of up to 56% in the fruit fly (Libert *et al.*, 2007). The gene with the most ubiquitous expression was *FBgn0086768* (Protein-L-isoaspartate (D-aspartate) O-methyltransferase). This gene is involved in protein repair mechanisms. Overexpression of this gene extended lifespan in flies by 32-39% under certain conditions (Chavous *et al.*, 2001).

We found that most currently known age-associated genes in the fruit fly are ubiquitously expressed (median S_i=0.28). It is possible, however, that longevity-associated pathways involve far more tissue-specific genes than Figure 6.1 indicates since tissue-specific genes are less well studied (Figure 6.5). An investigation of age-associated genes in several fruit fly body parts (Zhan *et al.*, 2007) also showed that most of the genes are broadly expressed, but that some of the genes are highly specific (Figure 6.1, Boxes 5-11). The authors of the study state that only 3% to 10% of age-related genes in any given tissue overlapped with those in any other tissue. This lack of overlap across genelists is partly explained by the tissue-specific age-related genes in each tissue, that by definition have low chances to overlap with the genes from another tissue. Again, the ordered list of fruit fly genes according to their tissue specificity could be used to remove genes with a certain tissue specificity from the tissue lists.

Chapter 7

Final Remarks and Future Work

7.1 Conclusions

In this thesis I described my investigations of applying machine learning methods to high throughput experimental and predicted biological data. This work made three novel contributions based on the systematic analysis of publicly archived data of protein sequences, three dimensional structures, gene expression and functional annotations: (a) remote homology detection based on amino acid sequences and secondary structures; (b) the analysis of tissue-specific gene expression for predictive signals in the sequence and secondary structure of the resulting protein product; and (c) a study of ageing in the fruit fly, a commonly used model organism, in which tissue specific and whole-organism gene expression changes are contrasted. The conclusions of these studies are summarised in the following, and future directions are given subsequently.

7.1.1 Remote homology detection using a kernel method that combines sequence and secondary structure similarity scores

In the first part of this thesis, we have developed a kernel-based remote homology detection method that allows for a combination of sequence and secondary structure similarity scores. We studied its performance to predict superfamily membership as defined by the SCOP database. We showed that a kernel method that combines sequence similarity scores with predicted secondary structure similarity scores performs similar to a classifier that uses scores calculated from sequences and true secondary structures, but performs better than a sequence-only based classifier and achieved a better mean than recently published results on the same data-set i.e. the *GPkernel* which in turn was compared to other methods such as *SVM-pairwise* method, the mismatch kernel and a PSI-BLAST based approach. Our method can be tuned to re-weight the influence of the scores, and it is widely applicable because alignment scores and secondary structures are readily computable. We note several important points about this work. First of all, we note that the observation that secondary structures provide complementary information to amino acid sequence is not new; in fact, as stated in the chapter 3.1, this has been shown by many other researchers. The difference to most other works lies in the use of SVMs and kernels thus investigating performance in a discriminative setting in contrast to instance-based and generative models - which has not yet been done in a comprehensive study. Even though it is known that secondary structures can improve remote homology detection methods, many recent methods do not use this kind of information. Thus, we hope this work gives a refreshing view on this issue and will encourage others to integrate secondary structures in their methods. Further we note that we have used the SCOP database, driven by the need to compare our results with previously published results. The manual annotations in SCOP, particularly at the higher levels of the hierarchy, do indeed use knowledge of secondary structures as an important source of information. Secondary structure assignment algorithms are all trained on the relatively small set of proteins for which structures have been determined, and SCOP domains are a subset of these. While the inevitable bias arising from these facts does not negate our conclusion, caution must be exercised in how far one might generalise our findings. The performance increases we see are small, and, in problems posed on the SCOP database we are operating at very high levels of accuracy. Still, for comparisons against other work and for reliabilities of annotation we needed to work with this database. In future work, we suggest to move away from this and to formulate sequence classification problems using other databases. Finally we note that the SVMs used in this work were trained to solve two-class problems. To be able to classify new proteins, which are not part of the benchmark sets, the method needs to be extended to solve multi-class problems. One of the most widely used approaches to solve multi-class problems is the one-against-all classification, in which a new instance is tested against all binary SVMs (102 for the SCOP version used here), and the classifier which outputs the largest score is chosen. In this work our intention was to explore if secondary structure inclusion can increase classifier performance - to establish this, in comparison to already published work, we limited the experiments to solving two-class problems, and performance of new instances has yet to be determined.

7.1.2 Tissue specificity of gene expression is correlated with the sequence and secondary structure of resulting protein product

The second part of this thesis concerns an investigation of the predictability of gene specificity based on the amino acid content associated with gene protein products, their predicted secondary structures and various genomic features in the fruit fly *Drosophila melanogaster* and the mouse *Mus Musculus*. Gene specificity can be predicted at better than random rates using all classifiers tested on most benchmark sets suggesting the existence of useful signals at these levels. The classifier based on amino acid percentages combined with genomic features performed best overall. It also compared favourably with a classifier previously published (De Ferrari and Aitken, 2006). We conclude that tissue specificity of gene expression is correlated with the sequence and secondary structure of the resulting protein product. Shannon's information theory provides a clearly defined statistical framework that has previously been proven valuable in several genomic applications. Here, we applied it to mouse and fruit fly expression profiles to obtain lists of genes ordered according to their tissue specificity and used it to investigate tissue specificity in these organisms. We concentrated on the fruit fly because it is the least studied organism (Table 3.1) for which good data have recently become available. Further studies are required to investigate if there is an evolutionary conserved correlation of amino acid sequences and tissue specificity, and if such a correlation could be used to predict this information for less well studied organisms. We found that the amino acid asparagine discriminated best between broadly expressed and tissue specific genes in both the fruit fly and the mouse and it would be interesting to follow up this finding in future studies. We showed that a classifier trained on mouse and tested on fly genes, or vice versa, performed better than random on a number of benchmark sets. The number of genes used in these benchmarks can be systematically optimised to further improve performance. There are several other areas for improvement and future research to extend this work. Recent technological advances that allow faster and cheaper DNA sequencing and transcriptional profiling (RNA-Seq) are likely to produce high-quality data that could be used to refine many of the approaches used in this work (Wang *et al.*, 2009c). For instance, RNA-Seq provides a better estimate of absolute expression level (Fu *et al.*, 2009) that in turn give a better estimate of true tissue-specificity. This thesis has taken some steps in the development of a computational method which can serve as a stable platform for further research on tissue-specificity when large scale RNA-Seq data becomes available.

7.1.3 Analysis of the tissue-specific contribution to whole body RNA transcript profiles in *Drosophila Melanogaster*

Finally we used the same computational approach as above to partition genes according to their tissue specificity in the fruit fly and used it to clarify tissue-specific fly transcripts and gene expression in the ageing fly, and in general. Based on an information theoretical approach, we investigated how to utilise FlyAtlas, a microarray-based atlas of gene expression in multiple adult tissues, to delineate tissue-specific from ubiquitous expression in whole-fly experiments. We began by taking the sorted list of fruit fly genes according to their degree of tissue specificity introduced in the previous chapter, obtained from the FlyAtlas gene expression profiles. We then used the defined tissue specificity to determine the capacity of Affymetrix high-density oligonucleotide whole-genome microarrays to capture tissue-specific age-associated changes in whole-fly samples. Importantly, we found that genes with tissue-specific expression are associated with higher fold changes amongst significantly differentially expressed genes and a lower mean expression signal. This indicates that changes in tissue-specific expression might be easier to detect than expression changes of broadly expressed genes when using whole-fly arrays. We also described how filtering genes with tissue-specific expression from data from a whole-fly ageing experiment affects data analysis and the derivation of meaningful information from the data. The significance of several age-related GO terms was increased after removing tissue-specific differentially expressed genes. This is due to a bias in GO annotation towards broadly expressed genes, and to differences in function of broadly and tissue-specifically expressed genes. This study was complemented by an analysis of the tissue specificity of age-associated genes in the fly. We found that most known age-associated genes are broadly expressed. As before, the future availability of RNA-Seq data is expected to be useful to validate some of the results of this work.

7.2 Future directions

Ideas for extensions of this work are given in the following paragraphs.

7.2.1 Joint alignments

In Chapters 4 and 5 we used a joint sequence and secondary structure approach to predict remote homology and tissue specificity. Future work could include the development of a method that better captures the relationship between sequence and secondary structure. A joint alignment between these two entities might be a step in this direction. Several previous studies have examined the improvement of pairwise sequence alignments by incorporating secondary structure information. In particular, it has been demonstrated that sequence alignments can be improved by limiting the number of gaps in the regions of secondary structures (Barton and Sternberg, 1987; Gerstein and Levitt, 1996; Lesk *et al.*, 1986). In these studies it has been shown that positions in an alignment that correspond to α-helices or β-strands are less likely to be affected by gaps. These studies demonstrated on overall improvement in alignment accuracy by limiting the number of gaps in regions of secondary structures, but they focused on a small number of model proteins. For instance, Barton and Sternberg (1987) considered five pairs of structurally homologous proteins, while Lesk *et al.* (1986) looked at proteins within the globin and serine proteinase families. Gerstein and Levitt (1996) also investigated a small number of proteins.

Multiple sequence alignments have also been shown to be improved by making use of secondary structures (Elofsson, 2002). For example, von Ohsen *et al.* (2004) combine amino acid profile-profile scores with weighted secondary structure profile-

profile scores to compute the final alignment score. They build a frequency profile for the target sequence over the amino acid alphabet and over the three-state secondary structure alphabet using the PSIPRED program. A secondary structure similarity matrix (Kawabata and Nishikawa, 2000) is used to compute the secondary structure profile-profile scoring term while the BLOSUM62 matrix is used to compute the amino acid profile-profile score. The authors use this profile–profile alignment approach in their software termed *Arby* which is a server for protein structure prediction based on its sequence. According to the CAFASP3 experiment (Critical Assessment of Fully Automated Structure Prediction), the server is one of the most sensitive methods for predicting the structure of single domain proteins. Chung and Yona (2004) also showed that integration of primary and secondary structure information can substantially improve detection of relationships between remotely related protein families. Their method augments sequence profile columns using PSIPRED secondary structure predictions and assesses statistical similarity using information theoretical principles.

Another example where a similar method is used is PRALINE which is a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information (Simossis and Heringa, 2005). PRALINE makes a profile-profile alignment with PSI-BLAST profiles used as templates. The profile can be complemented with a secondary structure prediction in an attempt to improve the alignment accuracy. A choice of seven different secondary structure prediction programs is provided that can be used individually or in combination as a consensus for integrating structural information into the alignment process. A different scoring scheme is used for profile positions with matching secondary structure elements than for positions that show mismatching residues. The authors report that the use of the secondary structure information significantly improves the PRALINE alignment quality.

The above mentioned methods make use of secondary structures to guide the alignments. Alignment methods that integrate three-dimensional structure have also been developed. For instance, the 3D-Coffee (Poirot *et al.*, 2004) method, or its newer version named Expresso (Armougom *et al.*, 2006) make use of PDB structures to assemble a structure-based multiple sequence alignment. Providing the appropriate structural information is available, Expresso is significantly more accurate than regular homology based methods and its alignments are often indistinguishable from reference structure based alignments. FUGUE (Shi *et al.*, 2001) is a program for recognising distant homologues by sequence-structure comparison. It utilises environment-specific substitution tables and structure-dependent gap penalties, where scores for amino acid matching and insertions or deletions are evaluated depending on the local environment of each amino acid residue in a known structure. Given a query sequence, FUGUE scans a database of structural profiles, calculates the sequence-structure compatibility scores and produces a list of potential homologues and alignments. JOY (Mizuguchi *et al.*, 1998a) is a program for annotating protein sequence alignments with three dimensional structural features. It was developed to help understand the conservation of amino acids in their specific local environments. HOMSTRAD (HOMologous STRucture Alignment Database) (de Bakker *et al.*, 2001; Mizuguchi *et al.*, 1998b; Stebbings and Mizuguchi, 2004) is a database of multiple alignments, created using the three dimensional structure as a guide (Mizuguchi *et al.*, 1998a). The alignments are annotated with JOY in a format that represents the local structural environment of each amino acid residue.

Even though some automatic methods are available that incorporate secondary structure information in alignments, the most frequently used alignment methods do not use this kind of information. We think there might be further improvements to the works mentioned above, and hence we started to investigate how weighted finite state machines (WFSMs) could be used to create such a method.

In the following a possible approach is detailed that could be used to generate such an alignment. The approach was developed during the course of this PhD study, but since a complete study is missing, this is given as an idea for future directions rather than an own chapter.

WSFMs, or more specifically weighted finite-state acceptors and transducers, are generic devices for modelling sequences of symbols. On an abstract level, acceptors represent a sequence of input symbols, while transducers encode a mapping between input and output sequences. Weights such as match or mismatch probability can be assigned to each transition. Regular expressions and HMMs, which are used for a wide range of applications in bioinformatics, are specialised cases of WFSMs. In this preliminary study, WFSM were trained to perform pairwise sequence alignments guided by secondary structure knowledge. The AT&T FSM library was used (Mehryar, 1997), which has been successfully applied to many natural language-processing tasks (Mohri *et al.*, 2002) and also to some bioinformatics problems (Cortes *et al.*, 2004). A conclusive, peer-reviewed publication on this software package in sequence analysis applications is, to the best of our knowledge, not available. The library and its potential to model sequence related algorithms was introduced in the tutorial 'Weighted Finite-State Transducers in Computational Biology' at the 13th Annual International Conference on Intelligent Systems for Molecular Biology (Cortes and Mohri, 2005).

WFSM were used to calculate an alignment which allows for the combination of the sequences and their corresponding secondary structures. Specifically we used acceptors to model sequences and secondary structures. We used transducers to define weights from one amino acid to another and from amino acids to secondary structure elements. Figure 7.1a, demonstrates how a standard sequence alignment can be calculated using WFSMs. The user provides two input sequences and weights for amino acid matches, mismatches and gaps. These inputs are represented by two unweighted acceptors (S1, S2) and one weighted transducer (T). Their graphical representation is shown in Figure 7.1. A high weight in T indicates that this transition is likely to occur and vice versa. For example, a transition from *A:A (-0.13)* is more likely than the introduction of a gap, *A:e (-3)*. In order to calculate an alignment, the above FSMs are composed. Composing FSMs means taking the output from the one FSM and to match it with the input of another FSM. For example, the first output of (S1) is an *A*. It matches all three inputs of (T), i.e. *A:A, A:S, A:V*. The resulting transition consists of the input of the first FSM and the output of the second FSM; i.e. composing of (S1) and (T) results in transitions: *A:A, A:S, A:V*. The resulting, temporary FSM, represents all transitions between the given sequence and an unknown sequence, which are, in principle, possible. This intermediate FSM, is further composed with (S2). The same procedure is done for all the other states. *A:e* introduces a gap in S1. Gaps can be introduced in S2 as well, but the example is greatly simplified. The resulting WFSM defines all possible paths between the two sequences. The path with the highest weight, is the most likely one and it is selected as the alignment.

Figure 7.1b illustrates how the same principles can be used to calculate an integrated alignment. The algorithm stays the same, the difference lies in the input that the user gives to the program. More precisely six WFSMs are composed in the following order

$$S_1 \circ T_1 \circ SS_1 \circ T_2 \circ SS_2 \circ T_3 \circ T_4 \circ S_2 \tag{7.1}$$

S_1 and S_2 in the equation above represent the amino acid sequence of protein 1 and 2 respectively. SS_1 and SS_2 represent the secondary structure sequence of protein 1 and 2 respectively. T_1 defines transition weights from an amino acid symbol to a secondary structure symbol. T_2 represents transition weights from one secondary structure element to another. T_3 defines transition weights from a secondary structure element to an amino acid. T_4 defines transition weights between amino acids. The user provides sequence and secondary structure, as well as four files which define transitions probabilities or weights between amino acids and secondary structure. If a lot of weight is given to matching secondary structure, the resulting alignment is different to the alignment shown in Figure 7.1a.

Figure 7.2 shows an real world example of such an integrated alignment.

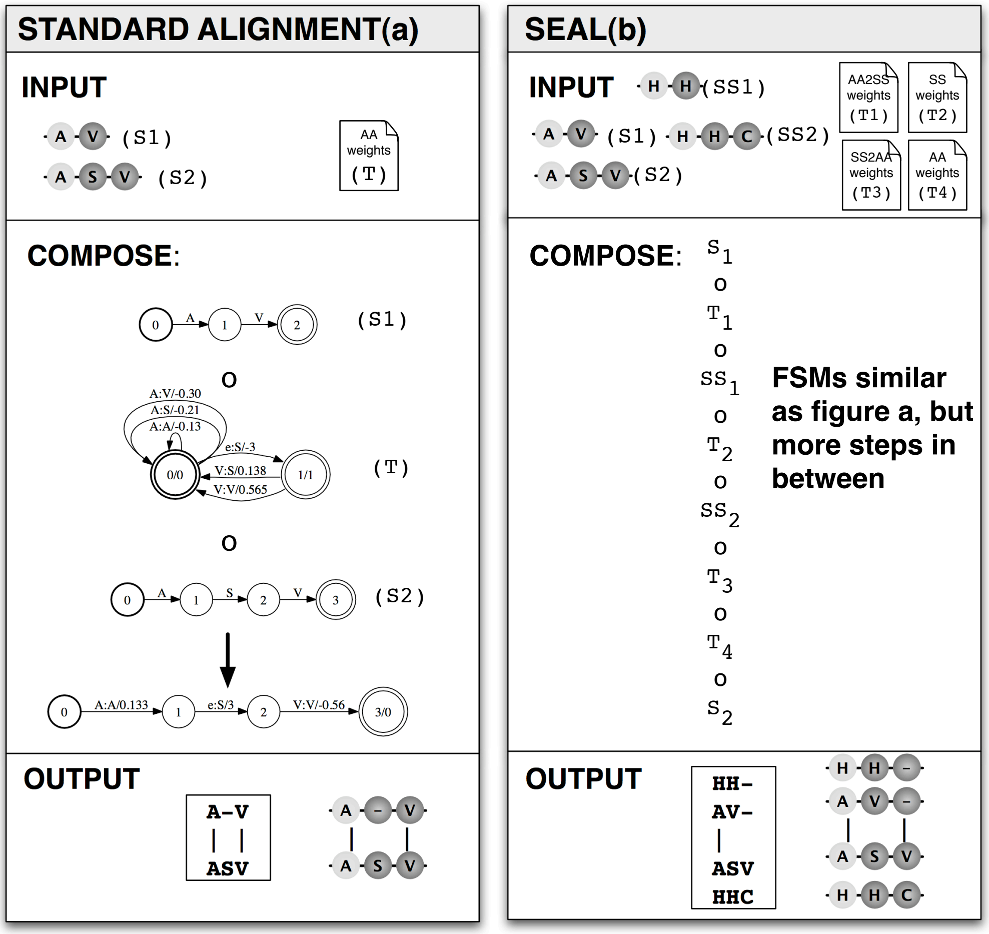


Figure 7.1: **Schematic view of weighted finite state machines used for pairwise alignments.** The sequences S1, S2, SS1 and SS2 are modeled as acceptors. The "weight" files in the figure define transition probabilities between one amino acid residue or secondary structure symbol to the next. Weighted transducers (T) are used to connect two sequences depending on the defined transition probabilities. **a)** WFSMs are used to align two sequences S1 and S2. **b)** Joint sequence and secondary structure alignment. WFSMs are used to align two sequences S1 and S2 and two secondary structure sequences SS1 and SS2. The method was named SEAL.

7.2.2 Predicting tissue specificity for other organisms

Considering the success of discriminating tissue-specific and broadly expressed genes within organism, future work might include the prediction of tissue specificity of genes

```
a) Global Sequence Alignment (EMBOSS)

P02186      ---------LLLHHHHHHHHHHHHHHHTTHHHHHHHHHHHHHHHLGGGGG
            ---------GLSDGEWELVLKTWGKVEADIPGHGETVFVRLFTGHPETLE
                      | ||      | | |              |     |
Q7SID0      PIIDQGPLPTLTDGDKKAINKIWPKIYKEYEQYSLNILLRFLKCFPQAQA
            LLLLSSSLLLLLHHHHHHHHHHHHHHHTTHHHHHHHHHHHHHHHLGGGGG

P02186      GLTTTTTLLSHHHHHTLHHHHHHHHHHHHHHHHHHHTTTLLH---HHHHH
            KFDKFKHLKTEGEMKASEDLKKQGVTVLTALGGILKKKGHHE---AEIQP
             | ||   |           | | |        |       |
Q7SID0      SFPKFSTKKS--NLEQDPEVKHQAVVIFNKVNEIINSMDNQEEIIKSLKD
            GLTTTTTLLS--LGGGLHHHHHHHHHHHHHHHHHHTTTTLHHHHHHHHHH

P02186      HHHHHHHTSLLLHHHHHHHHHHHHHHHHHHSTTTSLHHHHHHHHHHHHHH
            LAQSHATKHKIPIKYLEFISDAIIHVLQSKHPAEFGADAQGAMKKALELF
            | | | |  |         |               ||           |
Q7SID0      LSQKHKTVFKVDSIWFKELSSIFVSTIDG------GAEFEKLFSIICILL
            HHHHHHHTSLLLTTHHHHHHHHHHHHTTL------LHHHHHHHHHHHHHH

P02186      HHHHHHHHHHTTSLL
            RNDIAAKYKELGFQG
            |
Q7SID0      RSAY-----------
            HTTL-----------

b) Sequence/Secondary Structure Alignment

P02186      L---------LLHHHHHHHHHHHHHHHTTHHHHHHHHHHHHHHHLGGGGG
            G---------LSDGEWELVLKTWGKVEADIPGHGETVFVRLFTGHPETLE
                      | ||      | | |              |     |
Q7SID0      PIIDQGPLPTLTDGDKKAINKIWPKIYKEYEQYSLNILLRFLKCFPQAQA
            LLLLSSSLLLLLHHHHHHHHHHHHHHHTTHHHHHHHHHHHHHHHLGGGGG


P02186      GLTTTTTLLSHHHHHTLHHHHHHHHHHHHHHHHHHHTTTL---LHHHHHH
            KFDKFKHLKTEGEMKASEDLKKQGVTVLTALGGILKKKGH---HEAEIQP
             | ||   |   |       | | |        |
Q7SID0      SFPKFSTKKSNLE--QDPEVKHQAVVIFNKVNEIINSMDNQEEIIKSLKD
            GLTTTTTLLSLGG--GLHHHHHHHHHHHHHHHHHHTTTTLHHHHHHHHHH


P02186      HHHHHHHTSLLLHHHHHHHHHHHHHHHHHHSTTTSLHHHHHHHHHHHHHH
            LAQSHATKHKIPIKYLEFISDAIIHVLQSKHPAEFGADAQGAMKKALELF
            | | | |  |         |               ||
Q7SID0      LSQKHKTVFKVDSIWFKELSSIFVST------IDGGAE---------FEK
            HHHHHHHTSLLLTTHHHHHHHHHHHH------TTLLHH---------HHH


P02186      HHHHHHHHHHTTSLL
            RNDIAAKYKELGFQG
               |
Q7SID0      LFSIICILLRSA--Y
            HHHHHHHHHHTT--L
```

Figure 7.2: **Sequence alignment alone and augmented with secondary structure.** First graphic shows a standard sequence alignment. A secondary structure element is attributed to each amino acid. The grey boxes highlight areas where there is a mismatch of secondary structure elements. The algorithm employed to create the second graphic uses secondary structure information. If we search in UniProtKB for P02186 (MYG_ELEMA) for relevant hits using BLAST, we could not find Q7SID0 (GLBF1_EPTBU) within a significant E value 10^{-05} despite high simiarity in their secondary structure sequence (see Chapter 3.1 where these proteins are discussed further). The pipe symbols indicate an alignment match between two columns.

in other model organisms. For instance the worm, *C. elegans*, is a popular model organism, but tissue-specific information is only available for part of its genome. Despite the variety of techniques available and the number of studies performed thus far, our understanding of tissue-specific expression in *C. elegans* is not yet complete; most genes have not been analysed at the single-gene level, nor under diverse conditions and developmental stages (Chikina *et al.*, 2009). The fruit fly and mouse models trained in this work could potentially be used to infer tissue specificity for these worm data. In Chapter 5 we described how the amino acid asparagine was the best discriminator between tissue-specific and broadly expressed genes. It would be interesting to investigate this in other organisms, to see this is an evolutionary conserved signal.

7.2.3 Multi-view learning

In Chapter 4 and 5, prediction models have been built that include all the variables available, without taking into consideration that the data sets were comprised of multiple feature sets from diverse domains often referred to as *views*. Consider the collection of protein domains belonging to a particular superfamily used in Chapter 4. The available information about the protein domains can be organised in the following two views: the sequence alignment scores and the secondary structure alignment scores. It is of great interest to develop a model that provides insight into the underlying relationship amongst these two views, potentially identifying interactions between them, and also to assess their predictive capabilities. In Chapter 4, we investigated how giving different weights to the two views affects the classifier performance that allowed to investigate how complementary sequence and secondary structure information were in this problem. A technique that can further help to answer the above questions, and also improve predictive performance, is multi-view learning. Multi-view learning methods have been shown to be advantagous to learning with only a single view (Blum and Mitchell, 1998), especially in cases were the weakness of one view complements the strength of the other. Multi-view learning methods exploit view redundancy to learn from partially labeled data. In the multi-view learning paradigm, the input variable is partitioned into two different views X1 and X2 and there is a target variable Y of interest. The underlying assumption is that either view alone is sufficient to predict the target Y accurately. This provides a natural semi-supervised learning setting in which unlabeled data can be used to eliminate hypothesis from either view, whose predictions tend to disagree with predictions based on the other view. Multi-view learning has been applied in bioinformatics (Culp *et al.*, 2009; Scheffer and Krogel, 2004; Yamanishi *et al.*, 2004), and could be exploited on the problems of remote homology detection and prediction of tissue specificity to both increase accuracy and gain a better understanding of the interrelationships of the data.

Appendix A

PSI-BLAST output (a.1.1.2)

In the following we report the results of the PSI-BLAST experiment discussed in Chapter 4.2.5 for all the 68 positive test sequences in the benchmark set for the SCOP family a.1.1.2. The lines starting with Query= indicate which family member was used for the PSI-BLAST search. Subsequent lines show all PSI-BLAST hits to SCOP domains outside the query family a.1.1.2. The numbers at the end of the lines represent the PSI-BLAST scores and E values, respectively. For the evaluation in Chapter 4.2.5, only the first (best) hit were considered. For the test sequences below a total of 66 true positives and a total of 2 negatives were counted if all hits are considered independent of the E value.

```
Query=  d1jl7a_ a.1.1.2 (A:) Glycera globin (Marine bloodworm (Glyceradibranchiata))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.0    0.17
lcl|d1h10a_  d1h10a_ b.55.1.1 (A:) Rac-alpha serine/threonine kin...  27.6    8.8
Query=  d1vhba_ a.1.1.2 (A:) Bacterial dimeric hemoglobin (Vitreoscillastercoraria)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.1    0.093
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  27.9    6.2
Query=  d2gdm__ a.1.1.2 (-) Leghemoglobin Yellow lupin (Lupinus luteus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  42.7    2e-04
Query=  d2hbg__ a.1.1.2 (-) Glycera globin (Marine bloodworm (Glyceradibranchiata))
Query=  d1gcvb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Houndshark (Mustelusgriseus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  30.6    1.1
Query=  d1cg5b_ a.1.1.2 (B:) Hemoglobin, beta-chain (Cartilaginous fishakaei (Dasyatis akajei))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.2    0.35
Query=  d1a4fa_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Bar-headed goose(Anser indicus))
```

```
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.4    0.15
Query=  d1spgb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Teleost fish(Leiostomus xanthurus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  30.1    1.7
Query=  d1hlm__ a.1.1.2 (-) Hemoglobin, different isoforms (Sea cucumber(Caudina (Molpadia) arenicola))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  39.3    0.003
lcl|d1h10a_  d1h10a_ b.55.1.1 (A:) Rac-alpha serine/threonine kin...  27.7    7.8
Query=  d1or4a_ a.1.1.2 (A:) Heme-based aerotactic transducer HemAT, sensordomain (Bacillus subtilis)
lcl|d1ngka_  d1ngka_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  60.3    1e-09
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  57.6    8e-09
lcl|d1idra_  d1idra_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  53.3    2e-07
lcl|d1dlwa_  d1dlwa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  47.9    8e-06
lcl|d1dlya_  d1dlya_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  46.0    3e-05
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.1    0.43
Query=  d1g08a_ a.1.1.2 (A:) Hemoglobin, alpha-chain Cow (Bos taurus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  28.4    4.9
Query=  d1d8ua_ a.1.1.2 (A:) Non-symbiotic plant hemoglobin (Rice (Oryzasativa))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  41.2    7e-04
Query=  d1mbs__ a.1.1.2 (-) Myoglobin Common seal (Phoca vitulina)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  38.1    0.007
Query=  d1h97a_ a.1.1.2 (A:) Trematode hemoglobin/myoglobin (Paramphistomumepiclitum)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  45.3    4e-05
lcl|d1ql3a_  d1ql3a_ a.3.1.1 (A:) Cytochrome c552 (Paracoccus den...  32.6    0.24
Query=  d1mba__ a.1.1.2 (-) Myoglobin Sea hare (Aplysia limacina)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  42.3    3e-04
Query=  d1hdsa_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Deer (Odocoileusvirginianus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.6    0.22
Query=  d1a4fb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Bar-headed goose(Anser indicus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.7    0.26
Query=  d1a9we_ a.1.1.2 (E:) Hemoglobin, beta-chain (Human (Homo sapiens),embryonic gower II)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.2    0.096
Query=  d1cqxa1 a.1.1.2 (A:1-150) Flavohemoglobin, N-terminal domain(Alcaligenes eutrophus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.9    0.22
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  32.5    0.26
lcl|d1h10a_  d1h10a_ b.55.1.1 (A:) Rac-alpha serine/threonine kin...  27.5    8.6
Query=  d1uc3a_ a.1.1.2 (A:) Lamprey globin (River lamprey (Lampetrafluviatilis))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  30.7    0.91
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  30.3    1.4
Query=  d1a6m__ a.1.1.2 (-) Myoglobin Sperm whale (Physeter catodon)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  31.9    0.37
Query=  d1hbra_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Chicken (Gallusgallus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.4    0.064
Query=  d1v75b_ a.1.1.2 (B:) Hemoglobin, beta-chain (Aldabra giant tortoise(Geochelone gigantea))
```

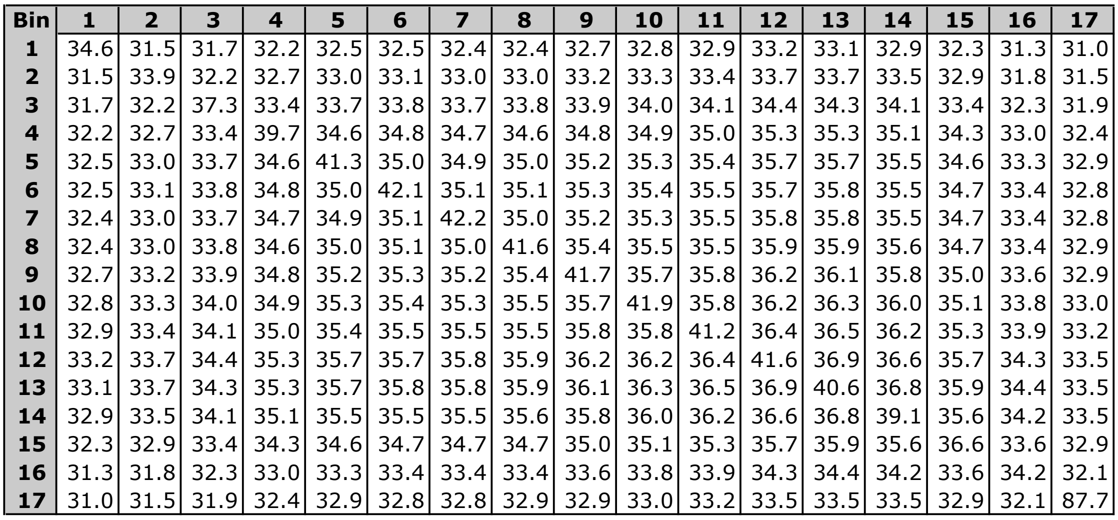
```
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  35.7    0.027
Query=  d1fhja_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Maned wolf(Chrysocyon brachyurus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  39.5    0.002
Query=  d1gjna_ a.1.1.2 (A:) Myoglobin Horse (Equus caballus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  31.1    0.68
Query=  d1cg5a_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Cartilaginous fishakaei (Dasyatis akajei))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.3    0.15
Query=  d1jl6a_ a.1.1.2 (A:) Glycera globin (Marine bloodworm (Glyceradibranchiata))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  31.4    0.63
lcl|d1h10a_  d1h10a_ b.55.1.1 (A:) Rac-alpha serine/threonine kin...  27.6    7.7
Query=  d1fsla_ a.1.1.2 (A:) Leghemoglobin (Soybean (Glycine max), isoformA)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.0    0.19
Query=  d3sdha_ a.1.1.2 (A:) Hemoglobin I (Ark clam (Scapharcainaequivalvis))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  39.2    0.003
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  28.0    5.7
Query=  d2mm1__ a.1.1.2 (-) Myoglobin Human (Homo sapiens)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  35.3    0.044
Query=  d1irda_ a.1.1.2 (A:) Hemoglobin, alpha-chain Human (Homo sapiens)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.5    0.13
Query=  d1hdsb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Deer (Odocoileusvirginianus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  31.0    0.69
Query=  d1emy__ a.1.1.2 (-) Myoglobin Asian elephant (Elephas maximus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.3    0.33
Query=  d2lhb__ a.1.1.2 (-) Lamprey globin (Sea lamprey (Petromyzonmarinus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.2    0.34
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  28.7    3.4
Query=  d1hlb__ a.1.1.2 (-) Hemoglobin, different isoforms (Sea cucumber(Caudina (Molpadia) arenicola))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  37.6    0.009
Query=  d1eca__ a.1.1.2 (-) Erythrocruorin (Midge (Chironomus thummithummi), fraction III)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.6    0.054
Query=  d1hbrb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Chicken (Gallusgallus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  36.5    0.014
Query=  d1lht__ a.1.1.2 (-) Myoglobin (Loggerhead sea turtle (Carettacaretta))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  31.2    0.66
Query=  d1ch4a_ a.1.1.2 (A:) Chimeric hemoglobin beta-alpha (Synthetic,based on Homo sapiens sequence)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.4    0.13
Query=  d1it2a_ a.1.1.2 (A:) Hagfish hemoglobin (Inshore hagfish(Eptatretus burgeri))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  45.7    3e-05
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  29.9    1.7
Query=  d1qpwa_ a.1.1.2 (A:) Hemoglobin, alpha-chain Pig (Sus scrofa)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  28.9    3.1
Query=  d1gvha1 a.1.1.2 (A:1-146) Flavohemoglobin, N-terminal domain(Escherichia coli)
```

```
lcl|d1rtxa_  d1rtxa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin ...  30.3    1.1
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  28.8    3.9
lcl|d1fmja_  d1fmja_ c.37.1.5 (A:) Retinol dehydratase (Fall army...  28.4    4.9
Query=  d1g08b_ a.1.1.2 (B:) Hemoglobin, beta-chain Cow (Bos taurus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  36.4    0.016
Query=  d1irdb_ a.1.1.2 (B:) Hemoglobin, beta-chain Human (Homo sapiens)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  37.5    0.008
Query=  d1mwca_ a.1.1.2 (A:) Myoglobin Pig (Sus scrofa)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.1    0.097
Query=  d1iwha_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Horse (Equuscaballus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  36.1    0.024
Query=  d1ew6a_ a.1.1.2 (A:) Dehaloperoxidase (Marine worm (Amphitriteornata))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  45.0    4e-05
Query=  d1v75a_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Aldabra gianttortoise (Geochelone gigantea))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.2    0.076
Query=  d1jebb_ a.1.1.2 (B:) Hemoglobin, beta-chain Mouse (Mus musculus)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  28.8    3.1
Query=  d1la6b_ a.1.1.2 (B:) Hemoglobin, beta-chain (Fish (Trematomusnewnesi))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  28.8    3.8
Query=  d1itha_ a.1.1.2 (A:) Hemoglobin Innkeeper worm (Urechis caupo)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.6    0.066
Query=  d1i3da_ a.1.1.2 (A:) Hemoglobin, beta-chain (Human fetus (Homosapiens), gamma-chain)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.4    0.14
Query=  d1jeba_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Human (Homo sapiens),zeta isoform)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.1    0.093
Query=  d1b0b__ a.1.1.2 (-) Hemoglobin I Clam (Lucina pectinata)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.1    0.093
Query=  d1gcva_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Houndshark (Mustelusgriseus))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  36.1    0.024
lcl|d1h10a_  d1h10a_ b.55.1.1 (A:) Rac-alpha serine/threonine kin...  30.3    1.3
Query=  d1myt__ a.1.1.2 (-) Myoglobin Yellowfin tuna (Thunnus albacares)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  37.6    0.008
Query=  d1outa_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Trout (Oncorhynchusmykiss))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.2    0.35
Query=  d1oj6a_ a.1.1.2 (A:) Neuroglobin Human (Homo sapiens)
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.7    0.25
Query=  d1qpwb_ a.1.1.2 (B:) Hemoglobin, beta-chain Pig (Sus scrofa)
Query=  d1scta_ a.1.1.2 (A:) Hemoglobin I (Ark clam (Scapharcainaequivalvis))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.6    0.061
Query=  d1s5xb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Emerald rockcod(Pagothenia bernacchii))
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  30.3    1.3
Query=  d1spga_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Teleost fish(Leiostomus xanthurus))
```

```
lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.9    0.056

Query=  d1ash__ a.1.1.2 (-) Ascaris hemoglobin, domain 1 (Pig roundworm(Ascaris suum))

lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  34.2    0.085

Query=  d1sctb_ a.1.1.2 (B:) Hemoglobin I (Ark clam (Scapharcainaequivalvis))

lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  39.2    0.003

Query=  d1iwhb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Horse (Equuscaballus))

lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  31.8    0.45

Query=  d1fhjb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Maned wolf (Chrysocyonbrachyurus))

lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  33.4    0.16

Query=  d1outb_ a.1.1.2 (B:) Hemoglobin, beta-chain (Trout (Oncorhynchusmykiss))

lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  32.7    0.24

Query=  d1la6a_ a.1.1.2 (A:) Hemoglobin, alpha-chain (Antarctic fish(Trematomus newnesi))

lcl|d1kr7a_  d1kr7a_ a.1.1.4 (A:) Nerve tissue mini-hemoglobin (n...  35.3    0.035
```

