Mining Protein Databases using Machine Learning Techniques

Renata da Silva Camargo[1,3] and Mahesan Niranjan[1,2]

[1]Department of Computer Science, The University of Sheffield, Regent Court, Sheffield, UK

[2]School of Electronics and Computer Science, University of Southampton

Summary

With a large amount of information relating to proteins accumulating in databases widely available online, it is of interest to apply machine learning techniques that, by extracting underlying statistical regularities in the data, make predictions about the functional and evolutionary characteristics of unseen proteins. Such predictions can help in achieving a reduction in the space over which experiment designers need to search in order to improve our understanding of the biochemical properties. Previously it has been suggested that an integration of features computable by comparing a pair of proteins can be achieved by an artificial neural network, hence predicting the degree to which they may be evolutionary related and homologous.

We compiled two datasets of pairs of proteins, each pair being characterised by seven distinct features. We performed an exhaustive search through all possible combinations of features, for the problem of separating remote homologous from analogous pairs, we note that significant performance gain was obtained by the inclusion of sequence and structure information. We find that the use of a linear classifier was enough to discriminate a protein pair at the family level. However, at the superfamily level, to detect remote homologous pairs was a relatively harder problem. We find that the use of nonlinear classifiers achieve significantly higher accuracies.

In this paper, we compare three different pattern classification methods on two problems formulated as detecting evolutionary and functional relationships between pairs of proteins, and from extensive cross validation and feature selection based studies quantify the average limits and uncertainties with which such predictions may be made. Feature selection points to a "knowledge gap" in currently available functional annotations. We demonstrate how the scheme may be employed in a framework to associate an individual protein with an existing family of evolutionarily related proteins.

1 Introduction

Information relating to proteins, such as their amino acid sequence, three dimensional atomic structures and annotations of potential functions obtained from biochemical experiments, is accumulating in publicly available databases at a very rapid rate. With the availability and large amount of data comes the natural question of the type of inferences we can make about newly discovered proteins such as their structural, functional and evolutionary characteristics. Such data mining or machine learning approaches have attracted much interest recently and have been applied to a range of problems in bioinformatics and computational biology. Automatic predictions of functional properties and relationships, derived from model based inference, can

[3]To whom correspondence should be addressed. E-mail: renata.camargo05@ntlworld.com

be very useful in many areas. These include the experimental search space in drug discovery projects and automatic functional annotations performed by teams of curators. One does not immediately conclude that an automatic predictor is to replace experimental or expert determination of biological properties, but what should be seen as important is the potential reduction in the space to search, in order to design experiments, generate hypotheses and detect and correct errors. Such research is nowadays widespread in many topics in bioinformatics that involve high throughput experiments.

In this paper we focus on a particular problem to do with proteins, that of modelling the functional and evolutionary relationships between pairs of proteins based on a number of features that can be computed by comparing their sequences, structures and information pertaining to their biological functions contained in databases. Holm *et al.* [1], continued in Dietman *et al.* [2], formulate the problem of detecting evolutionary relationships between proteins as a pattern classification problem. We start from this formulation and pursue a computational study that leads to several interesting observations. Specifically, Dietmann *et al.* use an artificial neural network to approximate the probability that two given proteins are homologous. They use the Structural Classification of Proteins (SCOP) database [3], in which proteins are grouped in a hierarchy based on their evolutionary relationships, to define protein pairs that are homologous. A set of features are computed to characterise pairs of proteins from their sequences, molecular structures and annotations about them in databases. A multi-layer perceptron (MLP) classifier is trained on this data to yield a score between zero and one representing the confidence with which one can assert that two given proteins are homologous.

While our work starts from the above work of Dietmann *et al.* [2], we make a number of observations not noted in their paper. Firstly, we compare three different classifiers and establish their relative performances, measured in terms of ROC curves and present associated uncertainties on this figures of merit. Secondly, we have carried out an exhaustive search through all subsets of features and established their relative importance. Such a search enables us to pinpoint what we will call a "knowledge gap" in information available in databases of functional annotation of proteins. Thirdly, we have extended the formulation, from making predictions about a pair of proteins, to ask how might one use this framework to assign a new protein to a specific family in the SCOP hierarchy. We are able to demonstrate that generating a hypothesis about which family a single new protein might belong to, can indeed be teased out of the predictions made from pair-wise classifications by majority voting.

