BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii19–ii25
doi:10.1093/bioinformatics/bti1102

Databases

Predicting post-synaptic activity in proteins with data mining

Gisele L. Pappa[1], Anthony J. Baines[2] and Alex A. Freitas[1,*]

[1]Computing Laboratory, University of Kent, Canterbury CT2 7NF, UK and [2]Department of Biosciences, University of Kent, Canterbury CT2 7NJ, UK

ABSTRACT

Summary: The bioinformatics problem being addressed in this paper is to predict whether or not a protein has post-synaptic activity. This problem is of great intrinsic interest because proteins with post-synaptic activities are connected with functioning of the nervous system. Indeed, many proteins having post-synaptic activity have been functionally characterized by biochemical, immunological and proteomic exercises. They represent a wide variety of proteins with functions in extracellular signal reception and propagation through intracellular apparatuses, cell adhesion molecules and scaffolding proteins that link them in a web. The challenge is to automatically discover features of the primary sequences of proteins that typically occur in proteins with post-synaptic activity but rarely (or never) occur in proteins without post-synaptic activity, and vice-versa. In this context, we used data mining to automatically discover classification rules that predict whether or not a protein has post-synaptic activity. The discovered rules were analysed with respect to their predictive accuracy (generalization ability) and with respect to their interestingness to biologists (in the sense of representing novel, unexpected knowledge).

Contact: A.A.Freitas@kent.ac.uk

1 INTRODUCTION

One of the great challenges of our era is to predict the functions of proteins based on their primary sequence. This is a very difficult problem, since the relationship between protein sequence and function is very complex (Gerlt and Babbitt, 2000; Devos and Valencia, 2000; Nagl, 2003). Indeed, although there is a vast amount of data stored in protein databases, there is still a large gap between the huge amount of data about protein sequences and the knowledge necessary for understanding the process of protein folding and associated protein functions.

Intuitively, however, protein databases contain important 'hidden relationships' (knowledge) between protein sequence and protein function. There is a clear and urgent motivation for discovering this hidden knowledge from protein databases for a number of reasons, such as a better understanding of diseases, designing more effective medical drugs, etc. This creates both a need and an opportunity to apply data mining techniques to the problem of automatically discovering knowledge from protein databases.

Data mining is a multi-disciplinary field, which consists of using methods of several research areas (arguably, mainly machine learning and statistical pattern recognition) to extract interesting knowledge from real-world datasets (Witten and Frank, 2000; Fayyad *et al*., 1996).

*To whom correspondence should be addressed.

This paper proposes a data mining approach to the problem of predicting whether or not a protein has post-synaptic activity, based on features of the protein's primary sequence. The proposed approach will be described later, in Sections 3 and 4. In this Introduction we only emphasize a major difference between the proposed data mining approach and a more traditional bioinformatics approach for predicting protein function, as follows.

In general, the approach most used to predict the function of a new protein—for which we know only its sequence—consists of performing a similarity search in a protein database. In essence, the program finds the most similar protein(s) to the new protein, and if that similarity is higher than a threshold, the function of the most similar protein is transferred to the new protein. Although this approach is very useful in many cases, it also has some limitations, as follows.

First, it is well-known that two proteins might have very similar sequences and perform different functions, or have very different sequences and perform the same or similar function (Syed and Yona, 2003; Gerlt and Babbitt, 2000). Second, the proteins being compared may be similar in regions of the sequence that are not determinants of function (Schug *et al*., 2002). Third, the prediction of function is based only on sequence similarity, ignoring many relevant biochemical properties of proteins (Karwath and King, 2002; Syed and Yona, 2003). Fourth, it does not produce a model for predicting function, and so it does not give insights into the relationship between the sequence, biochemical properties and function of proteins.

Another approach consists of inducing, from protein data, a model describing (in a very summarized form) the data, so that new proteins can be classified by the model. This is the data mining approach followed in this project. We emphasize that this model-induction approach aims at complementing—rather than replacing—the conventional similarity-based approach. In any case, the model-induction approach followed in this project has two important advantages (King *et al*., 2001). First, it can predict the function of a new protein even in the absence of sequence similarity between that protein and other proteins with known function. Second, if the discovered model is expressed in a comprehensible form (which is the case in this research, where knowledge is expressed by intuitively comprehensible IF-THEN rules), it can be used by biologists, to give new insights and possibly suggest new biological experiments. Indeed, in this research the rules discovered by a data mining algorithm were not only automatically evaluated with respect to their predictive accuracy—as is usual in data mining—but also manually interpreted in the context of relevant biochemical knowledge, in order to determine how interesting they were with respect to providing novel insights—unknown in the current literature—about the relationship between some protein sequence patterns and post-synaptic activity. This two-criteria evaluation reinforces the inter-disciplinary

