Discovering Multi-Label Hierarchical Classification Rules for Protein Function Prediction

Roberto Alves¹, Myriam Delgado¹, Fernando Camargo¹, Elaine Benelli², Alex Freitas³

¹Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, UFTPR
Av. Sete de Setembro, 3165, CEP: 80230-901, Curitiba – PR – Brazil

²Departamento de Bioquímica e Biologia Molecular, UFPR
Cx. Postal 19046, CEP: 81531-990, Curitiba – PR – Brazil

³Computing Laboratory and Centre for BioMedical Informatics, University of Kent
CT2 7NF, Canterbury, U.K.
ralves2006@gmail.com, fernubio@yahoo.com, myriamdelg@utfpr.edu.br
benelli@ufpr.br, A.A.Freitas@kent.ac.uk

Abstract. This work proposes the use of an Artificial Immune System (AIS) for predicting biological protein functions described in the gene ontology. The main challenge of this work is to discover a set of classification rules that are both hierarchical and multi-label, in the sense that a single classification rule can assign to a protein several classes (labels) in several levels of the gene ontology's hierarchy. The performance of the proposed algorithm is evaluated considering 3 protein data sets.

1. Introduction

This work addresses the multi-label hierarchical classification task of data mining, where the goal is to discover a classification model that predicts one or more classes – at each hierarchical level – for an example (data instance). Note that in a conventional single-label flat classification task, the classification model must predict just one class for an example. Multi-label hierarchical classification is considerably more difficult than single-label classification, because there are many more classes to be assigned to an example and the hierarchical level of each class must be also considered.

In essence, Bioinformatics involves the development of computational methods (including data mining methods) for the analysis of biological data. This paper focuses on a major bioinformatics problem, viz. the prediction of protein functions from protein sequence data. This is a very challenging problem in general, due to the very complex relationship between a protein's primary sequence and its biological function [Devos, and Valencia, 2000].

Protein functions are comprehensively defined in a structured, standardized dictionary of terms called the Gene Ontology (GO) [The Gene Ontology Consortium, 2004]. GO actually consists of 3 separate "domains" (very different types of GO terms): molecular function, biological process and cellular component. The GO is structurally organized in the form of a direct acyclic graph (DAG), where each GO term represents a node of the hierarchical structure. A "child" node can have one or more parent nodes in the DAG. This work aims at assigning GO terms (classes) to proteins, where each term
represents a biological function (process). The data mining method used in this paper discovers knowledge interpretable by the user, in the form of IF-THEN classification rules, unlike other methods proposed in the literature, whose classification model is typically a "black box" which normally does not provide any insight to the user about interesting hidden relationships in the data.

2. Multi-label Hierarchical Classification with an Artificial Immune System

The algorithm used in this paper is called MHC-AIS (Multi-label Hierarchical Classification with an Artificial Immune System). MHC-AIS is an instance of the relatively new computational intelligence paradigm of AIS [De Castro and Timmis 2002]. The training phase of the algorithm is performed by two major procedures, called Sequential Covering (SC) and Rule Evolution (RE) procedures, The former is usual in induction rules algorithms [Witten e Frank 2005], whilst the latter is often used in evolutionary algorithms, e.g. AIS based on the clonal selection theory. The SC procedure iteratively calls the RE procedure until (almost) all “antigens” (proteins, examples) are covered by the discovered rules. The RE procedure essentially evolves artificial “antibodies” (multi-label classification rules, in our case) that are used to classify antigens. Each antibody (candidate classification rule) consists of two parts: the rule antecedent (IF part), represented by a vector of conditions (attribute-value pairs), and the rule consequent (THEN part), represented by a subset of predicted classes. In this work the classes correspond to GO terms denoting protein functions (see Section 3 for a practical example). The details of MHC-AIS will be described in another paper.

This paper focuses instead on the application of the algorithm to a challenging bioinformatics problem. To place this algorithm in the context of the broader hierarchical classification literature, hierarchical classification methods can be divided in two approaches: local (including the top-down) and global (or big-bang) approaches. A local approach builds separate classifiers for each internal node of a hierarchy. In a global approach only one classifier is built to discriminate all classes in a hierarchy simultaneously. MHC-AIS belongs to the latter approach.

In biological databases a protein is annotated only with its most specific GO term. Given the semantics of the GO’s functional hierarchy, this implicitly means the protein also contains all the functional classes of its ancestral GO terms in the GO's DAG. Hence, MHC-AIS explicitly assigns to each antigen (protein) both its most specific class(es) (GO term(s)) and all its ancestral classes. MHC-AIS also considers the semantics of the GO’s functional hierarchy when creating classification rules – i.e., it guarantees that, if a rule predicts a given GO term, all its ancestral GO terms are also predicted by the rule.

3. Computational Results

MHC-AIS is evaluated considering 3 protein data sets. The first one contains 610 DNA-binding proteins, which are involved in gene expression [Alberts et al. 2002]. The second one contains 1411 ATPases, which catalyze ATP hydrolysis [Alberts et al. 2002]. The third one results from the union of the former two data sets. These datasets contain 11, 15 and 24 attributes, respectively. The predictor attributes in all data sets are mainly PROSITE patterns – a kind of biological motif well-known in bioinformatics [Hulo et al. 2006] – but each data set also has two other attributes, the sequence length and molecular weight. PROSITE patterns are binary attributes, indicating whether or not
the pattern occurs in a given protein’s sequence of amino acids, whilst sequence length and molecular weight are continuous attributes. The number of GO terms (classes to be predicted) in each data set is 38, 57 and 78, respectively.