There have been other similar attempts to use machine learning methods to automatically associate proteins to SCOP families. Several authors have looked at purely sequence based methods (*e.g.* the seminal work of Jaakkola *et al.* [4]) and have derived an extensive body of methods to essentially learn distances in the sequence space that take into account the probabilistic distributions of protein families. The use of hidden Markov models and kernels of various types fall into this category of work. Alternatively researchers have sought to induce simple rules from the data [5] by the selection and refinement of sequence based information. The "SCOPmap" work [6] is also a related approach that attempts to combine diverse sources of information and Chi *et al.* [7] have developed a protein fold classification system.

In this paper we report on two classification problems set up (*Materials and Methods*): detecting if a pair of proteins is evolutionarily related (same SCOP family) and that of detecting remote homology (same SCOP superfamily) using public biological database. Figure 1 defines these associations.

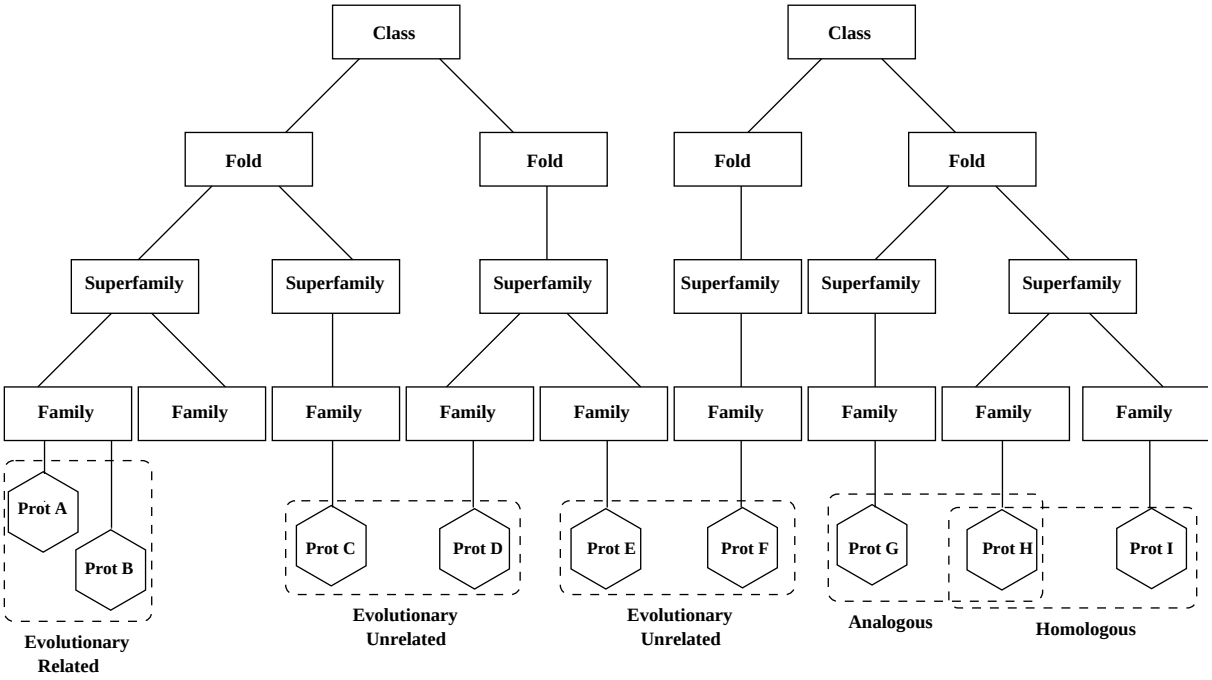


Figure 1: Schematic diagram of the SCOP hierarchical classification of proteins and the selection of evolutionarily related, unrelated, homologous and analogous proteins for the two classification problems considered in this paper

2 Methods

2.1 Data

We set up two classification problems: evolutionarily related vs unrelated protein pairs and homologous vs analogous protein pairs using the SCOP database, which is a hierarchical organisation of proteins whose three-dimensional structures have been experimentally determined and their evolutionary origins reasonably well understood. The hierarchy is in four levels, with the nodes labelled *class*, *fold*, *superfamily* and *family*. Proteins grouped at the family level are known to have a common evolutionary origin while at the higher superfamily level it is claimed that they "probably have evolved from a common ancestor". At the fold level proteins have structural similarities and class refers to broad secondary structural features that are common. We selected $N = 27,724$ unique protein chains, and compared the classification index numbers of each protein N_i, at family and superfamily levels, against $N-1$ proteins, with N_i left out. For multi-domains proteins we considered the first matched domain. After the all index comparison, we selected only the protein pairs that have the structure alignment available in the FSSP database [8], which we relied heavily on to extract features used in the classification experiments (see later).