Appendix B

Additional figures and information on Chapter 5



Bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	34.6	31.5	31.7	32.2	32.5	32.5	32.4	32.4	32.7	32.8	32.9	33.2	33.1	32.9	32.3	31.3	31.0
2	31.5	33.9	32.2	32.7	33.0	33.1	33.0	33.0	33.2	33.3	33.4	33.7	33.7	33.5	32.9	31.8	31.5
3	31.7	32.2	37.3	33.4	33.7	33.8	33.7	33.8	33.9	34.0	34.1	34.4	34.3	34.1	33.4	32.3	31.9
4	32.2	32.7	33.4	39.7	34.6	34.8	34.7	34.6	34.8	34.9	35.0	35.3	35.3	35.1	34.3	33.0	32.4
5	32.5	33.0	33.7	34.6	41.3	35.0	34.9	35.0	35.2	35.3	35.4	35.7	35.7	35.5	34.6	33.3	32.9
6	32.5	33.1	33.8	34.8	35.0	42.1	35.1	35.1	35.3	35.4	35.5	35.7	35.8	35.5	34.7	33.4	32.8
7	32.4	33.0	33.7	34.7	34.9	35.1	42.2	35.0	35.2	35.3	35.5	35.8	35.8	35.5	34.7	33.4	32.8
8	32.4	33.0	33.8	34.6	35.0	35.1	35.0	41.6	35.4	35.5	35.5	35.9	35.9	35.6	34.7	33.4	32.9
9	32.7	33.2	33.9	34.8	35.2	35.3	35.2	35.4	41.7	35.7	35.8	36.2	36.1	35.8	35.0	33.6	32.9
10	32.8	33.3	34.0	34.9	35.3	35.4	35.3	35.5	35.7	41.9	35.8	36.2	36.3	36.0	35.1	33.8	33.0
11	32.9	33.4	34.1	35.0	35.4	35.5	35.5	35.5	35.8	35.8	41.2	36.4	36.5	36.2	35.3	33.9	33.2
12	33.2	33.7	34.4	35.3	35.7	35.7	35.8	35.9	36.2	36.2	36.4	41.6	36.9	36.6	35.7	34.3	33.5
13	33.1	33.7	34.3	35.3	35.7	35.8	35.8	35.9	36.1	36.3	36.5	36.9	40.6	36.8	35.9	34.4	33.5
14	32.9	33.5	34.1	35.1	35.5	35.5	35.5	35.6	35.8	36.0	36.2	36.6	36.8	39.1	35.6	34.2	33.5
15	32.3	32.9	33.4	34.3	34.6	34.7	34.7	34.7	35.0	35.1	35.3	35.7	35.9	35.6	36.6	33.6	32.9
16	31.3	31.8	32.3	33.0	33.3	33.4	33.4	33.4	33.6	33.8	33.9	34.3	34.4	34.2	33.6	34.2	32.1
17	31.0	31.5	31.9	32.4	32.9	32.8	32.8	32.9	32.9	33.0	33.2	33.5	33.5	33.5	32.9	32.1	87.7

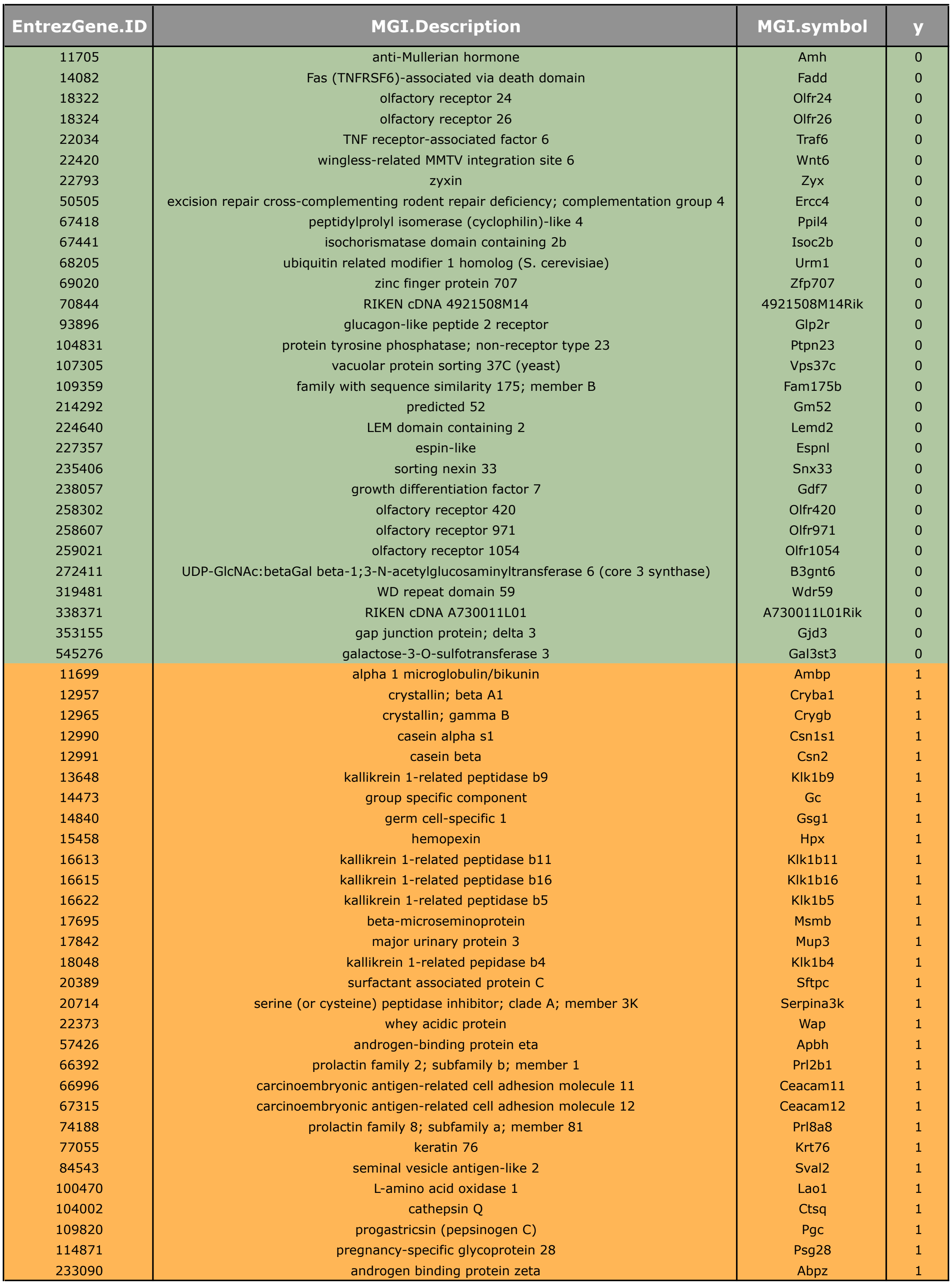
Figure B.1: **Mean similarity scores within and across bins (fruit fly)**. The mean Smith-Waterman similarity scores resulting from pairwise alignments between genes assigned to the same and different bins are shown. Bin 1 contains the most tissue-specific genes while bin 17 contains the most broadly expressed genes. Tissue specificity increases with the bin number. The mean values are highest within the bins, however these also contain scores resulting from self-alignments.

Bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	90.0	39.5	33.4	34.3	33.1	32.3	35.1	31.6	31.7	32.9	31.2	32.6	32.2	31.9	32.8	32.0	32.4	33.1	31.6	32.3	32.7	31.8	33.7	32.7	32.6	32.0	33.1	31.8	32.1	32.4	33.1
2	39.5	52.6	34.6	32.9	34.0	33.5	34.1	32.5	33.4	33.4	32.2	32.9	33.0	32.6	33.0	32.9	33.3	33.2	32.5	32.9	33.8	32.7	33.2	33.1	33.1	33.5	32.9	32.6	32.9	32.9	33.6
3	33.4	34.6	53.2	34.1	34.9	34.9	33.8	32.7	32.8	33.7	32.5	33.9	33.8	32.8	33.0	33.1	34.0	33.6	33.2	33.5	34.1	33.3	33.2	33.7	33.6	34.1	33.2	33.5	33.2	33.3	33.9
4	34.3	32.9	34.1	55.7	34.2	32.9	34.2	32.3	32.5	33.1	32.0	32.6	33.4	32.7	33.3	32.7	33.9	33.7	32.8	33.1	33.8	33.1	32.7	32.8	32.9	33.5	33.3	33.0	33.3	33.6	33.6
5	33.1	34.0	34.9	34.2	58.3	34.9	34.1	32.7	33.1	34.4	33.0	34.3	34.0	32.9	33.5	33.8	34.7	33.6	33.2	34.1	34.1	33.6	33.4	33.7	33.9	33.8	33.5	33.3	33.9	33.6	33.9
6	32.3	33.5	34.9	32.9	34.9	56.2	34.1	33.5	33.3	34.6	32.4	34.1	33.9	33.6	33.1	33.4	34.7	34.0	33.0	33.8	33.8	33.3	32.9	33.8	34.0	33.8	33.4	33.4	33.4	33.6	34.0
7	35.1	34.1	33.8	34.2	34.1	34.1	66.1	32.7	33.1	34.3	32.7	33.3	33.4	33.7	34.2	33.9	35.0	34.3	33.9	33.8	34.4	33.8	33.8	33.9	34.1	34.2	34.3	33.9	33.8	34.0	34.8
8	31.6	32.5	32.7	32.3	32.7	33.5	32.7	57.5	32.5	33.2	32.1	32.4	32.6	32.7	32.5	33.5	33.5	33.4	32.9	33.6	33.7	33.1	32.8	33.0	33.2	33.4	33.4	33.1	33.3	33.3	33.7
9	31.7	33.4	32.8	32.5	33.1	33.3	33.1	32.5	64.6	33.9	32.6	33.4	35.2	34.0	33.3	33.7	34.0	33.9	33.9	34.7	35.6	33.8	33.7	33.8	33.5	34.3	33.6	33.8	34.0	34.1	34.7
10	32.9	33.4	33.7	33.1	34.4	34.6	34.3	33.2	33.9	61.5	33.1	34.1	33.8	34.5	34.3	34.7	34.4	34.9	33.9	34.3	34.6	34.4	33.6	34.0	34.1	34.3	34.9	33.9	34.2	34.6	34.5
11	31.2	32.2	32.5	32.0	33.0	32.4	32.7	32.1	32.6	33.1	60.1	33.3	33.4	33.4	33.4	33.4	33.7	33.9	33.6	33.8	34.5	33.6	33.2	34.1	33.5	34.0	33.9	33.7	33.5	34.3	33.8
12	32.6	32.9	33.9	32.6	34.3	34.1	33.3	32.4	33.4	34.1	33.3	66.3	34.4	34.0	33.5	33.6	34.2	35.2	34.4	34.9	34.8	34.5	33.7	34.7	34.0	35.4	34.4	34.1	34.3	34.5	35.0
13	32.2	33.0	33.8	33.4	34.0	33.9	33.4	32.6	35.2	33.8	33.4	34.4	62.7	34.1	34.1	33.7	34.1	34.4	34.0	34.4	35.0	33.9	34.3	34.0	33.9	35.1	34.4	34.4	34.5	34.7	34.7
14	31.9	32.6	32.8	32.7	32.9	33.6	33.7	32.7	34.0	34.5	33.4	34.0	34.1	75.9	34.0	34.0	34.4	34.8	34.3	34.3	35.2	34.2	33.7	34.1	34.0	34.9	34.3	34.1	34.9	35.3	34.9
15	32.8	33.0	33.0	33.3	33.5	33.1	34.2	32.5	33.3	34.3	33.4	33.5	34.1	34.0	60.8	33.7	34.1	34.3	34.1	34.3	34.8	34.6	33.8	34.0	33.9	35.3	34.2	34.1	34.6	34.3	34.5
16	32.0	32.9	33.1	32.6	33.8	33.4	33.9	33.5	33.7	34.7	33.4	33.6	33.7	34.0	33.7	60.9	34.4	34.4	34.2	34.5	34.9	34.8	33.7	34.5	34.4	34.9	34.5	34.2	34.3	34.6	35.0
17	32.4	33.3	34.0	33.9	34.7	34.7	35.0	33.5	34.0	34.4	33.7	34.2	34.1	34.4	34.1	34.4	75.2	35.0	34.7	35.0	35.7	35.3	34.2	34.8	35.1	35.3	35.1	34.8	35.1	35.2	35.5
18	33.1	33.2	33.6	33.7	33.6	34.0	34.3	33.4	33.9	34.9	33.9	35.2	34.4	34.8	34.3	34.4	35.0	68.8	34.9	35.6	35.7	36.7	34.3	35.0	34.9	35.7	35.3	35.1	35.5	35.8	35.7
19	31.7	32.5	33.2	32.8	33.2	33.0	33.9	32.9	33.9	33.9	33.6	34.4	34.0	34.3	34.1	34.2	34.7	34.9	71.1	34.7	35.3	35.0	34.1	34.7	34.6	35.3	35.2	34.7	34.5	35.0	35.5
20	32.3	32.9	33.5	33.1	34.1	33.8	33.8	33.6	34.7	34.3	33.8	34.9	34.4	34.3	34.3	34.5	35.0	35.6	34.7	69.6	35.6	35.2	34.3	34.8	34.8	35.6	35.1	35.2	35.2	35.3	35.4
21	32.7	33.8	34.1	33.8	34.1	33.8	34.4	33.7	35.6	34.6	34.5	34.8	35.0	35.2	34.8	34.9	35.7	35.7	35.3	35.6	76.5	35.6	35.1	35.6	35.3	36.7	35.7	35.7	35.9	35.9	36.1
22	31.8	32.7	33.3	33.1	33.6	33.3	33.8	33.1	33.8	34.4	33.6	34.5	33.9	34.2	34.6	34.8	35.3	36.7	35.0	35.2	35.6	70.0	34.2	34.8	34.6	35.3	34.9	34.8	35.0	35.5	36.0
23	33.7	33.2	33.2	32.7	33.4	32.9	33.8	32.8	33.7	33.6	33.2	33.7	34.3	33.7	33.8	33.7	34.2	34.3	34.1	34.3	35.1	34.2	71.5	34.2	34.2	35.2	34.4	34.3	34.4	34.5	34.9
24	32.7	33.1	33.7	32.8	33.7	33.8	33.9	33.0	33.8	34.0	34.1	34.7	34.0	34.1	34.0	34.5	34.8	35.0	34.7	34.8	35.6	34.8	34.2	62.4	34.8	35.4	35.0	34.9	35.3	35.2	35.2
25	32.6	33.1	33.6	32.9	33.9	34.0	34.1	33.2	33.5	34.1	33.5	34.0	33.9	34.0	33.9	34.4	35.1	34.9	34.6	34.8	35.3	34.6	34.2	34.8	64.0	35.3	35.1	34.8	35.2	34.9	35.7
26	32.0	33.5	34.1	33.5	33.8	33.8	34.2	33.4	34.3	34.3	34.0	35.4	35.1	34.9	35.3	34.9	35.3	35.7	35.3	35.6	36.7	35.3	35.2	35.4	35.3	79.8	35.6	35.7	35.8	35.7	36.2
27	33.1	32.9	33.2	33.3	33.5	33.4	34.3	33.4	33.6	34.9	33.9	34.4	34.4	34.3	34.2	34.5	35.1	35.3	35.2	35.1	35.7	34.9	34.4	35.0	34.8	35.6	78.6	35.2	35.2	35.6	35.8
28	31.8	32.6	33.5	33.0	33.3	33.4	33.9	33.1	33.8	33.9	33.7	34.1	34.4	34.1	34.1	34.2	34.8	35.1	34.7	35.2	35.7	34.8	34.3	34.9	34.8	35.7	35.2	71.4	35.3	36.2	35.6
29	32.1	32.9	33.2	33.3	33.9	33.4	33.8	33.3	34.0	34.2	33.5	34.3	34.5	34.9	34.6	34.3	35.1	35.5	34.5	35.2	35.9	35.0	34.4	35.3	35.0	35.8	35.2	35.3	66.3	35.6	35.8
30	32.4	32.9	33.3	33.6	33.6	33.6	34.0	33.3	34.1	34.6	34.1	34.5	34.9	35.3	34.3	34.6	35.2	35.8	35.0	35.3	35.9	35.5	34.5	35.2	34.9	35.7	35.6	36.2	35.6	64.4	36.0
31	32.6	33.6	33.7	33.4	33.7	33.8	34.6	33.4	34.1	34.3	33.8	34.8	34.5	34.6	34.3	34.8	35.2	35.5	36.1	35.2	35.9	35.8	34.6	35.2	35.4	36.2	35.6	35.4	35.6	35.6	68.7
32	32.5	33.2	33.4	33.3	33.2	33.5	33.7	33.2	33.6	34.0	33.8	34.2	34.5	35.2	33.9	33.9	34.5	34.8	34.5	35.1	35.3	34.7	34.1	34.6	34.4	35.8	35.0	35.4	34.7	35.5	35.5
33	32.0	32.5	32.9	32.8	33.1	33.3	33.3	32.8	33.5	33.5	33.1	33.7	33.7	33.7	33.4	33.9	34.5	34.6	34.5	34.6	34.9	34.7	34.4	34.4	34.4	35.1	34.7	34.6	34.7	35.0	35.0
34	32.1	33.4	33.6	33.5	34.1	34.0	34.6	33.7	34.1	34.3	33.9	34.2	34.4	34.4	34.5	34.9	35.3	35.2	35.1	35.4	35.9	35.0	34.5	35.1	35.2	36.1	35.4	35.3	35.5	35.6	35.7
35	32.0	32.9	33.4	33.4	33.5	33.6	34.0	33.4	34.0	34.1	33.9	34.1	34.5	34.7	34.3	34.6	34.8	35.0	35.3	35.2	35.7	34.8	34.4	34.9	35.1	36.0	35.5	35.4	34.9	35.5	36.0
36	32.0	33.2	33.6	33.7	34.2	33.9	34.3	33.5	33.9	34.6	34.4	34.8	34.9	34.8	34.4	34.5	35.7	35.5	34.9	35.5	36.0	34.8	34.8	35.5	34.8	35.6	35.5	35.3	35.4	35.6	35.9
37	32.2	33.1	33.6	33.5	33.6	33.7	34.0	33.2	34.0	34.4	33.8	34.6	34.5	34.8	34.7	34.6	35.1	35.4	35.2	35.5	35.7	35.6	34.6	35.2	35.0	36.3	35.2	35.3	35.8	35.6	36.1
38	31.3	32.3	32.7	32.6	33.0	32.9	33.7	32.7	33.5	33.7	33.3	33.8	33.8	34.6	33.7	34.1	34.4	34.4	34.4	34.7	35.1	34.2	34.1	34.6	34.0	35.4	34.9	34.4	34.7	35.0	35.5
39	32.1	33.0	33.7	33.8	34.6	33.9	34.4	33.5	34.3	34.5	33.8	34.8	34.4	35.4	34.3	34.5	35.5	35.5	35.0	35.6	35.7	34.9	34.7	34.9	34.9	35.6	35.5	35.1	36.2	35.8	35.7
40	32.1	33.3	33.3	33.2	33.8	33.5	34.2	33.4	34.1	34.2	33.9	34.8	34.4	34.3	34.2	34.7	35.1	35.2	35.0	35.3	35.7	35.3	34.5	35.5	35.0	35.7	35.8	35.5	35.7	35.7	35.6
41	32.2	33.2	33.5	33.4	33.8	34.0	34.2	33.5	34.6	34.5	34.2	34.5	34.5	34.6	34.6	34.9	35.4	35.6	35.2	35.3	36.5	35.6	34.7	35.4	35.3	36.0	36.0	35.3	35.7	35.9	36.2
42	31.3	32.2	32.8	32.5	33.0	32.8	33.2	32.4	33.1	33.3	33.1	33.4	33.4	33.6	33.5	33.8	34.1	34.5	34.1	34.4	34.9	34.6	33.6	34.3	34.1	35.1	34.4	34.5	34.8	34.7	34.7
43	32.4	33.6	33.6	33.6	34.1	34.0	34.3	33.6	34.6	34.5	34.0	34.4	34.5	34.8	34.5	34.8	35.5	35.5	35.5	35.4	36.3	35.5	35.0	35.5	35.1	36.2	35.5	35.5	35.6	35.9	36.1
44	32.0	32.9	33.2	33.2	33.5	33.3	33.9	33.0	34.0	34.0	33.8	34.1	34.3	34.1	34.1	34.3	34.8	35.2	35.1	34.9	35.8	34.9	34.7	34.9	34.6	35.5	35.0	35.1	35.1	35.4	35.5
45	32.2	33.0	33.3	33.0	33.5	33.5	34.0	33.1	33.8	34.2	34.0	34.0	34.2	34.2	34.4	34.3	34.8	35.0	35.1	34.8	35.6	35.1	34.5	35.4	34.7	35.6	35.3	35.1	35.4	35.3	35.4
46	31.6	32.5	33.0	33.0	33.4	33.2	33.9	33.2	33.6	33.9	33.7	34.1	34.0	34.4	33.9	34.3	34.7	34.9	35.0	34.9	35.4	34.7	34.3	34.6	34.7	35.5	35.6	35.1	34.9	35.2	35.4
47	32.2	33.0	33.5	34.0	33.6	33.4	34.2	33.2	34.1	34.1	33.8	34.1	34.6	34.6	34.3	34.8	35.0	35.1	35.0	35.0	35.6	35.2	34.4	34.9	35.0	35.6	35.0	35.2	35.3	35.6	35.6
48	32.4	33.1	33.7	33.4	33.9	34.0	34.0	33.3	34.4	34.1	33.9	34.3	35.2	34.4	34.3	34.7	35.2	35.1	35.1	35.2	35.9	35.1	34.9	35.2	34.8	35.8	35.6	35.2	35.4	35.6	35.7
49	31.9	32.7	33.1	33.3	33.4	33.2	33.8	33.0	33.9	33.9	33.8	34.1	34.3	34.5	34.1	34.4	34.9	34.9	35.1	35.0	35.7	34.9	34.5	34.9	34.5	35.8	36.2	35.1	35.2	35.7	35.6
50	31.8	32.8	33.2	33.0	33.5	33.3	33.8	33.0	33.7	33.9	33.9	34.3	34.2	34.2	34.0	34.3	35.1	34.9	34.9	35.1	35.6	35.0	34.4	34.9	34.5	35.7	35.0	34.9	35.0	35.2	35.4
51	31.6	32.5	32.9	32.8	33.3	33.0	33.4	32.8	33.5	33.7	33.4	33.9	33.8	33.9	33.7	33.9	34.8	34.7	34.4	34.8	35.4	34.6	34.2	34.5	34.3	35.6	34.7	34.9	35.2	35.0	35.1
52	31.8	32.7	33.1	33.2	33.5	33.5	33.8	33.2	33.9	34.0	33.7	34.1	34.4	34.3	34.4	34.3	35.1	35.1	34.7	35.0	35.7	34.9	34.4	34.8	34.5	35.8	35.1	35.1	35.2	35.5	35.8
53	31.7	32.7	33.1	33.0	33.4	33.2	33.8	33.0	33.8	33.9	33.6	34.1	34.2	34.2	33.9	34.3	34.8	35.0	34.8	35.0	35.6	35.0	34.3	35.0	34.6	36.0	35.0	35.0	35.7	35.4	35.5
54	31.9	33.0	33.3	33.2	33.5	33.5	33.9	33.1	33.9	34.0	33.6	34.1	34.2	34.6	34.0	34.1	34.9	35.0	34.8	34.9	35.7	35.1	34.4	34.8	34.6	35.5	35.1	35.2	35.3	35.4	35.4
55	31.6	32.4	32.8	32.9	33.1	32.9	33.4	32.6	33.7	33.7	33.4	33.7	33.9	33.9	33.6	33.9	34.6	34.7	34.4	34.7	35.6	34.6	33.9	34.5	34.2	35.2	34.7	34.8	35.1	35.1	35.1
56	32.1	32.9	33.3	33.3	33.5	33.5	33.9	33.3	34.5	34.2	33.9	34.1	34.3	34.3	34.1	34.4	35.1	35.1	35.2	35.3	35.7	35.2	34.4	35.1	34.8	35.7	35.4	35.3	35.7	36.2	35.9
57	31.5	32.4	32.8	33.0	33.1	33.1	33.4	32.8	34.2	33.7	33.3	33.7	33.9	33.9	33.7	33.9	34.6	34.6	34.4	34.7	35.3	34.5	34.1	34.5	34.2	35.3	34.7	34.7	35.5	35.6	35.4
58	31.6	32.7	32.8	33.1	33.1	33.0	33.3	32.7	34.4	33.7	33.3	33.6	33.7	33.9	33.4	33.7	34.4	34.3	34.2	34.4	35.1	34.4	33.7	34.3	34.2	35.1	34.4	34.5	35.5	35.4	35.4
59	31.5	32.2	32.5	33.1	32.8	32.7	33.1	32.5	34.6	33.6	33.0	33.4	33.5	33.6	33.3	33.5	34.2	34.1	34.0	34.4	34.8	34.4	33.7	34.3	33.8	34.8	34.2	34.3	35.7	35.6	35.4
60	31.3	32.0	32.2	33.1	32.5	32.4	32.9	32.4	34.8	33.4	32.8	33.1	33.3	33.4	32.9	33.2	33.8	33.7	33.5	34.1	34.4	34.1	33.2	33.8	33.4	34.3	34.0	33.8	35.4	35.5	35.3

Figure B.2: **Mean similarity scores within and across groups (mouse part 1)**. The mean Smith-Waterman similarity scores resulting from pairwise alignments between genes assigned to the same and different bins are shown. Bin 1 contains the most tissue-specific genes while bin 60 contains the most broadly expressed genes. Tissue specificity increases with the bin number. Scores resulting from self-alignments were not considered in the calculation of the mean values. The table is continued on the next page. The mean values are highest within the bins, however these also contain scores resulting from self-alignments.