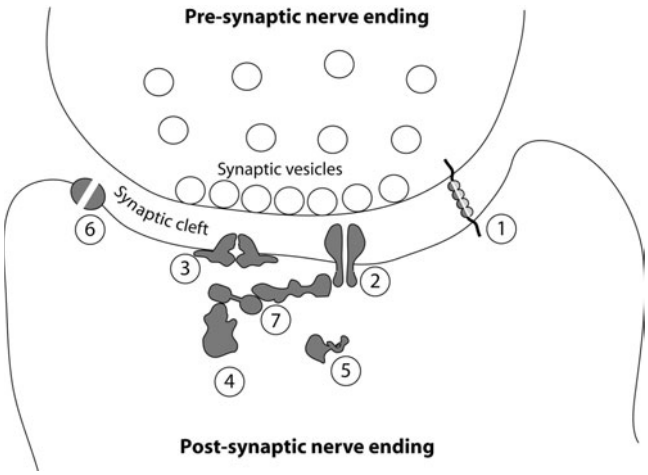

Fig. 1. Main elements involved in pre-synaptic and post-synaptic activity. A synapse is the point where two nerve cells communicate with each other by transmission of a chemical known as a neurotransmitter. The main elements found in synapses are shown in Figure 1. The cells are held together by cell adhesion molecules (1). In the cell where the signal is coming from (the pre-synaptic cell) neurotransmitters are stored in bags called synaptic vesicles. When signals are to be transmitted from the pre-synaptic cell to the post-synaptic cell, synaptic vesicles fuse with the pre-synaptic membrane and release their contents into the synaptic cleft between the cells. The transmitters then diffuse within the cleft, and some of them meet a post-synaptic receptor (2), which recognises them as a signal. This activates the receptor, which then transmits the signal on to other signalling components such as voltage-gated ion channels (3), protein kinases (4) and phosphatases (5). To ensure that the signal is terminated and to clear up residual neurotransmitters after the signal has terminated, transporters (6) remove neurotransmitters from the cleft. Within the post-synaptic cell, the signalling apparatus is organised by various scaffolding proteins (7).

nature of this project, which is, of course, a desirable feature in a bioinformatics project.

The remainder of this paper is organised as follows: Section 2 describes the target biological problem; Section 3 describes the preparation of the dataset for mining purposes, by using protein data available in UniProt/SwissProt and Prosite; Section 4 discusses the proposed data mining approach and the corresponding analysis of results; finally, Section 5 reports the conclusions and future research directions.

2 THE TARGET BIOLOGICAL PROBLEM

In essence, post-synaptic sites represent points where one nerve cell receives signals from another. As indicated in Figure 1, multiple types of proteins are expected to be found at these sites for reception and propagation of signals, and for joining the two nerve cells to each other. Note that Figure 1 is a very minimal summary of the types of proteins found in postsynaptic sites.

The bioinformatics problem being addressed in this paper is to predict whether or not a protein has post-synaptic activity. This problem is of great intrinsic interest because proteins with post-synaptic activities are connected with functioning of the nervous system. Indeed, many proteins having post-synaptic activity have been functionally characterized by biochemical, immunological and proteomic exercises [see e.g. Husi *et al*. (2000) and Walikonis *et al*. (2000)], and are now extensively catalogued and annotated in the Uniprot/SwissProt database. They represent a wide variety of proteins with functions in extracellular signal reception and propagation through intracellular apparatuses, cell adhesion molecules and scaffolding proteins that link them in a web.

The challenge is to automatically discover features of proteins' primary sequences that typically occur in proteins with post-synaptic activity but rarely (or never) occur in proteins without post-synaptic activity, and vice-versa. These discovered features constitute the essence of the knowledge discovered by a data mining algorithm. If the algorithm is successful in discovering knowledge with a high predictive power, that knowledge can be used to accurately discriminate between the two classes of proteins. In addition, and most important, the fact that in this project discovered knowledge will be expressed in a comprehensible form—as mentioned in the Introduction—represents potentially valuable knowledge by itself, because such knowledge can potentially give new insights to biologists about which sequence features are predictive of post-synaptic activity.

3 METHODS

The goal of this project is to predict—with a data mining algorithm—whether or not a protein has post-synaptic activity. The algorithm is used to discover interesting relationships between sequence features that are often present in post-synaptic proteins but usually absent in proteins without post-synaptic proteins, and vice-versa. Therefore, in order to discover this kind of knowledge, we need not only a set of post-synaptic proteins but also a control set of proteins which do not have post-synaptic activity.