Figure 1 shows a schematic diagram of the SCOP database and the choice of class labels. For the first data set, we defined a pair of proteins to be evolutionarily related if they belonged to the same SCOP family and found $21,800$ pairs in this category for which we could compute the discriminant features (see later). $85,954$ pairs of counter examples were generated as evolutionarily unrelated, taken from the same class in SCOP hierarchy, but each having a different fold. A further $67,498$ pairs were found from proteins taken from different branches at the class level.

The second problem consisted of 7,982 homologous pairs, which were defined as belonging to the same superfamily, but not members of the same family. The negative examples for this problem were analogues, defined as having the same fold, but not the same superfamily. We constructed features for 11,179 such protein pairs.

2.2 Features

For each protein pair as described above, we obtained seven characteristics, again following the work of Holm *et al.* [1], that contain information about their evolutionary and functional relationships. The features used are: structure similarity, sequence family overlap, enzyme class, site overlap, function preference, keyword overlap and sequence identity. These are described in the following paragraphs:

Structure similarity is a continuous valued feature obtained by a measure of how well two three dimensional structures align. Structure alignment algorithms such as Dali [9] enable structure similarity computation as a Z-score statistic taken over all alignments. Pre-computed alignments for pairs of proteins whose structures have been determined are available in the FSSP database and were used in this work.

Sequence family overlap is a binary feature that defines whether the protein pair shares overlap between their lists of homologous proteins, which is set to value one; otherwise the feature is set to value zero. The lists of homologous proteins were extracted from the HSSP database [10].

Enzyme class is a discrete feature, which defines if the pair of proteins under consideration is classified as sharing a functional similarity in terms of the biochemical reaction they catalyse. We looked at the first level of enzyme classification number and labelled the protein pairs to have a value one for this feature (*e.g.* `1ljr` which has EC number 2.5.1.18 and `1e2a`, whose EC number is 2.7.1.69) and zero otherwise.

Site overlap is a binary feature that takes the value one should the two proteins share a common active site to which a ligand molecule is known to bind. That is given by site annotation extracted from Swiss-Prot (Bairoch *et al.* [11], Release 39) database and ligand crystal structure information extracted from PDB database as follows: Firstly, we looked at each element in the list of homologous proteins for a pair and we selected the site annotation [1] and the amino acid coordinates. We searched for sequence identity in the list of homologous proteins and we considered only the amino acids with variability smaller than ten. Then, we compare the conserved amino acids with the ones related to the annotation and verify whether there exist matches, meaning site overlap. Secondly, for each protein in a pair, we selected the homologous and structurally similar protein list from FSSP database. Additionally, for each element in the lists, we searched for the ligand atoms, at the PDB database. With that information we calculated the distance between ligand molecules and amino acids in the protein, by using the tool Ligand Protein Contact [12]. Only residues in which the contact distance is smaller than four angstroms were considered [1, 12]. We searched for potential functional amino acids, by looking for sequence identity in the three-dimensional alignment between the protein and each homologue. Then we checked whether any of the identical amino acids had the same residue type and position as those found in contact with a ligand molecule. Finally, we concluded that a site overlap existed if the ligand is in contact with at least two amino acids.

Function preference is a continuous valued feature reflecting a measure of amino acid sequence conservation and clustering of conserved residues. We computed this feature by starting with the sequences of the protein pairs and following precisely the steps detailed in Holm *et al.* [1].

Keyword overlap is also a continuous valued feature obtained by a measure of frequency of common function information between a pair of proteins. We started by selecting 813 unique keywords from Swiss-Prot database. For each protein in the pair, we looked for their homologous proteins listed in the HSSP database, by reading the HSSP data file and selecting the PDB codes and Swiss-Prot identifiers. With these lists of identifiers, we searched from the field keyword (KW) in Swiss-Prot for all keywords related to each identifier in the list of homologous proteins, generating a large set of keywords. We then calculated the frequency in which each keyword in the list occurs generating a vector in keyword space. The relative keyword frequency is given by normalising this vector by its length. Then, the keyword overlap is defined as the dot product of their relative keyword frequency vectors.

