Characterisation of FAD-family Folds using a Machine Learning Approach

Aik Choon TAN, David GILBERT and Andrew TUSON

Bioinformatics Research Group
Department of Computing
School of Informatics
City University
Northampton Square
EC1V 0HB
London, U.K.
Email: {a.c.tan, drg, andrewt}@soi.city.ac.uk

Abstract

Flavin adenine dinucleotide (FAD) and its derivatives play a crucial role in biological processes. They are major organic cofactors and electron carriers in both enzymatic activities and biochemical pathways. We have analysed the relationships between sequence and structure of FAD-containing proteins using a machine learning approach. Decision trees were generated using the C4.5 algorithm as a means of automatically generating rules from biological databases (TOPS, CATH and PDB). These rules were then used as background knowledge for an ILP system to characterise the four different classes of FAD-family folds classified in Dym and Eisenberg (2001). These FAD-family folds are: glutathione reductase (GR), ferredoxin reductase (FR), *p*-cresol methylhydroxylase (PCMH) and pyruvate oxidase (PO). Each FAD-family was characterised by a set of rules. The "knowledge patterns" generated from this approach are a set of rules containing conserved sequence motifs, secondary structure sequence elements and folding information. Every rule was then verified using statistical evaluation on the measured significance of each rule. We show that this machine learning approach is capable of learning and discovering interesting patterns from large biological databases and can generate "knowledge patterns" that characterise the FAD-containing proteins, and at the same time classify these proteins into four different families.

Keywords: *flavin adenine dinucleotide (FAD); protein structure-sequence-function; machine learning; decision tree; inductive logic programming; knowledge discovery in biological databases.*

1. Introduction

It is generally believed that the three-dimensional structure of a protein is determined by its amino acid sequence. On the other hand, similar protein folds can have very different sequences (Doolittle, 1986). Although some similar protein structures may have low sequence similarity, the consensus sequence pattern exhibited in the particular protein family usually plays an important role in protein function or structure properties. In the post-genome era, one of the 'holy-grails' for the bioinformatics community is to predict the structure and function of a protein from its amino acid sequence. Current factual biological databases are overwhelmed by experimental data, this has motivated the first step to understanding the complex relation between protein sequence, structure and function.

In this paper, we characterise FAD-binding proteins into four different families using a machine learning approach. We have adapted the machine learning algorithm C4.5 (Quinlan, 1993) which outputs a decision tree that is equivalent to a set of symbolic rules. We induced the decision trees in a parallel fashion to output a decision forest which contains rules from different biological databases. These rules were then converted into background knowledge for the second learning system CProgol4.4 (Muggleton, 2001), which derived a rule-set output which was more accurate and comprehensible than the conventional combined data single tree approach.

2. FAD-(Flavin Adenine Dinucleotide) binding protein families

In this study, we focused on flavin adenine dinucleotide (FAD) binding proteins. This is because FAD, and other cofactors such as nicotinamide adenine dinucleotide (NADH) and adeninosine triphosphate (ATP), appear in many biological processes and represent the major fuel molecules in the cell. The main function of flavin-binding proteins is to carry out redox reactions in the cell. The unique structure of flavin enables it to take up one or two electrons from the substrate. The intermediate radical state of flavin (FADH•) is able to react with the most powerful oxidising agent in biological systems: molecular oxygen. This has made flavin different from the other coenzymes (NADH and ATP) and thus it plays an important role in cell metabolism.

Using the Combinatorial Extension (CE) program, Dym and Eisenberg (2001) have classified the FAD-binding proteins into four different structural families. They characterised each family using several conserved sequence and structure motifs, which involves in the cofactor binding site. The four major structural families are:

(a) Glutathione reductase (GR) which adopts a Rossmann fold with xhxhGxGxxGxxxhxxh(x)8hxhE(D) as the most conserved sequence motif. All the family members share the same 3-D structure.

