An Assessment of Feature Relevance in Predicting Protein Function from Sequence

Ali Al-Shahib, Chao He, Aik Choon Tan, Mark Girolami, and David Gilbert

Bioinformatics Research Centre, Department of Computing Science
University of Glasgow, Glasgow G12 8QQ, UK
{alshahib,chaohe,actan,girolami,drg}@dcs.gla.ac.uk

Abstract. Improving the performance of protein function prediction is the ultimate goal for a bioinformatician working in functional genomics. The classical prediction approach is to employ pairwise sequence alignments. However this method often faces difficulties when no statistically significant homologous sequences are identified. An alternative way is to predict protein function from sequence-derived features using machine learning. In this case the choice of possible features which can be derived from the sequence is of vital importance to ensure adequate discrimination to predict function. In this paper we have shown that carefully assessing the discriminative value of derived features by performing feature selection improves the performance of the prediction classifiers by eliminating irrelevant and redundant features. The subset selected from available features has also shown to be biologically meaningful as they correspond to features that have commonly been employed to assess biological function.

1 Introduction

Performing protein sequence comparison to achieve homology generally indicates similarity in function and structure. The standard way of predicting the function of a newly sequenced protein is to use sequence comparison tools, such as BLAST [1] that can identify the most similar proteins and use their function to infer that of the new sequence. However this method often fails for low sequence similarity proteins. Thus, other alternative techniques such as predicting protein function from microarray expression analysis [3], protein secondary structure [4] and protein sequence features [5, 8] have been proposed. King et al. [5] show that using physical and chemical properties directly derived from protein sequence provides a novel way of predicting protein function with reasonable accuracy. Different from direct sequence comparisons, this method allows an appropriate classification algorithm, which is learned from appropriate discriminative features, to be used as a discrimination function which maps the protein sequence to a biological function. However, the question is how to obtain those most appropriate and discriminative features from all available features for higher performance of the prediction system.

The elimination of irrelevant and redundant features in the data results in many advantages. First, it enables the classification system to achieve good or even better performance with a selected feature subset, thus reducing the computational costs and avoiding the dimensional curse generally faced by machine learning [2, 10]. Secondly, it helps the human expert to focus on a relevant subset of features, hence providing useful biological knowledge [10]. The most obvious way of performing feature selection is to manually select the biologically most relevant features in the dataset. However, this is not always practical, as many bioinformatics data sets could be associated with large numbers of features and it would be time consuming to manually perform feature selection. Also, there could be some hidden features that one could not possibly recognize the importance by just visualising the dataset. Thus we need an effective, fast and biologically reliable automatic method.

In this study, we employed the theoretically sound and practically feasible *filter* and *wrapper* feature subset selection methods [2] to eliminate irrelevant and redundant features in our feature set, and utilised naive Bayes and decision tree classifiers to asses the accuracy of the prediction as a result of the feature selection. We have shown that performing automatic feature selection when predicting protein function from sequence improves the performance of the predictive classifier compared with considering a full set of derived features. In addition, we have found that as a result of the feature selection, biologically relevant features were chosen that indicates the importance of feature selection in this task.

2 Methodology

2.1 Data Collection and Pre-processing

We have populated a database containing protein sequence information for seven sexually transmitted disease (STD) causing bacteria. These bacteria were chosen because the long term goal of this research is to predict large numbers of novel proteins in these bacteria, in order to further understand the pathogenicity of these organisms. The proteins and their functional annotation were obtained from the Los Alamos Laboratory[1]. By utilising a variety of bioinformatics tools, diverse sequence related features were obtained, which include those derived from the distribution of amino acids (e.g. amino acid composition, length of protein, molecular weight) and those derived from the properties associated with the molecular composition of the protein (e.g. pI, Hydropathicity, aliphatic index). More features such as structural and phylogenetic predictions/hypotheses could have been extracted, but first we want to focus our attention on sequence data alone in order to understand the possible limits of prediction accuracy when forced to rely on the information available within sequence alone. Further work considering additional biologically useful features will be carried out in the future. The entire data set contains 5,149 proteins represented as 33-dimensional feature vectors from 13 different functional classes. The description and sample size of each class are shown in Tab. 1. We have performed a linear normalisation (standardisation) on the data to rescale each feature to mean of 0 and standard deviation of 1.

[1] http://www.stdgen.lanl.gov/

Table 1. Functional classification of proteins according to the classification in [7].