Sequence identity is a continuous valued feature obtained by a measure of how similar is the amino acid sequence alignment of a pair of proteins. Pre-computed alignments for pairs of proteins are available in the FSSP database.

Histograms showing the distribution are available at supplementary material accompanying this paper.

2.3 Classifiers

For statistical pattern classification we used three algorithms: Perceptron, Multi-layer Perceptron and Support Vector Machines. The Perceptron is a linear classifier which imposes a hyperplane class boundary in the space of features, seven dimensional geometric space when we design classifiers using all the features as inputs. The MLP and SVM classifiers are capable of forming complex boundaries in the feature space, and are known to achieve generalisation (or the ability to perform classification on data outside the training set) in different ways. Complexity of an MLP is controlled by cross validation and early stopping during gradient descent training, while the SVM seeks to maximise a margin between the class boundary and correctly classified examples. MLPs are approximations to posterior probabilities of class membership and thus their solution is generally influenced by the locations of *all* the data in the distribution, while SVM solutions are set by data that lie *close to the class boundary*, emphasising the nature of the classification problem. Due to these known differences, we applied three algorithms on the two tasks to compare their performances and to determine the limits of statistical classification on these tasks.

Perceptron and Multi-layer Perceptron were implemented in C^{++} and for Support Vector Machines we used the package `SVMlight` [13]. Classifier architectures used were as follows: single layer perceptron like algorithm with hyperbolic tangent activation function; multi-layer perceptron with six hidden units, hyperbolic tangent activation function for all units, learning rate equal to 0.02 and momentum of 0.7. For the SVM we used the radial basis functions as kernels and optimised the kernel width and margin parameters by cross validation.

We made use of cross validation, not only to set the hyper-parameters of the classifiers (such as the number of MLP hidden units), but to assess the uncertainties associated with the fact that

the data we have is a finite sample from some underlying distribution in the space of features. Along with these uncertainties, outliers in the data cause variation in the performance on a hold out set. Thus we randomly partitioned the data into 30 training-validation-test splits, preserving the relative ratios of positives and negatives in each of these groups. In reporting results we quote test set performance and standard deviations computed over the 30 test sets.

3 Results and Discussion

3.1 Classification of pairs of proteins

We find that the simple perceptron classifier achieves very high accuracy for the task of detecting membership of the same SCOP family. The gain of using a nonlinear kernel classifier is marginal. Hence we did not pursue implementing an MLP for this problem. Note that individual features are all very good discriminants for this problem, following from the fact that proteins assigned to a specific family in SCOP are those that are well understood to have known common evolutionary origins. We also note that the gain in using a combination of all seven features is only marginal when compared to the best individual feature.

	Homologous vs. Analogous						Related vs. Unrelated			
	Perceptron		MLP		SVM		Perceptron		SVM	
Input features	mean	st. dev.	mean	st. dev.	mean	st. dev.	mean	st. dev.	mean	st. dev.
Structure Similarity	50.97	6.03	60.85	0.60	60.84	0.63	97.4	0.82	97.7	0.04
Sequence Family Overlap	52.47	8.79	58.82	0.77	59.62	0.56	98.05	0.05	98.05	0.05
Enzyme Class	57.46	8.95	64.99	0.47	64.93	0.65	82.10	14.24	87.56	0.10
Site Overlap	50.46	8.59	58.33	0.60	58.54	0.61	75.57	32.09	91.35	0.09
Function Preference	54.54	7.14	58.34	0.60	58.46	0.60	86.22	20.04	92.07	0.09
Keyword Overlap	59.73	12.05	70.73	0.41	70.63	0.43	87.79	20.54	92.55	0.09
Sequence Identity	51.20	1.19	66.96	0.55	66.83	0.65	97.56	0.02	97.91	0.05
All Seven Features	59.49	0.70	75.91	0.72	75.16	0.49	98.20	0.05	98.66	0.03

Table 1: Recognition rates (percentage) obtained for different classifiers on the two problems.

Classifer	Evolutionary Related vs Unrelated	Homologous vs Analogous
Perceptron	0.993	0.787
Multi-layer Perceptron	N/A	0.818
Support Vector Machine	0.987	0.778

Table 2: Areas under the ROC curves.

