On the Importance of Comprehensible Classification Models for Protein Function Prediction

Alex A. Freitas, Daniela C. Wieser, and Rolf Apweiler

Abstract—The literature on protein function prediction is currently dominated by works aimed at maximizing predictive accuracy, ignoring the important issues of validation and interpretation of discovered knowledge, which can lead to new insights and hypotheses that are biologically meaningful and advance the understanding of protein functions by biologists. The overall goal of this paper is to critically evaluate this approach, offering a refreshing new perspective on this issue, focusing not only on predictive accuracy but also on the comprehensibility of the induced protein function prediction models. More specifically, this paper aims to offer two main contributions to the area of protein function prediction. First, it presents the case for discovering comprehensible protein function prediction models from data, discussing in detail the advantages of such models, namely, increasing the confidence of the biologist in the system's predictions, leading to new insights about the data and the formulation of new biological hypotheses, and detecting errors in the data. Second, it presents a critical review of the pros and cons of several different knowledge representations that can be used in order to support the discovery of comprehensible protein function prediction models.

Index Terms—Biology, classifier design and evaluation, induction, machine learning.

1 Introduction

There is a general trend in the bioinformatics literature—probably influenced by a similar trend in the machine learning literature—of evaluating the quality of a classification model mainly in terms of predictive accuracy. An evidence of this trend is the large number of works performing protein function prediction with data mining methods such as support vector machines or neural networks—see [30], [66], [69], and [70] for a few examples. These methods are usually very effective in terms of predictive accuracy, but they are "black-box" methods that provide little biologically meaningful explanation for their prediction and give little new insight about the data or the application domain to biologists.

The appropriateness of this black-box approach depends on the application and the interest of the user of the machine learning/data mining system. In many bioinformatics applications such as protein function prediction, ideally, the discovered model should be interpreted and validated in the context of current biological knowledge, as will be discussed in detail later.

This paper aims at offering two main contributions to the bioinformatics literature. First, it presents the case for discovering comprehensible classification models—particularlyin the field of protein function prediction—that are not only accurate but also interpretable by the user. More precisely, this paper discusses the advantages of models for protein function prediction that can be understood by users, such as increasing the confidence of the biologist in the system's predictions, leading to new insights about the data and the formulation of new biological hypotheses, and detecting errors in the data.

The second main contribution of this paper is a critical review of the pros and cons of different knowledge representations particularly suitable for supporting the discovery of comprehensible knowledge in the context of protein function prediction.

Although the importance of intelligible protein function prediction models has been pointed out by a few other authors [9], [24], [63] in the context of specific research projects, to the best of our knowledge, this is the first paper to present both a detailed discussion of the case for comprehensible protein function prediction models and a review of the advantages and disadvantages of different knowledge representations that can be used to obtain such models.

This review paper seems timely because, as stated earlier, the majority of the bioinformatics community is currently focusing on maximizing predictive accuracy in their predictions, ignoring important issues about biological interpretation of computational predictions. For instance, in general, reviews of automated methods for protein function prediction—including extensive reviews such as [18] and [54]—do not discuss the need or motivation for producing comprehensible protein function prediction models. Bock and Gough [4] briefly acknowledge the importance of model comprehensibility, but there is no detailed discussion of the motivation for discovering intelligible models or

- *A.A. Freitas is with the Computing Laboratory, University of Kent, CT2 7NF Canterbury, UK. E-mail: A.A.Freitas@kent.ac.uk.*
- *D.C. Wieser and R. Apweiler are with the European Bioinformatics Institute, Hinxton, CB10 1SD Cambridge, UK. E-mail: {dwieser, apweiler}@ebi.ac.uk.*

Manuscript received 21 Sept. 2007; revised 18 Feb. 2008; accepted 4 May 2008; published online 15 May 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2007-09-0119. Digital Object Identifier no. 10.1109/TCBB.2008.47.