Bin	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	32.5	32.0	32.1	32.0	32.0	32.2	31.3	32.1	32.1	32.2	31.3	32.4	32.0	32.1	31.6	32.2	32.4	31.9	31.8	31.6	31.8	31.7	31.9	31.5	32.0	31.6	31.6	31.6	31.3
2	33.2	32.5	33.4	32.9	33.2	33.1	32.3	33.0	33.3	33.2	32.2	33.6	32.9	33.0	32.5	32.9	33.1	32.7	32.9	32.5	32.8	32.7	33.0	32.4	32.9	32.4	32.4	32.3	32.0
3	33.4	32.9	33.6	33.4	33.6	33.6	32.7	33.7	33.3	33.5	32.8	33.6	33.2	33.3	33.0	33.5	33.7	33.1	33.2	32.9	33.1	33.1	33.3	32.9	33.3	32.9	32.8	32.5	32.2
4	33.3	32.8	33.5	33.4	33.7	33.5	32.6	33.8	33.2	33.4	32.5	33.6	33.2	33.0	33.0	34.0	33.4	33.3	33.0	32.8	33.2	33.0	33.2	32.9	33.3	33.1	33.2	33.1	33.1
5	33.2	33.1	34.1	33.5	34.2	33.6	33.0	34.6	33.8	33.8	33.0	34.1	33.5	33.5	33.4	33.6	33.9	33.4	33.5	33.3	33.5	33.4	33.5	33.0	33.5	33.1	33.1	32.8	32.5
6	33.5	33.3	34.0	33.6	33.9	33.7	32.9	33.7	33.5	34.0	32.8	34.0	33.3	33.5	33.2	33.4	34.0	33.2	33.3	33.0	33.6	33.3	33.7	32.9	33.5	33.1	33.0	32.8	32.4
7	33.7	33.3	34.1	34.0	34.3	34.0	33.5	34.4	34.2	34.2	33.2	34.3	33.9	34.0	33.9	34.2	34.0	33.8	33.8	33.4	34.1	33.8	33.9	33.4	33.9	33.7	33.3	33.1	32.9
8	33.2	32.8	33.7	33.4	34.0	33.2	32.7	33.5	33.4	33.5	32.4	33.6	33.0	33.1	33.2	33.2	33.3	33.0	33.0	32.8	33.2	33.0	33.1	32.6	33.3	32.8	32.8	32.5	32.4
9	33.6	33.5	34.1	34.0	33.9	34.0	33.5	34.3	34.1	34.6	33.1	34.6	34.0	33.8	33.6	34.1	34.4	34.0	33.7	33.5	33.9	34.0	33.9	33.7	34.5	34.2	34.4	34.6	34.8
10	34.0	33.5	34.3	34.1	34.6	34.4	33.7	34.5	34.2	34.5	33.3	34.5	34.0	34.2	33.9	34.1	34.1	33.9	33.9	33.7	34.0	33.9	34.0	33.7	34.2	33.7	33.7	33.6	33.4
11	33.8	33.1	34.0	33.9	34.4	33.8	33.3	33.8	33.9	34.2	33.1	34.0	33.8	34.0	33.7	33.8	33.9	33.8	33.9	33.4	33.7	33.6	33.6	33.4	33.9	33.4	33.2	33.1	32.8
12	34.2	33.7	34.2	34.1	34.8	34.6	33.8	34.8	34.8	34.5	33.4	34.4	34.1	34.0	34.1	34.1	34.3	34.1	34.3	33.9	34.1	34.1	34.1	33.7	34.1	33.9	33.6	33.4	33.1
13	34.5	33.7	34.4	34.5	34.9	34.5	33.8	34.4	34.4	34.5	33.4	34.5	34.3	34.2	34.0	34.6	35.2	34.3	34.2	33.8	34.4	34.2	34.2	33.8	34.3	33.9	33.7	33.6	33.3
14	35.2	33.7	34.4	34.7	34.8	34.8	34.6	34.7	34.3	34.6	33.6	34.8	34.1	34.2	34.4	34.6	34.4	34.5	34.2	33.9	34.3	34.2	34.6	33.9	34.3	34.0	33.9	33.7	33.4
15	33.9	33.4	34.5	34.3	34.4	34.7	33.7	34.3	34.2	34.6	33.5	34.5	34.1	34.4	33.9	34.3	34.3	34.1	34.0	33.7	34.1	33.9	34.0	33.6	34.1	33.7	33.5	33.3	33.0
16	33.9	33.9	34.9	34.6	34.5	34.6	34.1	34.5	34.7	34.9	33.8	34.8	34.4	34.3	34.3	35.0	34.6	34.4	34.3	33.9	34.3	34.3	34.1	33.9	34.4	34.0	33.7	33.5	33.3
17	34.5	34.5	35.3	34.8	35.7	35.1	34.4	35.5	35.1	35.4	34.1	35.5	34.8	34.8	34.7	35.0	35.2	34.9	35.1	34.8	35.2	34.8	34.9	34.6	35.1	34.8	34.4	34.2	33.9
18	34.8	34.6	35.2	35.0	35.5	35.4	34.4	35.5	35.2	35.6	34.5	35.5	35.2	35.0	34.9	35.1	35.1	34.9	34.9	34.7	35.1	35.0	35.0	34.9	35.1	34.7	34.3	34.2	33.8
19	34.5	34.5	35.1	35.3	34.9	35.2	34.4	35.0	35.0	35.2	34.1	35.5	35.1	35.1	35.0	35.0	35.1	35.1	34.9	34.4	34.7	34.8	34.8	34.4	35.1	34.6	34.2	34.1	33.5
20	35.1	34.6	35.6	35.2	35.5	35.5	34.7	35.6	35.3	35.3	34.4	35.4	34.9	35.0	34.9	35.0	35.2	35.0	35.1	34.8	35.0	35.0	34.9	34.7	35.1	34.7	34.5	34.5	34.1
21	35.3	34.9	35.9	35.7	36.0	36.0	35.1	35.7	35.7	36.5	34.9	36.3	35.8	35.6	35.4	35.6	35.9	35.7	35.6	35.4	35.8	35.6	35.7	35.4	35.7	35.2	35.1	34.8	34.4
22	34.7	34.7	35.0	34.8	34.8	35.6	34.2	34.9	35.3	35.6	34.6	35.5	34.9	35.1	34.7	35.2	35.1	34.9	35.0	34.6	35.0	34.9	35.1	34.5	35.2	34.7	34.4	34.5	33.9
23	34.1	34.4	34.5	34.4	34.8	34.6	34.1	34.7	34.5	34.7	33.6	35.0	34.7	34.5	34.2	34.4	34.8	34.5	34.4	34.2	34.4	34.3	34.3	33.9	34.4	34.0	33.7	33.7	33.2
24	34.6	34.4	35.1	34.9	35.6	35.2	34.6	34.9	35.5	35.4	34.3	35.5	34.9	35.4	34.6	34.9	35.2	34.9	34.9	34.5	34.9	35.1	34.8	34.5	35.0	34.6	34.4	34.2	33.8
25	34.4	34.4	35.2	35.1	34.8	35.0	34.0	34.9	35.0	35.3	34.1	35.1	34.6	34.7	34.7	35.0	34.8	34.5	34.5	34.3	34.6	34.6	34.6	34.2	34.8	34.2	34.2	33.9	33.5
26	35.8	35.1	36.1	36.0	35.6	36.3	35.4	35.6	35.7	35.8	35.1	36.2	35.5	35.6	35.5	35.6	35.8	35.8	35.7	35.6	35.7	35.9	35.5	35.2	35.7	35.3	35.2	34.9	34.4
27	35.0	34.7	35.4	35.5	35.5	35.2	34.9	35.5	35.1	36.0	34.4	35.5	35.0	35.3	35.6	35.0	35.6	35.3	35.0	34.7	35.2	35.0	35.1	34.7	35.4	34.7	34.5	34.3	34.0
28	35.4	34.6	35.3	35.4	35.3	35.3	34.4	35.1	35.5	35.3	34.5	35.5	35.1	35.4	35.1	35.2	35.2	35.1	34.9	34.9	35.1	35.0	35.2	34.8	35.2	34.9	34.5	34.3	33.8
29	34.7	34.6	35.5	34.9	35.4	35.8	34.7	36.2	35.7	35.7	34.8	35.6	35.1	35.4	34.9	35.3	35.4	35.3	35.0	35.2	35.4	35.7	35.3	35.1	35.7	35.5	35.6	35.6	35.4
30	35.5	35.0	35.6	35.5	35.6	35.6	35.0	35.8	35.7	35.9	34.7	35.9	35.4	35.3	35.2	35.6	35.6	35.8	35.2	35.0	35.5	35.5	35.5	35.1	35.9	35.6	35.5	35.6	35.5
31	35.3	34.8	35.3	36.0	35.4	35.8	35.1	35.6	35.4	36.0	34.6	35.9	35.4	35.2	35.4	35.5	35.5	35.5	35.4	35.0	35.6	35.4	35.3	35.0	35.8	35.5	35.3	35.4	35.3
32	64.1	34.6	35.1	34.9	34.9	34.9	34.8	35.2	34.9	35.5	34.2	35.0	34.9	34.9	34.8	34.9	35.3	35.1	35.1	34.4	35.1	34.9	35.2	34.7	35.9	35.8	35.7	36.2	36.6
33	34.6	61.3	34.7	34.6	34.5	35.0	34.2	35.1	34.8	35.2	33.8	35.2	34.5	34.5	34.4	34.6	34.8	35.0	35.1	34.2	34.6	34.5	34.6	34.4	35.0	34.7	34.6	34.9	34.8
34	35.1	34.7	63.1	35.3	35.5	36.6	35.0	35.5	35.5	35.7	34.8	36.1	35.5	35.7	35.4	35.5	35.8	35.5	35.3	35.3	35.3	35.8	35.5	35.3	35.6	35.4	35.1	34.9	34.4
35	34.9	34.6	35.3	69.1	35.1	35.3	34.8	35.3	35.2	35.5	34.6	35.4	35.6	35.2	35.0	35.3	35.5	35.1	35.1	35.0	35.2	35.0	35.0	34.6	35.3	34.7	34.5	34.2	33.8
36	34.9	34.5	35.5	35.1	72.0	35.6	35.0	35.8	35.3	35.9	34.3	35.5	35.7	35.3	35.4	35.6	35.7	35.3	35.5	35.2	35.4	35.3	35.3	34.9	35.5	35.0	34.7	34.5	34.2
37	34.9	35.0	36.6	35.3	35.6	58.7	35.1	35.5	35.8	35.9	34.9	35.9	35.3	35.6	35.3	35.5	35.6	35.6	35.5	35.2	35.5	35.8	35.5	35.4	36.0	35.8	35.7	35.7	35.4
38	35.0	34.2	35.0	34.8	35.0	35.1	60.1	35.3	34.7	35.8	34.0	35.3	34.7	34.8	34.9	35.0	35.0	35.3	35.0	34.6	35.1	35.1	35.1	34.8	36.2	36.1	36.3	36.9	37.4
39	35.2	35.1	35.5	35.3	35.8	35.5	35.3	65.0	35.9	36.0	34.7	35.7	35.5	35.2	35.4	35.6	35.6	35.6	35.5	35.1	35.7	35.4	35.5	35.5	36.2	36.1	36.1	36.6	36.6
40	34.9	34.8	35.5	35.2	35.3	35.8	34.7	35.9	57.3	35.6	34.6	35.8	35.5	35.3	35.5	35.6	35.5	35.3	35.4	34.9	35.2	35.4	35.1	34.9	35.3	35.0	34.7	34.4	34.0
41	35.5	35.2	35.7	35.5	35.9	35.9	35.8	36.0	35.6	58.9	35.0	36.0	35.6	35.7	35.6	35.6	36.0	36.3	35.8	35.3	35.8	35.6	35.8	35.6	36.8	36.6	36.9	37.6	38.2
42	34.2	33.8	34.8	34.6	34.3	34.9	34.0	34.7	34.6	35.0	51.1	34.9	34.6	34.7	34.2	34.6	34.9	34.5	34.5	34.3	34.5	34.7	34.5	34.5	34.9	34.7	34.5	34.5	34.4
43	35.0	35.2	36.2	35.4	35.5	35.9	35.3	35.7	35.8	36.0	34.9	56.4	35.5	35.7	35.3	35.7	35.9	35.6	35.5	35.4	35.6	35.8	35.7	35.3	35.6	35.5	35.0	35.0	34.3
44	34.9	34.5	35.5	35.6	35.7	35.3	34.7	35.5	35.5	35.6	34.6	35.5	51.6	35.4	34.9	35.4	35.5	35.2	35.1	34.8	35.3	35.1	35.2	34.9	35.5	35.0	34.8	34.8	34.6
45	34.9	34.5	35.7	35.2	35.3	35.6	34.8	35.2	35.3	35.7	34.7	35.7	35.5	51.7	35.1	35.4	35.5	35.3	35.2	35.0	35.2	35.5	35.3	35.0	35.6	35.2	34.8	34.7	34.3
46	34.8	34.4	35.4	35.0	35.4	35.3	34.9	35.4	35.5	35.6	34.2	35.3	34.9	35.3	50.0	35.2	35.2	35.2	35.0	34.8	35.2	35.2	35.1	34.7	35.4	35.0	34.8	34.8	34.5
47	34.9	34.6	35.5	35.3	35.6	35.5	34.8	35.6	35.6	35.6	34.6	35.7	35.4	35.4	35.2	51.5	35.6	35.4	35.2	35.0	35.4	35.4	35.5	35.1	35.7	35.6	35.4	35.4	35.3
48	35.3	34.8	35.8	35.5	35.7	35.6	35.1	35.6	35.5	36.0	34.9	35.9	35.5	35.5	35.2	35.6	50.5	35.8	35.6	35.3	35.6	35.9	35.6	35.4	35.9	35.8	35.5	35.5	35.4
49	35.1	35.0	35.5	35.1	35.3	35.6	35.3	35.6	35.2	36.3	34.5	35.6	35.2	35.3	35.2	35.4	35.8	48.2	35.6	35.2	35.6	35.7	35.7	35.6	36.9	36.8	37.0	37.9	38.5
50	35.1	35.1	35.3	35.1	35.7	35.5	35.0	35.5	35.4	35.8	34.5	35.5	35.1	35.2	35.0	35.2	35.6	35.6	46.6	35.0	35.4	35.4	35.4	35.0	35.9	35.5	35.4	35.6	35.5
51	34.4	34.2	35.3	35.0	35.2	35.2	34.6	35.1	34.9	35.3	34.3	35.4	34.8	35.1	34.8	35.0	35.3	35.2	35.0	44.4	35.1	35.4	35.2	35.1	35.6	35.4	35.2	35.3	35.0
52	35.0	34.6	35.3	35.2	35.4	35.4	35.1	35.7	35.2	35.8	34.4	35.6	35.2	35.2	35.0	35.4	35.6	35.6	35.4	35.0	46.4	35.4	35.6	35.2	36.3	36.0	36.0	36.3	36.4
53	34.9	34.5	35.7	35.0	35.3	35.8	35.1	35.4	35.4	35.6	34.7	35.8	35.1	35.5	35.2	35.4	35.8	35.7	35.4	35.5	35.5	44.2	35.8	35.7	36.4	36.3	36.2	36.5	36.4
54	35.3	34.6	35.5	35.0	35.3	35.5	35.1	35.5	35.2	35.8	34.5	35.7	35.2	35.3	35.1	35.5	35.7	35.7	35.4	35.2	35.6	35.8	43.7	35.6	36.7	36.5	36.5	37.2	37.4
55	34.7	34.3	35.3	34.6	35.2	35.4	34.9	35.4	34.9	35.6	34.5	35.3	34.9	35.0	34.6	35.1	35.4	35.6	35.1	35.2	35.3	35.9	35.6	41.9	36.8	36.9	37.2	38.1	38.8
56	35.9	35.1	35.6	35.3	35.5	36.0	36.2	36.3	35.3	36.8	34.9	35.6	35.5	35.6	35.4	35.7	35.9	36.9	35.9	35.5	36.3	36.5	36.8	36.8	44.9	39.9	40.9	43.1	45.5
57	35.6	34.7	35.4	34.9	35.0	35.7	36.1	36.1	34.9	36.6	34.6	35.5	35.0	35.1	35.0	35.5	35.7	36.7	35.5	35.4	36.0	36.2	36.5	36.9	39.9	45.1	42.0	44.8	47.8
58	35.7	34.7	35.1	34.5	34.7	35.6	36.2	36.1	34.7	36.9	34.5	35.1	34.9	34.8	34.8	35.4	35.5	37.0	35.4	35.2	35.9	36.2	36.5	37.2	40.9	42.1	46.6	47.6	51.8
59	36.2	34.8	34.9	34.2	34.6	35.6	36.9	36.5	34.4	37.5	34.5	34.8	34.9	34.6	34.7	35.4	35.5	37.8	35.6	35.2	36.2	36.5	37.1	38.1	43.1	44.8	47.6	54.5	58.8
60	36.5	34.7	34.4	33.8	34.3	35.3	37.4	36.6	33.9	38.2	34.4	34.3	34.6	34.2	34.5	35.4	35.4	38.4	35.4	35.0	36.4	36.4	37.4	38.8	45.4	47.7	51.8	58.8	71.7

Figure B.3: **Mean similarity scores within and across groups (mouse part 2)**. The mean Smith-Waterman similarity scores resulting from pairwise alignments between genes assigned to the same and different bins are shown. Bin 1 contains the most tissue-specific genes while bin 60 contains the most broadly expressed genes. Tissue specificity increases with the bin number. The table is a continuation of the table on the previous page and includes mean scores for bins 32-60.



EntrezGene.ID	MGI.Description	MGI.symbol	y
11705	anti-Mullerian hormone	Amh	0
14082	Fas (TNFRSF6)-associated via death domain	Fadd	0
18322	olfactory receptor 24	Olfr24	0
18324	olfactory receptor 26	Olfr26	0
22034	TNF receptor-associated factor 6	Traf6	0
22420	wingless-related MMTV integration site 6	Wnt6	0
22793	zyxin	Zyx	0
50505	excision repair cross-complementing rodent repair deficiency; complementation group 4	Ercc4	0
67418	peptidylprolyl isomerase (cyclophilin)-like 4	Ppil4	0
67441	isochorismatase domain containing 2b	Isoc2b	0
68205	ubiquitin related modifier 1 homolog (S. cerevisiae)	Urm1	0
69020	zinc finger protein 707	Zfp707	0
70844	RIKEN cDNA 4921508M14	4921508M14Rik	0
93896	glucagon-like peptide 2 receptor	Glp2r	0
104831	protein tyrosine phosphatase; non-receptor type 23	Ptpn23	0
107305	vacuolar protein sorting 37C (yeast)	Vps37c	0
109359	family with sequence similarity 175; member B	Fam175b	0
214292	predicted 52	Gm52	0
224640	LEM domain containing 2	Lemd2	0
227357	espin-like	Espnl	0
235406	sorting nexin 33	Snx33	0
238057	growth differentiation factor 7	Gdf7	0
258302	olfactory receptor 420	Olfr420	0
258607	olfactory receptor 971	Olfr971	0
259021	olfactory receptor 1054	Olfr1054	0
272411	UDP-GlcNAc:betaGal beta-1;3-N-acetylglucosaminyltransferase 6 (core 3 synthase)	B3gnt6	0
319481	WD repeat domain 59	Wdr59	0
338371	RIKEN cDNA A730011L01	A730011L01Rik	0
353155	gap junction protein; delta 3	Gjd3	0
545276	galactose-3-O-sulfotransferase 3	Gal3st3	0
11699	alpha 1 microglobulin/bikunin	Ambp	1
12957	crystallin; beta A1	Cryba1	1
12965	crystallin; gamma B	Crygb	1
12990	casein alpha s1	Csn1s1	1
12991	casein beta	Csn2	1
13648	kallikrein 1-related peptidase b9	Klk1b9	1
14473	group specific component	Gc	1
14840	germ cell-specific 1	Gsg1	1
15458	hemopexin	Hpx	1
16613	kallikrein 1-related peptidase b11	Klk1b11	1
16615	kallikrein 1-related peptidase b16	Klk1b16	1
16622	kallikrein 1-related peptidase b5	Klk1b5	1
17695	beta-microseminoprotein	Msmb	1
17842	major urinary protein 3	Mup3	1
18048	kallikrein 1-related pepidase b4	Klk1b4	1
20389	surfactant associated protein C	Sftpc	1
20714	serine (or cysteine) peptidase inhibitor; clade A; member 3K	Serpina3k	1
22373	whey acidic protein	Wap	1
57426	androgen-binding protein eta	Apbh	1
66392	prolactin family 2; subfamily b; member 1	Prl2b1	1
66996	carcinoembryonic antigen-related cell adhesion molecule 11	Ceacam11	1
67315	carcinoembryonic antigen-related cell adhesion molecule 12	Ceacam12	1
74188	prolactin family 8; subfamily a; member 81	Prl8a8	1
77055	keratin 76	Krt76	1
84543	seminal vesicle antigen-like 2	Sval2	1
100470	L-amino acid oxidase 1	Lao1	1
104002	cathepsin Q	Ctsq	1
109820	progastricsin (pepsinogen C)	Pgc	1
114871	pregnancy-specific glycoprotein 28	Psg28	1
233090	androgen binding protein zeta	Abpz	1

Figure B.4: **List of mouse genes with broad and tissue-specific expression**. The 30 top/bottom genes are shown that led to a prediction performance of AUC=1 as described in Chapter 5.2.1.1. The values 0 and 1 in the 'y' column indicate broad (green) and tissue-specific (orange) expression, respectively. Data were mapped using Ensembl 58.

	Amino Acid (percentage)																				
EntrezGene.ID	Ap	Cp	Dp	Ep	Fp	Gp	Hp	Ip	Kp	Lp	Mp	Np	Pp	Qp	Rp	Sp	Tp	Vp	Wp	Yp	y
11705	10.65	2.17	3.25	6.32	1.99	8.66	1.62	1.26	0.72	16.97	1.26	2.35	10.11	5.05	6.86	6.32	6.14	5.23	1.81	1.26	0
14082	7.32	1.46	6.83	9.76	2.44	4.88	1.46	0.98	6.34	16.10	1.95	3.41	2.93	2.44	8.78	9.76	3.41	8.29	0.98	0.49	0
18322	6.71	4.15	2.88	1.60	6.39	3.83	3.19	7.99	4.47	12.46	4.47	3.51	5.11	0.96	3.51	8.95	6.71	8.63	0.32	4.15	0
18324	5.19	3.25	1.30	1.95	8.12	5.52	2.60	7.14	3.90	15.26	3.90	4.22	3.57	1.30	2.27	10.71	5.19	10.06	0.00	4.55	0
22034	6.23	6.04	4.53	7.92	3.58	4.91	3.96	5.28	4.34	8.87	3.40	4.15	5.66	6.23	6.23	6.42	4.53	4.91	0.38	2.45	0
22420	10.16	6.87	4.40	4.67	3.30	9.89	3.02	1.10	2.47	11.26	1.37	1.92	6.87	4.12	10.99	6.32	3.57	5.49	1.92	0.27	0
22793	8.16	4.08	3.90	5.67	3.55	6.03	2.13	2.13	5.14	5.50	1.77	2.13	18.97	8.69	3.19	6.38	4.26	6.38	0.18	1.77	0
50505	7.28	1.99	4.86	8.61	3.64	4.97	2.32	4.19	6.29	13.02	1.32	3.20	5.08	3.42	7.40	6.73	5.52	6.51	0.66	2.98	0
67418	3.25	1.83	8.94	10.98	4.27	5.69	2.64	5.89	11.38	5.69	1.42	3.46	4.47	3.46	6.30	6.30	4.47	5.28	0.61	3.66	0
67441	6.67	1.90	3.81	5.24	3.81	6.19	1.90	5.24	5.71	16.19	3.81	1.43	5.71	8.57	4.29	7.62	4.29	6.67	0.00	0.95	0
68205	4.95	0.99	7.92	7.92	3.96	9.90	1.98	6.93	5.94	16.83	0.99	2.97	4.95	5.94	2.97	2.97	0.99	6.93	2.97	0.99	0
69020	4.73	6.21	2.96	6.80	5.33	9.17	5.92	2.37	7.40	8.28	0.89	2.66	4.14	5.33	10.95	6.51	5.03	2.07	1.78	1.48	0
70844	4.11	4.11	2.74	3.42	9.59	6.85	2.74	2.74	3.42	13.70	2.05	3.42	5.48	4.11	6.16	8.22	6.16	4.79	2.05	4.11	0
93896	4.88	2.73	2.54	5.47	6.45	6.05	3.52	3.71	5.08	15.04	1.76	3.32	3.52	3.52	5.66	8.59	5.27	6.05	3.52	3.32	0
104831	8.92	0.89	4.08	6.21	2.48	5.67	2.96	2.48	4.49	11.17	2.19	2.13	12.88	7.98	4.79	7.03	4.26	6.32	0.53	2.54	0
107305	5.68	0.28	1.99	8.81	2.27	7.95	0.85	1.14	3.41	10.23	1.99	1.42	19.89	7.67	5.40	7.10	4.55	4.83	0.57	3.98	0
109359	6.02	0.96	5.06	7.95	4.58	4.10	2.89	4.34	4.34	6.51	1.69	5.06	5.30	6.99	6.51	13.73	4.58	5.30	0.24	3.86	0
214292	5.83	3.08	3.08	3.08	4.54	4.54	2.59	5.67	2.27	12.97	1.78	3.73	8.10	4.70	4.86	12.16	6.81	4.86	2.11	3.24	0
224640	9.98	1.57	5.09	8.02	2.35	6.85	1.76	2.94	4.11	11.55	1.76	1.96	7.44	2.54	10.96	8.22	3.13	4.70	3.13	1.96	0
227357	10.95	2.29	5.57	7.26	2.89	9.25	3.78	1.99	2.39	12.44	1.89	1.39	8.66	4.68	7.06	6.27	4.18	4.38	1.59	1.09	0
235406	6.27	0.87	6.62	6.45	5.92	7.49	4.01	3.66	5.57	8.54	3.48	2.96	5.05	5.92	6.10	8.19	3.83	4.70	1.22	3.14	0
238057	15.18	2.39	4.77	4.77	2.39	14.10	2.39	2.17	1.30	9.33	1.08	1.52	6.51	2.17	10.85	7.59	3.69	4.99	1.52	1.30	0
258302	7.79	4.05	1.87	3.12	8.41	4.36	3.12	10.90	2.18	12.15	3.12	2.49	4.05	1.87	3.74	8.10	7.17	6.85	1.56	3.12	0
258607	3.91	2.93	1.30	3.26	7.17	4.23	2.61	9.12	3.91	16.61	4.23	4.23	3.91	1.30	1.95	11.40	5.86	6.84	0.33	4.89	0
259021	4.49	2.56	3.21	2.56	7.05	3.53	2.24	11.86	3.21	15.38	4.17	3.85	3.53	1.92	3.53	8.33	5.45	7.69	0.64	4.81	0
272411	8.44	2.30	3.58	4.35	5.63	5.88	4.35	1.28	2.30	13.81	1.79	2.30	8.18	5.63	7.42	7.16	3.58	7.67	1.53	2.81	0
319481	6.25	3.02	5.65	5.14	3.93	5.54	3.23	3.53	4.74	8.47	1.51	3.93	6.15	4.84	6.55	9.98	5.65	6.75	2.32	2.82	0
338371	6.27	2.35	5.49	6.27	2.75	9.80	3.53	3.14	4.31	10.98	1.18	1.57	8.63	7.45	7.45	5.88	2.75	8.24	0.78	1.18	0
353155	11.51	3.60	3.24	4.32	5.40	8.27	2.88	2.16	1.80	14.03	1.08	0.36	9.35	3.96	7.19	6.83	2.52	7.19	1.44	2.88	0
545276	10.90	1.86	3.71	5.57	4.41	3.94	3.25	3.02	2.78	12.30	2.55	2.78	8.58	3.94	10.90	5.57	4.41	4.64	1.16	3.71	0
11699	6.02	4.58	4.01	8.02	3.44	8.88	1.72	5.73	5.73	8.60	1.72	4.01	3.72	4.30	5.16	7.45	6.59	4.01	1.43	4.87	1
12957	3.72	3.72	3.26	8.37	4.19	7.91	3.26	6.98	4.19	3.72	3.26	4.65	4.65	7.91	5.58	7.44	5.12	2.79	4.19	5.12	1
12965	1.71	5.14	6.86	5.71	6.29	7.43	2.29	3.43	1.71	6.29	4.57	2.86	3.43	5.71	10.86	9.14	2.29	3.43	2.29	8.57	1
12990	10.54	0.64	2.56	5.43	4.47	0.00	1.92	2.88	4.15	11.50	3.83	4.79	5.75	16.93	2.56	10.86	3.19	4.47	0.64	2.88	1
12991	7.79	0.87	1.73	4.76	3.46	0.87	2.60	4.76	3.90	14.72	1.73	3.46	9.09	12.99	0.87	10.82	4.76	9.52	0.00	1.30	1
13648	5.75	3.83	5.36	5.36	3.07	8.81	3.45	4.60	8.05	11.11	1.53	4.98	6.13	1.92	3.07	6.13	5.36	5.75	1.92	3.83	1
14473	6.30	5.88	4.20	9.24	4.20	2.73	1.47	2.52	7.14	10.92	2.94	3.36	5.67	4.62	3.99	9.24	7.56	4.62	0.00	3.36	1
14840	7.72	3.70	2.78	6.17	5.56	5.86	2.47	2.78	3.40	12.04	3.70	2.47	5.25	5.56	3.70	10.49	6.79	4.63	2.47	2.47	1
15458	6.96	2.83	5.22	4.57	4.57	8.70	3.26	2.61	4.57	8.48	0.87	4.13	7.39	2.61	5.65	9.13	5.43	5.65	3.91	3.48	1
16613	4.98	3.83	5.75	3.83	2.68	8.43	3.07	5.75	6.90	9.96	2.68	5.36	8.05	3.07	2.30	5.75	5.75	6.90	2.68	2.30	1
16615	4.98	4.21	6.51	4.21	3.83	8.05	1.53	4.98	7.28	11.49	3.07	4.21	7.66	3.45	1.53	5.75	5.36	7.28	2.68	1.92	1
16622	6.13	3.83	7.66	3.83	3.83	8.05	2.68	5.75	6.90	9.96	2.30	5.75	7.66	3.07	1.53	5.36	4.60	6.13	2.68	2.30	1
17695	5.31	8.85	8.85	3.54	3.54	2.65	1.77	3.54	7.08	5.31	4.42	7.96	4.42	1.77	3.54	7.96	8.85	5.31	3.54	1.77	1
17842	4.35	2.72	4.35	13.59	4.89	4.35	2.72	9.24	5.98	12.50	3.80	6.52	1.09	1.63	3.80	5.43	4.89	3.80	0.54	3.80	1
18048	5.08	3.91	7.81	5.86	2.73	6.64	2.73	4.30	5.47	10.94	2.73	5.08	8.20	3.52	1.95	6.64	5.47	5.08	2.73	3.13	1
20389	7.25	2.07	3.63	6.22	2.59	6.74	2.07	5.18	4.66	10.88	4.66	0.52	7.25	3.11	4.66	9.84	5.70	8.81	0.00	4.15	1
20714	7.89	0.72	5.26	8.13	5.74	4.78	1.91	6.46	6.94	11.24	3.59	4.31	4.07	5.26	2.63	6.22	6.46	5.74	0.48	2.15	1
22373	5.22	11.19	1.49	6.72	2.24	5.97	0.00	5.97	3.73	8.21	5.22	5.97	8.21	5.22	2.99	7.46	5.97	7.46	0.75	0.00	1
57426	6.45	4.30	5.38	11.83	3.23	6.45	2.15	6.45	9.68	13.98	1.08	3.23	4.30	0.00	1.08	3.23	8.60	5.38	0.00	3.23	1
66392	4.82	2.19	4.39	6.14	3.95	2.63	1.32	3.95	4.82	12.72	3.51	5.70	1.32	4.82	4.82	11.84	10.09	6.58	1.32	3.07	1
66996	4.29	1.65	4.62	3.63	2.64	4.62	3.30	6.27	5.61	12.87	1.32	4.95	5.28	3.96	3.63	9.90	9.57	7.26	2.31	2.31	1
67315	3.33	1.67	4.33	4.67	3.67	5.33	2.67	5.67	5.33	11.67	3.67	5.00	4.67	5.00	4.67	9.33	7.00	7.67	1.67	3.00	1
74188	5.39	2.49	2.90	7.47	6.22	2.07	2.90	7.47	7.47	13.28	2.07	4.15	4.15	2.90	5.39	9.54	5.81	3.32	2.07	2.90	1
77055	6.23	1.52	4.04	6.57	3.87	13.30	0.67	3.87	5.89	7.91	2.02	3.54	1.35	6.73	5.72	16.16	3.70	4.88	0.17	1.85	1
84543	6.25	3.47	3.47	4.17	3.47	3.47	1.39	6.25	6.94	9.03	4.86	6.25	5.56	5.56	0.00	8.33	6.25	9.72	1.39	4.17	1
100470	7.65	1.53	4.40	4.78	3.44	7.27	2.87	5.93	7.07	10.52	1.91	3.25	4.59	3.25	4.97	8.60	6.31	6.12	1.15	4.40	1
104002	6.12	2.33	3.79	5.83	4.08	9.33	1.46	4.66	5.54	7.58	2.62	7.58	4.66	2.04	5.25	6.41	4.96	7.29	3.50	4.96	1
109820	4.34	1.79	3.57	5.10	4.85	11.48	0.77	4.08	2.81	11.22	2.81	3.83	4.59	6.89	1.53	9.95	6.12	6.63	1.53	6.12	1
114871	4.45	1.69	2.75	5.30	3.18	5.72	2.54	4.66	4.03	11.86	1.91	4.66	3.39	4.66	5.93	9.11	9.32	8.47	1.69	4.66	1
233090	7.14	2.68	1.79	8.93	4.46	7.14	2.68	6.25	8.04	13.39	1.79	0.89	3.57	5.36	0.89	6.25	8.04	4.46	0.89	5.36	1

Figure B.5: **Amino acid matrix for mouse genes with broad and tissue-specific expression**. The amino acid matrix is shown for the 30 top/bottom genes that led to a prediction performance of AUC=1 as described in Chapter 5.2.1.1. The values 0 and 1 in the 'y' column indicate broad (green) and tissue-specific (orange) expression, respectively. The amino acid content is given as percentage (Xp) where X represents the amino acid.

A)

tissue-specific genes

EntrezGeneId	100470	104002	109820	114871	11699	12957	12965	12990	12991	13648	14473	14840	15458	16613	16615	16622	17695	17842	18048	20389	20714	22373	233090	57426	66392	66996	67315	74188	77055	84543
100470	2715	34	32	36	42	33	34	26	31	31	41	31	31	43	31	33	28	28	33	32	31	34	35	27	33	38	38	40	34	26
104002	34	1875	32	35	36	35	33	27	25	34	27	26	32	42	33	43	29	26	41	27	30	29	28	27	27	37	35	25	30	28
109820	32	32	2068	37	31	35	28	33	29	31	29	43	32	31	30	30	26	36	30	40	37	29	32	24	29	30	29	30	43	26
114871	36	35	37	2447	33	26	26	34	30	28	40	29	34	29	33	30	26	32	32	25	29	27	32	30	31	429	425	42	33	28
11699	42	36	31	33	1877	30	37	28	28	25	29	28	35	32	32	33	29	78	28	24	52	34	29	40	34	26	25	29	33	29
12957	33	35	35	26	30	1214	249	29	27	30	33	32	29	30	32	34	43	25	33	28	38	29	25	22	28	25	26	26	30	33
12965	34	33	28	26	37	249	988	35	27	26	31	45	25	30	21	33	37	30	29	26	28	35	29	24	25	32	33	30	33	20
12990	26	27	33	34	28	29	35	1568	47	30	33	47	30	26	31	26	30	38	29	28	39	28	24	24	30	28	29	26	31	24
12991	31	25	29	30	28	27	27	47	1141	36	32	36	26	36	26	30	20	28	32	34	27	37	30	22	35	31	27	28	41	30
13648	31	34	31	28	25	30	26	30	36	1425	34	30	26	1096	964	1101	29	28	1027	28	26	34	27	34	26	26	30	32	32	25
14473	41	27	29	40	29	33	31	33	32	34	2508	32	42	30	30	32	34	28	34	31	39	30	46	36	31	35	30	42	36	28
14840	31	26	43	29	28	32	45	47	36	30	32	1722	29	34	33	29	38	33	30	26	30	41	31	36	31	32	27	35	30	29
15458	31	32	32	34	35	29	25	30	26	26	42	29	2558	31	39	29	32	24	27	29	32	30	37	25	33	28	27	34	31	30
16613	43	42	31	29	32	30	30	26	36	1096	30	34	31	1436	1023	1163	33	33	1064	26	30	28	26	21	28	28	34	31	33	32
16615	31	33	30	33	32	32	21	31	26	964	30	33	39	1023	1422	1010	27	33	907	28	28	27	28	29	23	35	40	31	28	30
16622	33	43	30	30	33	34	33	26	30	1101	32	29	29	1163	1010	1439	30	32	1150	33	27	33	35	33	28	27	27	31	33	31
17695	28	29	26	26	29	43	37	30	20	29	34	38	32	33	27	30	644	22	29	27	25	26	29	22	31	30	21	29	24	32
17842	28	26	36	32	78	25	30	38	28	28	28	33	24	33	33	32	22	951	28	27	30	28	24	31	26	27	26	27	30	28
18048	33	41	30	32	28	33	29	29	32	1027	34	30	27	1064	907	1150	29	28	1416	27	27	33	27	26	26	28	31	35	33	31
20389	32	27	40	25	24	28	26	28	34	28	31	26	29	26	28	33	27	27	27	982	34	29	23	25	28	27	26	25	33	26
20714	31	30	37	29	52	38	28	39	27	26	39	30	32	30	28	27	25	30	27	34	2117	26	32	33	27	29	24	27	30	26
22373	34	29	29	27	34	29	35	28	37	34	30	41	30	28	27	33	26	28	33	29	26	733	27	26	24	23	25	28	26	26
233090	35	28	32	32	29	25	29	24	30	27	46	31	37	26	28	35	29	24	27	23	32	27	581	33	25	24	21	29	38	22
57426	27	27	24	30	40	22	24	24	22	34	36	36	25	21	29	33	22	31	26	25	33	26	33	485	26	28	32	28	29	21
66392	33	27	29	31	34	28	25	30	35	26	31	31	33	28	23	28	31	26	26	28	27	24	25	26	1154	30	27	116	50	27
66996	38	37	30	429	26	25	32	28	31	26	35	32	28	28	35	27	30	27	28	27	29	23	24	28	30	1581	1024	32	35	24
67315	38	35	29	425	25	26	33	29	27	30	30	27	27	34	40	27	21	26	31	26	24	25	21	32	27	1024	1562	35	26	24
74188	40	25	30	42	29	26	30	26	28	32	42	35	34	31	31	31	29	27	35	25	27	28	29	28	116	32	35	1257	39	32
77055	34	30	43	33	33	30	33	31	41	32	36	30	31	33	28	33	24	30	33	33	30	26	38	29	50	35	26	39	2977	33
84543	26	28	26	28	29	33	20	24	30	25	28	29	30	32	30	31	32	28	31	26	26	26	22	21	27	24	24	32	33	753

tissue-specific genes

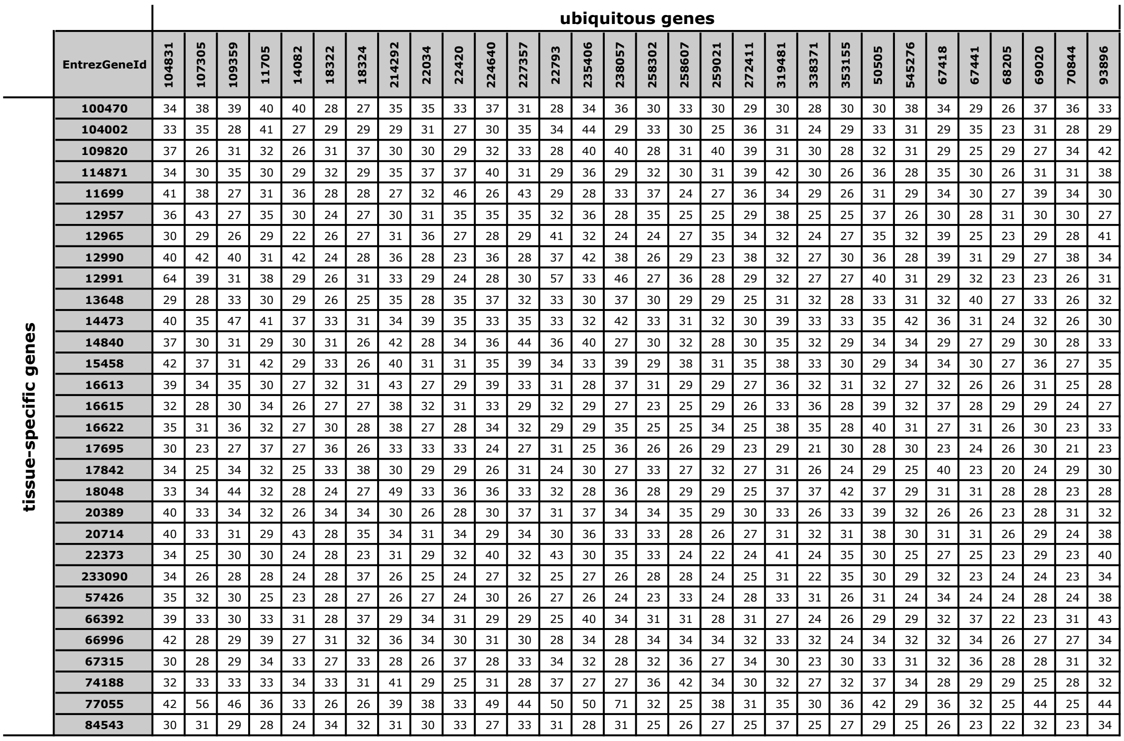
B)

ubiquitous genes

EntrezGeneId	104831	107305	109359	11705	14082	18322	18324	214292	22034	22420	224640	227357	22793	235406	238057	258302	258607	259021	272411	319481	338371	353155	50505	545276	67418	67441	68205	69020	70844	93896
104831	8881	104	44	53	34	35	31	39	41	39	49	58	124	41	44	42	40	33	46	36	43	50	38	40	43	45	30	40	36	38
107305	104	1899	36	54	43	28	24	32	34	42	56	44	83	33	38	25	34	23	35	41	36	57	39	43	40	32	26	26	27	32
109359	44	36	2138	31	30	28	30	30	41	34	33	41	37	32	36	25	29	33	24	47	27	32	39	45	37	32	26	29	30	38
11705	53	54	31	2897	31	35	31	38	34	38	47	44	40	36	94	34	30	30	35	41	37	44	48	34	33	38	26	37	31	32
14082	34	43	30	31	1021	31	36	33	30	42	28	32	29	41	30	28	29	35	27	36	34	33	37	31	31	29	27	32	25	31
18322	35	28	28	35	31	1623	709	36	28	31	29	29	29	27	29	616	776	663	30	33	28	36	31	29	24	29	27	28	25	34
18324	31	24	30	31	36	709	1564	35	36	31	28	29	28	30	27	563	957	771	29	30	28	31	33	31	33	28	26	23	28	34
214292	39	32	30	38	33	36	35	3269	31	27	31	34	35	31	34	34	41	34	37	40	35	37	39	36	31	31	30	30	28	36
22034	41	34	41	34	30	28	36	31	2862	34	28	38	36	32	36	29	33	33	31	46	31	32	35	37	36	30	25	48	26	37
22420	39	42	34	38	42	31	31	27	34	1993	45	46	45	34	35	33	27	28	34	34	29	67	41	36	29	26	28	33	41	39
224640	49	56	33	47	28	29	28	31	28	45	2698	46	42	37	42	31	26	31	36	36	45	42	35	34	40	32	32	34	42	44
227357	58	44	41	44	32	29	29	34	38	46	46	5353	47	37	54	30	32	33	45	41	40	47	40	45	35	33	29	40	34	33
22793	124	83	37	40	29	29	28	35	36	45	42	47	3115	35	34	33	32	26	34	39	47	49	38	35	35	40	25	45	31	36
235406	41	33	32	36	41	27	30	31	32	34	37	37	35	3047	38	31	26	28	35	41	35	42	35	32	34	39	27	33	27	31
238057	44	38	36	94	30	29	27	34	36	35	42	54	34	38	2420	32	27	27	39	31	28	43	42	35	31	30	27	32	36	31
258302	42	25	25	34	28	616	563	34	29	33	31	30	33	31	32	1671	586	498	27	32	32	38	29	34	27	33	26	29	27	39
258607	40	34	29	30	29	776	957	41	33	27	26	32	32	26	27	586	1560	767	27	34	29	33	30	29	31	26	27	26	26	39
259021	33	23	33	30	35	663	771	34	33	28	31	33	26	28	27	498	767	1583	28	32	27	42	38	30	30	22	26	26	29	40
272411	46	35	24	35	27	30	29	37	31	34	36	45	34	35	39	27	27	28	2082	38	38	26	36	33	27	34	23	28	35	38
319481	36	41	47	41	36	33	30	40	46	34	36	41	39	41	31	32	34	32	38	5334	28	40	49	30	38	30	28	35	30	35
338371	43	36	27	37	34	28	28	35	31	29	45	40	47	35	28	32	29	27	38	28	1350	32	30	41	34	30	25	32	22	30
353155	50	57	32	44	33	36	31	37	32	67	42	47	49	42	43	38	33	42	26	40	32	1478	34	34	29	26	29	25	28	35
50505	38	39	39	48	37	31	33	39	35	41	35	40	38	35	42	29	30	38	36	49	30	34	4656	33	41	36	30	34	33	38
545276	40	43	45	34	31	29	31	36	37	36	34	45	35	32	35	34	29	30	33	30	41	34	33	2272	35	39	27	36	34	28
67418	43	40	37	33	31	24	33	31	36	29	40	35	35	34	31	27	31	30	27	38	34	29	41	35	2613	34	33	36	29	34
67441	45	32	32	38	29	29	28	31	30	26	32	33	40	39	30	33	26	22	34	30	30	26	36	39	34	1049	23	27	24	30
68205	30	26	26	26	27	27	26	30	25	28	32	29	25	27	27	26	27	26	23	28	25	29	30	27	33	23	531	26	26	26
69020	40	26	29	37	32	28	23	30	48	33	34	40	45	33	32	29	26	26	28	35	32	25	34	36	36	27	26	1895	24	33
70844	36	27	30	31	25	25	28	28	26	41	42	34	31	27	36	27	26	29	35	30	22	28	33	34	29	24	26	24	796	32
93896	38	32	38	32	31	34	34	36	37	39	44	33	36	31	31	39	39	40	38	35	30	35	38	28	34	30	26	33	32	2746

ubiquitous genes

Figure B.6: **Sequence similarities within groups**. All-against-all sequence similarity scores within the classes are shown. The 30 top/bottom genes are listed that led to a prediction performance of AUC=1 as described in Chapter 5.2.1.1. Note that these scores were not used for the classifier achieving the AUC of 1, but the amino acid contents presented in Figure B.5 instead. The presented scores wore log normalised before used in any of the classifiers. Yellow cells flag alignment scores > 100. **A)** All-against-scores for the 30 most tissue-specific mouse genes **B)** All-against-scores for the 30 most broadly expressed mouse genes.