Table 1 shows recognition average accuracies and standard deviations computed over 30 random partitions for different classifiers on the two problems of homologous vs. analogous pairs and evolutionary related vs. unrelated pairs. Figure 3 shows these performances in the form of receiver operating characteristic (ROC) curves. Such a performance graph shows the balance one can strike between different types of errors and the area under this graph, being an

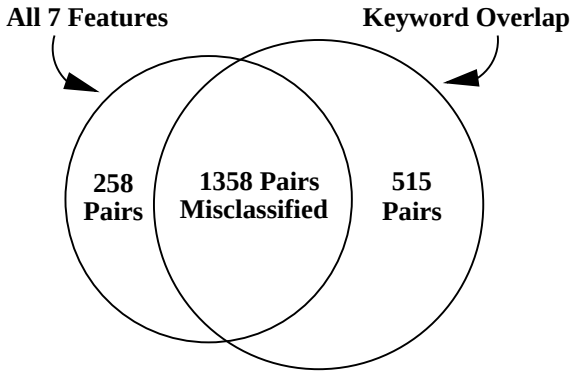


Figure 2: Venn diagram of classification errors, comparing errors made by a classifier with the best single feature (*keyword overlap*) with the errors made by classifier using all the features. The overlapping proteins are misclassified by both sets of features. Those that are misclassified by *keyword overlap*, but are correctly classified by the inclusion of additional computable features are of potential interest as annotation knowledge gaps.

integral over all possible false positive operating points is an effective statistical measure of performance differences, see Table 2.

The second problem of detecting remote homologues, or common membership at the superfamily level, however, is a relatively harder problem. Here, the distributions of the data in the seven dimensional space appears a complex one. We find that the use of nonlinear classifiers achieve significantly higher accuracies. Further using all the seven features as inputs significantly outperforms the best individual feature. We also find that the best individual discriminant for this problem is *keyword overlap*, suggesting that functional annotation about these proteins contains very good information about them. More importantly, the inclusion of other structure and sequence based computable features enhances the accuracy with which we can group these proteins, over and above what keywords in annotations suggest. We may regard this as a "knowledge gap" in what is currently known about proteins in databases such as Swiss-Prot [11, 14]. We took one of the partitions of the data and listed the protein pairs on which the classifiers made errors. The overlap is shown as a Venn diagram in Figure 2. To verify this assertion we took twenty pairs of proteins that were misclassified using the keyword overlap feature but were correctly classified when structure and sequence information were included.

Table 3 shows 20 homologous protein pairs that were misclassified in our experiments using keyword overlap only, but were correctly classified when information about sequence and structure were included. The table shows re-computed values of this feature using the most recent version of Swiss-Prot (version 50.5). We note that there is systematic increase in the scores for these protein pairs and that for the threshold we used for this classifier, 0.6, several of the 20 pairs would be correctly detected as homologous had we used the recent database of annotations. Note as an aside, the objective here is not to demonstrate an advantage in a statistical sense. In employing such a decision support system, one is unlikely to be interested in an overall performance of a fully automatic system. Instead, and as discussed later, we might be interested in the use of this to help detect anomalies in manually constructed databases of annotations.

We further note that the pattern classification tasks considered here are characterised by *a priori* imbalances in the available data. For the problem of detecting protein pairs that belong to the same family, the number of counter examples we have vastly exceeds the number of positive

		Keyword Overlap Score	
Protein 1	Protein 2	Swiss-Prot 39	Swiss-Prot 50.5
1a7w	1tafA	0.188	0.707
1a7w	1bh9B	0.188	0.743
1aoiB	1tafA	0	0.200
1aoiB	1bh9B	0	0.213
1a8l	1gp1A	0	0.410
1a8l	1qq2A	0	0.640
1fo5	1gp1A	0	0.380
1fo5	1qq2A	0	0.927
1erv	1gp1A	0.011	0.216
1erv	1g7eA	0.172	0.353
1erv	1b9yC	0.002	0.049
1erv	1qgvA	0	0.116
1qfnA	1qq2A	0.02	0.787
1a8y	1g7eA	0	0.103
1thx	1a8y	0.023	0.050
1erv	1a8y	0.002	0.035
1thx	1b9yC	0.003	0.068
1thx	1g7rA	0.080	0.367
1b9yC	1qgvA	0	0.026
1bjx	1b9yC	0.015	0.019

Table 3: Keyword overlap scores for 20 homologous protein pairs that were misclassified using this feature only.

examples. Thus a classifier that does not detect any pair of proteins as belonging to the same family will have a baseline performance of 86.0%. Similarly, for the problem of detecting homologous proteins, the random classifier can be expected to have an accuracy of 58%. The discrimination obtained with individual features is generally greater than these figures on both problems. The exceptions to this are the performance for discrete features (*e.g.* site overlap), where perceptron-like classifiers are not entirely appropriate.