(b) Ferredoxin reductase (FR), having a cylindrical β-domain as the central structure with RxYS(T) as the most conserved sequence motif.
(c) p-cresol methylhydroxylase (PCMH) consisting of two α+β sub-domains and the most conserved sequence motif being P(x)6G(A)xN.
(d) Pyruvate oxidase (PO) with a structure consisting of five parallel β-strands interspersed by α-helices that lie on both sides of the β-sheet with KxLxxLxxxL(x)6S(T)(x)6GxV as the most conserved sequence motif.

3. Machine learning approach

In most application domains we are more concerned with the predictive accuracy than the explanatory power of the learning output. This is not the case in bioinformatics because we believe that the comprehension of a rule (pattern) is as important as the accuracy of the learning system. Biology is an "understanding" orientated subject, and because biological databases are accumulating vast amount of data, human experts find it harder to identify the relationship between properties of the data (e.g. protein structure-sequence-function). Thus, we agree with Muggleton et. al. (1998) when comparing the performance of learning system in a bioinformatics context:

"If the predictive accuracies of two hypotheses are statistically equivalent then the hypothesis with better explanatory power will be preferred. Otherwise the one with higher accuracy will be preferred. (Muggleton et. al. 1998)".

The first machine learning algorithm we chose to adapt for this work was C4.5 release 8 (Quinlan, 1993), the successor of the ID3 learning algorithm (Quinlan, 1986). The decision tree algorithm is well known for its robustness, and learning efficiency with learning time complexity of $O(n\log_2 n)$ where n is the number of attributes. The output of the algorithm is a decision tree, which can be converted into a set of symbolic rules (**IF…THEN…**). The symbolic rules can be directly interpreted and compared with existing biological knowledge. Thus, decision trees have high expressive power in their patterns (rules).

The second learning system applied in this study was CProgol 4.4 (Muggleton, 2001), which is an inductive learning programming (ILP) program. ILP algorithms take examples *E* of a concept (e.g. FAD-binding protein family), together with background knowledge *B* (e.g. relation between a consensus sequence motif and structure) and construct a hypothesis *H* which explains *E* in terms of *B*. The hypothesis *H* can be translated into an (**IF…THEN…**) rule-set that can characterise *E* with relation to *B*. Thus, we believe that our final rule-set can have a high comprehensibility and accuracy in characterising protein (e.g. the FAD-binding families).

4. Methodology

The approach applied here is to derive *n* decision trees from *n* data sets. Each decision tree was induced from an individual data set. These decision trees individually represent different information (e.g. sequence, structure and function) for the FAD-binding protein families in *m* data sets. After growing the *m* decision trees individually, they must be combined to produce some final rules that characterise the protein families. In our approach, we used C4.5 to grow the *m* decision trees, and C4.5 rules to extract the rules from the *m* trees. These rules were used as the additional background knowledge for the ILP system when learning on the training examples. ILP outputs a final rule-set (hypothesis) that has high comprehensibility and accuracy in characterising the protein examples. The rule set will contain relationships between sequence, structure and function.

The data-set that we used in this study consisted of 42 fad-binding proteins which have 3 different attribute properties (*m*=3 in this study). The properties are sequence motifs, structure motifs and enzymatic functional classes. For each family, we divided the training set into positive and negative examples then learned a rule from that group (e.g. for GR family, GR will be the TPs; FR, PCMH and PO are the TNs).

Our research methodology can be summarised as the following steps:

Step 1: Select the target data sets (sequence, structures, function).
Step 2: Clean up the target data sets.
Step 3: Divide-and-conquer search strategy.
Step3a: ***Divide***: For every data set, grow an individual decision tree.
Step3b: ***Extract***: For every decision tree, derive the rules from the trees.
Step3c: ***Merge***: Use the rules induced from the various data sets as the background knowledge for the ILP system.
Step3d:***Conquer***: Apply ILP to "combine" the rules.
Step4: Evaluate the goodness of the rules.
Step5: Repeat step 3 and 4 to obtain the "best" rule-set that characterises the FAD-binding families.

The goodness of the rules can be evaluated using the confusion matrix in table 1. We evaluate the statistical significant of the rule-set by measuring their sensitivity (Sn)[1], specificity (Sp)[2], coefficient correlation (cc)[3], positive predictive value (PPV)[4].