		ubiquitous genes																													
	EntrezGeneId	104831	107305	109359	11705	14082	18322	18324	214292	22034	22420	224640	227357	22793	235406	238057	258302	258607	259021	272411	319481	338371	353155	50505	545276	67418	67441	68205	69020	70844	93896
tissue-specific genes	100470	34	38	39	40	40	28	27	35	35	33	37	31	28	34	36	30	33	30	29	30	28	30	30	38	34	29	26	37	36	33
	104002	33	35	28	41	27	29	29	29	31	27	30	35	34	44	29	33	30	25	36	31	24	29	33	31	29	35	23	31	28	29
	109820	37	26	31	32	26	31	37	30	30	29	32	33	28	40	40	28	31	40	39	31	30	28	32	31	29	25	29	27	34	42
	114871	34	30	35	30	29	32	29	35	37	37	40	31	29	36	29	32	30	31	39	42	30	26	36	28	35	30	26	31	31	38
	11699	41	38	27	31	36	28	28	27	32	46	26	43	29	28	33	37	24	27	36	34	29	26	31	29	34	30	27	39	34	30
	12957	36	43	27	35	30	24	27	30	31	35	35	35	32	36	28	35	25	25	29	38	25	25	37	26	30	28	31	30	30	27
	12965	30	29	26	29	22	26	27	31	36	27	28	29	41	32	24	24	27	35	34	32	24	27	35	32	39	25	23	29	28	41
	12990	40	42	40	31	42	24	28	36	28	23	36	28	37	42	38	26	29	23	38	32	27	30	36	28	39	31	29	27	38	34
	12991	64	39	31	38	29	26	31	33	29	24	28	30	57	33	46	27	36	28	29	32	27	27	40	31	29	32	23	23	26	31
	13648	29	28	33	30	29	26	25	35	28	35	37	32	33	30	37	30	29	29	25	31	32	28	33	31	32	40	27	33	26	32
	14473	40	35	47	41	37	33	31	34	39	35	33	35	33	32	42	33	31	32	30	39	33	33	35	42	36	31	24	32	26	30
	14840	37	30	31	29	30	31	26	42	28	34	36	44	36	40	27	30	32	28	30	35	32	29	34	34	29	27	29	30	28	33
	15458	42	37	31	42	29	33	26	40	31	31	35	39	34	33	39	29	38	31	35	38	33	30	29	34	34	30	27	36	27	35
	16613	39	34	35	30	27	32	31	43	27	29	39	33	31	28	37	31	29	29	27	36	32	31	32	27	32	26	26	31	25	28
	16615	32	28	30	34	26	27	27	38	32	31	33	29	32	29	27	23	25	29	26	33	36	28	39	32	37	28	29	29	24	27
	16622	35	31	36	32	27	30	28	38	27	28	34	32	29	29	35	25	25	34	25	38	35	28	40	31	27	31	26	30	23	33
	17695	30	23	27	37	27	36	26	33	33	33	24	27	31	25	36	26	26	29	23	29	21	30	28	30	23	24	26	30	21	23
	17842	34	25	34	32	25	33	38	30	29	29	26	31	24	30	27	33	27	32	27	31	26	24	29	25	40	23	20	24	29	30
	18048	33	34	44	32	28	24	27	49	33	36	36	33	32	28	36	28	29	29	25	37	37	42	37	29	31	31	28	28	23	28
	20389	40	33	34	32	26	34	34	30	26	28	30	37	31	37	34	34	35	29	30	33	26	33	39	32	26	26	23	28	31	32
	20714	40	33	31	29	43	28	35	34	31	34	29	34	30	36	33	33	28	26	27	31	32	31	38	30	31	31	26	29	24	38
	22373	34	25	30	30	24	28	23	31	29	32	40	32	43	30	35	33	24	22	24	41	24	35	30	25	27	25	23	29	23	40
	233090	34	26	28	28	24	28	37	26	25	24	27	32	25	27	26	28	28	24	25	31	22	35	30	29	32	23	24	24	23	34
	57426	35	32	30	25	23	28	27	26	27	24	30	26	27	26	24	23	33	24	28	33	31	26	31	24	34	24	24	28	24	38
	66392	39	33	30	33	31	28	37	29	34	31	29	29	25	40	34	31	31	28	31	27	24	26	29	29	32	37	22	23	31	43
	66996	42	28	29	39	27	31	32	36	34	30	31	30	28	34	28	34	34	34	32	33	32	24	34	32	32	34	26	27	27	34
	67315	30	28	29	34	33	27	33	28	26	37	28	33	34	32	28	32	36	27	34	30	23	30	33	31	32	36	28	28	31	32
	74188	32	33	33	33	34	33	31	41	29	25	31	28	37	27	27	36	42	34	30	32	27	32	37	34	28	29	29	25	28	32
	77055	42	56	46	36	33	26	26	39	38	33	49	44	50	50	71	32	25	38	31	35	30	36	42	29	36	32	25	44	25	44
	84543	30	31	29	28	24	34	32	31	30	33	27	33	31	28	31	25	26	27	25	37	25	27	29	25	26	23	22	32	23	34

Figure B.7: **Sequence similarities across groups**. Sequence similarity scores across the classes are shown. The 30 top/bottom genes are listed that led to a prediction performance of AUC=1 as described in Chapter 5.2.1.1. Note that these scores were not used for the classifier achieving the AUC of 1, but the amino acid contents presented in Figure B.5 instead. The presented scores wore log normalised before used in any of the classifiers.

Appendix C

Q values for GO terms discussed in Chapter 6

Table C.1: Supplementary data for Chapter 2.2.2 (downregulated categories). Overrepresented GO terms in the tissue specificity bins, or their combinations, are shown. Four different scores are presented: (1) EASE score (2) p values resulting from Fisher's exact test (3) q values estimated from the EASE scores (4) q values estimated from the p values in (2). All terms reported achieve an EASE score < 0.05. BP= Biological Process; CC=Cellular Component; MF=Molecular Function.

GO term	**(1)**	**(2)**	**(3)**	**(4)**
	EASE	**Fisher's**	**q value**	**q value**
	score	**p value**	**using (1)**	**using (2)**
Bin all				
BP GO:0010004 gastrulation involving germ band extension	4.50E-04	4.13E-05	1.44E-01	4.91E-03
BP GO:0042594 response to starvation	4.82E-04	1.22E-05	1.44E-01	4.33E-03
CC GO:0042600 chorion	8.37E-04	5.84E-05	1.55E-01	5.20E-03
MF GO:0005213 structural constituent of chorion	1.04E-03	3.81E-05	1.55E-01	4.91E-03
BP GO:0001708 cell fate specification	2.43E-03	4.30E-04	3.03E-01	2.78E-02
MF GO:0003899 DNA-directed RNA polymerase activity	4.81E-03	5.83E-04	5.38E-01	3.19E-02
BP GO:0008286 insulin receptor signaling pathway	7.81E-03	6.81E-04	6.18E-01	3.40E-02
BP GO:0046112 nucleobase biosynthetic process	8.53E-03	2.86E-04	6.28E-01	2.07E-02
BP GO:0009888 tissue development	1.03E-02	4.87E-03	6.28E-01	8.66E-02
MF GO:0005160 transforming growth factor beta receptor binding	1.47E-02	7.16E-04	7.94E-01	3.40E-02
MF GO:0005158 insulin receptor binding	1.86E-02	1.05E-03	9.48E-01	3.99E-02
MF GO:0003676 nucleic acid binding	1.94E-02	1.38E-02	9.48E-01	1.42E-01
BP GO:0007219 Notch signaling pathway	2.10E-02	4.01E-03	9.48E-01	7.51E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	2.15E-02	2.84E-03	9.48E-01	6.12E-02
CC GO:0044452 nucleolar part	2.70E-02	1.91E-03	1.00E+00	5.05E-02
BP GO:0007179 transforming growth factor beta receptor signaling pathway	3.38E-02	5.39E-03	1.00E+00	9.12E-02
BP GO:0000915 cytokinesis; contractile ring formation	3.40E-02	2.81E-03	1.00E+00	6.12E-02

Continued on next page

Table C.1 – continued from previous page

CC GO:0005886 plasma membrane	4.10E-02	2.42E-02	1.00E+00	1.42E-01
BP GO:0045793 positive regulation of cell size	4.55E-02	4.49E-03	1.00E+00	8.19E-02
Bin 1				
BP GO:0007306 eggshell chorion formation	2.52E-02	4.27E-04	1.11E-01	1.80E-03
MF GO:0005213 structural constituent of chorion	4.76E-02	1.15E-03	1.11E-01	2.65E-03
CC GO:0042600 chorion	4.92E-02	1.64E-03	1.11E-01	2.65E-03
Bin 2				
CC GO:0042600 chorion	4.97E-02	5.67E-03	1.00E+00	8.84E-02
Bin 3				
Bin 4				
Bin 5				
Bin 6				
Bin 7				
Bin 8				
Bin 9				
Bin 10				
MF GO:0004888 transmembrane receptor activity	3.53E-02	4.07E-03	1.00E+00	4.17E-02
Bin 11				
Bin 12				
Bin 13				
CC GO:0005886 plasma membrane	2.14E-02	4.98E-03	1.00E+00	5.38E-02
MF GO:0004672 protein kinase activity	3.84E-02	7.26E-03	1.00E+00	5.38E-02
Bin 14				
Bin 15				
MF GO:0003676 nucleic acid binding	3.40E-02	1.67E-02	1.00E+00	1.20E-01
BP GO:0000915 cytokinesis; contractile ring formation	3.68E-02	3.63E-04	1.00E+00	2.36E-02
Bin 16				
BP GO:0031887 lipid particle transport along microtubule	3.07E-02	2.58E-04	1.00E+00	1.59E-02
BP GO:0035152 regulation of tube architecture; open tracheal system	4.58E-02	7.67E-04	1.00E+00	2.36E-02
Bin s1				
CC GO:0042600 chorion	5.76E-04	4.21E-05	4.42E-02	1.97E-03
MF GO:0005213 structural constituent of chorion	6.20E-04	2.16E-05	4.42E-02	1.97E-03
BP GO:0007306 eggshell chorion formation	1.40E-03	7.01E-05	7.49E-02	2.46E-03
Bin s2				
MF GO:0005213 structural constituent of chorion	7.37E-04	2.55E-05	5.05E-02	2.98E-03
CC GO:0042600 chorion	9.74E-04	7.60E-05	5.05E-02	2.98E-03
BP GO:0007306 eggshell chorion formation	3.86E-03	2.76E-04	1.37E-01	7.71E-03
Bin s3				
MF GO:0005213 structural constituent of chorion	6.41E-04	2.06E-05	7.87E-02	4.10E-03
CC GO:0042600 chorion	7.20E-04	4.99E-05	7.87E-02	4.10E-03
BP GO:0007306 eggshell chorion formation	4.15E-03	3.01E-04	1.86E-01	1.06E-02
Bin s4				
CC GO:0042600 chorion	6.05E-04	3.93E-05	7.75E-02	2.99E-03
MF GO:0005213 structural constituent of chorion	6.43E-04	2.04E-05	7.75E-02	2.99E-03
BP GO:0007306 eggshell chorion formation	8.22E-03	7.79E-04	3.96E-01	2.38E-02
Bin s5				

Continued on next page

Table C.1 – continued from previous page

CC GO:0042600 chorion	5.26E-04	3.24E-05	9.34E-02	3.59E-03
MF GO:0005213 structural constituent of chorion	7.35E-04	2.40E-05	9.79E-02	3.59E-03
BP GO:0007306 eggshell chorion formation	9.33E-03	9.14E-04	4.15E-01	3.50E-02
Bin s6				
CC GO:0042600 chorion	4.75E-04	2.82E-05	1.27E-01	3.21E-03
MF GO:0005213 structural constituent of chorion	8.67E-04	2.96E-05	1.27E-01	3.21E-03
BP GO:0007306 eggshell chorion formation	9.59E-03	9.41E-04	5.61E-01	4.28E-02
Bin s7				
CC GO:0042600 chorion	8.24E-04	5.71E-05	1.95E-01	6.06E-03
MF GO:0005213 structural constituent of chorion	1.05E-03	3.83E-05	1.95E-01	6.06E-03
BP GO:0007306 eggshell chorion formation	1.51E-02	1.75E-03	7.00E-01	6.20E-02
Bin s8				
CC GO:0042600 chorion	8.43E-04	5.95E-05	1.78E-01	7.76E-03
MF GO:0005213 structural constituent of chorion	9.47E-04	3.30E-05	1.78E-01	7.76E-03
BP GO:0007306 eggshell chorion formation	1.83E-02	2.29E-03	1.00E+00	8.13E-02
BP GO:0008283 cell proliferation	4.94E-02	1.54E-02	1.00E+00	1.98E-01
Bin s9				
CC GO:0042600 chorion	8.54E-04	6.11E-05	2.34E-01	8.79E-03
MF GO:0005213 structural constituent of chorion	8.73E-04	2.94E-05	2.34E-01	8.79E-03
BP GO:0007306 eggshell chorion formation	2.47E-02	3.49E-03	1.00E+00	1.25E-01
Bin s10				
CC GO:0042600 chorion	9.79E-04	7.21E-05	3.23E-01	1.06E-02
MF GO:0005213 structural constituent of chorion	1.05E-03	3.76E-05	3.23E-01	1.06E-02
BP GO:0007306 eggshell chorion formation	3.48E-02	5.63E-03	1.00E+00	1.55E-01
Bin s11				
CC GO:0042600 chorion	8.11E-04	5.65E-05	3.80E-01	8.96E-03
MF GO:0005213 structural constituent of chorion	1.05E-03	3.74E-05	3.80E-01	8.96E-03
BP GO:0007306 eggshell chorion formation	3.89E-02	6.59E-03	1.00E+00	1.66E-01
Bin s12				
CC GO:0042600 chorion	8.37E-04	5.85E-05	4.98E-01	9.02E-03
MF GO:0005213 structural constituent of chorion	1.49E-03	6.29E-05	5.92E-01	9.02E-03
BP GO:0009888 tissue development	3.93E-02	1.94E-02	1.00E+00	1.37E-01
BP GO:0051707 response to other organism	4.67E-02	1.41E-02	1.00E+00	1.37E-01
Bin s13				
CC GO:0042600 chorion	6.93E-04	4.58E-05	4.21E-01	8.11E-03
MF GO:0005213 structural constituent of chorion	1.26E-03	4.99E-05	4.21E-01	8.11E-03
BP GO:0042594 response to starvation	6.54E-03	1.76E-04	1.00E+00	2.04E-02
BP GO:0046112 nucleobase biosynthetic process	9.64E-03	3.45E-04	1.00E+00	2.55E-02
CC GO:0044452 nucleolar part	9.99E-03	3.67E-04	1.00E+00	2.55E-02
MF GO:0003899 DNA-directed RNA polymerase activity	4.29E-02	4.08E-03	1.00E+00	1.17E-01
BP GO:0048741 skeletal muscle fiber development	4.69E-02	8.57E-03	1.00E+00	1.31E-01
Bin s14				
CC GO:0042600 chorion	6.63E-04	4.31E-05	3.10E-01	6.32E-03
MF GO:0003899 DNA-directed RNA polymerase activity	1.07E-03	8.08E-05	3.18E-01	9.44E-03
MF GO:0005213 structural constituent of chorion	1.14E-03	4.33E-05	3.18E-01	6.32E-03
BP GO:0042594 response to starvation	5.97E-03	1.53E-04	9.14E-01	1.49E-02

Continued on next page

Table C.1 – continued from previous page

CC GO:0055029 nuclear DNA-directed RNA polymerase complex	6.53E-03	5.27E-04	9.14E-01	3.08E-02
BP GO:0046112 nucleobase biosynthetic process	8.81E-03	3.01E-04	1.00E+00	2.51E-02
MF GO:0003676 nucleic acid binding	2.04E-02	1.40E-02	1.00E+00	1.49E-01
CC GO:0044452 nucleolar part	2.22E-02	1.39E-03	1.00E+00	6.77E-02
CC GO:0005886 plasma membrane	3.29E-02	1.88E-02	1.00E+00	1.49E-01
Bin s15				
BP GO:0042594 response to starvation	4.42E-04	1.09E-05	1.99E-01	4.43E-03
CC GO:0042600 chorion	5.43E-04	3.34E-05	1.99E-01	4.43E-03
MF GO:0005213 structural constituent of chorion	9.77E-04	3.53E-05	2.85E-01	4.43E-03
MF GO:0003899 DNA-directed RNA polymerase activity	4.48E-03	5.32E-04	8.22E-01	4.18E-02
BP GO:0046112 nucleobase biosynthetic process	8.06E-03	2.62E-04	1.00E+00	2.75E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.89E-02	2.38E-03	1.00E+00	7.86E-02
MF GO:0003676 nucleic acid binding	2.09E-02	1.48E-02	1.00E+00	1.35E-01
CC GO:0044452 nucleolar part	2.46E-02	1.67E-03	1.00E+00	7.68E-02
CC GO:0005886 plasma membrane	3.52E-02	2.01E-02	1.00E+00	1.35E-01
BP GO:0045793 positive regulation of cell size	4.31E-02	4.14E-03	1.00E+00	1.18E-01
Bin u1				
BP GO:0031887 lipid particle transport along microtubule	2.96E-02	2.40E-04	1.00E+00	1.46E-02
BP GO:0035152 regulation of tube architecture; open tracheal system	4.42E-02	7.14E-04	1.00E+00	2.17E-02
Bin u2				
MF GO:0003676 nucleic acid binding	3.24E-02	1.85E-02	1.00E+00	4.51E-02
BP GO:0031887 lipid particle transport along microtubule	3.45E-02	3.11E-04	1.00E+00	2.37E-02
Bin u3				
Bin u4				
BP GO:0042594 response to starvation	3.89E-03	8.10E-05	1.00E+00	1.37E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.09E-02	1.13E-03	1.00E+00	5.24E-02
MF GO:0003899 DNA-directed RNA polymerase activity	1.10E-02	1.16E-03	1.00E+00	5.24E-02
BP GO:0008340 determination of adult life span	3.28E-02	5.31E-03	1.00E+00	9.31E-02
MF GO:0003676 nucleic acid binding	3.90E-02	2.57E-02	1.00E+00	1.08E-01
CC GO:0005886 plasma membrane	3.92E-02	1.62E-02	1.00E+00	9.31E-02
MF GO:0004672 protein kinase activity	4.57E-02	1.61E-02	1.00E+00	9.31E-02
Bin u5				
BP GO:0042594 response to starvation	1.56E-04	2.45E-06	7.77E-02	4.56E-04
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.19E-02	1.28E-03	1.00E+00	3.98E-02
MF GO:0003899 DNA-directed RNA polymerase activity	1.42E-02	1.64E-03	1.00E+00	4.69E-02
CC GO:0044452 nucleolar part	1.79E-02	1.03E-03	1.00E+00	3.98E-02
BP GO:0008286 insulin receptor signaling pathway	2.00E-02	1.24E-03	1.00E+00	3.98E-02
BP GO:0045793 positive regulation of cell size	2.77E-02	2.09E-03	1.00E+00	5.51E-02
BP GO:0001558 regulation of cell growth	4.10E-02	3.92E-03	1.00E+00	7.81E-02
BP GO:0008340 determination of adult life span	4.70E-02	8.78E-03	1.00E+00	7.81E-02
MF GO:0003676 nucleic acid binding	4.94E-02	3.40E-02	1.00E+00	1.06E-01
Bin u6				
BP GO:0042594 response to starvation	1.94E-04	3.27E-06	1.10E-01	6.87E-04
MF GO:0003899 DNA-directed RNA polymerase activity	2.46E-03	2.48E-04	5.89E-01	1.82E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.40E-02	1.59E-03	1.00E+00	4.88E-02
CC GO:0044452 nucleolar part	2.00E-02	1.22E-03	1.00E+00	4.88E-02

Continued on next page

Table C.1 – continued from previous page

MF GO:0003676 nucleic acid binding	2.63E-02	1.79E-02	1.00E+00	8.19E-02
BP GO:0008286 insulin receptor signaling pathway	2.72E-02	2.00E-03	1.00E+00	5.68E-02
BP GO:0045793 positive regulation of cell size	3.16E-02	2.56E-03	1.00E+00	6.74E-02
Bin u7				
BP GO:0042594 response to starvation	1.94E-04	3.26E-06	1.15E-01	8.15E-04
MF GO:0003899 DNA-directed RNA polymerase activity	2.43E-03	2.44E-04	7.23E-01	1.82E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.31E-02	1.46E-03	1.00E+00	4.81E-02
CC GO:0044452 nucleolar part	1.91E-02	1.14E-03	1.00E+00	4.81E-02
CC GO:0005886 plasma membrane	2.00E-02	9.19E-03	1.00E+00	8.03E-02
BP GO:0008286 insulin receptor signaling pathway	2.71E-02	2.00E-03	1.00E+00	6.12E-02
MF GO:0003676 nucleic acid binding	3.50E-02	2.44E-02	1.00E+00	8.64E-02
BP GO:0045793 positive regulation of cell size	3.64E-02	3.20E-03	1.00E+00	7.62E-02
Bin u8				
BP GO:0042594 response to starvation	1.93E-04	3.24E-06	1.15E-01	9.66E-04
MF GO:0003899 DNA-directed RNA polymerase activity	2.40E-03	2.40E-04	7.18E-01	2.01E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.26E-02	1.37E-03	1.00E+00	5.21E-02
CC GO:0044452 nucleolar part	1.85E-02	1.09E-03	1.00E+00	5.21E-02
MF GO:0003676 nucleic acid binding	2.42E-02	1.66E-02	1.00E+00	9.27E-02
BP GO:0008286 insulin receptor signaling pathway	2.70E-02	1.98E-03	1.00E+00	6.73E-02
CC GO:0005886 plasma membrane	3.11E-02	1.52E-02	1.00E+00	9.27E-02
BP GO:0045793 positive regulation of cell size	3.62E-02	3.18E-03	1.00E+00	8.88E-02
BP GO:0007219 Notch signaling pathway	4.29E-02	7.64E-03	1.00E+00	9.27E-02
Bin u9				
BP GO:0042594 response to starvation	2.30E-04	4.09E-06	1.47E-01	2.05E-03
MF GO:0003899 DNA-directed RNA polymerase activity	3.22E-03	3.49E-04	9.92E-01	3.46E-02
BP GO:0046112 nucleobase biosynthetic process	5.08E-03	1.20E-04	9.92E-01	2.01E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.45E-02	1.66E-03	1.00E+00	5.95E-02
CC GO:0044452 nucleolar part	2.05E-02	1.26E-03	1.00E+00	5.95E-02
BP GO:0008286 insulin receptor signaling pathway	3.01E-02	2.34E-03	1.00E+00	7.83E-02
MF GO:0003676 nucleic acid binding	3.13E-02	2.20E-02	1.00E+00	1.05E-01
BP GO:0051707 response to other organism	3.70E-02	1.05E-02	1.00E+00	1.05E-01
BP GO:0045793 positive regulation of cell size	4.03E-02	3.75E-03	1.00E+00	1.04E-01
Bin u10				
BP GO:0042594 response to starvation	2.34E-04	4.20E-06	1.55E-01	2.41E-03
MF GO:0003899 DNA-directed RNA polymerase activity	3.51E-03	3.89E-04	1.00E+00	4.05E-02
BP GO:0046112 nucleobase biosynthetic process	7.60E-03	2.41E-04	1.00E+00	4.05E-02
MF GO:0003676 nucleic acid binding	9.49E-03	6.38E-03	1.00E+00	1.24E-01
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.47E-02	1.69E-03	1.00E+00	6.93E-02
CC GO:0044452 nucleolar part	2.07E-02	1.28E-03	1.00E+00	6.93E-02
BP GO:0008286 insulin receptor signaling pathway	3.05E-02	2.39E-03	1.00E+00	9.11E-02
BP GO:0045793 positive regulation of cell size	4.08E-02	3.81E-03	1.00E+00	1.24E-01
Bin u11				
BP GO:0042594 response to starvation	4.19E-04	1.01E-05	2.84E-01	3.07E-03
MF GO:0003899 DNA-directed RNA polymerase activity	3.68E-03	4.13E-04	9.97E-01	4.95E-02
BP GO:0046112 nucleobase biosynthetic process	7.79E-03	2.50E-04	1.00E+00	3.78E-02
MF GO:0003676 nucleic acid binding	1.16E-02	7.89E-03	1.00E+00	1.32E-01

Continued on next page

Table C.1 – continued from previous page

CC GO:0005886 plasma membrane	1.21E-02	5.95E-03	1.00E+00	1.32E-01
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.49E-02	1.72E-03	1.00E+00	6.96E-02
CC GO:0044452 nucleolar part	2.08E-02	1.29E-03	1.00E+00	6.96E-02
BP GO:0008286 insulin receptor signaling pathway	3.13E-02	2.47E-03	1.00E+00	8.33E-02
BP GO:0045793 positive regulation of cell size	4.18E-02	3.95E-03	1.00E+00	1.14E-01
Bin u12				
BP GO:0042594 response to starvation	4.15E-04	9.99E-06	2.85E-01	2.83E-03
MF GO:0003899 DNA-directed RNA polymerase activity	4.20E-03	4.91E-04	9.61E-01	5.56E-02
BP GO:0046112 nucleobase biosynthetic process	7.74E-03	2.47E-04	1.00E+00	4.67E-02
MF GO:0003676 nucleic acid binding	1.12E-02	7.66E-03	1.00E+00	1.26E-01
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.51E-02	1.76E-03	1.00E+00	6.65E-02
CC GO:0005886 plasma membrane	1.98E-02	1.03E-02	1.00E+00	1.26E-01
CC GO:0044452 nucleolar part	2.11E-02	1.32E-03	1.00E+00	6.65E-02
BP GO:0008286 insulin receptor signaling pathway	3.11E-02	2.45E-03	1.00E+00	7.72E-02
BP GO:0007283 spermatogenesis	3.42E-02	9.48E-03	1.00E+00	1.26E-01
BP GO:0045793 positive regulation of cell size	4.15E-02	3.91E-03	1.00E+00	1.11E-01
Bin u13				
BP GO:0042594 response to starvation	4.09E-04	9.79E-06	2.84E-01	3.08E-03
MF GO:0003899 DNA-directed RNA polymerase activity	4.12E-03	4.79E-04	1.00E+00	5.89E-02
BP GO:0046112 nucleobase biosynthetic process	7.66E-03	2.43E-04	1.00E+00	4.79E-02
MF GO:0003676 nucleic acid binding	9.40E-03	6.39E-03	1.00E+00	1.28E-01
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.55E-02	1.81E-03	1.00E+00	7.13E-02
CC GO:0044452 nucleolar part	2.14E-02	1.35E-03	1.00E+00	7.13E-02
BP GO:0009888 tissue development	3.50E-02	1.75E-02	1.00E+00	1.28E-01
CC GO:0005886 plasma membrane	3.98E-02	2.23E-02	1.00E+00	1.28E-01
BP GO:0008286 insulin receptor signaling pathway	4.11E-02	3.85E-03	1.00E+00	1.26E-01
BP GO:0045793 positive regulation of cell size	4.11E-02	3.85E-03	1.00E+00	1.26E-01
BP GO:0007283 spermatogenesis	4.23E-02	1.24E-02	1.00E+00	1.28E-01
Bin u14				
BP GO:0042594 response to starvation	4.19E-04	1.01E-05	2.97E-01	3.77E-03
MF GO:0003899 DNA-directed RNA polymerase activity	4.22E-03	4.93E-04	1.00E+00	7.44E-02
BP GO:0046112 nucleobase biosynthetic process	7.78E-03	2.49E-04	1.00E+00	5.01E-02
MF GO:0003676 nucleic acid binding	1.02E-02	6.99E-03	1.00E+00	1.33E-01
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.65E-02	1.98E-03	1.00E+00	7.96E-02
CC GO:0044452 nucleolar part	2.24E-02	1.44E-03	1.00E+00	7.96E-02
CC GO:0005886 plasma membrane	3.04E-02	1.68E-02	1.00E+00	1.33E-01
BP GO:0009888 tissue development	4.13E-02	2.11E-02	1.00E+00	1.33E-01
BP GO:0045793 positive regulation of cell size	4.17E-02	3.94E-03	1.00E+00	1.33E-01
BP GO:0008286 insulin receptor signaling pathway	4.74E-02	4.84E-03	1.00E+00	1.33E-01
Bin u15				
BP GO:0042594 response to starvation	4.43E-04	1.09E-05	3.25E-01	5.14E-03
MF GO:0003899 DNA-directed RNA polymerase activity	4.40E-03	5.20E-04	1.00E+00	7.86E-02
BP GO:0046112 nucleobase biosynthetic process	8.07E-03	2.63E-04	1.00E+00	5.30E-02
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	1.85E-02	2.30E-03	1.00E+00	9.93E-02
MF GO:0003676 nucleic acid binding	1.88E-02	1.32E-02	1.00E+00	1.38E-01
CC GO:0044452 nucleolar part	2.42E-02	1.62E-03	1.00E+00	9.93E-02

Continued on next page

Table C.1 – continued from previous page

CC GO:0005886 plasma membrane	3.35E-02	1.91E-02	1.00E+00	1.38E-01
CC GO:0042600 chorion	4.00E-02	3.66E-03	1.00E+00	1.23E-01
BP GO:0045793 positive regulation of cell size	4.32E-02	4.15E-03	1.00E+00	1.27E-01

Table C.2: Supplementary data for Chapter 2.2.2 (upregulated categories). Overrepresented GO terms in the tissue specificity bins, or their combinations, are shown. Four different scores are presented: (1) EASE score (2) p values resulting from Fisher's exact test (3) q values estimated from the EASE scores (4) q values estimated from the p values in (2). All terms reported achieve an EASE score < 0.05. BP= Biological Process; CC=Cellular Component; MF=Molecular Function.