3.2 Feature Selection

In pattern classification problems, described by a number of features, it is often desirable to explore their relative importance. Often it is possible for discriminant information to lie in a small subspace of the features. When the number of features is high the search for an optimum subspace is difficult due to the combinatorial search required and one resorts to suboptimal sequential forward selection or backward elimination procedures (*e.g.* Lovell *et al.* [15]). Since we only have a total of seven features we carried out an exhaustive search through all possible combinations of features.

Table 4 shows the best subsets of features and the corresponding accuracies for SVM classifiers in these subspaces. For the easier problem of detecting if a pair of proteins is from the same family, the search does not give us any additional insight. The gain over the best single feature

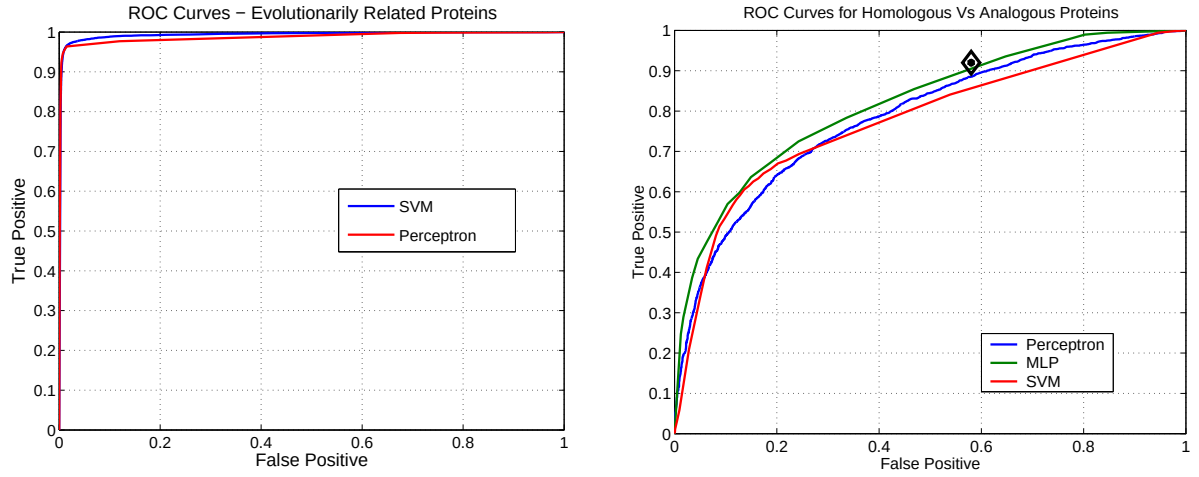
	Evolutionary Related Protein Pairs		Homologous Protein Pairs	
No. Features	SVM Accuracy	Feature Combination	SVM Accuracy	Feature Combination
1	98.0%	{Seq. Family Overlap}	70.6%	{Keyword Overlap}
2	98.2%	{Structure Similarity, Sequence Identity}	72.7%	{Keyword Overlap, Sequence Identity}
3	98.6%	{Structure Similarity, Sequence Identity, Seq. Family Overlap}	74.1%	{Keyword Overlap, Sequence Identity, Structure Similarity}
4	98.6%	{Structure Similarity, Sequence Identity, Seq. Family Overlap, Enzyme Class}	75.1%	{Keyword Overlap, Sequence Identity, Enzyme Class, Structure Similarity}
5	98.7%	{Structure Similarity, Sequence Identity, Seq. Family Overlap, Enzyme Class, Site Overlap}	75.2%	{Keyword Overlap, Sequence Identity, Enzyme Class, Structure Similarity, Seq. Family Overlap}
6	98.7%	{Structure Similarity, Sequence Identity, Seq. Family Overlap, Enzyme Class, Site Overlap, Keyword Overlap}	75.2%	{Keyword Overalp, Sequence Identity, Enzyme Class, Structure Similarity, Seq. Family Overlap, Site Overlap}

