

# Protein Function Prediction: Application of a Propositional Rules Learning System to a set of Human Protein Sequences

Manuel J. Gómez<sup>(1)</sup>, Francisco Javier Guijarro<sup>(1)</sup>, Ramón P. Otero<sup>(2)</sup>, Lars J. Jensen<sup>(3)</sup>,

Søren Brunak<sup>(3)</sup>, and Alfonso Valencia<sup>(1)</sup>

<sup>(1)</sup>Centro Nacional de Biotecnología, CSIC, Madrid, Spain.  
mjgommo, gijaro, valencia@cnb.uam.es

<sup>(2)</sup>Universidade da Coruña, Coruña, Spain  
otero@dc.fi.udc.es

<sup>(3)</sup>Center for Biological Sequence Analysis, DTU, Lyngby, Denmark  
brunak@cbs.dtu.dk

**Keywords.** Protein function; Genome annotation; Propositional rules learning system; Machine learning.

## Introduction

Protein function prediction is a key aspect of genome annotation that mainly relies on detecting near or remote homology, by sequence similarity. When no clear homologs can be detected, other function prediction methods are required, such as those based on genome comparisons that detect similarity of phylogenetic profiles, gene neighborhoods and gene fusions. These methods have become possible with the availability of multiple complete genome sequences.

Another method to predict protein function that does not rely on sequence similarity was developed by Jensen *et al* [1]. Starting from a set of human protein sequences, a number of protein features that could be either calculated from the sequence, such as the isoelectric point, or predicted, such as the secondary structure, were deduced. Then, an array of tens of artificial neural networks (ANN), was trained in the association between protein properties and functional classes. These classes were part of two basic classification schemes: the Euclid system [2] and the Enzyme Commission system, limited to the first digit (EC1d). By using a bootstrapping strategy to combine the features, it was possible to identify which features were more discriminatory for the different classes. Also, the method was implemented as a web server (<http://www.cbs.dtu.dk/services/ProtFun>).

In the present work, we have used the same human protein data set, to associate protein features with functional classes with a different machine learning algorithm, *C4.5* [3]. *C4.5* builds decision trees and generates propositional rules that can be used as predictors. More importantly, examination of the rules generated by *C4.5* can produce some insight on the characteristics of proteins that belong to the same functional class, since the rules may correspond to real biological regularities. This is an advantage over ANN approaches, which are considered to behave as "black boxes" in the sense that it is difficult to extract information about the links between attributes and classes. In addition to the Euclid and EC1d classifications, we have also used the molecular function ontology scheme from the GO project. We have also compared the predictive performance of *C4.5* when applied in the context of global classification strategies (in which training is performed by providing examples of all classes) and binary classification strategies (in which training is performed independently for each class, providing positive and negative examples for each of them). Finally, we have manually combined the attributes that represent protein features (obtaining what we call the *processed* data set) in an attempt to generate more clear rules.

Then, the objective of this work is triple: first, we want to compare the predictive performance of two machine learning approaches: ANNs and propositional rule learning; second, we want to establish which training strategy (binary / global), classification scheme (Euclid / EC1d / GO) and data set level of detail (raw / processed) are more appropriated to achieve good predictive performances in the context of our data; and third, we want to take advantage of the capacity of *C4.5* to generate rules that can represent links between protein features and protein functions.

## Results and discussion

From the tests enumerated above, it was concluded that the best combination for achieving improved prediction accuracies would imply using the original information in the raw-data data set, classifying proteins according to the GO classification scheme, and using a binary classification strategy, conditions that resulted in a general prediction accuracy of 70%. Even in these conditions, the predictive performance of the ANNs was better than that of our *C4.5* generated predictor. However, a number of functional classes, in particular

those related with transport and binding, were predicted by the *C4.5* system at reasonable levels of accuracy (84%). The relative defect in predictive performance of our system could be compensated by some properties of propositional rule learning approaches that make them very interesting, such as reduced computational times, easiness of use and, more importantly, the fact that associations between attributes and classes are expressed in a form that can be explored and analyzed by human end users. In this sense, and although the use of the raw-data data set implied the generation of rules with more obscure propositions, it was possible to extract simple rules with biological meaning.

In a second approach, the rule sets obtained were analyzed to determine the distribution of feature utilization in rules, to identify the features that are more relevant for function determination. Although it became clear that the most relevant features vary from class to class, it was possible to observe some general trends: features related with secondary structure, sub-cellular localization and the presence or absence of transmembrane segments and signal peptides, appeared frequently in rules for all classes. The same features were considered also relevant by Jensen *et al*, although the approach to identify them was different [1]. There were, also, some interesting exceptions to those trends. For example, the feature "Isoelectric point", which was quite irrelevant for most classes and classification schemes, appeared frequently in rules that define Translation proteins, in the Euclid classification scheme. An example of such a rule would be:

"*expasy\_pi > 10 -> class Translation [89.5%]*"

that is interpreted as "if the isoelectric point is higher than 10, then the protein is involved in translation, with a likelihood of 89,5%". This rule has a clear biological meaning, since most proteins belonging to this class interact with RNA molecules (for example, ribosomal proteins) and, therefore, must be rich in basic amino acids. Interestingly, the "Isoelectric point" feature was not identified as relevant for classifying any protein from the data set in the previous work [1], presumably because its correlation with charge related features.

Rule based systems have been applied to similar problems before. DesJardins *et al* used three different machine learning techniques to induce classifiers that would predict whether a protein is an enzyme and, if that were the case, to predict its Enzyme Commission class at the first digit and second digit levels [4]. King *et al* used a combination of inductive logic programming clustering and rule learning for predicting functional class from sequence [5,6]. As in previous studies [4], we find that the combination of attributes to generate a processed-data data set, characterized by simpler, more understandable attributes, resulted in a decrease in predictive performance, illustrating the problem of developing the appropriate technique for attribute clustering. We are in the process of overcoming this problem by the use of ILP (inductive logic programming) algorithms, that have been already applied to preprocess the data submitted to *C4.5* and to construct new attributes by combination of others [5,6].

Our results are complementary to those previous studies. We explore the possibility of analyzing feature usage distribution in rules to identify biologically relevant characteristics of the different functional classes. Rules referring to the more relevant features were assumed to be the most informative, and this criteria was used to extract key rules and their associated features. For the future we will explore the implementation of the system as a tool to provide molecular biologists, with functional class predictions, together with the additional information about the rules, most relevant features and relations in the protein space.

#### Acknowledgements

We thank Ramon Alonso-Allende for his help with graphics, and Federico Abascal and Jose Maria Fernandez for help with Perl programming. MJG is recipient of an I3P contract from the Spanish CSIC. JG is beneficiary of a fellowship from the MCYT. The project was in part supported by a EC grant (TEMBLOR).

#### References

- [1] Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. (2002) Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol.* 319(5):1257-65.
- [2] Tamames J, Ouzounis C, Casari G, Sander C, Valencia A. (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics.* 14(6):542-3.
- [3] Quinlan, J.R. (1993) *C4.5: Programms for Machine Learning.* Morgan Kaufmann, San Francisco, CA.
- [4] DesJardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA. (1997) Prediction of Enzyme Classification from Protein Sequence without the use of Sequence Similarity. *ISMB*, 1997: 92-98.
- [5] King RD, Karwath A, Clare A, Dehaspe L. (2000) Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining. *Yeast.* 17(4):283-93.
- [6] King RD, Karwath A, Clare A, Dehaspe L. (2001) The utility of different representations of protein sequence for predicting functional class. *Bioinformatics.* 17(5):445-54.