---

[1] $Sn = \frac{TP}{TP + FN}, 0 \le Sn \le 1$

[2] $Sp = \frac{TN}{TN + FP}, 0 \le Sp \le 1$

[3] $cc = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(FN + TN)}}, -1 \le cc \le 1$

Table 1: Confusion matrix showing true positives, true negatives, false positives and false negatives covered by the rules.

	Positive Examples	**Negative Examples**
Covered	True Positive (TP)	False Positive (FP)
Not covered	False Negative (FN)	True Negative (TN)

5. Results and Discussions

We performed two different experiments. The first experiment used our combination method, while the second experiment was conventional data combination, where different data sets were combined into one large table and a single tree induced from the data. The results from these two experiments are shown in table 2 (our combination approach) and table 3 (data combination single tree approach). The rule-set of our approach are shown in figure 1 at the protein topological level with the colouring secondary structure elements represents the sequence-structure relationship.

From table 2 and table 3 we can see that our approach has more accuracy over the GR and FR family and the rule set produced was more comprehensible. This is because our rule-set included structural information that helps to distinguish the noises in the training set. Therefore, our approach increases Sp, cc and PPV of the rules. Furthermore, our rules also indicate the location of the sequence motif in the protein structure, which we believed it is more meaningful for the biologists in understanding the protein sequence-structure relationship. For example, in the case of class GR in our method produces a rule set that can be translated into natural language as follows:

If *the protein has a sequence motif* $GxG(x)_2G(x)_{16\text{-}19}[DE]$ *in* β_1-α_1-β_2 *of the 3-layer* β-β-α *sandwich structure and carried out oxidoreductases reaction* ***Then*** *it is GR family.*

The data combination approach rule only identified the sequence motif for this family. The rule from the data combination approach can be translated into:

If *the protein has a sequence motif* $GxG(x)_2G(x)_{16\text{-}19}[DE]$ ***Then*** *it is class GR.*

Thus the rule induced from the second experiment is less informative compared to our approach. This is because in data combination, the learning algorithm only finds the shortest rule that discriminates between two classes and ignores other 'important' features in the data. For this example, the sequence are more conserved in the family members because these motifs involved in the cofactor binding site. Although the conventional method does

[4] $PPV = \frac{TP}{TP + FP}$

successfully classify the proteins using the most discriminative patterns, it ignores other (structure and function) information in the input data-set. The results will be less useful for knowledge acquisition purposes, because biologists tend to prefer more comprehensive rules that can help them to explain the complex relationship between protein sequence-structure-function properties.

The other advantage of our method compared to data combination using a single-tree is that our approach reduces the learning time complexity and memory space requirements of the algorithm. Most of the learning algorithms receive the input file as a flat file format, thus if the input data-set is very large, the memory space will be taken up by the input files. Thus, the computational demands will increase. At the same time, the learning time for the algorithm will decrease if the input n is a large value. In our approach, we divide the various data sources into sub-tables (n/m), the learning time complexity is $O(mlog_2(n/m))$ + ILP learning time, and thus retain the fast efficient learning time of the learning algorithms (decision trees).

Table 2: Rule-set generated from the combination of decision trees and ILP.

	Rule	Sn	Sp	cc	PPV
GR	Class('GR',A):-protein(A,B,C,D),Sequence(B,GxG(x)$_2$G(x)$_{16-19}$[DE]),Structure(C,bbasandwich),Has_seq(strand1_helix1_strand2,B,C),Function(D,oxidoreductases).	1.0	1.0	1.0	1.0
FR	Class('FR',A):-protein(A,B,C,D),Sequence(B,RxY[ST]), Structure(C,betabarrel),Has_seq(strand4,B,C), Function(D,oxidoreductases).	1.0	1.0	1.0	1.0
PCMH	Class('PCMH',A):-protein(A,B,C,D),Sequence(B,[AP](x)$_{6-8}$[AG]xN),Structure(C,abasandwich),Has_seq(helix1_strand2, B,C),Function(D,oxidoreductases).	1.0	1.0	1.0	1.0
PO	Class('PO',A):-protein(A,B,C,D),Sequence(B,KxL(x)$_2$(x)$_3$L), Structure(C,abasandwich),has_seq(helix1_strand2,B,C), Function(D,oxidoreductases).	1.0	1.0	1.0	1.0

Table 3: Rule-set generated from single decision trees derived from data combination.