GO term	**(1)**	**(2)**	**(3)**	**(4)**
	EASE	**Fisher's**	**q value**	**q value**
	score	**p value**	**using (1)**	**using (2)**
Bin all				
BP GO:0006508 proteolysis	1.90E-20	6.90E-21	6.05E-18	2.20E-18
CC GO:0005792 microsome	4.79E-09	8.10E-10	5.39E-07	9.12E-08
MF GO:0004263 chymotrypsin activity	3.75E-08	2.08E-09	3.02E-06	2.09E-07
MF GO:0004497 monooxygenase activity	3.79E-08	9.41E-09	3.02E-06	7.50E-07
MF GO:0004295 trypsin activity	6.43E-08	1.22E-08	4.57E-06	9.31E-07
BP GO:0006869 lipid transport	5.27E-07	9.87E-08	2.72E-05	5.11E-06
CC GO:0005777 peroxisome	5.68E-07	6.49E-08	2.74E-05	3.88E-06
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	1.69E-06	2.28E-07	7.72E-05	1.09E-05
MF GO:0020037 heme Binding	1.91E-06	5.27E-07	8.19E-05	2.15E-05
MF GO:0008061 chitin Binding	2.33E-06	3.98E-07	9.49E-05	1.75E-05
CC GO:0005764 lysosome	3.66E-05	3.83E-06	1.23E-03	1.38E-04
MF GO:0005506 iron ion Binding	4.47E-05	1.69E-05	1.45E-03	5.39E-04
BP GO:0051189 prosthetic group metabolic process	2.11E-04	5.48E-05	6.42E-03	1.59E-03
BP GO:0008202 steroid metabolic process	2.83E-04	1.09E-04	8.33E-03	2.97E-03
BP GO:0006118 electron transport	3.28E-04	1.68E-04	9.53E-03	4.39E-03
BP GO:0006013 mannose metabolic process	4.07E-04	2.46E-05	1.16E-02	7.73E-04
MF GO:0004558 alpha-glucosidase activity	4.37E-04	4.09E-05	1.23E-02	1.22E-03
MF GO:0004559 alpha-mannosidase activity	4.70E-04	2.94E-05	1.30E-02	9.06E-04
MF GO:0004867 serine-type endopeptidase inhibitor activity	7.08E-04	1.98E-04	1.88E-02	5.04E-03
MF GO:0004035 alkaline phosphatase activity	8.72E-04	6.80E-05	2.26E-02	1.91E-03
MF GO:0004806 triacylglycerol lipase activity	9.19E-04	1.62E-04	2.35E-02	4.30E-03
MF GO:0042708 elastase activity	1.48E-03	1.38E-04	3.64E-02	3.73E-03
MF GO:0005344 oxygen transporter activity	1.63E-03	9.16E-05	3.84E-02	2.54E-03
MF GO:0008010 structural constituent of chitin-based larval cuticle	4.39E-03	7.53E-04	9.76E-02	1.57E-02
MF GO:0008970 phospholipase A1 activity	5.10E-03	4.73E-04	1.08E-01	1.11E-02
MF GO:0004179 membrane alanyl aminopeptidase activity	5.81E-03	1.07E-03	1.15E-01	1.97E-02
MF GO:0004182 carboxypeptidase A activity	5.81E-03	1.07E-03	1.15E-01	1.97E-02
BP GO:0006032 chitin catabolic process	6.31E-03	9.40E-04	1.17E-01	1.82E-02
BP GO:0001501 skeletal development	6.51E-03	1.24E-03	1.19E-01	2.20E-02

Continued on next page

Table C.2 – continued from previous page

MF GO:0004568 chitinase activity	7.18E-03	1.11E-03	1.29E-01	1.98E-02
MF GO:0016490 structural constituent of peritrophic membrane	7.18E-03	1.11E-03	1.29E-01	1.98E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	1.06E-02	2.77E-03	1.81E-01	4.61E-02
MF GO:0050809 diazepam Binding	1.09E-02	7.57E-04	1.83E-01	1.57E-02
BP GO:0008652 amino acid biosynthetic process	1.10E-02	3.98E-03	1.83E-01	6.24E-02
MF GO:0019204 nucleotide phosphatase activity	1.29E-02	3.40E-03	2.13E-01	5.42E-02
BP GO:0042049 cell acyl-CoA homeostasis	1.64E-02	1.46E-03	2.58E-01	2.57E-02
MF GO:0000062 acyl-CoA Binding	1.78E-02	1.64E-03	2.76E-01	2.83E-02
MF GO:0004364 glutathione transferase activity	1.84E-02	5.23E-03	2.79E-01	7.88E-02
BP GO:0005992 trehalose biosynthetic process	2.03E-02	6.10E-04	3.04E-01	1.34E-02
MF GO:0008336 gamma-butyrobetaine dioxygenase activity	2.16E-02	6.68E-04	3.17E-01	1.44E-02
MF GO:0016401 palmitoyl-CoA oxidase activity	2.16E-02	6.68E-04	3.17E-01	1.44E-02
MF GO:0009055 electron carrier activity	2.18E-02	1.25E-02	3.17E-01	1.48E-01
MF GO:0008533 astacin activity	2.67E-02	3.06E-03	3.78E-01	5.01E-02
MF GO:0005319 lipid transporter activity	3.15E-02	7.61E-03	4.36E-01	9.97E-02
MF GO:0008431 vitamin E Binding	3.74E-02	5.13E-03	5.01E-01	7.79E-02
MF GO:0005549 odorant Binding	3.84E-02	1.52E-02	5.02E-01	1.74E-01
MF GO:0005529 sugar Binding	3.85E-02	1.30E-02	5.02E-01	1.51E-01
BP GO:0048066 pigmentation during development	4.20E-02	1.58E-02	5.36E-01	1.76E-01
BP GO:0006800 oxygen and reactive oxygen species metabolic process	4.86E-02	1.74E-02	6.01E-01	1.87E-01
Bin 1				
BP GO:0006508 proteolysis	2.06E-02	1.01E-02	3.91E-01	1.40E-01
Bin 2				
BP GO:0006508 proteolysis	2.40E-13	4.56E-14	3.30E-11	6.28E-12
MF GO:0004295 trypsin activity	2.04E-03	2.60E-04	3.30E-02	4.21E-03
Bin 3				
Bin 4				
Bin 5				
Bin 6				
Bin 7				
MF GO:0004867 serine-type endopeptidase inhibitor activity	4.64E-03	2.09E-04	5.79E-01	2.61E-02
Bin 8				
Bin 9				
Bin 10				
Bin 11				
MF GO:0005506 iron ion Binding	2.99E-03	3.54E-04	9.17E-01	7.59E-02
CC GO:0005777 peroxisome	2.28E-02	7.77E-04	1.00E+00	8.34E-02
Bin 12				
CC GO:0005777 peroxisome	1.93E-02	8.48E-04	9.65E-01	1.31E-01
BP GO:0006118 electron transport	2.66E-02	5.42E-03	9.65E-01	2.13E-01
Bin 13				
CC GO:0005792 microsome	4.78E-02	4.71E-03	1.00E+00	2.04E-01
CC GO:0031966 mitochondrial membrane	4.78E-02	4.71E-03	1.00E+00	2.04E-01
Bin 14				
Bin 15				
Bin 16				
Bin s1				

Continued on next page

Table C.2 – continued from previous page

BP GO:0006508 proteolysis	2.55E-14	6.55E-15	1.98E-12	4.77E-13
MF GO:0004295 trypsin activity	2.69E-04	3.74E-05	5.44E-03	7.59E-04
MF GO:0004263 chymotrypsin activity	9.87E-04	1.31E-04	1.80E-02	2.27E-03
MF GO:0004806 triacylglycerol lipase activity	1.40E-03	6.60E-05	2.42E-02	1.27E-03
MF GO:0004035 alkaline phosphatase activity	1.64E-02	1.07E-03	2.14E-01	1.69E-02
MF GO:0019204 nucleotide phosphatase activity	1.64E-02	1.07E-03	2.14E-01	1.69E-02
BP GO:0001501 skeletal development	1.83E-02	1.23E-03	2.14E-01	1.85E-02
Bin s2				
BP GO:0006508 proteolysis	2.67E-15	7.15E-16	5.38E-13	1.44E-13
MF GO:0004295 trypsin activity	3.71E-05	5.45E-06	1.40E-03	2.36E-04
MF GO:0004263 chymotrypsin activity	1.59E-03	2.47E-04	4.81E-02	7.86E-03
MF GO:0004806 triacylglycerol lipase activity	8.36E-03	1.10E-03	2.02E-01	2.75E-02
MF GO:0019204 nucleotide phosphatase activity	1.11E-02	1.13E-03	2.50E-01	2.75E-02
CC GO:0005764 lysosome	2.42E-02	2.02E-03	4.19E-01	4.36E-02
MF GO:0004035 alkaline phosphatase activity	3.38E-02	4.08E-03	5.11E-01	6.86E-02
BP GO:0007498 mesoderm development	3.42E-02	9.17E-03	5.11E-01	1.32E-01
BP GO:0001501 skeletal development	3.53E-02	4.32E-03	5.11E-01	7.06E-02
MF GO:0008970 phospholipase A1 activity	4.69E-02	3.53E-03	6.38E-01	6.40E-02
Bin s3				
BP GO:0006508 proteolysis	3.69E-18	8.99E-19	8.88E-16	2.16E-16
MF GO:0004295 trypsin activity	4.77E-07	5.12E-08	2.46E-05	3.08E-06
MF GO:0004263 chymotrypsin activity	4.88E-04	6.26E-05	1.94E-02	2.66E-03
MF GO:0004806 triacylglycerol lipase activity	3.73E-03	4.05E-04	9.74E-02	1.08E-02
CC GO:0005764 lysosome	3.78E-03	2.31E-04	9.74E-02	6.96E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	1.23E-02	1.55E-03	2.78E-01	3.73E-02
MF GO:0004035 alkaline phosphatase activity	1.96E-02	1.96E-03	3.66E-01	4.29E-02
BP GO:0001501 skeletal development	2.00E-02	2.02E-03	3.66E-01	4.29E-02
MF GO:0019204 nucleotide phosphatase activity	2.31E-02	3.84E-03	3.97E-01	7.29E-02
MF GO:0008533 astacin activity	3.06E-02	1.91E-03	4.89E-01	4.29E-02
BP GO:0006013 mannose metabolic process	3.11E-02	1.95E-03	4.89E-01	4.29E-02
MF GO:0004558 alpha-glucosidase activity	3.83E-02	5.70E-03	5.65E-01	9.52E-02
Bin s4				
BP GO:0006508 proteolysis	3.42E-19	8.28E-20	1.43E-16	3.46E-17
MF GO:0004295 trypsin activity	1.04E-06	1.32E-07	6.24E-05	9.19E-06
MF GO:0004263 chymotrypsin activity	2.85E-04	3.35E-05	1.08E-02	1.33E-03
MF GO:0004806 triacylglycerol lipase activity	1.60E-03	1.87E-04	4.94E-02	6.00E-03
CC GO:0005764 lysosome	2.94E-03	1.67E-04	8.20E-02	5.57E-03
BP GO:0001501 skeletal development	3.66E-03	2.77E-04	9.87E-02	8.27E-03
MF GO:0004035 alkaline phosphatase activity	4.05E-03	3.15E-04	1.06E-01	9.07E-03
MF GO:0004558 alpha-glucosidase activity	9.10E-03	1.05E-03	1.90E-01	2.51E-02
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	9.10E-03	1.05E-03	1.90E-01	2.51E-02
MF GO:0019204 nucleotide phosphatase activity	9.96E-03	1.64E-03	2.03E-01	3.44E-02
BP GO:0006013 mannose metabolic process	2.36E-02	1.33E-03	4.30E-01	2.99E-02
BP GO:0006665 sphingolipid metabolic process	2.36E-02	1.33E-03	4.30E-01	2.99E-02
MF GO:0008533 astacin activity	2.52E-02	1.45E-03	4.48E-01	3.18E-02
BP GO:0006869 lipid transport	2.80E-02	3.74E-03	4.78E-01	6.80E-02
MF GO:0008970 phospholipase A1 activity	3.02E-02	4.14E-03	5.05E-01	7.36E-02
Bin s5				

Continued on next page

Table C.2 – continued from previous page

BP GO:0006508 proteolysis	6.64E-18	1.76E-18	2.18E-15	5.78E-16
MF GO:0004295 trypsin activity	5.63E-06	9.04E-07	3.69E-04	6.35E-05
CC GO:0005764 lysosome	8.65E-05	3.57E-06	4.25E-03	1.95E-04
MF GO:0004263 chymotrypsin activity	2.16E-04	2.43E-05	9.65E-03	1.09E-03
MF GO:0004806 triacylglycerol lipase activity	2.59E-03	3.61E-04	7.96E-02	1.18E-02
BP GO:0001501 skeletal development	3.03E-03	2.18E-04	9.02E-02	7.94E-03
MF GO:0004035 alkaline phosphatase activity	3.44E-03	2.56E-04	9.95E-02	8.97E-03
MF GO:0004558 alpha-glucosidase activity	7.77E-03	8.61E-04	2.01E-01	2.42E-02
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	7.77E-03	8.61E-04	2.01E-01	2.42E-02
MF GO:0019204 nucleotide phosphatase activity	1.41E-02	2.65E-03	3.46E-01	6.21E-02
BP GO:0006013 mannose metabolic process	2.10E-02	1.12E-03	4.70E-01	2.98E-02
BP GO:0006665 sphingolipid metabolic process	2.10E-02	1.12E-03	4.70E-01	2.98E-02
MF GO:0008533 astacin activity	2.27E-02	1.25E-03	4.86E-01	3.15E-02
BP GO:0006869 lipid transport	2.43E-02	3.09E-03	5.07E-01	6.89E-02
MF GO:0008970 phospholipase A1 activity	4.56E-02	7.91E-03	8.79E-01	1.52E-01
MF GO:0004559 alpha-mannosidase activity	4.90E-02	5.32E-03	9.26E-01	1.14E-01
Bin s6				
BP GO:0006508 proteolysis	9.88E-19	2.77E-19	5.69E-16	1.59E-16
MF GO:0004295 trypsin activity	1.12E-05	1.94E-06	8.31E-04	1.68E-04
MF GO:0004263 chymotrypsin activity	6.09E-05	6.27E-06	3.69E-03	4.01E-04
CC GO:0005764 lysosome	2.79E-04	1.78E-05	1.53E-02	1.02E-03
BP GO:0001501 skeletal development	7.31E-04	4.54E-05	3.51E-02	2.38E-03
MF GO:0004806 triacylglycerol lipase activity	2.64E-03	3.68E-04	9.81E-02	1.46E-02
MF GO:0004035 alkaline phosphatase activity	3.48E-03	2.59E-04	1.25E-01	1.11E-02
CC GO:0005792 microsome	4.63E-03	1.05E-03	1.48E-01	3.37E-02
MF GO:0004558 alpha-glucosidase activity	7.87E-03	8.72E-04	2.38E-01	3.05E-02
BP GO:0008652 amino acid biosynthetic process	1.24E-02	2.27E-03	3.49E-01	6.39E-02
MF GO:0019204 nucleotide phosphatase activity	1.43E-02	2.69E-03	3.91E-01	7.38E-02
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	1.50E-02	2.20E-03	4.02E-01	6.34E-02
BP GO:0006665 sphingolipid metabolic process	2.11E-02	1.13E-03	4.99E-01	3.51E-02
MF GO:0004497 monooxygenase activity	2.12E-02	8.25E-03	4.99E-01	1.70E-01
MF GO:0008533 astacin activity	2.29E-02	1.26E-03	5.17E-01	3.73E-02
BP GO:0006013 mannose metabolic process	4.57E-02	4.82E-03	9.46E-01	1.18E-01
MF GO:0008970 phospholipase A1 activity	4.60E-02	7.98E-03	9.46E-01	1.70E-01
Bin s7				
BP GO:0006508 proteolysis	3.06E-18	9.53E-19	3.87E-15	1.21E-15
MF GO:0004295 trypsin activity	6.92E-05	1.60E-05	4.57E-03	1.01E-03
MF GO:0004263 chymotrypsin activity	7.23E-05	7.60E-06	4.57E-03	5.06E-04
CC GO:0005764 lysosome	3.44E-04	2.27E-05	1.81E-02	1.31E-03
BP GO:0001501 skeletal development	8.36E-04	5.34E-05	4.06E-02	2.94E-03
BP GO:0008652 amino acid biosynthetic process	2.67E-03	4.61E-04	1.13E-01	2.01E-02
MF GO:0004035 alkaline phosphatase activity	3.79E-03	2.88E-04	1.45E-01	1.30E-02
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	5.02E-03	6.51E-04	1.81E-01	2.65E-02
MF GO:0004806 triacylglycerol lipase activity	5.35E-03	9.10E-04	1.88E-01	3.39E-02
CC GO:0005792 microsome	7.23E-03	2.19E-03	2.23E-01	6.74E-02
MF GO:0004558 alpha-glucosidase activity	8.53E-03	9.65E-04	2.45E-01	3.47E-02
MF GO:0019204 nucleotide phosphatase activity	1.56E-02	3.00E-03	4.03E-01	9.04E-02
MF GO:0004497 monooxygenase activity	2.02E-02	8.78E-03	4.81E-01	1.54E-01

Continued on next page

Table C.2 – continued from previous page

MF GO:0004867 serine-type endopeptidase inhibitor activity	2.05E-02	6.43E-03	4.81E-01	1.35E-01
MF GO:0005344 oxygen transporter activity	2.41E-02	1.35E-03	5.25E-01	4.50E-02
MF GO:0008533 astacin activity	2.41E-02	1.35E-03	5.25E-01	4.50E-02
BP GO:0006869 lipid transport	3.89E-02	8.39E-03	7.02E-01	1.54E-01
BP GO:0006013 mannose metabolic process	4.86E-02	5.26E-03	7.81E-01	1.29E-01
BP GO:0006665 sphingolipid metabolic process	4.86E-02	5.26E-03	7.81E-01	1.29E-01
MF GO:0008970 phospholipase A1 activity	4.88E-02	8.63E-03	7.81E-01	1.54E-01
MF GO:0016490 structural constituent of peritrophic membrane	4.88E-02	8.63E-03	7.81E-01	1.54E-01
MF GO:0042708 elastase activity	4.88E-02	8.63E-03	7.81E-01	1.54E-01
Bin s8				
BP GO:0006508 proteolysis	1.94E-19	6.07E-20	1.30E-16	4.08E-17
MF GO:0004295 trypsin activity	5.95E-06	1.29E-06	5.00E-04	1.16E-04
MF GO:0004263 chymotrypsin activity	9.76E-05	1.15E-05	6.56E-03	8.12E-04
CC GO:0005764 lysosome	5.25E-04	4.27E-05	2.82E-02	2.49E-03
MF GO:0004806 triacylglycerol lipase activity	1.18E-03	1.68E-04	5.88E-02	8.37E-03
BP GO:0001501 skeletal development	1.50E-03	1.33E-04	6.72E-02	6.86E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	2.10E-03	2.79E-04	8.29E-02	1.25E-02
CC GO:0005792 microsome	2.71E-03	8.23E-04	9.86E-02	3.07E-02
BP GO:0008652 amino acid biosynthetic process	4.78E-03	9.78E-04	1.43E-01	3.31E-02
BP GO:0006869 lipid transport	5.51E-03	9.85E-04	1.54E-01	3.31E-02
MF GO:0004497 monooxygenase activity	6.05E-03	2.50E-03	1.56E-01	6.22E-02
MF GO:0004035 alkaline phosphatase activity	6.63E-03	7.01E-04	1.68E-01	2.94E-02
MF GO:0004867 serine-type endopeptidase inhibitor activity	1.15E-02	3.51E-03	2.49E-01	7.73E-02
MF GO:0008061 chitin Binding	1.15E-02	3.51E-03	2.49E-01	7.73E-02
MF GO:0004558 alpha-glucosidase activity	1.27E-02	1.78E-03	2.59E-01	4.75E-02
MF GO:0005344 oxygen transporter activity	2.05E-02	1.08E-03	3.88E-01	3.45E-02
MF GO:0042708 elastase activity	2.17E-02	3.78E-03	4.05E-01	8.19E-02
MF GO:0019204 nucleotide phosphatase activity	3.94E-02	1.05E-02	6.79E-01	1.66E-01
MF GO:0008970 phospholipase A1 activity	4.04E-02	6.70E-03	6.81E-01	1.13E-01
BP GO:0006013 mannose metabolic process	4.11E-02	4.13E-03	6.81E-01	8.55E-02
MF GO:0008533 astacin activity	4.44E-02	4.61E-03	7.19E-01	8.95E-02
MF GO:0016814 hydrolase activity; acting on carbon-nitrogen ...	4.44E-02	4.61E-03	7.19E-01	8.95E-02
Bin s9				
BP GO:0006508 proteolysis	2.98E-18	1.01E-18	1.40E-15	4.72E-16
MF GO:0004295 trypsin activity	1.16E-05	2.73E-06	8.59E-04	2.02E-04
MF GO:0004263 chymotrypsin activity	6.61E-05	7.37E-06	4.64E-03	5.18E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	1.36E-04	1.76E-05	8.34E-03	1.08E-03
CC GO:0005764 lysosome	1.97E-04	1.72E-05	1.07E-02	1.08E-03
BP GO:0006869 lipid transport	1.22E-03	2.17E-04	4.19E-02	8.46E-03
MF GO:0004806 triacylglycerol lipase activity	1.59E-03	2.49E-04	5.18E-02	9.38E-03
CC GO:0005792 microsome	2.33E-03	7.35E-04	6.96E-02	2.07E-02
BP GO:0001501 skeletal development	2.61E-03	2.93E-04	7.55E-02	1.03E-02
MF GO:0004497 monooxygenase activity	4.43E-03	1.83E-03	1.17E-01	4.35E-02
BP GO:0008652 amino acid biosynthetic process	4.59E-03	1.08E-03	1.19E-01	2.80E-02
MF GO:0008061 chitin Binding	4.77E-03	1.36E-03	1.21E-01	3.41E-02
MF GO:0004035 alkaline phosphatase activity	5.43E-03	5.46E-04	1.34E-01	1.70E-02
BP GO:0006013 mannose metabolic process	9.17E-03	7.16E-04	2.05E-01	2.07E-02
MF GO:0004558 alpha-glucosidase activity	1.05E-02	1.40E-03	2.30E-01	3.41E-02

Continued on next page

Table C.2 – continued from previous page

MF GO:0005344 oxygen transporter activity	1.81E-02	9.02E-04	3.62E-01	2.39E-02
MF GO:0004559 alpha-mannosidase activity	2.01E-02	2.40E-03	3.82E-01	5.53E-02
BP GO:0051189 prosthetic group metabolic process	2.51E-02	9.66E-03	4.58E-01	1.58E-01
MF GO:0042708 elastase activity	2.85E-02	5.61E-03	5.13E-01	9.95E-02
MF GO:0019204 nucleotide phosphatase activity	3.23E-02	8.17E-03	5.61E-01	1.40E-01
MF GO:0008970 phospholipase A1 activity	3.48E-02	5.49E-03	5.73E-01	9.95E-02
MF GO:0004867 serine-type endopeptidase inhibitor activity	3.84E-02	1.51E-02	6.07E-01	2.18E-01
MF GO:0008533 astacin activity	3.94E-02	3.89E-03	6.08E-01	8.14E-02
MF GO:0016903 oxidoreductase activity; acting on the aldehyde . . .	3.94E-02	3.89E-03	6.08E-01	8.14E-02
Bin s10				
BP GO:0006508 proteolysis	1.22E-17	4.34E-18	6.47E-15	2.29E-15
MF GO:0004295 trypsin activity	1.88E-05	4.67E-06	1.42E-03	3.52E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	2.53E-05	3.38E-06	1.82E-03	2.98E-04
MF GO:0004263 chymotrypsin activity	4.30E-05	4.53E-06	2.62E-03	3.52E-04
CC GO:0005764 lysosome	1.62E-04	1.37E-05	7.55E-03	7.73E-04
CC GO:0005792 microsome	5.03E-04	1.48E-04	1.95E-02	5.32E-03
BP GO:0008652 amino acid biosynthetic process	5.93E-04	1.20E-04	2.14E-02	4.92E-03
BP GO:0006869 lipid transport	7.30E-04	1.36E-04	2.41E-02	5.15E-03
CC GO:0005777 peroxisome	9.52E-04	1.28E-04	2.93E-02	4.93E-03
MF GO:0008061 chitin Binding	1.77E-03	4.66E-04	5.10E-02	1.34E-02
MF GO:0004497 monooxygenase activity	2.18E-03	8.88E-04	5.85E-02	2.20E-02
MF GO:0004806 triacylglycerol lipase activity	3.17E-03	6.03E-04	7.98E-02	1.68E-02
BP GO:0001501 skeletal development	6.73E-03	1.06E-03	1.54E-01	2.49E-02
MF GO:0005506 iron ion Binding	7.03E-03	3.31E-03	1.57E-01	6.81E-02
BP GO:0006013 mannose metabolic process	7.75E-03	5.74E-04	1.71E-01	1.62E-02
MF GO:0004035 alkaline phosphatase activity	8.50E-03	1.07E-03	1.82E-01	2.49E-02
MF GO:0004558 alpha-glucosidase activity	8.50E-03	1.07E-03	1.82E-01	2.49E-02
BP GO:0006118 electron transport	1.22E-02	6.71E-03	2.42E-01	1.02E-01
MF GO:0020037 heme Binding	1.42E-02	6.37E-03	2.74E-01	1.00E-01
MF GO:0004559 alpha-mannosidase activity	1.69E-02	1.91E-03	3.05E-01	4.21E-02
BP GO:0006032 chitin catabolic process	2.15E-02	3.94E-03	3.60E-01	7.34E-02
MF GO:0004568 chitinase activity	2.34E-02	4.37E-03	3.82E-01	7.43E-02
MF GO:0042708 elastase activity	2.34E-02	4.37E-03	3.82E-01	7.43E-02
BP GO:0008202 steroid metabolic process	2.61E-02	1.31E-02	4.22E-01	1.85E-01
MF GO:0008970 phospholipase A1 activity	2.95E-02	4.41E-03	4.68E-01	7.43E-02
BP GO:0051189 prosthetic group metabolic process	3.35E-02	1.37E-02	5.16E-01	1.92E-01
MF GO:0004867 serine-type endopeptidase inhibitor activity	3.43E-02	1.47E-02	5.21E-01	2.01E-01
MF GO:0005344 oxygen transporter activity	3.45E-02	3.23E-03	5.21E-01	6.73E-02
Bin s11				
BP GO:0006508 proteolysis	9.07E-17	3.38E-17	5.12E-14	1.90E-14
CC GO:0005777 peroxisome	1.03E-05	1.06E-06	8.30E-04	9.92E-05
MF GO:0004263 chymotrypsin activity	2.09E-05	2.00E-06	1.36E-03	1.70E-04
MF GO:0004295 trypsin activity	2.39E-05	6.20E-06	1.50E-03	3.89E-04
BP GO:0006869 lipid transport	5.17E-05	8.17E-06	2.86E-03	4.76E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	6.45E-05	1.04E-05	3.22E-03	5.89E-04
CC GO:0005764 lysosome	2.02E-04	1.89E-05	8.77E-03	9.40E-04
MF GO:0004497 monooxygenase activity	8.99E-04	3.59E-04	2.98E-02	1.17E-02
CC GO:0005792 microsome	1.11E-03	3.59E-04	3.47E-02	1.17E-02

Continued on next page

Table C.2 – continued from previous page

MF GO:0004558 alpha-glucosidase activity	1.48E-03	1.47E-04	4.39E-02	5.78E-03
BP GO:0006118 electron transport	1.52E-03	7.58E-04	4.42E-02	2.00E-02
MF GO:0008061 chitin Binding	1.62E-03	4.61E-04	4.64E-02	1.35E-02
BP GO:0008652 amino acid biosynthetic process	1.95E-03	4.95E-04	5.51E-02	1.40E-02
MF GO:0004806 triacylglycerol lipase activity	4.40E-03	9.26E-04	1.14E-01	2.34E-02
BP GO:0001501 skeletal development	4.43E-03	6.39E-04	1.14E-01	1.74E-02
BP GO:0006013 mannose metabolic process	5.71E-03	3.86E-04	1.35E-01	1.23E-02
MF GO:0020037 heme Binding	5.74E-03	2.47E-03	1.35E-01	5.36E-02
MF GO:0005506 iron ion Binding	5.89E-03	2.80E-03	1.36E-01	5.85E-02
MF GO:0004035 alkaline phosphatase activity	5.97E-03	6.86E-04	1.36E-01	1.84E-02
MF GO:0005344 oxygen transporter activity	6.19E-03	4.28E-04	1.40E-01	1.29E-02
BP GO:0051189 prosthetic group metabolic process	8.86E-03	3.08E-03	1.85E-01	6.12E-02
MF GO:0004559 alpha-mannosidase activity	1.27E-02	1.31E-03	2.59E-01	3.17E-02
MF GO:0042708 elastase activity	1.68E-02	2.87E-03	3.23E-01	5.85E-02
MF GO:0008970 phospholipase A1 activity	2.24E-02	3.06E-03	4.03E-01	6.12E-02
BP GO:0006032 chitin catabolic process	2.32E-02	4.48E-03	4.03E-01	7.66E-02
MF GO:0004867 serine-type endopeptidase inhibitor activity	2.77E-02	1.15E-02	4.59E-01	1.73E-01
MF GO:0004568 chitinase activity	3.60E-02	8.11E-03	5.75E-01	1.29E-01
BP GO:0008202 steroid metabolic process	4.37E-02	2.35E-02	6.78E-01	2.83E-01
MF GO:0008533 astacin activity	4.92E-02	6.24E-03	7.50E-01	1.02E-01
Bin s12				
BP GO:0006508 proteolysis	3.12E-17	1.17E-17	1.87E-14	7.00E-15
CC GO:0005777 peroxisome	9.47E-07	9.88E-08	9.47E-05	9.88E-06
MF GO:0004295 trypsin activity	6.80E-06	1.61E-06	5.32E-04	1.45E-04
MF GO:0004263 chymotrypsin activity	7.98E-06	6.79E-07	5.87E-04	6.43E-05
BP GO:0006869 lipid transport	2.63E-05	5.26E-06	1.43E-03	3.15E-04
CC GO:0005792 microsome	5.23E-05	1.45E-05	2.48E-03	6.88E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	6.03E-05	1.01E-05	2.79E-03	5.34E-04
MF GO:0004497 monooxygenase activity	1.32E-04	4.79E-05	5.73E-03	1.96E-03
CC GO:0005764 lysosome	2.03E-04	1.99E-05	7.77E-03	8.95E-04
BP GO:0006118 electron transport	5.26E-04	2.57E-04	1.79E-02	7.97E-03
MF GO:0008061 chitin Binding	7.96E-04	2.10E-04	2.51E-02	6.87E-03
MF GO:0004558 alpha-glucosidase activity	8.32E-04	7.33E-05	2.58E-02	2.81E-03
MF GO:0020037 heme Binding	1.14E-03	4.46E-04	3.31E-02	1.25E-02
MF GO:0005506 iron ion Binding	1.84E-03	8.26E-04	5.10E-02	2.13E-02
MF GO:0004035 alkaline phosphatase activity	3.72E-03	3.80E-04	8.94E-02	1.14E-02
MF GO:0005344 oxygen transporter activity	4.17E-03	2.56E-04	9.88E-02	7.97E-03
MF GO:0004806 triacylglycerol lipase activity	4.67E-03	1.03E-03	1.08E-01	2.50E-02
MF GO:0004867 serine-type endopeptidase inhibitor activity	5.75E-03	2.00E-03	1.30E-01	4.23E-02
BP GO:0001501 skeletal development	6.72E-03	1.15E-03	1.49E-01	2.71E-02
BP GO:0051189 prosthetic group metabolic process	7.54E-03	2.59E-03	1.66E-01	5.13E-02
MF GO:0042708 elastase activity	1.08E-02	1.64E-03	2.28E-01	3.73E-02
BP GO:0008652 amino acid biosynthetic process	1.15E-02	3.97E-03	2.35E-01	7.60E-02
BP GO:0008202 steroid metabolic process	1.18E-02	5.82E-03	2.38E-01	9.62E-02
BP GO:0006013 mannose metabolic process	1.46E-02	1.74E-03	2.86E-01	3.89E-02
MF GO:0004559 alpha-mannosidase activity	1.55E-02	1.88E-03	2.94E-01	4.08E-02
MF GO:0008970 phospholipase A1 activity	1.55E-02	1.88E-03	2.94E-01	4.08E-02
BP GO:0006032 chitin catabolic process	2.22E-02	4.37E-03	3.77E-01	7.71E-02

Continued on next page

Table C.2 – continued from previous page

MF GO:0004568 chitinase activity	2.37E-02	4.76E-03	3.99E-01	8.16E-02
MF GO:0008533 astacin activity	3.72E-02	4.20E-03	5.77E-01	7.71E-02
MF GO:0009055 electron carrier activity	4.06E-02	2.30E-02	6.25E-01	2.49E-01
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	4.37E-02	1.34E-02	6.50E-01	1.89E-01
MF GO:0019204 nucleotide phosphatase activity	4.60E-02	1.37E-02	6.78E-01	1.90E-01
BP GO:0006635 fatty acid beta-oxidation	4.95E-02	1.01E-02	7.08E-01	1.49E-01
BP GO:0005992 trehalose biosynthetic process	4.96E-02	2.45E-03	7.08E-01	4.96E-02
Bin s13				
BP GO:0006508 proteolysis	5.61E-19	2.02E-19	2.60E-16	9.33E-17
MF GO:0004295 trypsin activity	8.75E-07	1.80E-07	8.52E-05	1.59E-05
BP GO:0006869 lipid transport	9.72E-07	1.67E-07	9.00E-05	1.54E-05
CC GO:0005777 peroxisome	1.38E-06	1.61E-07	1.11E-04	1.54E-05
MF GO:0004263 chymotrypsin activity	1.95E-06	1.39E-07	1.38E-04	1.52E-05
CC GO:0005792 microsome	1.51E-05	3.82E-06	7.98E-04	2.02E-04
MF GO:0004497 monooxygenase activity	2.92E-05	9.66E-06	1.42E-03	4.47E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	3.20E-05	5.14E-06	1.45E-03	2.51E-04
CC GO:0005764 lysosome	1.28E-04	1.21E-05	4.94E-03	5.21E-04
MF GO:0020037 heme Binding	3.07E-04	1.09E-04	1.01E-02	3.40E-03
MF GO:0008061 chitin Binding	3.19E-04	7.68E-05	1.03E-02	2.63E-03
MF GO:0004558 alpha-glucosidase activity	3.58E-04	2.66E-05	1.10E-02	1.07E-03
BP GO:0006118 electron transport	4.59E-04	2.25E-04	1.37E-02	6.60E-03
MF GO:0005506 iron ion Binding	8.68E-04	3.72E-04	2.51E-02	1.07E-02
MF GO:0004035 alkaline phosphatase activity	1.86E-03	1.60E-04	5.14E-02	4.87E-03
MF GO:0005344 oxygen transporter activity	2.34E-03	1.21E-04	6.10E-02	3.74E-03
MF GO:0004867 serine-type endopeptidase inhibitor activity	2.79E-03	8.96E-04	7.07E-02	2.18E-02
BP GO:0051189 prosthetic group metabolic process	2.92E-03	9.44E-04	7.10E-02	2.21E-02
MF GO:0004806 triacylglycerol lipase activity	3.41E-03	7.30E-04	8.10E-02	1.93E-02
BP GO:0001501 skeletal development	4.65E-03	7.53E-04	1.05E-01	1.94E-02
MF GO:0042708 elastase activity	5.58E-03	7.16E-04	1.24E-01	1.92E-02
BP GO:0008202 steroid metabolic process	6.98E-03	3.31E-03	1.54E-01	6.01E-02
BP GO:0006013 mannose metabolic process	8.46E-03	8.52E-04	1.82E-01	2.10E-02
MF GO:0004559 alpha-mannosidase activity	9.00E-03	9.23E-04	1.89E-01	2.19E-02
MF GO:0008970 phospholipase A1 activity	9.00E-03	9.23E-04	1.89E-01	2.19E-02
MF GO:0004179 membrane alanyl aminopeptidase activity	1.02E-02	2.05E-03	2.08E-01	3.91E-02
BP GO:0006032 chitin catabolic process	1.18E-02	1.97E-03	2.26E-01	3.85E-02
MF GO:0004568 chitinase activity	1.27E-02	2.16E-03	2.36E-01	4.08E-02
BP GO:0008652 amino acid biosynthetic process	1.36E-02	4.93E-03	2.47E-01	8.61E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	1.54E-02	4.24E-03	2.74E-01	7.55E-02
MF GO:0009055 electron carrier activity	1.97E-02	1.06E-02	3.41E-01	1.57E-01
MF GO:0019204 nucleotide phosphatase activity	2.98E-02	8.03E-03	4.83E-01	1.23E-01
MF GO:0005319 lipid transporter activity	3.20E-02	7.39E-03	5.07E-01	1.16E-01
BP GO:0006800 oxygen and reactive oxygen species metabolic process	3.61E-02	1.17E-02	5.52E-01	1.71E-01
BP GO:0005992 trehalose biosynthetic process	3.70E-02	1.55E-03	5.52E-01	3.30E-02
MF GO:0016401 palmitoyl-CoA oxidase activity	3.82E-02	1.63E-03	5.66E-01	3.43E-02
MF GO:0008533 astacin activity	3.94E-02	4.99E-03	5.79E-01	8.63E-02
Bin s14				
BP GO:0006508 proteolysis	9.85E-20	3.52E-20	3.10E-17	1.11E-17
MF GO:0004295 trypsin activity	2.26E-07	4.33E-08	2.03E-05	3.71E-06