In data mining terminology, the set of proteins with post-synaptic activity is called the set of positive examples, whereas the control set of proteins (without post-synaptic activity) is called the set of negative examples. The problem of finding sequence features that discriminate between these two kinds of proteins is then cast as a classification problem, where the goal is to predict the value of a class attribute for each example (protein) based on a set of predictor attributes for that example. The classes are whether or not a protein has post-synaptic activity, and the predictor attributes are mainly Prosite patterns, as described below.

More precisely, the dataset mined in this project was constructed in two phases. In the first phase we carefully selected relevant proteins from the UniProt database. UniProt (Universal Protein Resource) was created by the union of Swiss-Prot, TrEMBL and PIR databases, and is a repository for protein sequence and functional data (Uniprot, 2004, http://www.ebi.uniprot.org/index.shtml). UniProt was chosen as the source of the data to be mined because it is the world standard non-redundant, comprehensive protein sequence database. UniProt is divided in three database layers: UniParc (UniProt Archive), UniProt Knowledgebase (UniProt/SwissProt) and UniProt NREF (UniRef). In this research we have used UniProt/SwissProt, which is the richest annotated layer. One of the great advantages of this layer is the comprehensive annotation of SwissProt. This is curated to ensure minimal redundancy, accurate and comprehensive annotation of function, expression, sequence features (e.g. domain structures), literature references, links to other databases, etc.

The second phase of data preparation used the links from UniProt/SwissProt to the Prosite database, in order to retrieve information from Prosite that was used to create the predictor attributes. These two phases are described in detail next.

3.1 Phase 1: selecting positive and negative examples

Table 1 shows the queries submitted to UniProt/SwissProt in order to select the set of positive examples (proteins with post-synaptic activity) and the set of negative examples (proteins without post-synaptic activity). For each query, the table shows the specification of the query and the number of examples

Table 1. Queries submitted to UniProt/SwissProt to create the dataset

Query No.	Query specification	# Examples
Selection of positive examples		
1	Post-synaptic !toxin	356
	Total number of positive examples	356
Query No.	Query/Keywords	# Examples
Selection of negative examples		
2	Heart !(result of query 1)	3106
3	Cardiac !(results of queries 1, 2)	331
4	Liver !(results of queries 1, 2, 3)	2794
5	Hepatic !(results of queries 1, 2, 3, 4)	256
6	Kidney !(results of query 1, 2, 3, 4, 5)	988
	Total number of negative examples	7475

(proteins) returned by that query. The specification of each query consists of keywords and the logical operator 'NOT' (!).

Note that the queries did not specify any specific species, i.e. proteins from all species were considered. This was done in order to maximize the number of examples in the data to be mined. Positive examples were selected not only by the presence of the keyword 'post-synaptic' but also by the absence of the keyword 'toxin'. The reason for this latter criterion is that several entries in UniProt/SwissProt refer to the toxin α-latrotoxin. This protein acts post-synaptically, but it is not, of course, a post-synaptic protein.

The selection of the negative examples was more difficult, since of course the UniProt/SwissProt entries do not have an explicit 'not post-synaptic' keyword. The trivial 'solution' of retrieving all entries that do not explicitly have the keyword 'post-synaptic' is not satisfactory, for two reasons. First, this would produce a very large number of negative examples, which would be much larger than the number of positive examples. This would create a dataset with an extremely unbalanced class distribution, and it would be very difficult for the classification algorithm to discover rules that correctly classify the positive (post-synaptic) class. Second, and most important, many of the negative examples would be 'trivial', leading to uninteresting, trivial rules for the discrimination between positive and negative examples. For instance, there is no need to discover rules discriminating plant proteins from post-synaptic proteins, since it is obvious that plants do not have a nervous system. Hence, including plant proteins in the set of negative examples would only contribute to the discovery of uninteresting, trivial classification rules.

The goal is to select a set of negative examples where, although the proteins do not have post-synaptic activity, they have some characteristics that could be confused with some of the characteristics of post-synaptic proteins (the positive examples), making it difficult to discriminate between these two kinds of proteins. Intuitively, the higher the similarity between positive and negative examples, the harder the classification problem, and so the stronger the motivation to use a data mining algorithm to discover interesting classification rules representing novel knowledge to a biologist. (Recall that the ultimate goal is to automatically discover interesting, novel rules that provide new insight to biologists about which sequence features are most correlated with the presence or absence of post-synaptic activity.)

Note that, although the positive and negative examples must have enough similarity to lead to the discovery of interesting, non-trivial rules, they also must be different enough to allow the discovery of reliable classification rules for each class. Hence, the challenge is to find a good trade-off between these two goals.