Class ID	Class Name	Sample Size
1	Amino acid biosynthesis	231
2	Biosynthesis of cofactors, prosthetic groups, and carriers	264
3	Cell envelope	577
4	Cellular processes	409
5	Central intermediary metabolism	146
6	DNA metabolism	456
7	Energy metabolism	513
8	Fatty acid and phospholipid metabolism	155
9	Purines, pyrimidines, nucleosides, and nucleotides	261
10	Regulatory functions	250
11	Transcription	210
12	Translation	906
13	Transport and binding proteins	771

2.2 Feature Selection Methods

There are two common approaches for feature selection [2]: a *filter* evaluates features according to measures based general statistical characteristics of the data, while a *wrapper* uses the intended prediction algorithm itself to evaluate the usefulness of features. In this study, a variety of both filter and wrapper methods were examined. Experiments were carried out using the WEKA[2] environment [6].

Within the filter, four different search algorithms (forward, backward, genetic and ranker) along with two evaluation criteria (correlation based feature subset evaluation and relief attribute evaluation) were investigated. The selected feature subsets were further employed to devise a multi-class Gaussian naive Bayes classifier to predict protein function. Its performance was compared with that obtained using the full feature set.

For the wrapper feature selection approach, a genetic search algorithm was employed to generate feature subsets which were then evaluated by a decision tree using 5-fold cross-validation. To analyse the discriminatory power of the features that the wrapper selected, we induced decision tree classifiers from the full feature set and the selected feature subset respectively, and their performance was evaluated using 10-fold cross validation.

3 Experimental Results

Filter and wrapper methods have consistently selected common features[3], which illustrates the high discriminatory power. Experimental results on the full feature set as well as the selected subsets by the filter and the wrapper are shown in Fig. 1 and Fig. 2 respectively, which display the boxplots of coverage measurements (percentage of true positives over the sum of true positives and false

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

[3] Please refer to `http://www.brc.dcs.gla.ac.uk/~alshahib/fs.htm` for the entire list of selected features.

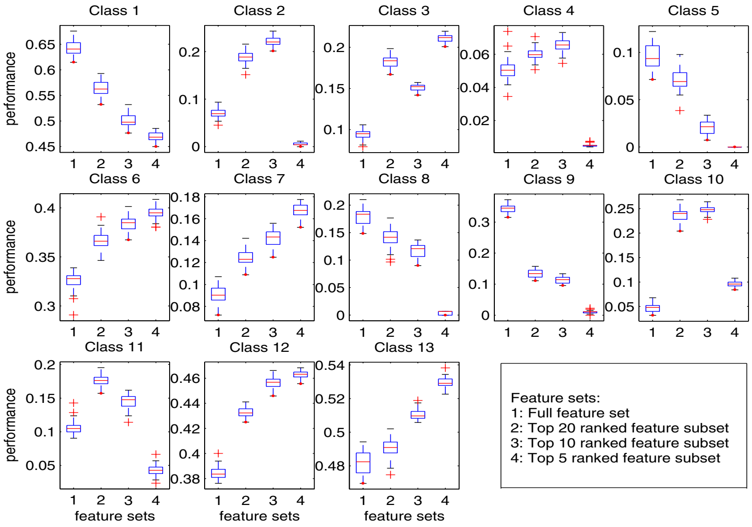

Fig. 1. Boxplot of prediction performance using feature sets selected by filter.

negatives) of 13 classes using 10-fold cross-validation. The performance of 9 out of 13 classes for the filter (using the top 20 and top 10 ranked feature subsets) and 7 out of 13 classes for the wrapper has been improved. Even when using just the top 5 filter-ranked feature subset, the performance of 6 out of 13 classes has been improved. It indicates that original features for most of the protein functional classes are redundant and irrelevant, thus using selected subsets improves performance. All of the 7 classes improved by the wrapper approach were also improved by the filter approach. The performance of four functional classes decreases using both feature selection approaches. They are class 1 (amino acid biosynthesis), 5 (central intermediary metabolism), 8 (fatty acid and phospholipids metabolism) and 9 (purines pyrimidines nucleosides and nucleotides). We argue that for those classes the discriminative information is provided by all features. This indicates that the importance of features differs when predicting different protein functions, which will be further discussed in the following section. Although feature selection improves the performance, the overall performance is still not optimal because it is difficult to predict protein function from amino acid sequence alone [5, 8] which consists of insufficient information for defining protein functions. Adopting additional features such as structural, gene expression or phylogenetic profile may improve the performance. Imbalanced data suffered by the classifier may be another reason (Further investigation addressing this problem is presently being undertaken).