Continued on next page

Table C.2 – continued from previous page

BP GO:0006869 lipid transport	4.01E-07	6.62E-08	3.22E-05	5.43E-06
MF GO:0004263 chymotrypsin activity	4.01E-07	2.38E-08	3.22E-05	2.36E-06
CC GO:0005792 microsome	7.19E-07	1.53E-07	4.84E-05	1.07E-05
CC GO:0005777 peroxisome	8.28E-07	9.55E-08	5.21E-05	7.20E-06
MF GO:0004497 monooxygenase activity	9.81E-07	2.72E-07	5.73E-05	1.61E-05
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	8.77E-06	1.24E-06	4.14E-04	6.01E-05
MF GO:0008061 chitin Binding	1.57E-05	2.94E-06	6.43E-04	1.29E-04
MF GO:0020037 heme Binding	3.22E-05	9.99E-06	1.24E-03	3.70E-04
MF GO:0005506 iron ion Binding	1.85E-04	7.28E-05	6.21E-03	2.29E-03
BP GO:0006118 electron transport	2.17E-04	1.04E-04	7.18E-03	3.17E-03
CC GO:0005764 lysosome	2.59E-04	3.04E-05	8.14E-03	1.02E-03
MF GO:0004035 alkaline phosphatase activity	8.54E-04	6.13E-05	2.48E-02	1.96E-03
MF GO:0004558 alpha-glucosidase activity	8.74E-04	9.39E-05	2.50E-02	2.90E-03
MF GO:0004867 serine-type endopeptidase inhibitor activity	8.93E-04	2.52E-04	2.51E-02	7.00E-03
BP GO:0051189 prosthetic group metabolic process	1.08E-03	3.15E-04	3.01E-02	8.37E-03
MF GO:0005344 oxygen transporter activity	1.23E-03	5.30E-05	3.31E-02	1.72E-03
BP GO:0008202 steroid metabolic process	1.38E-03	5.80E-04	3.65E-02	1.39E-02
MF GO:0004806 triacylglycerol lipase activity	1.57E-03	2.99E-04	4.01E-02	8.06E-03
MF GO:0042708 elastase activity	2.64E-03	2.83E-04	6.31E-02	7.74E-03
BP GO:0001501 skeletal development	4.00E-03	6.53E-04	9.08E-02	1.54E-02
BP GO:0006013 mannose metabolic process	4.50E-03	3.76E-04	1.01E-01	9.86E-03
MF GO:0004559 alpha-mannosidase activity	4.88E-03	4.17E-04	1.06E-01	1.03E-02
MF GO:0008970 phospholipase A1 activity	4.88E-03	4.17E-04	1.06E-01	1.03E-02
BP GO:0006032 chitin catabolic process	5.67E-03	7.86E-04	1.15E-01	1.65E-02
MF GO:0004568 chitinase activity	6.22E-03	8.83E-04	1.25E-01	1.83E-02
MF GO:0004179 membrane alanyl aminopeptidase activity	8.23E-03	1.62E-03	1.59E-01	3.07E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	1.45E-02	4.00E-03	2.60E-01	6.74E-02
BP GO:0008652 amino acid biosynthetic process	1.52E-02	5.70E-03	2.70E-01	9.11E-02
MF GO:0009055 electron carrier activity	1.67E-02	9.04E-03	2.84E-01	1.29E-01
MF GO:0005319 lipid transporter activity	2.14E-02	4.50E-03	3.51E-01	7.38E-02
MF GO:0008533 astacin activity	2.52E-02	2.66E-03	4.03E-01	4.78E-02
BP GO:0005992 trehalose biosynthetic process	2.65E-02	9.21E-04	4.15E-01	1.87E-02
MF GO:0019204 nucleotide phosphatase activity	2.66E-02	7.15E-03	4.15E-01	1.10E-01
MF GO:0004182 carboxypeptidase A activity	2.73E-02	6.19E-03	4.17E-01	9.65E-02
MF GO:0016401 palmitoyl-CoA oxidase activity	2.76E-02	9.82E-04	4.17E-01	1.97E-02
MF GO:0016490 structural constituent of peritrophic membrane	3.33E-02	6.27E-03	4.83E-01	9.70E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	4.19E-02	1.43E-02	5.98E-01	1.71E-01
MF GO:0005529 sugar Binding	4.63E-02	1.62E-02	6.43E-01	1.87E-01
BP GO:0048066 pigmentation during development	4.64E-02	1.77E-02	6.43E-01	2.00E-01
BP GO:0042049 cell acyl-CoA homeostasis	4.96E-02	3.41E-03	6.68E-01	5.96E-02
Bin s15				
BP GO:0006508 proteolysis	2.41E-19	8.85E-20	7.61E-17	2.79E-17
MF GO:0004295 trypsin activity	4.88E-08	8.53E-09	4.36E-06	7.68E-07
CC GO:0005792 microsome	7.17E-08	1.30E-08	5.65E-06	1.07E-06
MF GO:0004263 chymotrypsin activity	1.14E-07	5.86E-09	8.41E-06	5.53E-07
MF GO:0004497 monooxygenase activity	1.16E-07	2.86E-08	8.41E-06	2.08E-06
CC GO:0005777 peroxisome	3.73E-07	4.06E-08	2.07E-05	2.74E-06
BP GO:0006869 lipid transport	1.55E-06	2.99E-07	7.32E-05	1.49E-05

Continued on next page

Table C.2 – continued from previous page

MF GO:0008061 chitin Binding	3.34E-06	5.44E-07	1.44E-04	2.39E-05
MF GO:0020037 heme Binding	5.83E-06	1.61E-06	2.34E-04	6.35E-05
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	7.45E-06	1.07E-06	2.87E-04	4.61E-05
MF GO:0005506 iron ion Binding	7.98E-05	3.00E-05	2.69E-03	9.46E-04
CC GO:0005764 lysosome	2.11E-04	2.46E-05	6.63E-03	8.16E-04
MF GO:0004867 serine-type endopeptidase inhibitor activity	3.07E-04	7.72E-05	9.50E-03	2.27E-03
BP GO:0008202 steroid metabolic process	3.18E-04	1.20E-04	9.70E-03	3.44E-03
MF GO:0004558 alpha-glucosidase activity	4.26E-04	3.97E-05	1.26E-02	1.23E-03
MF GO:0004035 alkaline phosphatase activity	4.59E-04	2.86E-05	1.33E-02	9.15E-04
BP GO:0051189 prosthetic group metabolic process	5.34E-04	1.44E-04	1.53E-02	4.01E-03
MF GO:0005344 oxygen transporter activity	7.38E-04	2.76E-05	2.08E-02	8.98E-04
MF GO:0004806 triacylglycerol lipase activity	8.90E-04	1.56E-04	2.44E-02	4.26E-03
BP GO:0006118 electron transport	1.19E-03	6.36E-04	3.18E-02	1.35E-02
MF GO:0042708 elastase activity	1.45E-03	1.35E-04	3.71E-02	3.80E-03
BP GO:0006013 mannose metabolic process	2.73E-03	1.97E-04	6.44E-02	5.32E-03
MF GO:0004559 alpha-mannosidase activity	2.99E-03	2.22E-04	6.88E-02	5.90E-03
BP GO:0001501 skeletal development	3.79E-03	6.30E-04	8.62E-02	1.35E-02
MF GO:0004179 membrane alanyl aminopeptidase activity	4.28E-03	7.31E-04	9.13E-02	1.46E-02
BP GO:0006032 chitin catabolic process	4.54E-03	6.11E-04	9.13E-02	1.33E-02
MF GO:0008970 phospholipase A1 activity	5.01E-03	4.63E-04	9.84E-02	1.12E-02
MF GO:0004568 chitinase activity	5.05E-03	6.97E-04	9.84E-02	1.44E-02
MF GO:0019204 nucleotide phosphatase activity	8.54E-03	2.05E-03	1.57E-01	3.72E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	8.99E-03	2.28E-03	1.63E-01	3.98E-02
BP GO:0008652 amino acid biosynthetic process	9.79E-03	3.47E-03	1.74E-01	5.66E-02
MF GO:0004182 carboxypeptidase A activity	1.61E-02	3.19E-03	2.65E-01	5.24E-02
MF GO:0008533 astacin activity	1.76E-02	1.61E-03	2.86E-01	2.99E-02
BP GO:0005992 trehalose biosynthetic process	2.04E-02	6.12E-04	3.19E-01	1.33E-02
MF GO:0016401 palmitoyl-CoA oxidase activity	2.14E-02	6.59E-04	3.30E-01	1.38E-02
MF GO:0016490 structural constituent of peritrophic membrane	2.15E-02	3.52E-03	3.30E-01	5.69E-02
MF GO:0005319 lipid transporter activity	3.09E-02	7.43E-03	4.49E-01	1.00E-01
MF GO:0005529 sugar Binding	3.29E-02	1.07E-02	4.71E-01	1.32E-01
MF GO:0008431 vitamin E Binding	3.70E-02	5.04E-03	5.12E-01	7.94E-02
BP GO:0048066 pigmentation during development	3.77E-02	1.39E-02	5.13E-01	1.60E-01
MF GO:0050809 diazepam Binding	4.03E-02	2.46E-03	5.33E-01	4.20E-02
MF GO:0009055 electron carrier activity	4.08E-02	2.42E-02	5.35E-01	2.25E-01
BP GO:0006800 oxygen and reactive oxygen species metabolic process	4.33E-02	1.51E-02	5.61E-01	1.72E-01
MF GO:0004364 glutathione transferase activity	4.77E-02	1.52E-02	6.06E-01	1.73E-01
Bin u1				
Bin u2				
Bin u3				
Bin u4				
CC GO:0005792 microsome	2.15E-03	1.21E-04	2.65E-01	2.12E-02
BP GO:0006869 lipid transport	2.69E-03	2.85E-04	2.65E-01	2.68E-02
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	3.24E-02	5.21E-03	7.79E-01	1.17E-01
BP GO:0008202 steroid metabolic process	3.59E-02	6.05E-03	8.35E-01	1.18E-01
MF GO:0004095 carnitine O-palmitoyltransferase activity	3.99E-02	4.13E-04	8.44E-01	2.88E-02
Bin u5				
BP GO:0006869 lipid transport	1.54E-04	1.56E-05	1.97E-02	2.33E-03

Continued on next page

Table C.2 – continued from previous page

CC GO:0005777 peroxisome	5.48E-04	3.52E-05	4.10E-02	3.51E-03
MF GO:0004497 monooxygenase activity	6.37E-03	8.43E-04	2.59E-01	2.91E-02
MF GO:0020037 heme Binding	7.21E-03	9.91E-04	2.59E-01	3.17E-02
CC GO:0005792 microsome	1.37E-02	1.55E-03	4.17E-01	4.35E-02
BP GO:0006118 electron transport	2.94E-02	1.12E-02	7.33E-01	1.55E-01
BP GO:0006633 fatty acid biosynthetic process	4.59E-02	4.58E-03	9.47E-01	8.56E-02
Bin u6				
CC GO:0005777 peroxisome	1.03E-06	5.35E-08	1.47E-04	1.25E-05
BP GO:0006869 lipid transport	1.44E-04	1.79E-05	1.10E-02	1.36E-03
MF GO:0020037 heme Binding	2.18E-04	2.90E-05	1.46E-02	1.94E-03
MF GO:0005506 iron ion Binding	3.40E-04	6.74E-05	1.94E-02	3.25E-03
CC GO:0005792 microsome	1.00E-03	1.12E-04	4.38E-02	4.73E-03
MF GO:0004497 monooxygenase activity	1.26E-03	1.91E-04	5.34E-02	7.03E-03
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	2.19E-03	3.73E-04	7.32E-02	1.06E-02
BP GO:0006633 fatty acid biosynthetic process	2.78E-03	2.75E-04	7.91E-02	8.69E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	3.28E-03	3.79E-05	8.31E-02	2.27E-03
BP GO:0006118 electron transport	4.85E-03	1.78E-03	1.17E-01	3.69E-02
BP GO:0008202 steroid metabolic process	1.04E-02	2.50E-03	2.04E-01	4.46E-02
BP GO:0006635 fatty acid beta-oxidation	1.61E-02	1.85E-03	2.97E-01	3.70E-02
MF GO:0004364 glutathione transferase activity	1.78E-02	2.12E-03	3.22E-01	3.95E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	3.65E-02	5.90E-03	6.12E-01	8.45E-02
MF GO:0009055 electron carrier activity	3.82E-02	1.50E-02	6.24E-01	1.66E-01
BP GO:0006098 pentose-phosphate shunt	3.94E-02	3.44E-03	6.24E-01	5.36E-02
Bin u7				
CC GO:0005777 peroxisome	4.49E-07	2.64E-08	8.10E-05	4.76E-06
BP GO:0006869 lipid transport	1.65E-05	2.07E-06	1.39E-03	1.87E-04
MF GO:0020037 heme Binding	3.82E-05	5.38E-06	2.41E-03	3.11E-04
MF GO:0005506 iron ion Binding	5.33E-04	1.24E-04	2.04E-02	3.97E-03
CC GO:0005792 microsome	7.12E-04	9.56E-05	2.57E-02	3.77E-03
MF GO:0004497 monooxygenase activity	1.06E-03	1.85E-04	3.51E-02	5.19E-03
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	1.84E-03	3.57E-04	5.66E-02	9.02E-03
BP GO:0006032 chitin catabolic process	2.63E-03	1.29E-04	6.79E-02	3.97E-03
MF GO:0004568 chitinase activity	2.86E-03	1.44E-04	7.22E-02	4.33E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	4.46E-03	6.02E-05	1.08E-01	2.62E-03
BP GO:0006633 fatty acid biosynthetic process	5.01E-03	5.83E-04	1.09E-01	1.34E-02
BP GO:0006118 electron transport	5.36E-03	2.17E-03	1.15E-01	3.75E-02
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	6.73E-03	8.58E-04	1.42E-01	1.84E-02
MF GO:0004364 glutathione transferase activity	8.07E-03	1.09E-03	1.62E-01	2.30E-02
BP GO:0008202 steroid metabolic process	8.58E-03	2.26E-03	1.64E-01	3.85E-02
MF GO:0008061 chitin Binding	1.56E-02	1.71E-03	2.74E-01	3.08E-02
BP GO:0006635 fatty acid beta-oxidation	2.90E-02	4.17E-03	4.43E-01	6.55E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	4.15E-02	7.07E-03	5.97E-01	9.30E-02
MF GO:0008010 structural constituent of chitin-based larval cuticle	4.58E-02	4.24E-03	6.35E-01	6.55E-02
MF GO:0009055 electron carrier activity	4.85E-02	2.11E-02	6.58E-01	2.10E-01
Bin u8				
CC GO:0005777 peroxisome	1.02E-06	6.78E-08	1.00E-04	7.14E-06
BP GO:0006869 lipid transport	1.26E-06	1.51E-07	1.15E-04	1.38E-05
MF GO:0020037 heme Binding	2.08E-06	3.00E-07	1.59E-04	2.41E-05

Continued on next page

Table C.2 – continued from previous page

MF GO:0004497 monooxygenase activity	5.76E-05	9.75E-06	3.04E-03	4.95E-04
CC GO:0005792 microsome	7.93E-05	1.08E-05	3.68E-03	5.09E-04
MF GO:0005506 iron ion Binding	8.07E-05	1.91E-05	3.68E-03	8.25E-04
BP GO:0008202 steroid metabolic process	2.88E-04	6.72E-05	1.13E-02	2.05E-03
BP GO:0006032 chitin catabolic process	5.25E-04	2.74E-05	1.63E-02	9.62E-04
MF GO:0004568 chitinase activity	6.22E-04	3.40E-05	1.88E-02	1.13E-03
MF GO:0008061 chitin Binding	6.32E-04	5.73E-05	1.88E-02	1.82E-03
BP GO:0006118 electron transport	7.19E-04	2.73E-04	2.10E-02	7.06E-03
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	1.02E-03	2.03E-04	2.73E-02	5.46E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	2.25E-03	2.87E-04	5.22E-02	7.15E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	5.62E-03	8.54E-05	1.18E-01	2.54E-03
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	9.74E-03	1.42E-03	1.96E-01	2.82E-02
BP GO:0006633 fatty acid biosynthetic process	1.04E-02	1.51E-03	2.00E-01	2.87E-02
MF GO:0009055 electron carrier activity	1.28E-02	5.21E-03	2.36E-01	7.84E-02
MF GO:0004364 glutathione transferase activity	1.67E-02	2.83E-03	2.98E-01	4.67E-02
MF GO:0016885 ligase activity; forming carbon-carbon bonds	1.77E-02	7.99E-04	3.04E-01	1.74E-02
BP GO:0048066 pigmentation during development	2.20E-02	4.11E-03	3.55E-01	6.40E-02
MF GO:0005319 lipid transporter activity	2.59E-02	3.46E-03	4.03E-01	5.57E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	2.83E-02	5.73E-03	4.30E-01	8.01E-02
MF GO:0016229 steroid dehydrogenase activity	3.50E-02	2.62E-03	5.02E-01	4.43E-02
MF GO:0015020 glucuronosyltransferase activity	4.21E-02	6.98E-03	5.84E-01	9.04E-02
BP GO:0006635 fatty acid beta-oxidation	4.34E-02	7.32E-03	5.94E-01	9.12E-02
Bin u9				
CC GO:0005777 peroxisome	4.39E-08	2.84E-09	4.88E-06	3.42E-07
BP GO:0006869 lipid transport	2.20E-06	3.09E-07	1.44E-04	2.24E-05
MF GO:0020037 heme Binding	2.84E-06	5.03E-07	1.64E-04	2.84E-05
MF GO:0008061 chitin Binding	6.21E-06	4.97E-07	3.23E-04	2.84E-05
MF GO:0005506 iron ion Binding	9.91E-06	2.43E-06	4.77E-04	1.10E-04
MF GO:0004497 monooxygenase activity	1.54E-05	2.98E-06	6.76E-04	1.27E-04
CC GO:0005792 microsome	4.20E-05	6.89E-06	1.68E-03	2.77E-04
BP GO:0006118 electron transport	2.40E-04	9.55E-05	8.08E-03	2.60E-03
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	9.08E-04	1.98E-04	2.85E-02	4.85E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	9.82E-04	1.31E-04	3.02E-02	3.47E-03
BP GO:0008202 steroid metabolic process	1.11E-03	3.25E-04	3.03E-02	7.57E-03
BP GO:0006032 chitin catabolic process	1.13E-03	7.28E-05	3.03E-02	2.06E-03
BP GO:0006800 oxygen and reactive oxygen species metabolic process	1.20E-03	2.07E-04	3.15E-02	4.99E-03
MF GO:0004568 chitinase activity	1.25E-03	8.29E-05	3.18E-02	2.30E-03
MF GO:0004867 serine-type endopeptidase inhibitor activity	1.69E-03	3.12E-04	4.08E-02	7.38E-03
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	4.99E-03	7.96E-04	1.05E-01	1.62E-02
BP GO:0048066 pigmentation during development	5.17E-03	1.01E-03	1.05E-01	2.00E-02
BP GO:0005992 trehalose biosynthetic process	7.68E-03	1.37E-04	1.42E-01	3.48E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	8.10E-03	1.49E-04	1.48E-01	3.71E-03
MF GO:0009055 electron carrier activity	8.20E-03	3.53E-03	1.48E-01	5.21E-02
MF GO:0004364 glutathione transferase activity	1.06E-02	2.00E-03	1.80E-01	3.56E-02
MF GO:0005319 lipid transporter activity	1.06E-02	1.49E-03	1.80E-01	2.80E-02
MF GO:0019203 carbohydrate phosphatase activity	1.56E-02	5.72E-04	2.48E-01	1.23E-02
BP GO:0006635 fatty acid beta-oxidation	1.74E-02	2.92E-03	2.70E-01	4.40E-02
BP GO:0006633 fatty acid biosynthetic process	2.06E-02	3.67E-03	3.11E-01	5.25E-02

Continued on next page

Table C.2 – continued from previous page

BP GO:0046483 heterocycle metabolic process	2.20E-02	8.72E-03	3.28E-01	1.00E-01
MF GO:0016885 ligase activity; forming carbon-carbon bonds	2.51E-02	1.37E-03	3.63E-01	2.65E-02
MF GO:0015020 glucuronosyltransferase activity	2.65E-02	5.11E-03	3.79E-01	6.84E-02
MF GO:0003995 acyl-CoA dehydrogenase activity	2.83E-02	3.82E-03	3.94E-01	5.36E-02
MF GO:0016209 antioxidant activity	3.08E-02	6.25E-03	4.23E-01	8.21E-02
BP GO:0006012 galactose metabolic process	3.46E-02	2.44E-03	4.59E-01	4.20E-02
BP GO:0006508 proteolysis	3.73E-02	2.36E-02	4.90E-01	2.07E-01
BP GO:0019731 antibacterial humoral response	4.22E-02	9.61E-03	5.40E-01	1.08E-01
BP GO:0006725 aromatic compound metabolic process	4.30E-02	1.91E-02	5.46E-01	1.78E-01
MF GO:0016408 C-acyltransferase activity	4.92E-02	4.44E-03	6.11E-01	6.17E-02
Bin u10				
CC GO:0005777 peroxisome	1.85E-07	1.41E-08	1.67E-05	1.28E-06
MF GO:0004497 monooxygenase activity	4.01E-07	7.36E-08	3.18E-05	5.95E-06
CC GO:0005792 microsome	1.23E-06	1.86E-07	7.78E-05	1.19E-05
MF GO:0020037 heme Binding	1.40E-06	2.67E-07	7.78E-05	1.47E-05
MF GO:0004867 serine-type endopeptidase inhibitor activity	1.41E-06	1.83E-07	7.78E-05	1.19E-05
MF GO:0005506 iron ion Binding	7.24E-06	1.89E-06	3.53E-04	8.94E-05
BP GO:0006869 lipid transport	1.15E-05	1.92E-06	5.20E-04	8.94E-05
MF GO:0008061 chitin Binding	2.16E-05	2.09E-06	8.73E-04	9.45E-05
BP GO:0006118 electron transport	3.01E-05	1.16E-05	1.16E-03	4.36E-04
BP GO:0008202 steroid metabolic process	5.55E-04	1.68E-04	1.81E-02	5.15E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	2.25E-03	3.56E-04	6.09E-02	8.81E-03
BP GO:0006800 oxygen and reactive oxygen species metabolic process	2.30E-03	4.44E-04	6.09E-02	1.07E-02
BP GO:0006032 chitin catabolic process	2.59E-03	2.17E-04	6.12E-02	5.65E-03
MF GO:0004568 chitinase activity	2.88E-03	2.49E-04	6.60E-02	6.37E-03
BP GO:0006508 proteolysis	3.17E-03	1.82E-03	7.06E-02	3.40E-02
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	3.55E-03	9.52E-04	7.67E-02	1.98E-02
MF GO:0009055 electron carrier activity	7.15E-03	3.21E-03	1.39E-01	4.81E-02
BP GO:0051189 prosthetic group metabolic process	8.86E-03	1.94E-03	1.68E-01	3.40E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	9.12E-03	1.68E-03	1.71E-01	3.24E-02
BP GO:0005992 trehalose biosynthetic process	9.56E-03	1.92E-04	1.75E-01	5.65E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	1.01E-02	2.09E-04	1.83E-01	5.65E-03
BP GO:0048066 pigmentation during development	1.37E-02	3.34E-03	2.34E-01	4.81E-02
MF GO:0005344 oxygen transporter activity	1.95E-02	8.00E-04	3.07E-01	1.73E-02
MF GO:0019203 carbohydrate phosphatase activity	1.95E-02	8.00E-04	3.07E-01	1.73E-02
MF GO:0004364 glutathione transferase activity	1.95E-02	4.33E-03	3.07E-01	5.84E-02
BP GO:0006635 fatty acid beta-oxidation	2.52E-02	4.73E-03	3.91E-01	6.27E-02
MF GO:0005319 lipid transporter activity	2.77E-02	5.35E-03	4.12E-01	6.91E-02
BP GO:0046483 heterocycle metabolic process	2.80E-02	1.21E-02	4.12E-01	1.20E-01
BP GO:0006725 aromatic compound metabolic process	2.85E-02	1.29E-02	4.12E-01	1.26E-01
MF GO:0004246 peptidyl-dipeptidase A activity	3.12E-02	1.91E-03	4.35E-01	3.40E-02
MF GO:0016885 ligase activity; forming carbon-carbon bonds	3.12E-02	1.91E-03	4.35E-01	3.40E-02
BP GO:0042445 hormone metabolic process	3.53E-02	5.18E-03	4.84E-01	6.74E-02
MF GO:0003995 acyl-CoA dehydrogenase activity	3.80E-02	5.74E-03	5.05E-01	7.29E-02
MF GO:0015020 glucuronosyltransferase activity	3.81E-02	8.24E-03	5.05E-01	9.89E-02
BP GO:0006633 fatty acid biosynthetic process	4.02E-02	8.90E-03	5.20E-01	1.02E-01
BP GO:0006012 galactose metabolic process	4.26E-02	3.36E-03	5.46E-01	4.81E-02
CC GO:0005604 basement membrane	4.36E-02	7.28E-03	5.46E-01	8.89E-02

Continued on next page

Table C.2 – continued from previous page

MF GO:0016209 antioxidant activity	4.40E-02	1.00E-02	5.46E-01	1.06E-01
Bin u11				
MF GO:0004497 monooxygenase activity	7.92E-09	1.38E-09	1.30E-06	2.84E-07
CC GO:0005792 microsome	2.97E-08	4.19E-09	3.25E-06	4.30E-07
MF GO:0020037 heme Binding	9.62E-08	1.78E-08	7.17E-06	1.46E-06
MF GO:0005506 iron ion Binding	3.65E-07	9.08E-08	2.24E-05	5.32E-06
CC GO:0005777 peroxisome	3.69E-07	3.04E-08	2.24E-05	1.99E-06
BP GO:0006118 electron transport	4.54E-06	1.70E-06	2.13E-04	7.55E-05
MF GO:0008061 chitin Binding	5.67E-06	5.33E-07	2.59E-04	2.65E-05
MF GO:0004867 serine-type endopeptidase inhibitor activity	7.06E-06	1.09E-06	3.13E-04	4.97E-05
BP GO:0006869 lipid transport	2.26E-05	4.05E-06	9.29E-04	1.66E-04
BP GO:0008202 steroid metabolic process	6.21E-05	1.76E-05	2.49E-03	6.86E-04
MF GO:0009055 electron carrier activity	1.93E-03	8.22E-04	5.86E-02	1.87E-02
MF GO:0005344 oxygen transporter activity	2.32E-03	7.75E-05	6.53E-02	2.49E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	3.21E-03	5.44E-04	7.77E-02	1.33E-02
BP GO:0006032 chitin catabolic process	3.30E-03	2.96E-04	7.77E-02	7.72E-03
MF GO:0004568 chitinase activity	3.72E-03	3.45E-04	8.28E-02	8.84E-03
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	4.08E-03	7.61E-04	8.93E-02	1.78E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	4.74E-03	1.05E-03	9.98E-02	2.31E-02
BP GO:0051189 prosthetic group metabolic process	5.58E-03	1.28E-03	1.14E-01	2.65E-02
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	5.63E-03	1.62E-03	1.14E-01	3.17E-02
BP GO:0006508 proteolysis	6.09E-03	3.69E-03	1.22E-01	5.70E-02
CC GO:0005615 extracellular space	8.70E-03	6.85E-04	1.66E-01	1.63E-02
BP GO:0005992 trehalose biosynthetic process	1.09E-02	2.34E-04	1.99E-01	6.86E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	1.16E-02	2.58E-04	2.09E-01	7.42E-03
BP GO:0008652 amino acid biosynthetic process	1.48E-02	4.65E-03	2.59E-01	6.35E-02
MF GO:0016814 hydrolase activity; acting on carbon-nitrogen ...	1.98E-02	3.36E-03	3.12E-01	5.70E-02
BP GO:0048066 pigmentation during development	2.40E-02	6.64E-03	3.58E-01	8.45E-02
MF GO:0016646 oxidoreductase activity; acting on the CH-NH group ...	2.87E-02	3.76E-03	4.25E-01	5.70E-02
MF GO:0004364 glutathione transferase activity	2.96E-02	7.34E-03	4.27E-01	9.07E-02
BP GO:0006635 fatty acid beta-oxidation	3.13E-02	6.27E-03	4.42E-01	8.04E-02
MF GO:0005319 lipid transporter activity	3.47E-02	7.18E-03	4.82E-01	8.99E-02
MF GO:0004246 peptidyl-dipeptidase A activity	3.55E-02	2.34E-03	4.82E-01	4.27E-02
MF GO:0016885 ligase activity; forming carbon-carbon bonds	3.55E-02	2.34E-03	4.82E-01	4.27E-02
MF GO:0019203 carbohydrate phosphatase activity	3.55E-02	2.34E-03	4.82E-01	4.27E-02
MF GO:0008010 structural constituent of chitin-based larval cuticle	3.66E-02	5.35E-03	4.92E-01	7.14E-02
BP GO:0009063 amino acid catabolic process	4.19E-02	1.34E-02	5.46E-01	1.39E-01
MF GO:0003995 acyl-CoA dehydrogenase activity	4.53E-02	7.35E-03	5.86E-01	9.07E-02
BP GO:0006012 galactose metabolic process	4.81E-02	4.07E-03	6.12E-01	5.70E-02
Bin u12				
MF GO:0004497 monooxygenase activity	2.45E-09	4.39E-10	4.71E-07	1.13E-07
CC GO:0005792 microsome	1.41E-08	2.03E-09	1.53E-06	2.34E-07
MF GO:0020037 heme Binding	7.35E-08	1.40E-08	6.95E-06	1.28E-06
MF GO:0005506 iron ion Binding	3.92E-07	1.01E-07	2.71E-05	5.84E-06
CC GO:0005777 peroxisome	7.34E-07	6.60E-08	4.24E-05	4.24E-06
MF GO:0008061 chitin Binding	1.96E-06	1.88E-07	9.70E-05	1.02E-05
BP GO:0006118 electron transport	2.72E-06	1.03E-06	1.31E-04	4.94E-05
BP GO:0006869 lipid transport	9.75E-06	1.73E-06	4.33E-04	7.90E-05

Continued on next page

Table C.2 – continued from previous page

MF GO:0004867 serine-type endopeptidase inhibitor activity	2.38E-05	4.23E-06	1.03E-03	1.83E-04
BP GO:0008202 steroid metabolic process	5.73E-05	1.68E-05	2.16E-03	6.18E-04
MF GO:0005344 oxygen transporter activity	2.58E-04	7.24E-06	8.93E-03	2.98E-04
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	1.09E-03	1.84E-04	3.36E-02	5.69E-03
MF GO:0009055 electron carrier activity	1.90E-03	8.28E-04	5.31E-02	1.86E-02
BP GO:0006032 chitin catabolic process	3.79E-03	3.53E-04	9.11E-02	8.87E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	3.97E-03	7.04E-04	9.28E-02	1.67E-02
BP GO:0006508 proteolysis	4.12E-03	2.49E-03	9.38E-02	4.44E-02
MF GO:0004568 chitinase activity	4.34E-03	4.21E-04	9.76E-02	1.03E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	5.83E-03	1.34E-03	1.22E-01	2.88E-02
BP GO:0008652 amino acid biosynthetic process	6.61E-03	1.96E-03	1.35E-01	3.79E-02
BP GO:0051189 prosthetic group metabolic process	8.00E-03	1.97E-03	1.56E-01	3.79E-02
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	8.41E-03	2.58E-03	1.62E-01	4.49E-02
BP GO:0005992 trehalose biosynthetic process	1.17E-02	2.62E-04	2.14E-01	7.68E-03
BP GO:0048066 pigmentation during development	1.23E-02	3.32E-03	2.22E-01	5.37E-02
MF GO:0016401 palmitoyl-CoA oxidase activity	1.26E-02	2.93E-04	2.25E-01	8.32E-03
CC GO:0005615 extracellular space	1.30E-02	1.23E-03	2.25E-01	2.66E-02
MF GO:0005044 scavenger receptor activity	2.28E-02	4.04E-03	3.50E-01	5.76E-02
MF GO:0016814 hydrolase activity; acting on carbon-nitrogen ...	2.28E-02	4.04E-03	3.50E-01	5.76E-02
MF GO:0016490 structural constituent of peritrophic membrane	2.45E-02	2.94E-03	3.70E-01	4.81E-02
BP GO:0009063 amino acid catabolic process	2.83E-02	9.17E-03	4.19E-01	1.06E-01
MF GO:0016646 oxidoreductase activity; acting on the CH-NH group ...	3.21E-02	4.38E-03	4.68E-01	5.76E-02
MF GO:0004364 glutathione transferase activity	3.48E-02	9.00E-03	4.93E-01	1.05E-01
BP GO:0006635 fatty acid beta-oxidation	3.52E-02	7.34E-03	4.93E-01	8.88E-02
BP GO:0017143 insecticide metabolic process	3.58E-02	2.37E-03	4.93E-01	4.28E-02
MF GO:0004246 peptidyl-dipeptidase A activity	3.85E-02	2.65E-03	5.12E-01	4.49E-02
MF GO:0016885 ligase activity; forming carbon-carbon bonds	3.85E-02	2.65E-03	5.12E-01	4.49E-02
MF GO:0019203 carbohydrate phosphatase activity	3.85E-02	2.65E-03	5.12E-01	4.49E-02
MF GO:0005319 lipid transporter activity	3.97E-02	8.58E-03	5.25E-01	1.00E-01
MF GO:0008010 structural constituent of chitin-based larval cuticle	4.08E-02	6.23E-03	5.35E-01	7.67E-02
Bin u13				
MF GO:0004497 monooxygenase activity	2.33E-09	4.34E-10	4.61E-07	9.69E-08
CC GO:0005792 microsome	9.20E-09	1.37E-09	1.03E-06	1.63E-07
MF GO:0020037 heme Binding	6.32E-08	1.25E-08	5.64E-06	1.12E-06
MF GO:0005506 iron ion Binding	4.31E-07	1.16E-07	3.20E-05	8.24E-06
CC GO:0005777 peroxisome	1.66E-06	1.65E-07	9.32E-05	1.05E-05
BP GO:0006118 electron transport	3.61E-06	1.40E-06	1.90E-04	6.94E-05
BP GO:0006869 lipid transport	4.21E-06	7.45E-07	2.15E-04	3.80E-05
MF GO:0008061 chitin Binding	6.01E-06	6.74E-07	2.90E-04	3.54E-05
MF GO:0004867 serine-type endopeptidase inhibitor activity	4.19E-05	7.92E-06	1.70E-03	3.21E-04
BP GO:0008202 steroid metabolic process	1.51E-04	4.78E-05	5.28E-03	1.65E-03
MF GO:0005344 oxygen transporter activity	2.92E-04	8.48E-06	9.66E-03	3.36E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	3.17E-04	4.81E-05	1.03E-02	1.65E-03
BP GO:0051189 prosthetic group metabolic process	5.60E-04	1.23E-04	1.65E-02	3.33E-03
BP GO:0008652 amino acid biosynthetic process	9.12E-04	2.36E-04	2.62E-02	5.92E-03
BP GO:0006032 chitin catabolic process	1.02E-03	9.41E-05	2.67E-02	2.75E-03
MF GO:0004568 chitinase activity	1.20E-03	1.15E-04	3.05E-02	3.16E-03
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	1.43E-03	2.54E-04	3.46E-02	6.30E-03

Continued on next page

Table C.2 – continued from previous page

BP GO:0006508 proteolysis	1.59E-03	9.36E-04	3.73E-02	1.90E-02
MF GO:0009055 electron carrier activity	1.70E-03	7.51E-04	3.95E-02	1.56E-02
BP GO:0048066 pigmentation during development	5.49E-03	1.42E-03	1.09E-01	2.69E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	7.01E-03	1.67E-03	1.36E-01	3.14E-02
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	1.16E-02	3.75E-03	2.09E-01	6.14E-02
BP GO:0005992 trehalose biosynthetic process	1.25E-02	2.90E-04	2.21E-01	6.89E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	1.34E-02	3.22E-04	2.35E-01	7.57E-03
BP GO:0009063 amino acid catabolic process	1.58E-02	4.97E-03	2.70E-01	6.68E-02
CC GO:0005615 extracellular space	1.95E-02	2.20E-03	3.16E-01	3.89E-02
MF GO:0005044 scavenger receptor activity	2.53E-02	4.64E-03	3.94E-01	6.68E-02
MF GO:0016814 hydrolase activity; acting on carbon-nitrogen ...	2.53E-02	4.64E-03	3.94E-01	6.68E-02
MF GO:0016490 structural constituent of peritrophic membrane	2.67E-02	3.31E-03	4.11E-01	5.52E-02
MF GO:0005529 sugar Binding	2.93E-02	7.18E-03	4.47E-01	9.15E-02
MF GO:0016646 oxidoreductase activity; acting on the CH-NH group ...	3.49E-02	4.92E-03	5.19E-01	6.68E-02
BP GO:0006725 aromatic compound metabolic process	3.78E-02	1.89E-02	5.57E-01	1.97E-01
MF GO:0004364 glutathione transferase activity	3.92E-02	1.05E-02	5.69E-01	1.24E-01
BP GO:0006635 fatty acid beta-oxidation	3.92E-02	8.44E-03	5.69E-01	1.05E-01
MF GO:0004246 peptidyl-dipeptidase A activity	4.08E-02	2.90E-03	5.69E-01	4.93E-02
MF GO:0016885 ligase activity; forming carbon-carbon bonds	4.08E-02	2.90E-03	5.69E-01	4.93E-02
MF GO:0019203 carbohydrate phosphatase activity	4.08E-02	2.90E-03	5.69E-01	4.93E-02
MF GO:0005319 lipid transporter activity	4.39E-02	9.80E-03	6.02E-01	1.17E-01
MF GO:0008010 structural constituent of chitin-based larval cuticle	4.42E-02	6.99E-03	6.02E-01	8.98E-02
Bin u14				
MF GO:0004497 monooxygenase activity	3.70E-08	8.09E-09	4.29E-06	9.38E-07
CC GO:0005792 microsome	7.77E-08	1.35E-08	7.59E-06	1.31E-06
MF GO:0020037 heme Binding	8.98E-07	2.13E-07	5.96E-05	1.27E-05
CC GO:0005777 peroxisome	1.23E-06	1.36E-07	7.36E-05	9.31E-06
MF GO:0008061 chitin Binding	4.95E-06	6.19E-07	2.55E-04	3.19E-05
MF GO:0005506 iron ion Binding	6.41E-06	2.02E-06	3.22E-04	9.61E-05
BP GO:0006869 lipid transport	1.04E-05	1.99E-06	4.93E-04	9.61E-05
MF GO:0004867 serine-type endopeptidase inhibitor activity	3.00E-05	5.90E-06	1.33E-03	2.67E-04
BP GO:0006118 electron transport	6.45E-05	2.88E-05	2.72E-03	1.11E-03
BP GO:0051189 prosthetic group metabolic process	9.86E-05	2.06E-05	3.81E-03	8.48E-04
BP GO:0006508 proteolysis	1.29E-04	7.19E-05	4.90E-03	2.47E-03
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	1.64E-04	2.57E-05	5.97E-03	1.04E-03
MF GO:0005344 oxygen transporter activity	3.97E-04	1.25E-05	1.25E-02	5.53E-04
BP GO:0008202 steroid metabolic process	8.75E-04	3.24E-04	2.58E-02	8.46E-03
BP GO:0008652 amino acid biosynthetic process	1.71E-03	4.79E-04	4.61E-02	1.11E-02
BP GO:0006032 chitin catabolic process	2.12E-03	2.38E-04	5.10E-02	6.49E-03
MF GO:0004568 chitinase activity	2.50E-03	2.93E-04	5.96E-02	7.76E-03
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	2.86E-03	5.76E-04	6.72E-02	1.29E-02
MF GO:0005529 sugar Binding	5.36E-03	1.19E-03	1.17E-01	2.51E-02
MF GO:0009055 electron carrier activity	7.97E-03	4.01E-03	1.63E-01	6.76E-02
BP GO:0006800 oxygen and reactive oxygen species metabolic process	1.05E-02	2.72E-03	2.06E-01	5.21E-02
BP GO:0048066 pigmentation during development	1.15E-02	3.38E-03	2.22E-01	6.26E-02
MF GO:0004295 trypsin activity	1.24E-02	3.70E-03	2.38E-01	6.29E-02
BP GO:0005992 trehalose biosynthetic process	1.45E-02	3.65E-04	2.65E-01	9.14E-03
MF GO:0016401 palmitoyl-CoA oxidase activity	1.57E-02	4.08E-04	2.74E-01	9.76E-03