In this spirit, the negative examples were selected by using the keywords 'heart', 'cardiac', 'liver', 'hepatic' and 'kidney'. Consider, for instance, the query 4 in Table 1: 'Liver !(results of queries 1, 2, 3)'. This query means that all UniProt/ SwissProt entries that had the keyword 'liver' and were not already included in the results of the previous queries (1, 2, 3) were selected as negative examples. The latter selection criterion was necessary to avoid the possibility that two or more copies of the same data entry—i.e. a data entry having two or more of the above-mentioned keywords—were duplicated in the set of negative examples.

The previously-mentioned five keywords were chosen as the basis for obtaining negative examples for two reasons. First, it is known that, in general, proteins found in these sites do not present post-synaptic activity. Second, proteins found in these sites often have some characteristics similar to post-synaptic proteins. Indeed, in the context of this project, many of the same types of proteins represented at post-synaptic sites (kinases/phosphatases/channels, etc.) are present in abundance in heart, liver and kidney tissues.

It should be noted that the queries listed in Table 1 searched for the corresponding keywords in all the fields of the UniProt/ SwissProt entries. This means that the selection criteria are not perfect, since those keywords could be present in fields where the presence of the keyword does not mean that the protein has the corresponding function/characteristic. This potentially introduces some 'noise' in the dataset being mined. However, this was necessary, because queries searching for keywords only in the field 'KEYWORD' of UniProt/ SwissProt returned too few proteins. In any case, the amount of noise introduced by the imperfect selection criteria seems to be relatively small, since the data mining algorithm was able to discovery quite accurate classification rules, as will be shown later.

3.2 Phase 2: generating the predictor attributes

Once the set of examples to be mined has been selected from Uniprot/Swissprot, the next step was to generate a set of predictor attributes representing relevant properties of the sequences those proteins. The predictor attributes must have a good predictive power and facilitate easy interpretation by biologists. In this project we have focused mainly on generating attributes based on Prosite patterns associated with the proteins—a type of attribute satisfying both the previously-mentioned properties. The Prosite database stores significant patterns and profiles that help to identify the family of a new protein (Hulo *et al*., 2004). We decided to use attributes based only on Prosite patterns, and not Prosite profiles, for two reasons. First, the matching between a Prosite pattern and a protein can be exactly computed, producing a simple binary attribute—i.e. the pattern either occurs or does not occur in a given protein. This reduces the size of the search space for the data mining algorithm and simplifies the interpretation of the rule by biologists. By contrast, the matching between a Prosite profile and a protein is an approximate matching, and the data mining algorithm would have to search in a correspondingly much larger search space. Second, the use of both patterns and profiles would lead to a very large number of attributes, which would again expand the size of the search space, and so would significantly increase the risk of overfitting the induced model to the data. It should be noted that, even considering only the binary attributes derived from Prosite patterns, this led to 443 attributes (as explained next), which corresponds to a huge search space of size 2^{443}.

For each protein selected in phase 1 (regardless of the protein being a positive or a negative example), we retrieved all Prosite entry id's that occurred in the field database cross-references (DR) of UniProt/SwissProt. For each Prosite entry id, we 'followed the link' from UniProt/SwissProt to Prosite, in order to access information about that Prosite entry. Once this was done for all proteins selected in phase 1, we had a large set of Prosite entries. We then selected, for use as predictor attributes, the entries that:

(i) were marked as a 'pattern' in the ID line of that entry in the Prosite database;

(ii) were not commented (in the CC line of the Prosite database) as very general patterns (/SKIP_FLAG = TRUE)—it was necessary to exclude those patterns because they appear in almost all proteins, and so are not useful to discriminate between the two classes of proteins;

(iii) occurred in at least two proteins of the dataset being mined—this was necessary to remove extremely specific patterns, occurring in

just one protein of the dataset, which do not have any generalisation power.

Finally, each of the selected Prosite patterns was encoded as one binary attribute of the dataset being mined, taking on the value 'yes' or 'no' for each protein—indicating whether or not the pattern occurs in that protein, respectively.

Note that, a few proteins in the dataset to be mined did not have any Prosite pattern, i.e. they had the value 'no' for all predictor attributes based on Prosite patterns. These proteins were removed from the data to be mined.

In addition to the previously-described attributes based on Prosite patterns, we added to the dataset two simple predictor attributes derived directly from the proteins' sequences, namely the sequence length and the molecular weight of the protein (both attributes are available from the corresponding fields in UniProt/SwissProt entries). Other kinds of attributes will be considered in future research, but for now it is interesting to note that even the current set of predictor attributes is enough to discover quite accurate classification rules, as will be shown later.