Table 4: Search for combinations of features that maximize discrimination for the problems of detecting pairs of evolutionarily related proteins and homologous proteins.

of sequence family overlap is marginal. However, for the more difficult problem of separating remote homologues from analogous proteins, we note that significant performance gain is obtained by the inclusion of sequence and structure based information over the best single feature classifier, from 70.6% to 74.1%. Inclusion of further information (*e.g.* site overlap) does not lead to significant performance gains.

3.3 Blind test on individual unseen proteins

In order to evaluate if our method may be useful as an aid in automatically deriving homologous relationships of proteins newly included in databases, we randomly selected 23 proteins shown in Table 5 and removed all pairs with these from our datasets. The 406 protein pairs thus removed formed a test set, 297 of which were homologous and 109 were analogous. This table shows classification performance of the best classifier (MLP) on this new test set. The overall accuracy achieved on this test set is 72.7%. Note the way the data set was constructed results in several homologous pairs for the test proteins of interest and small numbers of analogous pairs.



(a) (b)

Figure 3: Performance of the classifiers shown as receiver operating characteristic (ROC) curves for the two problems considered: Evolutionary related vs unrelated, and homologous vs analogous. The diamond point on the latter graph indicates classification of pairs formed with protein `2hdda` **used in the test set (see text).**

This is because analogous pairs, forming the negative examples, span a much wider space (*i.e.* a large number of possibilities of forming such pairs exists). For the protein `2hdda`, for which there is a reasonable number of positive and negative examples, we computed the true positive and false positive rates, and superposed it on the ROC curves in Figure 3. It is seen that these are consistent with the ROC curve of the MLP.

We then classified each of these test protein pairs at the family level. Results of this classification are shown in Table 5. For classification at this level we chose the perceptron algorithm and set the classification threshold at a level at which the false positive rate was zero. Thus we get a measure of how accurately can we recognise membership of a family for a protein which is not in the training set while not generating any false positives. Taking the protein `1xel` as an example, we note that of the 20 proteins that have a common membership in family `C.2.1.2`, 12 of them are correctly classified by a perceptron. We can roughly say that `1xel` can be automatically labelled as belonging to this family with a confidence 0.6, noting that the training set contains no pairs of proteins involving `1xel`. This does not hold for protein `1cz1A`, however, for which homologous proteins can be detected with impressive accuracy, which is an atypical lucky example.

4 Conclusions

This work is an exploration of the degree to which statistical regularities in databases of protein information may be extracted, by the formulation of classification problems. We formulated two inference problems and carried out a thorough evaluation of different machine learning algorithms that form linear and complex class boundaries. The results quantify the extent to which inferences can be made by combining information stored in protein databases. Of particular note is that the comparison between single features and combination of features suggests the existence of a knowledge gap in the annotation of protein functions, in that more can be

		Family Level				Superfamily Level			
		Evol. Related		Evol. Unrelated		Homologous		Analogous	
Protein	SCOP ID	Pairs	Correct	Pairs	Correct	Pairs	Correct	Pairs	Correct
1dlyA	*A*.1.1.1	0	0	46	46	17	15	0	0
1vin	*A*.74.1.1	0	0	63	63	4	4	0	0
1jkw	*A*.74.1.1	2	2	46	46	4	3	0	0
1bu2A	*A*.74.1.1	3	3	61	61	4	3	0	0
1f5qB	*A*.74.1.1	2	2	51	51	4	4	0	0
2hddA	*A*.4.1.1	9	9	35	35	13	12	24	10
1nksA	*C*.37.1.1	0	0	147	147	39	31	0	0
1zin	*C*.37.1.1	2	2	87	87	33	25	0	0
1stmA	*B*.10.1.2	6	1	43	43	19	16	0	0
1dmr	*B*.52.2.2	4	4	74	74	4	1	2	2
1cz1A	*C*.1.8.3	4	4	25	25	14	13	66	66
1xel	*C*.2.1.2	20	12	153	153	28	19	0	0
1gnd	*C*.3.1.3	0	0	101	101	15	11	0	0
1dekA	*C*.37.1.1	3	1	136	136	33	12	0	0
1bjx	*C*.47.1.2	0	0	0	0	18	12	2	2
1bam	*C*.52.1.3	2	2	38	38	11	8	2	2
1a5r	*D*.15.1.1	2	0	5	5	5	0	10	10
1ecsA	*D*.32.1.2	5	5	6	6	5	3	0	0
1b66A	*D*.96.1.2	2	2	13	13	4	2	0	0
1f88A	*F*.2.1.1	0	0	28	28	12	1	0	0
2omf	*F*.4.3.1	7	7	21	21	4	4	3	3
1cmr	*G*.3.7.2	1	0	0	0	4	1	0	0
1txb	*G*.7.1.1	2	0	0	0	3	0	0	0