Continued on next page

Table C.2 – continued from previous page

CC GO:0005764 lysosome	1.62E-02	2.59E-03	2.76E-01	5.01E-02
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group ...	2.32E-02	8.45E-03	3.74E-01	1.11E-01
MF GO:0008431 vitamin E Binding	2.43E-02	2.81E-03	3.89E-01	5.32E-02
BP GO:0009063 amino acid catabolic process	2.67E-02	9.30E-03	4.24E-01	1.19E-01
CC GO:0016021 integral to membrane	3.04E-02	2.20E-02	4.62E-01	2.15E-01
CC GO:0005615 extracellular space	3.14E-02	4.33E-03	4.73E-01	7.11E-02
MF GO:0016814 hydrolase activity; acting on carbon-nitrogen ...	3.27E-02	6.49E-03	4.84E-01	8.90E-02
MF GO:0016490 structural constituent of peritrophic membrane	3.28E-02	4.41E-03	4.84E-01	7.11E-02
MF GO:0005549 odorant Binding	3.66E-02	1.11E-02	5.31E-01	1.33E-01
MF GO:0005044 scavenger receptor activity	3.97E-02	8.45E-03	5.67E-01	1.11E-01
MF GO:0042626 ATPase activity; coupled to transmembrane movement . . .	4.15E-02	2.09E-02	5.79E-01	2.06E-01
MF GO:0016646 oxidoreductase activity; acting on the CH-NH group ...	4.27E-02	6.52E-03	5.87E-01	8.90E-02
MF GO:0004246 peptidyl-dipeptidase A activity	4.72E-02	3.64E-03	6.26E-01	6.26E-02
MF GO:0016885 ligase activity; forming carbon-carbon bonds	4.72E-02	3.64E-03	6.26E-01	6.26E-02
MF GO:0019203 carbohydrate phosphatase activity	4.72E-02	3.64E-03	6.26E-01	6.26E-02
MF GO:0042708 elastase activity	4.72E-02	3.64E-03	6.26E-01	6.26E-02
Bin u15				
BP GO:0006508 proteolysis	3.18E-14	1.26E-14	8.53E-12	3.38E-12
MF GO:0004497 monooxygenase activity	9.06E-08	2.15E-08	7.09E-06	1.68E-06
CC GO:0005792 microsome	1.56E-07	2.85E-08	1.09E-05	1.98E-06
BP GO:0006869 lipid transport	5.13E-07	9.11E-08	2.92E-05	5.35E-06
CC GO:0005777 peroxisome	1.42E-06	1.59E-07	7.39E-05	8.53E-06
MF GO:0004295 trypsin activity	3.12E-06	6.04E-07	1.44E-04	2.91E-05
MF GO:0020037 heme Binding	5.50E-06	1.47E-06	2.30E-04	6.01E-05
MF GO:0008061 chitin Binding	6.24E-06	1.00E-06	2.55E-04	4.28E-05
MF GO:0005506 iron ion Binding	5.03E-05	1.80E-05	1.93E-03	6.76E-04
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	7.50E-05	1.18E-05	2.76E-03	4.50E-04
MF GO:0004867 serine-type endopeptidase inhibitor activity	9.50E-05	2.10E-05	3.32E-03	7.44E-04
BP GO:0008202 steroid metabolic process	2.56E-04	9.29E-05	8.43E-03	3.01E-03
BP GO:0006118 electron transport	4.41E-04	2.20E-04	1.40E-02	6.36E-03
MF GO:0005344 oxygen transporter activity	5.57E-04	1.93E-05	1.71E-02	7.09E-04
BP GO:0051189 prosthetic group metabolic process	6.52E-04	1.70E-04	1.94E-02	5.24E-03
MF GO:0004179 membrane alanyl aminopeptidase activity	2.18E-03	3.19E-04	5.59E-02	8.80E-03
BP GO:0006032 chitin catabolic process	3.23E-03	4.00E-04	7.21E-02	9.62E-03
MF GO:0004568 chitinase activity	3.68E-03	4.71E-04	8.03E-02	1.09E-02
CC GO:0005764 lysosome	4.61E-03	6.76E-04	9.60E-02	1.46E-02
BP GO:0008652 amino acid biosynthetic process	5.51E-03	1.81E-03	1.12E-01	3.39E-02
CC GO:0043190 ATP-Binding cassette (ABC) transporter complex	7.06E-03	1.71E-03	1.43E-01	3.24E-02
MF GO:0005529 sugar Binding	1.23E-02	3.25E-03	2.29E-01	5.70E-02
MF GO:0004035 alkaline phosphatase activity	1.44E-02	1.23E-03	2.57E-01	2.45E-02
MF GO:0004263 chymotrypsin activity	1.44E-02	1.23E-03	2.57E-01	2.45E-02
MF GO:0042708 elastase activity	1.44E-02	1.23E-03	2.57E-01	2.45E-02
MF GO:0016490 structural constituent of peritrophic membrane	1.68E-02	2.55E-03	2.87E-01	4.56E-02
BP GO:0005992 trehalose biosynthetic process	1.75E-02	4.85E-04	2.93E-01	1.10E-02
MF GO:0016401 palmitoyl-CoA oxidase activity	1.86E-02	5.30E-04	3.08E-01	1.18E-02
MF GO:0005319 lipid transporter activity	2.33E-02	5.21E-03	3.71E-01	8.15E-02
BP GO:0048066 pigmentation during development	2.51E-02	8.55E-03	3.94E-01	1.07E-01
BP GO:0006800 oxygen and reactive oxygen species metabolic process	2.64E-02	8.25E-03	3.98E-01	1.05E-01

Continued on next page

Table C.2 – continued from previous page

MF GO:0009055 electron carrier activity	2.76E-02	1.56E-02	4.08E-01	1.65E-01
MF GO:0004558 alpha-glucosidase activity	3.06E-02	3.87E-03	4.45E-01	6.54E-02
MF GO:0008431 vitamin E Binding	3.06E-02	3.87E-03	4.45E-01	6.54E-02
MF GO:0050809 diazepam Binding	3.51E-02	1.99E-03	4.86E-01	3.63E-02
MF GO:0004364 glutathione transferase activity	3.52E-02	1.05E-02	4.86E-01	1.24E-01
MF GO:0016811 hydrolase activity; acting on carbon-nitrogen . . .	3.69E-02	1.36E-02	5.02E-01	1.48E-01
BP GO:0001501 skeletal development	3.93E-02	8.26E-03	5.31E-01	1.05E-01
MF GO:0016814 hydrolase activity; acting on carbon-nitrogen ...	4.32E-02	9.36E-03	5.67E-01	1.13E-01
BP GO:0009063 amino acid catabolic process	4.81E-02	1.88E-02	6.14E-01	1.83E-01
CC GO:0005615 extracellular space	4.92E-02	8.36E-03	6.24E-01	1.06E-01

Appendix D

Additional results from Chapter 6

Table D.1: Supplementary data for Chapter 2.2.2. GO terms that are over-represented in the up-/downregulated genes with a q value < 0.05 are shown, whereas the q values were estimated from the EASE scores. The column 'frequency' indicates the number of tissue specificity bins the GO term was found to be significant in. We expect 33 × 0.05 = 1.65 false positive GO terms in this list.

GO term	frequency
BP GO:0001501 skeletal development	2
BP GO:0006013 mannose metabolic process	1
BP GO:0006032 chitin catabolic process	3
BP GO:0006118 electron transport	14
BP GO:0006508 proteolysis	20
BP GO:0006800 oxygen and reactive oxygen species metabolic process	1
BP GO:0006869 lipid transport	19
BP GO:0008202 steroid metabolic process	11
BP GO:0008652 amino acid biosynthetic process	3
BP GO:0051189 prosthetic group metabolic process	6
CC GO:0005764 lysosome	12
CC GO:0005777 peroxisome	18
CC GO:0005792 microsome	17
CC GO:0042600 chorion	1
CC GO:0043190 ATP-binding cassette (ABC) transporter complex	2
MF GO:0004035 alkaline phosphatase activity	3
MF GO:0004263 chymotrypsin activity	16
MF GO:0004295 trypsin activity	18
MF GO:0004497 monooxygenase activity	15
MF GO:0004558 alpha-glucosidase activity	6
MF GO:0004559 alpha-mannosidase activity	1
MF GO:0004568 chitinase activity	3
MF GO:0004806 triacylglycerol lipase activity	5
MF GO:0004867 serine-type endopeptidase inhibitor activity	10
MF GO:0005213 structural constituent of chorion	1

Continued on next page

Table D.1 – continued from previous page

MF GO:0005344 oxygen transporter activity	7
MF GO:0005506 iron ion binding	14
MF GO:0008061 chitin binding	14
MF GO:0009055 electron carrier activity	1
MF GO:0016616 oxidoreductase activity; acting on ..	2
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	12
MF GO:0020037 heme binding	15
MF GO:0042708 elastase activity	2

Table D.2: Supplementary data for Chapter 2.2.2. GO terms that are over-represented in the up-/downregulated genes with a q value < 0.05 are shown, whereas the q values were estimated from the Fisher p values. The column 'frequency' indicates the number of tissue specificity bins the GO term was found to be significant in. We expect 77*0.05 = 3.85 false positive GO terms in this list.

GO term	frequency
BP GO:0001501 skeletal development	15
BP GO:0001708 cell fate specification	1
BP GO:0005992 trehalose biosynthetic process	12
BP GO:0006012 galactose metabolic process	2
BP GO:0006013 mannose metabolic process	11
BP GO:0006032 chitin catabolic process	13
BP GO:0006118 electron transport	16
BP GO:0006508 proteolysis	22
BP GO:0006633 fatty acid biosynthetic process	3
BP GO:0006635 fatty acid beta-oxidation	2
BP GO:0006665 sphingolipid metabolic process	3
BP GO:0006800 oxygen and reactive oxygen species metabolic process	5
BP GO:0006869 lipid transport	21
BP GO:0007306 eggshell chorion formation	7
BP GO:0008202 steroid metabolic process	13
BP GO:0008286 insulin receptor signaling pathway	2
BP GO:0008652 amino acid biosynthetic process	9
BP GO:0010004 gastrulation involving germ band extension	1
BP GO:0017143 insecticide metabolic process	1
BP GO:0031887 lipid particle transport along microtubule	3
BP GO:0035152 regulation of tube architecture; open tracheal system	2
BP GO:0042049 cell acyl-CoA homeostasis	1
BP GO:0042594 response to starvation	16
BP GO:0046112 nucleobase biosynthetic process	9
BP GO:0048066 pigmentation during development	3
BP GO:0051189 prosthetic group metabolic process	10
CC GO:0005615 extracellular space	3
CC GO:0005764 lysosome	16
CC GO:0005777 peroxisome	18

Continued on next page

Table D.2 – continued from previous page

CC GO:0005792 microsome	22
CC GO:0042600 chorion	17
CC GO:0043190 ATP-binding cassette (ABC) transporter complex	10
CC GO:0044452 nucleolar part	4
CC GO:0055029 nuclear DNA-directed RNA polymerase complex	4
MF GO:0000062 acyl-CoA binding	1
MF GO:0003676 nucleic acid binding	1
MF GO:0003899 DNA-directed RNA polymerase activity	10
MF GO:0004035 alkaline phosphatase activity	16
MF GO:0004095 carnitine O-palmitoyltransferase activity	1
MF GO:0004179 membrane alanyl aminopeptidase activity	5
MF GO:0004182 carboxypeptidase A activity	1
MF GO:0004246 peptidyl-dipeptidase A activity	4
MF GO:0004263 chymotrypsin activity	17
MF GO:0004295 trypsin activity	18
MF GO:0004364 glutathione transferase activity	4
MF GO:0004497 monooxygenase activity	19
MF GO:0004558 alpha-glucosidase activity	13
MF GO:0004559 alpha-mannosidase activity	7
MF GO:0004568 chitinase activity	13
MF GO:0004806 triacylglycerol lipase activity	16
MF GO:0004867 serine-type endopeptidase inhibitor activity	13
MF GO:0004888 transmembrane receptor activity	1
MF GO:0005158 insulin receptor binding	1
MF GO:0005160 transforming growth factor beta receptor binding	1
MF GO:0005213 structural constituent of chorion	17
MF GO:0005319 lipid transporter activity	1
MF GO:0005344 oxygen transporter activity	15
MF GO:0005506 iron ion binding	15
MF GO:0005529 sugar binding	1
MF GO:0008010 structural constituent of chitin-based larval cuticle	1
MF GO:0008061 chitin binding	17
MF GO:0008336 gamma-butyrobetaine dioxygenase activity	1
MF GO:0008533 astacin activity	7
MF GO:0008970 phospholipase A1 activity	5
MF GO:0009055 electron carrier activity	4
MF GO:0016229 steroid dehydrogenase activity	1
MF GO:0016401 palmitoyl-CoA oxidase activity	14
MF GO:0016490 structural constituent of peritrophic membrane	3
MF GO:0016616 oxidoreductase activity; acting on the CH-OH group of donors; NAD ..	7
MF GO:0016799 hydrolase activity; hydrolyzing N-glycosyl compounds	22
MF GO:0016885 ligase activity; forming carbon-carbon bonds	6
MF GO:0019203 carbohydrate phosphatase activity	5
MF GO:0019204 nucleotide phosphatase activity	4
MF GO:0020037 heme binding	16
MF GO:0042708 elastase activity	6
MF GO:0050809 diazepam binding	3

Glossary

α-helix A coiled secondary structure of a polypeptide chain formed by hydrogen bonding between amino acids separated by 3.6 residues. 9

β-sheet Two or more polypeptide chains that run alongside each other and are linked in a regular manner by hydrogen bonds between the main chain C=O and N-H groups. The R-groups (side chains) of neighbouring residues point in opposite directions; β-sheets can be parallel, anti-parallel or mixed. 9

3_{10} helix A rare type of secondary structure found in proteins. The amino acids in are arranged in a right-handed helical structure. Each amino acid corresponds to a 120° turn in the helix. 10, 54

acceptor (finite state) A finite state acceptor is an Finite State Machine (FSM) with no outputs. 151

alternative splicing The generation of different mRNAs by varying the pattern of pre-mRNA splicing. 6

amino acid The subunits from which proteins are assembled. Each amino acid consists of an amino functional group, and a carboxyl acid group, and differs from other amino acids by the composition of an R group. 5

angstrom (Å) A unit of measure. One angstrom is 10^{-10} metres. Often used to indicate structural similarity between two proteins (see RMSD. A similarity below 3 angstroms indicates a strong structural similarity. 33, 207

Area Under the Curve (AUC) An indication of the diagnostic accuracy of a ROC curve. AUC values closer to 1 indicate the method reliably distinguishes among the positive and the negative class, whereas values at 0.5 indicate that the predictor is no better than random. 27, 82

ArrayExpress A public repository for transcriptomics data (`http://www.ebi.ac.uk/microarray-as/ae/`). 123

asparagine One of the 20 most common natural amino acids on earth. 96

ASTRAL The ASTRAL compendium provides databases and tools useful for analysing protein structures and their sequences. It is partially derived from the SCOP and the PDB databases. 54

bases The molecular building blocks of mRNA and RNA. These include adenine (A), cytosine (C), guanine (G), thymine (T), and (in RNA only) uracil (U). In mRNA, A attaches only to T, and C attaches only to G. In RNA, A attaches only to U, and C attaches only to G. 5

bioinformatics The merging field of biology, computer science and information technology with the goal of revealing new insights and principles in biology through the analysis of biological data using computers, machine learning and statistical techniques. 19

bit A binary digit, or the amount of information required to distinguish between two equally likely possibilities or choices. 23

BLAST Commonly used algorithm for searching databases for similar sequences. 33, 206

blastoderm A stage of insect embryogenesis in which a layer of nuclei or cells around the embryo surround an internal mass of yolk. 39

boxplot A graph summarising the distribution of a set of data values. 30

cis-regulatory A site on a DNA molecule that functions as a binding site for a sequence-specific DNA binding protein. The term cis indicates that protein binding to this site affects only nearby DNA sequences on the same molecule. 17

classification Assigning a class to a measurement. 19

coding sequence (CDS) The combination of exons on a gene. 6

codon The basic unit of the genetic code; one of the 64 nucleotide triplets that code for an amino acid or stop sequence. 5

coiled coil Stable rodlike quaternary protein structure formed by two or three α helices interacting with each other. Coiled coils are commonly found in fibrous proteins and transcription factors. 9

CpG island A stretch of DNA in which the frequency of the dinucleotide CG sequence is higher than in other DNA regions. 15, 43

curse of dimensionality The name given to algorithmic challenges posed by high-dimensional spaces necessary to map data with many features, in which the resulting exponential growth in hypervolumes means that the data inevitably will be distributed ever more sparsely. 23

cytoskeleton The internal scaffolding of cells. 14, 42

Dahomey A frequently used fruit fly stock that was originally collected in Dahomey, West Africa. 123

differentially expressed A gene is differentially expressed when its expression values under two or more conditions are statistically significantly different. 123

DNA (deoxyribonucleic acid) DNA is the carrier of genetic information. It consists of a sequence of hundreds of millions of nucleotides that code for proteins. 5, 199, 200, 202, 206

DNA methylation A natural regulatory process in the cell, which controls gene activity via the attachment of a methyl group to DNA. 15

domain Compact, globular regions of proteins that are the basic units of tertiary structure. 19, 41

dot-product An operation with two vectors that results in a scalar quantity. Also known as scalar or inner product. 20

downregulated gene A gene which has been observed to have lower expression (lower mRNA levels) in one sample compared to another sample (here wild-type fruit fly). 123

Drosophila melanogaster (D. melanogaster) A species of fruit fly commonly used as model organism in biology. 2

DSSP A database and program of secondary structure assignments for all protein entries in the PDB. 10, 54

E value (Expectation value) The number of different alignments with scores equivalent to or better than the observed score that are expected to occur in a database search by chance. The lower the E value, the more significant the score. 35

EMBOSS The European Molecular Biology Open Software Suite. EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. 84

enhancer A short regulatory DNA sequence that increases the level of expression of a gene. 17

enzyme A protein that functions as a catalyst. 16, 42

ester Any of a class of organic compounds that react with water to produce alcohols and organic or inorganic acids. 201

esterase An enzyme that splits esters into an acid and an alcohol in a chemical reaction with water called hydrolysis. A wide range of different esterases exist that differ in their substrate specificity, their protein structure, and their biological function. 112

exon A segment of a gene that contains a coding sequence. 5, 199, 203, 208

expression breadth The number of tissues in which a gene is expressed. 102, 114

extracellular matrix A network in an animal tissue which provides support to cells. 42

false positive rate The proportion of negative examples that are predicted positive. 25

feature The measurements which represent the data. Here, used as input for SVMs. 37, 204

Finite State Machine (FSM) A machine which can be totally described by a finite set of states, it being in one these at any one time, plus a set of rules which determine when it moves from one state to another. 197, 209

FlyBase A database for fruit fly genetics and molecular biology (`http://flybase.org/`). 123

fold change A way of describing now much larger or smaller one number is compared with another. 123, 126, 147

gcrma A background correction method that corrects for unspecific binding due to high GC content in the probe sets of on Affymetrix GeneChips. 84

GenAge A curated database of genes related to ageing (`http://genomics.senescence.info/genes/`). 123

gene A segment of mRNA that encodes a polypeptide chain or an RNA molecule. 1

gene expression The process by which DNA is translated into RNAs or proteins. 39

Gene Ontology (GO) A controlled vocabulary of terms relating to molecular function, biological process, or cellular components. It allows scientists to use consistent terminology when describing the roles of genes and proteins in cells. 41, 123

genetic code The correspondence between nucleotide triplets and amino acids in proteins. 5

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) A housekeeping gene that codes for an enzyme involved in glycolysis. 14

glycolysis The cellular degradation of the simple sugar glucose to yield ATP as an energy source. 14, 93

GNF The Genomics Institute of the Novartis Research Foundation. 75

hairpin A structural motif involving two β strands that look like a hairpin. A special case of a turn. 10

hemoglobin An oxygen-binding protein that and carries oxygen it from the lungs to the tissues. 33

heterozygote Heterozygous refers to having inherited different forms of a particular gene from each parent. A heterozygous genotype stands in contrast to a homozygous genotype, where an individual inherits identical forms of a particular gene from each parent. 123

hidden Markov model (HMM) A statistical model for an ordered sequence of variables. 33

housekeeping gene A constitutive gene that is transcribed at a relatively constant level across many or all known conditions. The housekeeping gene's products are typically needed for maintenance of the cell. It is generally assumed that their expression is unaffected by experimental conditions. Examples include actin and GAPDH. 13, 41

hydrophilic Having a strong affinity for water. 9

hydrophobic Lacking affinity for water. 9

in situ hybridization (ISH) A method that is used to label specific sequences of nucleic acids in cells or chromosomes. Commonly used to identify mRNA expression in tissues or whole organisms. ISH detects the formation of nucleic acid hybrid molecules between the target nucleic acid and a labelled probe that contains a complementary sequence. 14, 39

Insulin and Insulin-like growth factor signaling (IIS) A conserved signaling pathway from insects to humans. Growth factors, released after feeding, (insulin and insulin-like growth factor, IGF-1) stimulate receptors of this pathway and promote cellular and oraganismal anabolic growth. The IIS pathway is central to regulation of life span, metabolism, and the stress response. 123

intron A noncoding sequence that interrupts exons in a gene. 5

k-nearest neighbour A classification method that classifies an instance by calculating the distances between the instances and instances in the training data set. Then it assigns the instance to the class that is most common among its k-nearest neighbours, where k is an integer. 33

ligand A substance that is able to bind to another biomolecule (substrat/receptor) by means of intermolecular forces to form a complex by altering its chemical conformation (three-dimensional shape). 42

loess normalisation A computationally intensive method in which a polynomial regression is fitted to each point in the data and more weight is given to data nearer the point of interest. It is often applied to hybridization array data to remove differences in global signal intensity among data sets. 84

machine learning Computational approaches to learn new knowledge on the basis of observed examples. 1, 3

messenger RNA (mRNA) An RNA molecule that is complementary to one of the mRNA strands of a gene. It serves as a template for protein synthesis. 6, 197, 198, 200, 201, 203–210

methyl group A functional group consisting of one carbon and three hydrogen atoms: -CH_3. 200

microarray Microarrays are used for analysing the expression of thousands of genes simultaneously. 3, 39

mitochondrial electron transport chain A collective term describing the mitochondrial enzymes that are needed to generate the electron and proton 'gradient' that is used to generate ATP. 49

multi-view learning In a multi-view problem, one can partition the domain's features in subsets that are sufficient for learning the target concept. 155

myoglobin A protein occurring widely in muscle tissue as an oxygen carrier. It acts as an emergency oxygen store. 33

Naive Bayes A supervised classification algorithm that uses the Bayes rule to compute the fit between a new observation and some previously observed data. Bayes' rule expresses the conditional probability of the event A given the event B in terms of the conditional probability of the event B given the event A. 40, 92

neural network An analytic technique modelled after the processes of learning in the cognitive system and the neurological functions of the brain. A neural network is capable of predicting new observations from other observations after learning from existing data. 12

nuclear matrix The dense fibrillar network lying on the inner side of the nuclear membrane. 17, 43

nucleus The center of a cell, where the mRNA is contained. 17

null mutation A mutation that results in the complete loss of function of a gene product. 48, 123

oxidative stress Physiological stress on the body that is caused by the cumulative damage done by free radicals inadequately neutralized by antioxidants. Oxidative stress is held to be associated with ageing. 136

p value The probability, if the test statistic really were distributed as it would be under the null hypothesis, of observing a test statistic as extreme as, or more extreme than the one actually observed. The smaller the p value, the more strongly the test rejects the null hypothesis, that is, the hypothesis being tested. A value of 0.05 is a common significance level to which p values are compared. 28

paralogous genes Two genes at different chromosomal locations in the same organism that have structural similarities indicating that they derived from a common ancestral gene and have since diverged from the parent copy by mutation. The genes encoding myoglobin and hemoglobin are considered to be ancient paralogs. 111

PDB (Protein Data Bank) A repository for the three-dimensional structural data of proteins. 12, 33, 54, 198, 200

phosphodiester bond A bond between two sugar groups and a phosphate group. Such bonds form the sugar-phosphate-sugar backbone of mRNA and RNA. 16

pi helix (π helix) A type of secondary structure that is common in membrane proteins. 10, 54

polysome Complex of ribosomes for simultaneous translation of mRNA. 18

pre-mRNA The unspliced mRNA that contains all exons and introns. 197

probe A labeled, single-stranded DNA or RNA molecule of specific base sequence, that is used to detect the complementary base sequence by hybridization. At Affymetrix, probe refers to unlabeled oligonucleotides synthesized on a GeneChip probe array. 14, 15

probeset ID The Affymetrix probe-set identificator. 123

profile A table that lists the frequencies of each amino acid in each position of protein sequence. Frequencies are calculated from multiple alignments of sequences containing a domain of interest. 36, 38

promoter Area of DNA that regulates gene expression. 7, 15

protein A large molecule composed of amino acids. Proteins are required for the structure, function, and regulation of the body cells, tissues and organs. 1, 2, 202

PSI-BLAST (Position-Specific Iterated BLAST) An improved BLAST algorithm. 12, 36

quantiles The quantile of a distribution of values is a number xp such that a proportion p of the population values are less than or equal to xp. For example, the 25th percentile of a variable is a value (xp) such that 25% (p) of the values of the variable fall below that value. 84

Receiver Operating Characteristic (ROC) curve A ROC curve is a generalization of the set of potential combinations of sensitivity and specificity possible for predictors. A ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for the different possible cut-points of a diagnostic test. 25, 198, 210

receptor A molecule or surface in a cell that recognizes and binds to a specific messenger molecule, leading to a biological response. 42

regression Predicting the value of random variable y from measurement x. Regression generalizes classification since y can be any quantity, including a class index. Many classification algorithms can be understood as thresholding the output of a regression. 29

remote homology Evolutionary relationship between two proteins that do not display high sequence similarity. 1

ribosome A ribosome is a cellular structure made of RNA and protein that serves as the site for protein synthesis in the cell. The ribosome reads the sequence of the mRNA and translates the sequence of RNA bases into a sequence of amino acids. 6, 18, 41, 207

ribosome density Refers to the average number of ribosome bound per unit length of coding sequence. 18

ribosome occupancy The fraction of a given gene's transcripts associated with ribosomes. 18

RMSD (Root Mean Square Deviation) Measurement for protein structure similarity. Measured in angstroms. A similarity below 3 angstroms indicatesa strong structural similarity. 33, 197

RNA (Ribonucleic acid) A molecule similar to mRNA but single-stranded. An RNA strand has a backbone made of alternating sugar and phosphate groups. Attached to each sugar is one of four bases–adenine (A), uracil (U), cytosine (C), or guanine (G). Different types of RNA exist: messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Some small RNAs have been found to be involved in regulating gene expression. 5, 198, 201, 202, 204–209

RNA-Seq The use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content. 15, 148

Scaffold/Matrix Attachment Regions (S/MAR) Regulatory mRNA elements of the eukaryotic genome. These elements coordinate the expression of gene loci. Attachment of a genomic segment to the nuclear matrix places a gene in close proximity to its transcription factor, providing an essential step to expression. 43

SCOP (Structural Classification of Proteins) A structural classification of proteins database for the investigation of sequences and structures. 19, 33, 54, 198

secondary structure The structure of a protein created by the formation of hydrogen bonds between amino acids. See α helix and β sheet. 197

semi-supervised learning A class of machine learning techniques that make use of both labeled and unlabeled data for training. 156

sensitivity See true positive rate. 206

simple sequence repeats (SSR) Tandem iterations of short oligonucleotides. 17

Smith-Waterman algorithm An algorithm to perform pairwise sequence alignments. 33

specificity The proportion of negatives in a binary classification test which are correctly identified. 206

splicing The process by which introns, non-coding regions, are excised out of the premature mRNA transcript and exons, coding regions, are joined together to generate mature mRNA. 5

Support vector machine (SVM) A learning algorithm that performs binary or multi-class supervised classification tasks. 2, 19

Swiss-Prot A curated protein sequence database. 35

synonymous codon usage Codons that are translated into the same amino acid. 42

t-test A statistical test that is used to find out if there is a significant difference between the means (averages) of two different groups. 58, 210

TATA box A DNA consensus sequence found in the promoters of many eukaryotic gene at about -25 nucleotides of the transcription start site. 43

transcription The process of copying information from mRNA into new strands of mRNA. 5

transcription factor A protein that binds to regulatory regions and helps control gene expression. 42, 208

Transcription factor binding sites (TFBS) Short sequence segments (≈10 bp) located near genes' transcription start sites and are recognized by respective transcription factors for gene regulation. 125, 139

transducer (finite state) A finite state transducer is a Finite State Machine (FSM) with both input and outputs. 151

TRANSFAC The TRANSFAC database contains data on transcription factors, their experimentelly-proven binding sites, and regulated genes. 125

transfer RNA (tRNA) A small RNA molecule that participates in protein synthesis. Each tRNA molecule has a trinucleotide region called the anticodon and a region for attaching a specific amino acid. During translation, each time an amino acid is added to the growing protein, a tRNA molecule forms base pairs with its complementary sequence on the mRNA molecule, ensuring that the appropriate amino acid is inserted into the protein. 6, 207

translation The synthesis of a polypeptide chain from an mRNA template. 6

true positive rate The percentage of instances with a particular value that are correctly identified as positive by a test. 25

turn A type of secondary structure. Often responsible for sharp bends and twists in other secondary structures. 10

UniProtKB (Universal Protein Knowledgebase) A repository for the collection of functional information on proteins, with accurate, consistent and rich annotation (`http://www.uniprot.org/`). 7, 33

upregulated gene A gene which has been observed to have higher expression (lower mRNA levels) in one sample compared to another sample (here wild-type fruit fly). 123

vertical averaging A method to combine several ROC curves. Vertical averaging takes vertical samples of the ROC curves for fixed FP rates and averages the corresponding TP rates. 26, 82

Weighted Finite State Machine (WFSM) See FSM. A weight (transition probability) is encoded in the machine. 69, 150

Wilcoxon test An alternative to the t-test for dependent samples. It is designed to test a hypothesis about the median of a population distribution. 58

Bibliography

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman (1990), Basic local alignment search tool., *J Mol Biol*, *215*(3), 403–410. (page 33, 74)

Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., *Nucleic Acids Res*, *25*(17), 3389–3402. (page 36)

Armougom, F., S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame (2006), Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee., *Nucleic Acids Res*, *34*(Web Server issue), W604–8. (page 150)

Ashburner, M., *et al.* (2000), Gene ontology: tool for the unification of biology., *Nat Genet*, *25*(1), 25–29. (page 123, 134)

Ayroles, J., *et al.* (2009), Systems genetics of complex traits in Drosophila melanogaster., *Nat Genet*, *41*(3), 299–307. (page 129)

Baker, E., and R. Hubbard (1984), Hydrogen bonding in globular proteins., *Prog Biophys Mol Biol*, *44*(2), 97–179. (page 9)

Baldi, P., and A. Long (2001), A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes., *Bioinformatics*, *17*(6), 509–519. (page 128)

Bartke, A. (2008), Impact of reduced insulin-like growth factor-1/insulin signaling on aging in mammals: novel findings., *Aging Cell*, *7*(3), 285–290. (page 48)

Barton, G., and M. Sternberg (1987), Evaluation and improvements in the automatic alignment of protein sequences., *Protein Eng*, *1*(2), 89–94. (page 69, 148)

Berman, H. (2008), The Protein Data Bank: a historical perspective., *Acta Crystallogr A*, *64*(Pt 1), 88–95. (page 33)

Berman, H., G. Kleywegt, H. Nakamura, J. Markley, and S. Burley (2010), Safeguarding the integrity of protein archive., *Nature*, *463*(7280), 425. (page 32)

Blum, A., and T. Mitchell (1998), Combining Labeled and Unlabeled Data with Co-Training, *Computational Learning Theory*, pp. 92–100. (page 155)

Bowie, J., R. Luthy, and D. Eisenberg (1991), A method to identify protein sequences that fold into a known three-dimensional structure., *Science*, *253*(5016), 164–170. (page 11)

Broughton, S., and L. Partridge (2009), Insulin/IGF-like signalling, the central nervous system and aging., *Biochem J*, *418*(1), 1–12. (page 39, 48)

Brown, P., and D. Botstein (1999), Exploring the new world of the genome with DNA microarrays., *Nat Genet*, *21*(1 Suppl), 33–37. (page 39)

Budovskaya, Y., K. Wu, L. Southworth, M. Jiang, P. Tedesco, T. Johnson, and S. Kim (2008), An elt-3/elt-5/elt-6 GATA transcription circuit guides aging in C. elegans., *Cell*, *134*(2), 291–303. (page 140)

Bystroff, C., and D. Baker (1998), Prediction of local structure in proteins using a library of sequence-structure motifs., *J Mol Biol*, *281*(3), 565–577. (page 37)

Cai, Z., Z. Mao, S. Li, and L. Wei (2006), Genome comparison using Gene Ontology (GO) with statistical testing., *BMC Bioinformatics*, *7*, 374. (page 35)

Cao, S., J. Dhahbi, P. Mote, and S. Spindler (2001), Genomic profiling of short- and long-term caloric restriction effects in the liver of aging mice., *Proc Natl Acad Sci U S A*, *98*(19), 10,630–10,635. (page 48, 49)

Chalfie, M., Y. Tu, G. Euskirchen, W. Ward, and D. Prasher (1994), Green fluorescent protein as a marker for gene expression., *Science*, *263*(5148), 802–805. (page 47)

Chandonia, J., G. Hon, N. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. Brenner (2004), The ASTRAL Compendium in 2004., *Nucleic Acids Res*, *32*(Database issue), D189–92. (page 54)

Chavous, D., F. Jackson, and C. O'Connor (2001), Extension of the Drosophila lifespan by overexpression of a protein repair methyltransferase., *Proc Natl Acad Sci U S A*, *98*(26), 14,814–14,818. (page 141)

Chih-Chung, C., and L. Chih-Jen (2001), LIBSVM: a library for support vector machines., *http://www.csie.ntu.edu.tw/~cjlin/libsvm*. (page 57)

Chikina, M., C. Huttenhower, C. Murphy, and O. Troyanskaya (2009), Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans., *PLoS Comput Biol*, *5*(6), e1000,417. (page 44, 46, 155)