After all this data preparation process, we ended up with a dataset composed by 4303 examples (260 belonging to the positive class and 4043 belonging to the negative class) and 445 predictor attributes. These 445 attributes include 443 Prosite patterns, the sequence length and the molecular weight of each protein.

4 RESULTS

In order to discover knowledge from the dataset described in the previous section, we have used the well-known C4.5Rules rule induction algorithm (Quinlan, 1993). This algorithm was chosen as the data mining algorithm in our experiments mainly because it produces comprehensible knowledge, represented by a set of high-level, easily-interpretable classification rules of the form: IF (conditions) THEN (class). This kind of rule has the intuitive meaning that, if an example (protein) satisfies the conditions in the rule antecedent, the example is assigned to the class predicted by the rule consequent. It should be noted that the comprehensibility of discovered knowledge is a very important issue in bioinformatics [see e.g. Mirkin and Ritter (2000), Clare and King (2002) and Sebban *et al*. (2002)], because the discovered knowledge should be interpreted and validated by biologists, rather than being blindly trusted as a 'black box'.

We used the default parameters of C4.5Rules. The classification rules discovered by C4.5Rules were evaluated according to two criteria, namely predictive accuracy and interestingness to biologists, as follows. Predictive accuracy was estimated by a well-known 10-fold cross-validation procedure (Witten and Frank, 2000), as usual in data mining. In essence, the dataset was divided into 10 partitions, with approximately the same number of examples (proteins) in each partition. In the i-th iteration, $i = 1, 2, \ldots, 10$, the i-th partition was used as the test set and the other 9 partitions were temporarily merged and used as the training set. In each iteration C4.5Rules discovered a rule set from the training set and used that rule set to classify examples in the test set (unseen during training), in order to evaluate the generalisation ability of discovered knowledge. The classification accuracy rate of the discovered rules can then be computed as the average accuracy rate over the 10 test sets, and this is the measure of predictive accuracy most popular in the literature. In our experiments, this produced an accuracy rate of 97.85%.

It should be noted, however, that in the context of this project this traditional measure of accurate rate is not a very effective one. The reason is that the class distribution is very unbalanced: only 6.4% of the examples have the positive class. Hence, as a baseline solution for this classification problem, the 'majority classifier'—which predicts the majority (negative) class for all examples—would trivially obtain an accuracy rate of 93.9%, without providing any insight about the relationship between the predictor attributes and the classes.

Therefore, we use a more 'demanding' measure of predictive accuracy, for which a high value can be obtained only by accurately classifying examples of both classes. The measure in question is the product: true positive rate (TPR) $\times$ true negative rate (TNR) (Hand, 1997). These terms (which are sometimes referred to as Sensitivity and Specificity, respectively) are defined as follows.

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \quad \text{TNR} = \text{TN}/(\text{TN} + \text{FP}),$$

where

- TP = number of true positives—i.e. the number of examples that were predicted as positive class by the discovered rule set, and indeed have the positive class;
- FN = number of false negatives—i.e. the number of examples that were predicted as negative class, but actually have the positive class;
- TN = number of true negatives—i.e. the number of examples that were predicted as negative class, and indeed have the negative class;
- FP = number of false positives—i.e. the number of examples that were predicted as positive class, but actually have the negative class.

In our experiments the average values (over the 10 iterations of the cross-validation procedure) of the TPR and TNR were 0.85 and 0.98 respectively, resulting in the final measure of predictive accuracy as TPR $\times$ TNR $= 0.84$ (with a standard deviation of 0.09). Note that, the baseline majority classifier obtains TPR $\times$ TNR $= 0 \times 93.9 = 0$, i.e. it is very strongly penalized (as it should be) for never predicting the positive class.

It should also be noted that, although the vast majority of the data mining literature focuses on measuring only the predictive accuracy of the discovered rules, the ultimate goal of data mining is to discover knowledge that is comprehensible and interesting (novel, unexpected) to the user (Fayyad *et al*., 1996; Han and Kamber, 2001). We emphasize that a very accurate rule will not be useful to the user if it represents a previously known pattern. Consider, for instance, the following hypothetical example. In a hospital's medical database a data mining algorithm could discover the rule: IF (patient is pregnant) THEN (patient's gender is female). This rule is extremely accurate, but it is also completely useless, since it represents an obvious pattern. As a real-world example of the difficult of discovering novel, unexpected rules, (Tsumoto, 2000) reports that, in experiments with two medical datasets, $<1\%$ of the discovered rules were found to be interesting or unexpected to medical experts.