	Rule	Sn	Sp	cc	PPV
GR	GxG(x)$_2$G(x)$_{16-19}$[DE]=yes $\rightarrow$ class GR	1.0	.97	.91	.86
FR	RxY[ST] = yes $\rightarrow$ class FR	1.0	.91	.91	.91
PCMH	[AP](x)$_{6-8}$[AG]xN = yes $\rightarrow$ class PCMH	1.0	1.0	1.0	1.0
PO	K(x)$_7$I(x)$_2$D(x)$_{10}$D = yes $\rightarrow$ class PO	1.0	1.0	1.0	1.0

6. Conclusions

When trying to learn from large and diverse data sets (e.g. biological databases) it is important to produce a rule-set that encapsulates all the information from different sources. In this work we investigated an approach where decision trees were derived individually from various data sets and the rules from the decision trees were given to the ILP as background knowledge. The ILP system induces rules by combining the additional background

knowledge and the training examples to produce a "knowledge" rule-set. We have shown that the rules produced from our approach are more informative than the conventional method. Our current work is to apply this approach to larger data sets in order to learn significant relationships between protein sequence-structure-function.

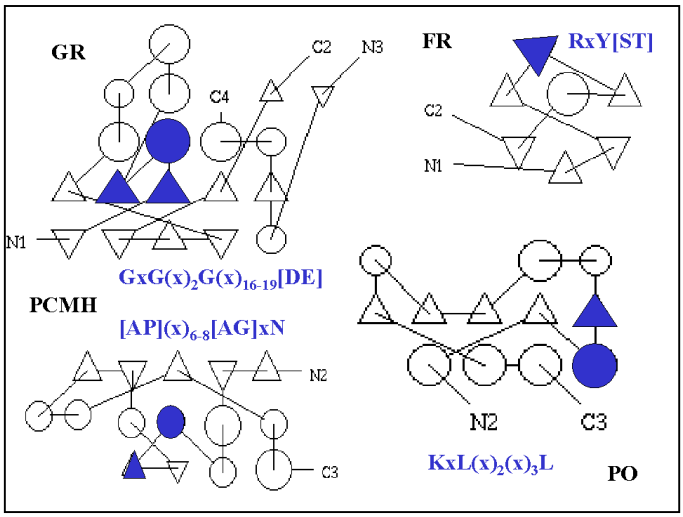

Figure 1: The TOPS (Westhead et. al., 1998) cartoons showing four FAD-binding families. The coloured SSEs are the sequence-structure relationship in the protein.

7. Acknowledgements

We would like to thank Eduardo Alonso, Olivier Sand, Gilleain Torrance, Ali Al-Shahib and Mallika Veeramalai for discussion. AC Tan's scholarships was funded by the Department of Computing, City University.

8. References

Doolittle, R.F. (1986). *Of URFs and ORFs: A primer on how to analyse derived amino acid sequences*. University Science Books, Mill Valley: CA.

Dym, O. and Eisenberg, D. (2001). Sequence-structure analysis of FAD-containing proteins. *Protein Science*, 10: 1712-1728.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Muggleton, S., Srinivasan, A., King, R.D. and Sternberg, M.J.E. (1998). Biochemical knowledge discovery using inductive logic programming. In H. Motoda (Ed.) *Proc. Of the First Conference on Discovery Science*, Berlin, Springer-Verlag.

Muggleton, S. (2001). CProgol4.4: a tutorial introduction. In *Inductive Logic Programming and Knowledge Discovery in Databases*. Springer-Verlag, To appear.

Quinlan, J.R. (1986). Induction on decision trees. *Machine Learning*, 1: 81-106.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo: C.A.

Westhead, D. R., Slidel, T.W.F., Flores, T. P. J. and Thornton, J. M. (1999). Protein structural topology: automated analysis, diagrammatic representation and database searching, *Protein Science*, 8: 897-904.