Chintapalli, V., J. Wang, and J. Dow (2007), Using FlyAtlas to identify better Drosophila melanogaster models of human disease., *Nat Genet*, *39*(6), 715–720. (page 39, 45, 48, 49, 50, 74, 129)

Chothia, C., and A. Lesk (1986), The relation between the divergence of sequence and structure in proteins., *EMBO J*, *5*(4), 823–826. (page 33)

Chou, P., and G. Fasman (1978), Prediction of the secondary structure of proteins from their amino acid sequence., *Adv Enzymol Relat Areas Mol Biol*, *47*, 45–148. (page 11)

Chung, R., and G. Yona (2004), Protein family comparison using statistical models and predicted structural information., *BMC Bioinformatics*, *5*, 183. (page 36, 149)

Clancy, D., D. Gems, L. Harshman, S. Oldham, H. Stocker, E. Hafen, S. Leevers, and L. Partridge (2001), Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein., *Science*, *292*(5514), 104–106. (page 48)

Cole, C., J. Barber, and G. Barton (2008), The Jpred 3 secondary structure prediction server., *Nucleic Acids Res*, *36*(Web Server issue), 197–201. (page 12)

Cortes, C., and M. Mohri (2005), Finite-State Transducers in Computational Biology., *Tutorial presented at the 13th Annual International Conference on Intelligent Systems for Molecular Biology.* (page 69, 151)

Cortes, C., P. Haffner, and M. Mohri (2004), Rational Kernels: Theory and Algorithms., *Journal of Machine Learning Research (JMLR)*, *5*, 1035–1062. (page 151)

Cowen, T. (2001), A heady message for lifespan regulation., *Trends Genet*, *17*(3), 109–113. (page 48)

Culp, M., G. Michailidis, and K. Johnson (2009), On multi-view learning with additive models., *Ann. Appl. Stat.*, *3*(1), 292–318. (page 156)

Curtis, R., B. Geesaman, and P. DiStefano (2005), Ageing and metabolism: drug discovery opportunities., *Nat Rev Drug Discov*, *4*(7), 569–580. (page 136)

Dabney, A., J. Storey, and G. Warnes (2010), qvalue: Q-value estimation for false discovery rate control, *R package*, *1.22.0.* (page 138)

de Bakker, P., A. Bateman, D. Burke, R. Miguel, K. Mizuguchi, J. Shi, H. Shirai, and T. Blundell (2001), HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families., *Bioinformatics*, *17*(8), 748–749. (page 150)

De Ferrari, L., and S. Aitken (2006), Mining housekeeping genes with a Naive Bayes classifier., *BMC Genomics*, *7*, 277. (page v, 26, 40, 46, 79, 92, 94, 95, 96, 113, 146)

de Magalhães, J., and O. Toussaint (2004), GenAge: a genomic and proteomic network map of human ageing., *FEBS Lett*, *571*(1-3), 243–247. (page 47, 123, 140)

Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel (2009), e1071: Misc Functions of the Department of Statistics (e1071)., *The R programming language*. (page 80)

Djawdan, M., A. Chippindale, M. Rose, and T. Bradley (1998), Metabolic reserves and evolved stress resistance in Drosophila melanogaster., *Physiol Zool*, *71*(5), 584–594. (page 137)

Dobson, R., P. Munroe, M. Caulfield, and M. Saqi (2009), Global sequence properties for superfamily prediction: a machine learning approach., *J Integr Bioinform*, *6*(1), 109. (page 38)

Durbin, R. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.*, 356 pp., Cambridge University Press, Cambridge. (page 33)

Duret, L., and D. Mouchiroud (2000), Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate., *Mol Biol Evol*, *17*(1), 68–74. (page 42)

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber (2005), BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis., *Bioinformatics*, *21*(16), 3439–3440. (page 125)

Edwards, A., S. Rollmann, T. Morgan, and T. Mackay (2006), Quantitative genomics of aggressive behavior in Drosophila melanogaster., *PLoS Genet*, *2*(9), e154. (page 129)

Eisenberg, E., and E. Levanon (2003), Human housekeeping genes are compact., *Trends Genet*, *19*(7), 362–365. (page 42, 46, 100, 129)

Elango, N., B. Hunt, M. Goodisman, and S. Yi (2009), DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera., *Proc Natl Acad Sci U S A*, *106*(27), 11,206–11,211. (page 16, 43)

Elofsson, A. (2002), A study on protein sequence alignment quality., *Proteins*, *46*(3), 330–339. (page 148)

Engelhardt, B., M. Jordan, K. Muratore, and S. Brenner (2005), Protein molecular function prediction by Bayesian phylogenomics., *PLoS Comput Biol*, *1*(5), e45. (page 35)

Eyrich, V., M. Marti-Renom, D. Przybylski, M. Madhusudhan, A. Fiser, F. Pazos, A. Valencia, A. Sali, and B. Rost (2001), EVA: continuous automatic evaluation of protein structure prediction servers., *Bioinformatics*, *17*(12), 1242–1243. (page 12)

Farre, D., N. Bellora, L. Mularoni, X. Messeguer, and M. Alba (2007), Housekeeping genes tend to show reduced upstream sequence conservation., *Genome Biol*, *8*(7), R140. (page 79)

Flicek, P., *et al.* (2010), Ensembl's 10th year., *Nucleic Acids Res*, *38*(Database issue), D557–62. (page 1)

Foret, S., R. Kucharski, Y. Pittelkow, G. Lockett, and R. Maleszka (2009), Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes., *BMC Genomics*, *10*, 472. (page 43, 103)

Fowlkes, C., *et al.* (2008), A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm., *Cell*, *133*(2), 364–374. (page 39)

Freilich, S., T. Massingham, S. Bhattacharyya, H. Ponsting, P. Lyons, T. Freeman, and J. Thornton (2005), Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins., *Genome Biol*, *6*(7), R56. (page 42)

Freilich, S., T. Massingham, E. Blanc, L. Goldovsky, and J. Thornton (2006), Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins., *Genome Biol*, *7*(10), R89. (page 42)

Freitas, A., D. Wieser, and R. Apweiler (2010), On the importance of comprehensible classification models for protein function prediction., *IEEE/ACM Trans Comput Biol Bioinform*, *7*(1), 172–182. (page 23)

Friedberg, I. (2006), Automated protein function prediction–the genomic challenge., *Brief Bioinform*, *7*(3), 225–242. (page 33)

Frishman, D., and P. Argos (1995), Knowledge-based protein secondary structure assignment., *Proteins*, *23*(4), 566–579. (page 10)

Frith, M., Y. Fu, L. Yu, J. Chen, U. Hansen, and Z. Weng (2004), Detection of functional DNA motifs via statistical over-representation., *Nucleic Acids Res*, *32*(4), 1372–1381. (page 125, 139)

Fu, C., M. Hickey, M. Morrison, R. McCarter, and E. Han (2006), Tissue specific and non-specific changes in gene expression by aging and by early stage CR., *Mech Ageing Dev*, *127*(12), 905–916. (page 48)

Fu, X., *et al.* (2009), Estimating accuracy of RNA-Seq and microarrays with proteomics., *BMC Genomics*, *10*, 161. (page 146)

Ganapathi, M., P. Srivastava, S. Das Sutar, K. Kumar, D. Dasgupta, G. Pal Singh, V. Brahmachari, and S. Brahmachari (2005), Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes., *BMC Bioinformatics*, *6*, 126. (page 17, 43, 46, 106)

Gardiner-Garden, M., and M. Frommer (1987), CpG islands in vertebrate genomes., *J Mol Biol*, *196*(2), 261–282. (page 16, 43)

Garnier, J., D. Osguthorpe, and B. Robson (1978), Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins., *J Mol Biol*, *120*(1), 97–120. (page 11)

Gerstein, M., and M. Levitt (1996), Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures., *Proc Int Conf Intell Syst Mol Biol*, *4*, 59–67. (page 69, 148)

Ginalski, K., J. Pas, L. Wyrwicz, M. von Grotthuss, J. Bujnicki, and L. Rychlewski (2003), ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure., *Nucleic Acids Res*, *31*(13), 3804–3807. (page 36)

Girardot, F., C. Lasbleiz, V. Monnier, and H. Tricoire (2006), Specific age-related signatures in Drosophila body parts transcriptome., *BMC Genomics*, *7*, 69. (page 47, 49, 131)

Gopal, S., *et al.* (2001), Homology-based annotation yields 1,042 new candidate genes in the Drosophila melanogaster genome., *Nat Genet*, *27*(3), 337–340. (page 35)

Grandori, C., *et al.* (2003), Werner syndrome protein limits MYC-induced cellular senescence., *Genes Dev*, *17*(13), 1569–1574. (page 140)

Greer, E., and A. Brunet (2008), Signaling networks in aging., *J Cell Sci*, *121*(Pt 4), 407–412. (page 48, 140)

Handstad, T., A. Hestnes, and P. Saetrom (2007), Motif kernel generated by genetic programming improves remote homology and fold detection., *BMC Bioinformatics*, *8*, 23. (page 26, 35, 52, 58, 64, 69, 70)

Hastings, K. (1996), Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families., *J Mol Evol*, *42*(6), 631–640. (page 41)

Herman, N., and T. Schneider (1992), High information conservation implies that at least three proteins bind independently to F plasmid incD repeats., *J Bacteriol*, *174*(11), 3558–3560. (page 24)

Holley, L., and M. Karplus (1989), Protein secondary structure prediction with a neural network., *Proc Natl Acad Sci U S A*, *86*(1), 152–156. (page 11)

Holzenberger, M., J. Dupont, B. Ducos, P. Leneuve, A. Geloen, P. Even, P. Cervera, and Y. Le Bouc (2003), IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice., *Nature*, *421*(6919), 182–187. (page 48)

Hong, M., A. Myers, P. Magnusson, and J. Prince (2008), Transcriptome-wide assessment of human brain and lymphocyte senescence., *PLoS ONE*, *3*(8), e3024. (page 48, 49, 131)

Hosack, D., G. J. Dennis, B. Sherman, H. Lane, and R. Lempicki (2003), Identifying biological themes within lists of genes with EASE., *Genome Biol*, *4*(10), R70. (page 124)

Hou, Y., W. Hsu, M. Lee, and C. Bystroff (2004), Remote homolog detection using local sequence-structure correlations., *Proteins*, *57*(3), 518–530. (page 37)

Houthoofd, K., B. Braeckman, I. Lenaerts, K. Brys, A. De Vreese, S. Van Eygen, and J. Vanfleteren (2002), Ageing is reversed, and metabolism is reset to young levels in recovering dauer larvae of C. elegans., *Exp Gerontol*, *37*(8-9), 1015–1021. (page 136)

Hua, S., and Z. Sun (2001), A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach., *J Mol Biol*, *308*(2), 397–407. (page 11)

Huang da, W., *et al.* (2007), DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists., *Nucleic Acids Res*, *35*(Web Server issue), W169–75. (page 124)

Hunt-Newbury, R., *et al.* (2007), High-throughput in vivo analysis of gene expression in Caenorhabditis elegans., *PLoS Biol*, *5*(9), e237. (page 48)

Jaakkola, T., M. Diekhans, and D. Haussler (2000), A discriminative framework for detecting remote protein homologies., *J Comput Biol*, *7*(1-2), 95–114. (page 35)

Jeffrey, G., and W. Saenger (1991), Hydrogen Bonding in Biological Structures, *Springer*. (page 9)

Jin, L., and R. Lloyd (1997), In situ hybridization: methods and applications., *J Clin Lab Anal*, *11*(1), 2–9. (page 47)

Joachims, T. (1999), *Making Large-Scale Support Vector Machine Learning Practical*, 169–184 pp., MIT Press. (page 22)

Jones, D. (1999a), Protein secondary structure prediction based on position-specific scoring matrices., *J Mol Biol*, *292*(2), 195–202. (page 9, 11, 54, 74)

Jones, D. (1999b), GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences., *J Mol Biol*, *287*(4), 797–815. (page 36)

Jones, K., and F. Robertson (1970), Localisation of reiterated nucleotide sequences in Drosophila and mouse by in situ hybridisation of complementary RNA., *Chromosoma*, *31*(3), 331–345. (page 39)

Kabsch, W., and C. Sander (1983a), Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features., *Biopolymers*, *22*(12), 2577–2637. (page 9, 10, 54)

Kabsch, W., and C. Sander (1983b), How good are predictions of protein secondary structure?, *FEBS Lett*, *155*(2), 179–182. (page 11)

Kadota, K., J. Ye, Y. Nakai, T. Terada, and K. Shimizu (2006), ROKU: a novel method for identification of tissue-specific genes., *BMC Bioinformatics*, *7*, 294. (page 24)

Kaeberlein, M., B. Jegalian, and M. McVey (2002), AGEID: a database of aging genes and interventions., *Mech Ageing Dev*, *123*(8), 1115–1119. (page 48)

Kawabata, T., and K. Nishikawa (2000), Protein structure comparison using the markov transition model of evolution., *Proteins*, *41*(1), 108–122. (page 149)

Kayo, T., D. Allison, R. Weindruch, and T. Prolla (2001), Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys., *Proc Natl Acad Sci U S A*, *98*(9), 5093–5098. (page 49)

Kim, S., and E. Rulifson (2004), Conserved mechanisms of glucose sensing and regulation by Drosophila corpora cardiaca cells., *Nature*, *431*(7006), 316–320. (page 137)

Kim, S., *et al.* (2005), Age-dependent changes of gene expression in the Drosophila head., *Neurobiol Aging*, *26*(7), 1083–1091. (page 47, 48, 49)

Klebes, A., B. Biehs, F. Cifuentes, and T. Kornberg (2002), Expression profiling of Drosophila imaginal discs., *Genome Biol*, *3*(8), 1–16. (page 47)

Kloczkowski, A., K. Ting, R. Jernigan, and J. Garnier (2002), Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence., *Proteins*, *49*(2), 154–166. (page 11)

Kolpakov, R., G. Bana, and G. Kucherov (2003), mreps: Efficient and flexible detection of tandem repeats in DNA., *Nucleic Acids Res*, *31*(13), 3672–3678. (page 17, 84)

Kretschmann, E., W. Fleischmann, and R. Apweiler (2001), Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT., *Bioinformatics*, *17*(10), 920–926. (page 33)

Landis, G., D. Abdueva, D. Skvortsov, J. Yang, B. Rabin, J. Carrick, S. Tavare, and J. Tower (2004), Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster., *Proc Natl Acad Sci U S A*, *101*(20), 7663–7668. (page 47, 48, 49)

Larracuente, A., T. Sackton, A. Greenberg, A. Wong, N. Singh, D. Sturgill, Y. Zhang, B. Oliver, and A. Clark (2008), Evolution of protein-coding genes in Drosophila., *Trends Genet*, *24*(3), 114–123. (page 41)

Lawson, M., and L. Zhang (2008), Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region., *Gene*, *407*(1-2), 54–62. (page 17, 43, 46, 105)

Lee, C., R. Weindruch, and T. Prolla (2000), Gene-expression profile of the ageing brain in mice., *Nat Genet*, *25*(3), 294–297. (page 49)

Lee, C., D. Allison, J. Brand, R. Weindruch, and T. Prolla (2002), Transcriptional profiles associated with aging and middle age-onset caloric restriction in mouse hearts., *Proc Natl Acad Sci U S A*, *99*(23), 14,988–14,993. (page 49)

Lehner, B., and A. Fraser (2004), Protein domains enriched in mammalian tissue-specific or widely expressed genes., *Trends Genet*, *20*(10), 468–472. (page 42)

Lein, E., *et al.* (2007), Genome-wide atlas of gene expression in the adult mouse brain., *Nature*, *445*(7124), 168–176. (page 39)

Lercher, M., A. Urrutia, and L. Hurst (2002), Clustering of housekeeping genes provides a unified model of gene order in the human genome., *Nat Genet*, *31*(2), 180–183. (page 129)

Lesk, A., M. Levitt, and C. Chothia (1986), Alignment of the amino acid sequences of distantly related proteins using variable gap penalties., *Protein Eng*, *1*(1), 77–78. (page 69, 148)

Leslie, C., E. Eskin, and W. Noble (2002), The spectrum kernel: a string kernel for SVM protein classification., *Pac Symp Biocomput*, pp. 564–575. (page 35)

Leslie, C., E. Eskin, A. Cohen, J. Weston, and W. Noble (2004), Mismatch string kernels for discriminative protein classification., *Bioinformatics*, *20*(4), 467–476. (page 35)

Levin, J., B. Robson, and J. Garnier (1986), An algorithm for secondary structure determination in proteins based on sequence similarity., *FEBS Lett*, *205*(2), 303–308. (page 11)

Liao, B., and J. Zhang (2006), Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution., *Mol Biol Evol*, *23*(6), 1119–1128. (page 129)

Liao, L., and W. Noble (2003), Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships., *J Comput Biol*, *10*(6), 857–868. (page 33, 35, 59)

Libert, S., J. Zwiener, X. Chu, W. Vanvoorhies, G. Roman, and S. Pletcher (2007), Regulation of Drosophila life span by olfaction and food-derived odors., *Science*, *315*(5815), 1133–1137. (page 141)

Lin, Y., L. Seroude, and S. Benzer (1998), Extended life-span and stress resistance in the Drosophila mutant methuselah., *Science*, *282*(5390), 943–946. (page 48)

Lodish, H., A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. T. Matsudaira (2007), *Molecular Cell Biology*, W.H.Freeman & Co Ltd. (page 9)

Loewenstern, D., and P. Yianilos (1999), Significantly lower entropy estimates for natural DNA sequences., *J Comput Biol*, *6*(1), 125–142. (page 24)

Lyko, F., B. Ramsahoye, and R. Jaenisch (2000), DNA methylation in Drosophila melanogaster., *Nature*, *408*(6812), 538–540. (page 102, 103)

Lyne, R., *et al.* (2007), FlyMine: an integrated database for Drosophila and Anopheles genomics., *Genome Biol*, *8*(7), R129. (page 112)

Magwire, M. (2007), Mutations increasing Drosophila melanogaster life span., *PhD dissertation.* (page 47, 48, 49, 129)

Margelevicius, M., and C. Venclovas (2010), Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison., *BMC Bioinformatics*, *11*(1), 89. (page 38)

Marsden, R., L. McGuffin, and D. Jones (2002), Rapid protein domain assignment from amino acid sequence using predicted secondary structure., *Protein Sci*, *11*(12), 2814–2824. (page 36)

Martin, D., M. Berriman, and G. Barton (2004), GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes., *BMC Bioinformatics*, *5*, 178. (page 35)

Martinez, O., and M. Reyes-Valdes (2008), Defining diversity, specialization, and gene specificity in transcriptomes through information theory., *Proc Natl Acad Sci U S A*, *105*(28), 9709–9714. (page 24, 76)

Martinez-Lara, E., *et al.* (2003), Glutathione S-transferase isoenzymatic response to aging in rat cerebral cortex and cerebellum., *Neurobiol Aging*, *24*(3), 501–509. (page 136)

Matys, V., *et al.* (2003), TRANSFAC: transcriptional regulation, from patterns to profiles., *Nucleic Acids Res*, *31*(1), 374–378. (page 125)

McDonald, I., and J. Thornton (1994), Satisfying hydrogen bonding potential in proteins., *J Mol Biol*, *238*(5), 777–793. (page 9)

McElwee, J., E. Schuster, E. Blanc, J. Thornton, and D. Gems (2006), Diapause-associated metabolic traits reiterated in long-lived daf-2 mutants in the nematode Caenorhabditis elegans., *Mech Ageing Dev*, *127*(5), 458–472. (page 48)

McElwee, J., *et al.* (2007), Evolutionary conservation of regulated longevity assurance mechanisms., *Genome Biol*, *8*(7), R132. (page 47, 48, 123, 124, 128)

McGuffin, L., and D. Jones (2002), Targeting novel folds for structural genomics., *Proteins*, *48*(1), 44–52. (page 36)

McGuffin, L., K. Bryson, and D. Jones (2001), What are the baselines for protein fold recognition?, *Bioinformatics*, *17*(1), 63–72. (page 36)

Mehryar, M. (1997), Finite-State Transducers in Language and Speech Processing, *Computational Linguistics*, *23*, 269–311. (page 151)

Miller, A. J. (1990), *Subset selection in regression. Monographs on statistics and applied probability.*, 229 pp., Chapman and Hall, London. (page 23, 94)

Miller, R. (2009), Cell stress and aging: new emphasis on multiplex resistance mechanisms., *J Gerontol A Biol Sci Med Sci*, *64*(2), 179–182. (page 48)

Mizuguchi, K., C. Deane, T. Blundell, M. Johnson, and J. Overington (1998a), JOY: protein sequence-structure representation and analysis., *Bioinformatics*, *14*(7), 617–623. (page 150)

Mizuguchi, K., C. Deane, T. Blundell, and J. Overington (1998b), HOMSTRAD: a database of protein structure alignments for homologous families., *Protein Sci*, *7*(11), 2469–2471. (page 150)

Mohri, M., F. Pereira, and M. Riley (2002), Weighted Finite-State Transducers in Speech Recognition, *Computer speech and Language*, *16*, 69–88. (page 151)

Montgomerie, S., S. Sundararaj, W. Gallin, and D. Wishart (2006), Improving the accuracy of protein secondary structure prediction using structural alignment., *BMC Bioinformatics*, *7*, 301. (page 12)

Morozova, T., R. Anholt, and T. Mackay (2007), Phenotypic and transcriptional response to selection for alcohol sensitivity in Drosophila melanogaster., *Genome Biol*, *8*(10), R231. (page 129)

Moult, J., K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano (2009), Critical assessment of methods of protein structure prediction - Round VIII., *Proteins*, *77 Suppl 9*, 1–4. (page 13)

Moustafa, A. (2007), JAligner: Open source Java implementation of Smith-Waterman., *http://jaligner.sourceforge.net/*. (page 54)

Murzin, A., S. Brenner, T. Hubbard, and C. Chothia (1995), SCOP: a structural classification of proteins database for the investigation of sequences and structures., *J Mol Biol*, *247*(4), 536–540. (page 33, 51)

Mutch, D., A. Berger, R. Mansourian, A. Rytz, and M. Roberts (2002), The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data., *BMC Bioinformatics*, *3*, 17. (page 128)

Nishikawa, K., and T. Ooi (1986), Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods., *Biochim Biophys Acta*, *871*(1), 45–54. (page 11)

Pages, H., P. Aboyoun, R. Gentleman, and S. DebRoy (2009), Biostrings: String objects representing biological sequences, and matching algorithms., *Bioconductor*. (page 80)

Park, S., and T. Prolla (2005), Gene expression profiling studies of aging in cardiac and skeletal muscles., *Cardiovasc Res*, *66*(2), 205–212. (page 48, 49)

Parkinson, H., *et al.* (2007), ArrayExpress–a public database of microarray experiments and gene expression profiles., *Nucleic Acids Res*, *35*(Database issue), D747–50. (page 123)

Partridge, L. (2008), Some highlights of research on aging with invertebrates, 2008., *Aging Cell*, *8*(5), 509–513. (page 19, 47)

Patient, S., D. Wieser, M. Kleen, E. Kretschmann, M. Jesus Martin, and R. Apweiler (2008), UniProtJAPI: a remote API for accessing UniProt data., *Bioinformatics*, *24*(10), 1321–1322. (page 75)

Pattison, J., L. Folk, R. Madsen, T. Childs, and F. Booth (2003), Transcriptional profiling identifies extensive downregulation of extracellular matrix gene expression in sarcopenic rat soleus muscle., *Physiol Genomics*, *15*(1), 34–43. (page 49)

Piper, M., and A. Bartke (2008), Diet and aging., *Cell Metab*, *8*(2), 99–104. (page 47, 136)

Piper, M., and L. Partridge (2007), Dietary restriction in Drosophila: delayed aging or experimental artefact?, *PLoS Genet*, *3*(4), e57. (page 47)

Pletcher, S., S. Libert, and D. Skorupa (2005), Flies and their golden apples: the effect of dietary restriction on Drosophila aging and age-dependent gene expression., *Ageing Res Rev*, *4*(4), 451–480. (page 47, 48, 49)

Plotkin, J., H. Robins, and A. Levine (2004), Tissue-specific codon usage and the expression of human genes., *Proc Natl Acad Sci U S A*, *101*(34), 12,588–12,591. (page 42)

Poirot, O., K. Suhre, C. Abergel, E. O'Toole, and C. Notredame (2004), 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment., *Nucleic Acids Res*, *32*(Web Server issue), W37–40. (page 150)

Pollastri, G., D. Przybylski, B. Rost, and P. Baldi (2002), Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles., *Proteins*, *47*(2), 228–235. (page 11)

Qian, L., and R. Bodmer (2009), Partial loss of GATA factor Pannier impairs adult heart function in Drosophila., *Hum Mol Genet*. (page 140)

Qian, N., and T. Sejnowski (1988), Predicting the secondary structure of globular proteins using neural network models., *J Mol Biol*, *202*(4), 865–884. (page 11)

Qin, X., S. Ahn, T. Speed, and G. Rubin (2007), Global analyses of mRNA translational control during early Drosophila embryogenesis., *Genome Biol*, *8*(4), R63. (page 107)

Quinlan, J. (1990), Induction of Decision Trees., *Readings in Machine Learning*. (page 33)

R, D. C. T. (2009), R: A Language and Environment for Statistical Computing. (page 125)

Raghava, G., and J. Han (2005), Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein., *BMC Bioinformatics*, *6*, 59. (page 44)

Rangwala, H., and G. Karypis (2005), Profile-based direct kernels for remote homology detection and fold recognition., *Bioinformatics*, *21*(23), 4239–4247. (page 35)

Rao, A., A. Hero, D. States, and J. Engel (2007), Motif discovery in tissue-specific regulatory sequences using directed information., *EURASIP J Bioinform Syst Biol*, p. 13853. (page 43, 46)

Reverter, A., A. Ingham, and B. Dalrymple (2008), Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes., *BioData Min*, *1*(1), 8. (page 129)

Ritchie, W., S. Granjeaud, D. Puthier, and D. Gautheret (2008), Entropy measures quantify global splicing disorders in cancer., *PLoS Comput Biol*, *4*(3), e1000,011. (page 24)

Rodwell, G., *et al.* (2004), A transcriptional profile of aging in the human kidney., *PLoS Biol*, *2*(12), e427. (page 49)

Rost, B., and C. Sander (1993), Prediction of protein secondary structure at better than 70% accuracy., *J Mol Biol*, *232*(2), 584–599. (page 11)

Rusch, D., *et al.* (2007), The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific., *PLoS Biol*, *5*(3), e77. (page 32)

Russell, R., and G. Barton (1994), Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility., *J Mol Biol*, *244*(3), 332–350. (page 36)

Sadreyev, R., and N. Grishin (2003), COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance., *J Mol Biol*, *326*(1), 317–336. (page 38)

Saigo, H., J. Vert, N. Ueda, and T. Akutsu (2004), Protein homology detection using string alignment kernels., *Bioinformatics*, *20*(11), 1682–1689. (page 35)

Salamov, A., and V. Solovyev (1995), Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments., *J Mol Biol*, *247*(1), 11–15. (page 11)

Salminen, A., J. Huuskonen, J. Ojala, A. Kauppinen, K. Kaarniranta, and T. Suuronen (2008), Activation of innate immunity system during aging: NF-kB signaling is the molecular culprit of inflamm-aging., *Ageing Res Rev*, *7*(2), 83–105. (page 140)

Scheffer, T., and M. Krogel (2004), *Multirelational learning, text mining, and semi-supervised learning for functional genomics.*, vol. 57(1/2), 61-81 pp., Springer. (page 156)

Schmitt, A., and H. Herzel (1997), Estimating the entropy of DNA sequences., *J Theor Biol*, *188*(3), 369–377. (page 24)

Schneider, T. (2000), Evolution of biological information., *Nucleic Acids Res*, *28*(14), 2794–2799. (page 24)

Schug, J., W. Schuller, C. Kappen, J. Salbaum, M. Bucan, and C. J. Stoeckert (2005), Promoter features related to tissue specificity as measured by Shannon entropy., *Genome Biol*, *6*(4), R33. (page 16, 24, 42, 46, 79, 101, 102)

Shakhnarovich, G., T. Darrell, and P. Indyk (2005), *Nearest-neighbor methods in learning and vision theory and practice.*, vol. Neural information processing series, MIT Press, Cambridge, Mass. (page 33)

Shannon, C. (1948), A Mathematical Theory of Communication., *Bell Sys. Tech. J.*, *27*, 379–423, 623–656. (page 23, 24)

Shi, J., T. Blundell, and K. Mizuguchi (2001), FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties., *J Mol Biol*, *310*(1), 243–257. (page 36, 150)

Simossis, V., and J. Heringa (2005), PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information., *Nucleic Acids Res*, *33*(Web Server issue), W289–94. (page 149)

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005), ROCR: visualizing classifier performance in R., *Bioinformatics*, *21*(20), 3940–3941. (page 58, 83)

Skorupa, D., A. Dervisefendic, J. Zwiener, and S. Pletcher (2008), Dietary composition specifies consumption, obesity, and lifespan in Drosophila melanogaster., *Aging Cell*, *7*(4), 478–490. (page 47)

Smith, T., and M. Waterman (1981), Identification of common molecular subsequences., *J Mol Biol*, *147*(1), 195–197. (page 33)

Soding, J. (2005), Protein homology detection by HMM-HMM comparison., *Bioinformatics*, *21*(7), 951–960. (page 36)

Stebbings, L., and K. Mizuguchi (2004), HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database., *Nucleic Acids Res*, *32*(Database issue), D203–7. (page 150)

Storey, D. (2003), The positive false discovery rate: A Bayesian interpretation and the q-value., *Annals of Statistics*, *31*, 2013–2035. (page 28)

Strait, B., and T. Dewey (1996), The Shannon information entropy of protein sequences., *Biophys J*, *71*(1), 148–155. (page 24)

Su, A., *et al.* (2004), A gene atlas of the mouse and human protein-encoding transcriptomes., *Proc Natl Acad Sci U S A*, *101*(16), 6062–6067. (page 39, 45, 75)

Tautz, D., and C. Pfeifle (1989), A non-radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene hunchback., *Chromosoma*, *98*(2), 81–85. (page 39)

The Arabidopsis Initiative (2000), Analysis of the genome sequence of the flowering plant Arabidopsis thaliana., *Nature*, *408*(6814), 796–815. (page 35)

The UniProt Consortium (2010), The Universal Protein Resource (UniProt) in 2010., *Nucleic Acids Res*, *38*(Database issue), D142–D148. (page 1, 7, 32, 75)

Thompson, F., G. Barker, T. Nolan, D. Gems, and M. Viney (2009), Transcript profiles of long- and short-lived adults implicate protein synthesis in evolved differences in ageing in the nematode Strongyloides ratti., *Mech Ageing Dev*, *130*(3), 167–172. (page 48)

Tissenbaum, H., and G. Ruvkun (1998), An insulin-like signaling pathway affects both longevity and reproduction in Caenorhabditis elegans., *Genetics*, *148*(2), 703–717. (page 48)

Tomancak, P., B. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. Celniker, and G. Rubin (2007), Global analysis of patterns of gene expression during Drosophila embryogenesis., *Genome Biol*, *8*(7), R145. (page 40)

Umeda-Kameyama, Y., M. Tsuda, C. Ohkura, T. Matsuo, Y. Namba, Y. Ohuchi, and T. Aigaki (2007), Thioredoxin suppresses Parkin-associated endothelin receptor-like receptor-induced neurotoxicity and extends longevity in Drosophila., *J Biol Chem*, *282*(15), 11,180–11,187. (page 141)

Vapnik, V. N. (1999), *The Nature of Statistical Learning Theory (Information Science and Statistics).*, Springer. (page 19, 33)

Vinogradov, A. (2004), Compactness of human housekeeping genes: selection for economy or genomic design?, *Trends Genet*, *20*(5), 248–253. (page 129)

von Ohsen, N., I. Sommer, R. Zimmer, and T. Lengauer (2004), Arby: automatic protein structure prediction using profile-profile alignment and confidence measures., *Bioinformatics*, *20*(14), 2228–2235. (page 148)

Wallqvist, A., Y. Fukunishi, L. Murphy, A. Fadel, and R. Levy (2000), Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases., *Bioinformatics*, *16*(11), 988–1002. (page 36)

Wang, Y., M. Jorda, P. Jones, R. Maleszka, X. Ling, H. Robertson, C. Mizzen, M. Peinado, and G. Robinson (2006), Functional CpG methylation system in a social insect., *Science*, *314*(5799), 645–647. (page 16, 102, 103)

Wang, Y., R. Sadreyev, and N. Grishin (2009a), PROCAIN: protein profile comparison with assisting information., *Nucleic Acids Res*, *37*(11), 3522–3530. (page 38)

Wang, Y., R. Sadreyev, and N. Grishin (2009b), PROCAIN server for remote protein sequence similarity search., *Bioinformatics*, *25*(16), 2076–2077. (page 38)

Wang, Z., M. Gerstein, and M. Snyder (2009c), RNA-Seq: a revolutionary tool for transcriptomics., *Nat Rev Genet*, *10*(1), 57–63. (page 15, 146)

Weissig, H., and P. Bourne (2002), Protein structure resources., *Acta Crystallogr D Biol Crystallogr*, *58*(Pt 6 No 1), 908–915. (page 12)

Welle, S., A. Brooks, J. Delehanty, N. Needler, and C. Thornton (2003), Gene expression profile of aging in human muscle., *Physiol Genomics*, *14*(2), 149–159. (page 49)

Whitehead, A., and D. Crawford (2006), Variation within and among species in gene expression: raw material for evolution., *Mol Ecol*, *15*(5), 1197–1211. (page 48)

Wieser, D., and M. Niranjan (2009), Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores., *In Silico Biol*, *9*(3), 89–103. (page 38)

Wilcox, J. (1993), Fundamental principles of in situ hybridization., *J Histochem Cytochem*, *41*(12), 1725–1733. (page 40)

Wu, C., *et al.* (2006), The Universal Protein Resource (UniProt): an expanding universe of protein information., *Nucleic Acids Res*, *34*(Database issue), D187–91. (page 33)

Wu, K., C. Grandori, M. Amacker, N. Simon-Vermot, A. Polack, J. Lingner, and R. Dalla-Favera (1999), Direct activation of TERT transcription by c-MYC., *Nat Genet*, *21*(2), 220–224. (page 140)

Yamanishi, Y., J. Vert, and M. Kanehisa (2004), Protein network inference from multiple genomic data: a supervised approach., *Bioinformatics*, *20 Suppl 1*, i363–70. (page 156)

Yang, J., A. Su, and W. Li (2005), Gene expression evolves faster in narrowly than in broadly expressed mammalian genes., *Mol Biol Evol*, *22*(10), 2113–2118. (page 42)

Yi, T., and E. Lander (1993), Protein secondary structure prediction using nearest-neighbor methods., *J Mol Biol*, *232*(4), 1117–1129. (page 11)

Zahn, J., *et al.* (2006), Transcriptional profiling of aging in human muscle reveals a common aging signature., *PLoS Genet*, *2*(7), e115. (page 48, 49)

Zhan, M., H. Yamaza, Y. Sun, J. Sinclair, H. Li, and S. Zou (2007), Temporal and spatial transcriptional profiles of aging in Drosophila melanogaster., *Genome Res*, *17*(8), 1236–1243. (page 49, 123, 131, 141)

Zhang, L., and W. Li (2004), Mammalian housekeeping genes evolve more slowly than tissue-specific genes., *Mol Biol Evol*, *21*(2), 236–239. (page 41)

Zhu, J., F. He, S. Hu, and J. Yu (2008), On the nature of human housekeeping genes., *Trends Genet.* (page 41, 43, 44, 46, 129)

Zid, B., A. Rogers, S. Katewa, M. Vargas, M. Kolipinski, T. Lu, S. Benzer, and P. Kapahi (2009), 4E-BP extends lifespan upon dietary restriction by enhancing mitochondrial activity in Drosophila., *Cell*, *139*(1), 149–160. (page 84, 107)

Zvelebil, M., G. Barton, W. Taylor, and M. Sternberg (1987), Prediction of protein secondary structure and active sites using the alignment of homologous sequences., *J Mol Biol*, *195*(4), 957–961. (page 11)

Zweig, M., and G. Campbell (1993), Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine., *Clin Chem*, *39*(4), 561–577. (page 26)

Index