Taking into account our ultimate goal of discovering novel, unexpected rules, the rules discovered by C4.5Rules were also manually evaluated with respect to how surprising they are, by comparison with current biochemical knowledge in the area. In other words, the goal of this evaluation is to determine the extent to which the discovered rules represent novel, unexpected knowledge, leading to novel insights about which Prosite patterns are most strongly associated with the presence or absence of post-synaptic activity in proteins.

In order to perform this evaluation, we need to re-visit Figure 1. Most of the types of proteins shown in Figure 1—objects (1)–(7) in

Table 2. Rules discovered by C4.5Rules

Id	Classification rule
32	IF (NEUROTR_ION_CHANNEL = yes) THEN (class = yes)
19	IF (CADHERIN_1 = yes) AND (920 < seq_length <= 1025) THEN (class = yes)
29	IF (GUANYLATE_KINASE_1 = yes) AND (78928 < mol_weigth <= 113386) THEN (class = yes)
34	IF (43_KD_POSTSYNAPTIC = yes) THEN (class = yes)
35	IF (NA_DICARBOXYL_SYMP_1 = yes) THEN (class = yes)
8	IF (CARBOXYLESTERASE_B_2 = yes) AND (seq_length > 828) THEN (class = yes)
33	IF (DYNAMIN = yes) THEN (class = yes)
6	IF (LIPASE_SER = yes) AND (seq_length > 699) THEN (class = yes)
10	IF (G_PROTEIN_RECEP_F1_1 = yes) AND (11287 < mol_weigth <= 14398) THEN (class = yes)
14	IF (C1Q = yes) AND (seq_length <= 194) THEN (class = yes)
23	IF (A4_EXTRA = yes) AND (BPTI_KUNITZ_1 = no) THEN (class = yes)
26	IF (PPTA = yes) AND (G_PROTEIN_RECEP_F2_1 = no) AND (seq_length > 895) THEN (class = yes)
17	IF (SER_THR_PHOSPHATASE = yes) AND (seq_length > 318) THEN (class= yes)
21	IF (G_PROTEIN_RECEP_F3_1 = yes) AND (mol_weight <= 114180) THEN (class = yes)
2	IF (C1Q = no) AND (EGF_1 = no) AND (GUANYLATE_KINASE_1 = no) AND (LIPASE_SER = no) AND (CARBOXYLESTERASE_B_2 = no) AND (SER_THR_PHOSPHATASE = no) AND (NA_DICARBOXYL_SYMP_1= no) AND (43_KD_POSTSYNAPTIC = no) AND (DYNAMIN = no) AND (A4_EXTRA = no) AND (NEUROTR_ION_CHANNEL = no) AND (G_PROTEIN_RECEP_F1_1 = no) AND (seq_length <= 895) THEN (class = no)
7	IF (C1Q = no) AND (GUANYLATE_KINASE_1 = no) AND (SER_THR_PHOSPHATASE = no) AND (NA_DICARBOXYL_SYMP_1= no) AND (KD_POSTSYNAPTIC = no) AND (A4_EXTRA = no) AND (NEUROTR_ION_CHANNEL = no) AND (G_PROTEIN_RECEP_F1_1 = no) AND (seq_length <= 828) THEN (class = no)
12	IF (SER_THR_PHOSPHATASE = no) AND (43_KD_POSTSYNAPTIC = no) AND (NEUROTR_ION_CHANNEL = no) AND (307 < seq_length <= 437) THEN (class = no)
4	IF (C1Q = no) AND (PPTA = no) AND (GUANYLATE_KINASE_1 = no) AND (LIPASE_SER = no) AND (CARBOXYLESTERASE_B_2 = no) AND (SER_THR_PHOSPHATASE = no) AND (NA_DICARBOXYL_SYMP_1= no) AND (43_KD_POSTSYNAPTIC = no) AND (DYNAMIN = no) AND (A4_EXTRA = no) AND (CADHERIN_1 = no) AND (NEUROTR_ION_CHANNEL = no) AND (G_PROTEIN_RECEP_F1_1 = no) AND (G_PROTEIN_RECEP_F3_1 = no) THEN (class = no)
16	IF (NEUROTR_ION_CHANNEL = no) AND (seq_length <= 318) THEN (class = no)
20	IF (seq_length > 1025) THEN (class = no)
	(default rule) IF (protein does not satisfy any of the above rules) THEN (class = no)

that figure—contain signatures within its sequence that can be recognised in specific Prosite patterns. The only exception is protein type (3), voltage-gated ion channels, for which there is no Prosite pattern. For each of the other protein types, relevant Prosite patterns include: (1) CADHERIN_1; (2) NEUROTR_ION_CHANNEL; (4) PROTEIN_KINASE_ST; (5) SER_THR_PHOSPHATASE; (6) NA_DICARBOXYL_SYMP_1, 43_ KD_POSTSYNAPTIC; (7) GUANYLATE_KINASE_1.