4 Discussion

Having shown in our experiments that employing selected feature subsets improved the prediction performance, we have further noticed that some impor-

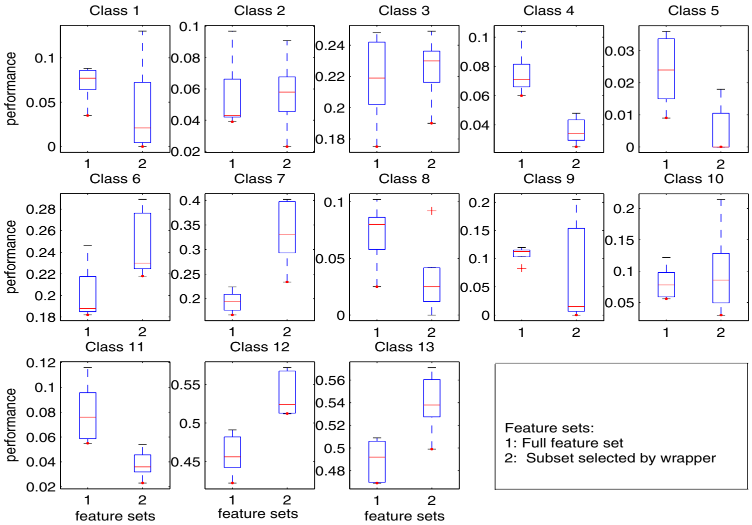

Fig. 2. Boxplot of prediction performance using feature sets selected by Wrapper.

tant features such as Isoelectric point (pI) and Grand Average of Hydropathicity (GRAVY) were commonly selected by both filter and wrapper methods. The filter consistently ranks them as the top 2 discriminative features. In the wrapper, the pI and GRAVY were constantly chosen as the root nodes of decision tree classifiers which indicates the importance of these features in predicting protein function. These selected features are indeed biologically meaningful.

The pI of a protein is the pH of a solution in which the protein has a net electrical charge of zero. The reason why the pI is a biologically relevant discriminatory feature in predicting particular functional classes is because it determines the functionally important charge status of a protein in a given environment, and certain functions critically depend on the net charge of the particular protein. Prominent examples are DNA replication (DNA metabolism), transcriptional and translational functional classes. All of these functions require an interaction with highly acidic nucleotide sequences (DNA or RNA), thus positively charged proteins (i.e. protein with a high pI) are needed. Moreover, certain pI values could be discriminatory for transport and binding proteins especially in pathogenic bacteria (such as the ones in our database). This is because bacterial pathogenic cells tend to acidify the environment by metabolic processes, thus their secreted proteins should have a low pI to function properly in acidic environment compared to intracellular proteins (about neutral environment).

The GRAVY value for a protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence [9]. The hydrophobicity of the cell membrane is a major factor in the transport of the metabolites to and out of the cell. It makes biological sense to associate discriminatory GRAVY values with predicting proteins with transport and binding functions. In addition, certain GRAVY values can also be discriminatory in predicting other classes that are hydrophobicity dependent: such as cell envelope, fatty acid and phospholipid metabolism and energy metabolism.

We believe that the reason why other features, such as molecular weight, protein length and atomic composition were not chosen as top features is because of the redundancy between these features, thus making it less possible for a classifier to discriminate between these features in predicting function. We have also understood from this feature selection process that in many cases more abstract features that have been combined at a high level such as pI and GRAVY are the most discriminatory features compared to low-level features such as composition of a single amino acid in a protein. In future work, we intend to combine similar but not redundant features for example the amino acids' glutamate and aspartate composition as one feature in predicting protein function.

5 Conclusion

In this paper we have reported the employment of theoretically sound and practically feasible filter and wrapper feature selection methods to investigate the feature relevance in predicting protein function from sequence. Our experimental results have shown that performing feature selection on protein chemical and physical sequence data improves the performance of the predictive classifier compared with considering a full set of derived features. We have also shown that the selected features are biologically meaningful to provide high discriminatory power of determining protein functions. We plan further work taking into account additional features such as structural, phylogenetic and expression information in selecting the biologically most relevant features. Other well-established machine learning algorithms such as Support Vector Machines will also be investigated.

References

1. Altschul, S.F., et al.: Basic local alignment search tool. J Mol Biol. **215** (1990) 403-10
2. Guyon, I., Elissee, A.: An introduction to variable and feature selection. JMLR, **3** (2003) 1157-1182
3. Pavlidis, P., et al.: Learning gene functional classifications from multiple data types. J Comp Biol, **9** (2002) 401-11
4. Rost, B.: Protein secondary structure prediction continues to rise, J Struc Biol, **134** (2001) 204-18
5. King, R.D., et al.: The utility of different representations of protein sequence for predicting functional class. Bioinformatics, **17** (2001) 445-54
6. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (2000)
7. Riley, M.:. Functions of the gene products of Escherichia coli. Microbiol Rev, **57:** (1993) 862-952
8. Jensen, R., et al.: Prediction of human protein function according to Gene Ontology categories. Bioinformatics, **19** (2003) 635-42
9. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J Mol Biol, **157** (1982) 105-32
10. Saeys Y, et al.: Fast feature Selection using a simple estimation of distribution algorithm: a case study on splice site prediction. Bioinformatics, **19** (2003) 179-88