Table 5: Classification results for the 406 unseen protein pairs, for both data sets.

inferred by the incorporation of structural and sequence information than can be gained from keywords used in functional annotations.

The conclusions from this paper suggest a number of possible ways in which the framework may be deployed. Firstly, it could be useful in providing a support tool to select candidate proteins for manual curation in constructing databases of protein function. Secondly, it may also be possible to employ such a machine learning framework to look for inconsistencies in manually constructed functional annotations, because they will show up as errors. We can thus construct multiple prediction systems as the two considered here and look for protein pairs that repeatedly show up as errors, thereby automatically highlighting erroneous annotations, the propagation of which has to be managed systematically in the coming years [16]. Finally yet another way of using a classification framework is one of detecting outliers. Li *et al.* [17] define and detect outliers with respect to a classification problem. Annotation errors may show up as outliers when one formulates a pattern recognition problem and these may be flagged as candidates for investigation by an expert.

Authors contributions

RC and MN contributed equally to the paper.

Acknowledgements

RC was funded by a grant from the Engineering and Physical Research Council (EPSRC) UK.

References

[1] Liisa Holm and Chris Sander. Decision support system for the evolutionary classification of protein structures. *Proceedings of the 5th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 5:140–146, 1997.

[2] Sabine Dietmann and Liisa Holm. Identification of homology in protein structure classification. *Nature Structural Biology*, 8(11):953–957, 2001.

[3] Alex G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[4] Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies. *Advances in Neural Information Processing Systems*, 11:487–493, 1999.

[5] Boris Mirkin and O. Ritter. *Genomics and Proteomics: Functional and Computational Aspects*, chapter A Feature Based Approach to Discrimination and Prediction of Protein Folding Groups, pages 157–177. Springer, 2000.

[6] Sara Cheek, Yuan Qi, S. Sri Krishna, Lisa N. Kinch, and Nick V. Grishin. SCOPmap: Automated assignment of protein structures to evolutionary superfamilies. *BMC Bioinformatics*, 5(197):197, 2004.

[7] Pin-Hao Chi, Chi-Ren Shyu, and D. Xu. A fast SCOP fold classification system using content-based E-predict algorithm. *BMC Bioinformatics*, 7(362):362, 2006.

[8] Liisa Holm and Chris Sander. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research*, 22:3600–3609, 1994.

[9] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–138, 1993.

[10] Chris Sander and Reinhard Schneider. Database of homology-derived protein structures. *Proteins: Structure, Function & Genetics*, 9:56–68, 1991.

[11] A. Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1):304–305, 2000.

[12] Vladimir Sobolev, Anatoli Sorokine, Jaime Prilusky, Enrique E. Abola, and Marvin Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332, 1999.

[13] T. Joachims. *Advances in Kernel Methods: Support Vector Learning*, chapter Making large-scale SVM learning practical, pages 169–184. MIT Press, 1999.

[14] Hellen Berman. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[15] D. R. Lovell, B. Rosario, Mahesan Niranjan, R. W. Prager, K. J. Dalton, R. Derom, and J. Chalmers. Design, construction and evaluation of systems to predict risk in obstetrics. *International Journal of Medical Informatics*, 46(3):159–173, 1997.

[16] Walter R. Gilks, Benjamin Audit, Daniela de Angelis, Sophia Tsoka, and Christos A. Ouzounis. Percolation of annotation errors through hierarchically structured protein sequence databases. *Mathematical Biosciences*, 193(2):223–234, 2005.

[17] Hongyu Li and Mahesan Niranjan. Outlier detection in benchmark classification tasks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5:557–560, 2006.