Since these are all expected, a particularly surprising rule would be one whose conditions (in the 'IF part' of the rule) referred to other Prosite patterns, which are considered unrelated to the presence or absence of post-synaptic activity. A rule could also be surprising even if it referred only to the above mentioned patterns, as long as the rule referred to an unexpected combination of those patterns.

Table 2 shows the complete set of discovered rules. Note that several discovered rules are 'expected', representing well-known patterns, and therefore not useful for a biologist expert in the field. For instance, Rule 32 is a typical example of an expected rule:
32: IF (NEUROTR_ION_CHANNEL = yes) THEN (class = yes).

Rule 32 reflects the abundance of ligand-gated ion channels (a type of neurotransmitter receptor that includes important glutamate, serotonin and acetylcholine receptors) at post-synaptic sites (protein type 2 in Fig. 1). This rule has an accuracy of 99.2%, which confirms our earlier remark that a rule can be very accurate but useless to the user, when the rule is pretty obvious like this one. (The accuracy of a rule is essentially measured by the conditional probability of the rule consequent given the rule antecedent. In other words, it is computed as the number of examples satisfying both the antecedent and the consequent of the rule divided by the number of examples satisfying the antecedent of the rule.) Other strongly expected rules include: Rule 19 (protein type 1 in Fig. 1); Rules 29 and 34 (protein type 7 in Fig. 1); Rule 35 (protein type 6 in Fig. 1).

Some rules that might be expected were not discovered by C4.5Rules. For example, several other ion channels (such as inwardly rectifying K+ channels) are associated with post-synaptic structures. However, the Prosite database—even though it is one of the most comprehensive databases of its type—does not contain a signature for these channels, so this represents a limitation of the predictor attributes that we have chosen to generate in this project.

Some expected rules have a very limited accuracy, in particular Rule 17 [IF (SER_THR_PHOSPHATASE = yes) AND (seq_length > 318) THEN (class = yes)], with accuracy = 31.4% and Rule 21 [IF (G_PROTEIN_RECEP_ F3_1 = yes) AND (mol_weight ≤ 114 180) THEN (class = yes)], also with accuracy =

31.4%. The low accuracy of these two rules comes from the fact that ser/thr phosphatases and G-protein coupled receptors are expressed in every human cell, and not just post-synaptically.

The unexpected rules are much more complicated, but they are very surprisingly accurate. Therefore, in general they represent interesting knowledge to biologists who are experts in post-synaptic proteins.

For example, Rule 7 states that if 8 specific Prosite signatures are absent, then the protein is not post-synaptic with 99.8% accuracy. Similarly, Rule 2 states the same thing with 12 Prosite signatures. These rules could not have been predicted a priori with just biological knowledge. What Rules 2 and 7 do is to take a number of Prosite signatures that appear in individual 'expected' rules and combine them in a way that says that, when none of those signatures is present, then the proteins are not post-synaptic. This has real utility in classifying novel proteins, excluding them from the post-synaptic class.

In order to better understand those rules, we also retrieved, from the dataset being mined, the proteins which are exceptions to each of those rules—i.e. proteins that satisfy the conditions in the rule antecedent but have a class different from the one predicted by the rule consequent. These exceptions are quite revealing.

An exception to Rule 7 is the SPOCK protein (Uniprot: TIC1_MOUSE). This is an extracellular matrix (ECM) protein that is associated with the post-synaptic area of pyramidal neurons. The signatures in Rule 7 are membrane-associated or cytoplasmic; thus they do not cover this ECM protein. The synaptic cleft is rather poor in ECM proteins, so most other ECM proteins (collagen etc.) are accurately included in this rule.

Rule 2 has some interesting exceptions, and again points to limitations in the method used to generate the predictor attributes. The protein b-Raf is a protein kinase that is ubiquitous in animal tissues, and has a role in mitogenic signalling. It is also found in synaptic structures. In neurons, it is thought to be part of the system that responds to growth factors such as NGF. In this sense it is not classically part of the system that responds to neurotransmitters, but rather has a role in development and maintenance of the nervous system. B-raf from human and mouse (Uniprot entries: BRAF_HUMAN and BRAF_MOUSE) are exceptions to Rule 2. This exception reflects both the ubiquity of b-Raf and the fact that it represents the nature of the signalling pathways it is involved in.