Recall that in data mining the discovered knowledge should be not only accurate, but also comprehensible to the user [Witten e Frank 2005]. In this spirit, the results can be evaluated according to two criteria, viz. the predictive accuracy and simplicity of the discovered rule set. In this paper, the predictive accuracy is evaluated by the F-measure, which involves computing the precision and recall of the discovered rule set on the test set (unseen during training). More precisely, the set of GO terms predicted for a test example \( t \), denoted \( \text{PredGO}(t) \), consists of the union of all GO terms in the consequent of all rules covering \( t \) – i.e. all rules whose conditions are satisfied by \( t \)’s attribute values. MHC-AIS computes the Precision and Recall for a test example \( t \) – denoted \( P(t) \) and \( R(t) \), respectively – as per equations (1) and (2), where \( \text{TrueGO}(t) \) is the set of true GO terms for test example \( t \).

\[
P(t) = \frac{|\text{PredGO}(t) \cap \text{TrueGO}(t)|}{|\text{PredGO}(t)|} \quad (1)
\]

\[
R(t) = \frac{|\text{PredGO}(t) \cap \text{TrueGO}(t)|}{|\text{TrueGO}(t)|} \quad (2)
\]

In other words, precision is the proportion of true classes among all predicted classes, whilst recall is the proportion of predicted classes among all true classes. The F-measure for a test example \( t \) is given by equation (3), the harmonic mean of \( P \) and \( R \).

\[
F(t) = \frac{(2 \times P(t) \times R(t))}{(1 + P(t) + R(t))} \quad (3)
\]

Finally, once \( P(t) \) and \( R(t) \) have been computed for each test example \( t \), the system computes the overall F-measure over the entire test set \( T \) by equation (4), where \(|T|\) denotes the cardinality of the test set \( T \).

\[
\text{Predictive Accuracy} = F(T) = \frac{\sum_{t \in T} F(t)}{|T|} \quad (4)
\]

The simplicity of the discovered knowledge was measured by the number of discovered rules and average number of rule conditions. The predictive accuracy and simplicity results are presented in Table 1, where the ”±” symbol represents the standard deviations considering the well-known 10-fold cross-validation procedure.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#attrib.</th>
<th>#classes</th>
<th>predictive accuracy</th>
<th>#rules</th>
<th>#conditions by rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA binding</td>
<td>11</td>
<td>38</td>
<td>94.39±0.79</td>
<td>15.30±0.39</td>
<td>3.80±0.10</td>
</tr>
<tr>
<td>ATPase</td>
<td>15</td>
<td>57</td>
<td>89.65±0.93</td>
<td>15.20±0.21</td>
<td>5.77±0.80</td>
</tr>
<tr>
<td>ATPase + DNA bind.</td>
<td>24</td>
<td>78</td>
<td>79.54±3.35</td>
<td>31.33±0.95</td>
<td>8.84±0.17</td>
</tr>
</tbody>
</table>

MHC-AIS obtained satisfactory results with respect to predictive accuracy, considering the complexity of the target problem. The worst accuracy was obtained for the ATPase + DNA-binding data set, whilst the best accuracy was obtained for the DNA-binding data set. A possible explanation for these results is that the DNA-binding data set has the smallest number of classes to be predicted (38) among the three data sets, whilst the ATPase + DNA-binding data set has the largest number of classes (78).

In any case, MHC-AIS succeeded in discovering relatively small rule sets –
particularly in the first two data sets, DNA-binding and ATPase – which facilitates the interpretation of the discovered knowledge by the user. As the third data set is the result of the former two data sets’ union, a larger rule set was discovered by MHC-AIS. Considering simplicity aspects, it is important to point out that, for all data sets, on average less than 40% of predictor attributes were represented in the rule antecedents. An example of a rule discovered rule in the DNA binding data set is presented below:

\[
\text{IF } (\text{PS00676} = 1) \text{ and } (\text{PS00688} = 1) \text{ and } (\text{SQ\_LEN} \geq 464) \text{ and } (\text{MOL\_WE} \geq 56098) \\
\text{THEN } (5488 \text{ or } 43167 \text{ or } 46872)
\]

The biological interpretation of this rule is: if a protein presents “sigma-54 interaction domain ATP-binding region B” and “sigma-54 interaction domain C-terminal part” signatures and “sequence length is greater than 464” and “molecular weight is greater than 56098” then the predicted classes (biological functions) are: “binding” (5488) or “ion binding” (43167) or “metal ion binding” (46872). It is essential to emphasize that the GO hierarchy was considered in the example rule consequent above, i.e. the true hierarchical path is $5488 \rightarrow 43167 \rightarrow 46872$ (from shallower to deeper nodes).

4. Conclusion and Future Work

This work described the application of an artificial immune system-based rule induction algorithm to a challenging and important problem in bioinformatics, namely the prediction of protein function. The problem is challenging, from a data mining point of view, because it is a multi-label and hierarchical classification problem, where many classes can be assigned to an example at each level of the class hierarchy. The obtained results were considered satisfactory in terms of both predictive accuracy and simplicity of the discovered rules.

Future work will involve mainly to compare the results of MHC-AIS with other multi-label hierarchical classification algorithms and to analyze – in collaboration with a biologist – how biologically relevant and useful the discovered rules are.

References