5 CONCLUSIONS

This paper proposed a data mining approach to generate comprehensible rules that predict whether or not a protein has post-synaptic activity, based on Prosite patterns occurring (or not) in the protein, as well as on a couple of simple protein properties computed directly from the protein's primary sequence (namely the sequence length and the molecular weight of the protein).

The discovered rules were evaluated with respect to both their predictive accuracy and their degree of surprisingness (unexpectedness) to the user. The discovered rules were very successful with respect to predictive accuracy. The main contribution of this paper, however, is the analysis of the rules with respect to their surprisingess. Although this is a very important issue in data mining (as discussed earlier), and particularly crucial in the context of scientific discovery, this issue is largely ignored in virtually all the literature about prediction of protein function from sequence.

From a biological perspective, the discovered rules overall reveal interesting features of this approach to mining functional data from Uniprot/SwissProt. A number of expected rules accurately predict some aspects of post-synaptic function. Other rules (unexpectedly) can exclude post-synaptic function with astonishing accuracy. Still other rules indicate the limitation of this approach. The lack of voltage-gated ion channel Prosite patterns (related to type 3 proteins in the notation of Fig. 1) reflects limitations in Prosite: future approaches to this problem will need to consider this. In the future we also plan to generate a more diverse set of predictor attributes, capturing information about other relevant properties of protein sequences.

A direction for future research would be to estimate the 'interestingness' of the discovered rules by using some data-driven rule interestingness measures proposed in the literature. Then we would be able to automatically rank the discovered rules according to those interestingness measures, and present the rules to the user in decreasing order of estimated interestingness. We could also measure the correlation between the value of those data-driven interestingness measures and the subjective, real interest of the rules to a biologist. This would allow us to evaluate how effective those data-driven interestingness measures are in the sense of being good estimators of the real human interest in the rules. It would also be interesting to analyse the rules discovered by other data mining algorithms.

ACKNOWLEDGEMENTS

G.L.P. is financially supported by CAPES (a Brazilian research-support agency), process number 1650-02-5.

Conflict of Interest: none declared.

REFERENCES

Clare,A. and King,R.D. (2002) Machine learning of functional class from phenotype data. *Bioinformatics*, **18**, 160–166.

Devos,D. and Valencia,A. (2000) Practical limits of functional prediction. *Proteins*, **41**, 98–107.

Fayyad,U.M., Piatetsky-Shapiro,G. and Smyth,P. (1996) From data mining to knowledge discovery: an overview. In: Fayyad,U.M., Piatetsky-Shapiro,G. and Smyth,P. (eds), *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT, pp. 1–34.

Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**(5), reviews 0005.1–reviews 0005.10.

Han,J. and Kamber,M. (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hand,D.J. (1997) *Construction and Assessment of Classification Rules*. Wiley.

Hulo,N. *et al*. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.

Husi,H. *et al*. (2000) Proteomic analysis of NMDA receptor–adhesion protein signaling complexes. *Nat. Neurosci.*, **3**, 661–669.

Karwath,A. and King,R.D. (2002) Homology induction: the use of machine learning to improve sequence similarity searches. *BMC Bioinformatics*, **3**, 11.

King,R.D. *et al*. (2001) The utility of different representations of protein sequence for predicting functional class. *Bioinformatics*, **17**, 445–454.

Mirkin,B. and Ritter,O. (2000) A feature-based approach to discrimination and prediction of protein folding groups. In: Suhai,S. (ed.), *Genomics and Proteomics: Functional and Computational Aspects*. Kluwer/Plenum, pp. 157–177.

Nagl,S. (2003) Function prediction from protein sequence. In Orengo,C.A., Jones,D.T. and Thornton,J.M. (eds), *Bioinformatics: Genes, Proteins & Computers*, Bios, pp. 65–79.

Quinlan,J.R. (1993) *C4.5: Machine Learning Programs*. Morgan Kaufmann.

Schug,J. *et al*. (2002) Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res.*, **12**, 648–655.

Sebban,M. *et al*. (2002) A data mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. *Bioinformatics*, **18**, 235–243.

Syed,U. and Yona,G. (2003) Using a mixture of probabilistic decision trees for direct prediction of protein function. In *Proceedings of the 2003 Conference on Research in Computational Molecular Biology (RECOMB-2003)*, Berlin, Germany, pp. 289–300.

Tsumoto,S. (2000) Clinical knowledge discovery in hospital information systems: two case studies. In *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases* (*PKDD-2000*), **LNAI** 1910, 652–656. Springer-Verlag.

Uniprot: The Universal Protein Resource. (2004)

Walikonis,I.R.S. *et al*. (2000) Identification of proteins in the postsynaptic density fraction by mass spectrometry. *J. Neurosci.*, **20**, 4069–4080.

Witten,H. and Frank,E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.