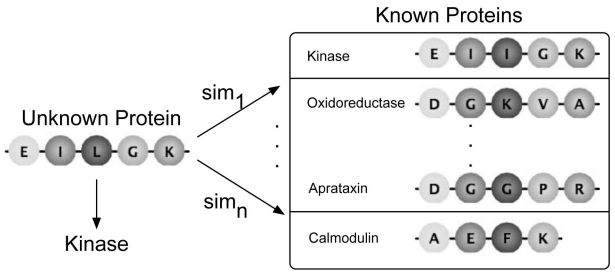

Fig. 1. Basic principle of IBL systems.

any discussion of the pros and cons of different knowledge representations.

The remainder of this paper is organized as follows: Section 2 discusses the differences between the instance-based learning (IBL) and the model induction approaches for protein function prediction. The paper focuses on model induction systems, but the comparison between these two kinds of systems provided in this section is useful to place the latter in the context of the broader literature on protein function prediction. Section 3 addresses in detail the case for comprehensible protein function prediction models, in contrast with the black-box approach that currently dominates the literature in this area. Section 4 discusses some kinds of knowledge representation that are more suitable for supporting the discovery of comprehensible predictive models, highlighting the pros and cons of different representations in this context. Finally, Section 5 presents the conclusions and suggests future research directions.

2 Instance-Based Learning versus Model Induction Approaches

2.1 The Instance-Based Learning Approach

A common method for predicting protein functions is based on transferring the functions of one or more homologous proteins identified on the basis of their close similarity to the target. This kind of method is hereafter referred to as the *IBL (or lazy learning) approach*, and its basic principle is demonstrated in Fig. 1. The system computes a similarity score $\mathrm{sim_i}$ between the unknown protein and each of the proteins with known function, where $i = 1, \ldots, n$, and n is the number of characterized proteins in the database. If the system finds a protein with a high similarity to the unknown protein, then the function of the former is transferred to the latter.

The Smith-Waterman algorithm and its heuristic counterpart BLAST are prominent examples of IBL systems. The E-value reported by BLAST takes into account matches, mismatches, insertions, and deletions in two protein sequences and indicates the extent to which two proteins are similar. A low E-value typically indicates high similarity and often prompts biologists to transfer functional annotation from a well-characterized protein to an unknown-function target protein.

In terms of machine learning paradigms, this kind of protein function prediction method can be considered as belonging to the IBL or Nearest Neighbor paradigm [26], sometimes called "lazy learning" [1]. The motivation for the latter name is that the actual learning is postponed to the moment when a new protein is to be classified. Methods such as the Smith-Waterman algorithm and BLAST share two core characteristics of IBL methods: 1) the training phase essentially consists of storing known-function proteins, and 2) the learning occurs in the testing phase, where an algorithm is used to identify the training sequence most similar to the target instance. This kind of method implicitly uses the training data itself as a "model" in a loose sense of the term, because it does not create any abstract model generalizing from the specific instances in the data.

One advantage of an IBL method is that the simplicity of its training phase makes it naturally incremental. That is, as more and more proteins with known function are added to the database, the training set is immediately expanded, which should in principle increase the predictive accuracy of new functional predictions.

Although IBL methods are useful and powerful in many cases, they also have limitations. First, it is well known that two proteins might have similar sequences but perform different functions or have different sequences and perform the same or a similar function [18], [20], [61]. For instance, in a case study about the classification of voltage-gated potassium channels into four different classes, Szafron et al. [62] pointed out that many of the sequences had close homologues in other classes, rather than their own classes. As another example, although proteins belonging to the G-protein-couple receptor (GPCR) superfamily have high structural homology, many members of that superfamily have a remarkably low degree of sequence similarity [14]. Some GPCRs—e.g., the histamine receptors—may bind the same ligand and the same G protein while having less than 25 percent sequence identity, whereas other GPCRs have a much higher degree of sequence similarity but unrelated functional classes. The case of GPCRs is particularly important since approximately 50 percent of the marketed medical drugs target GPCRs [34], [16]. One could argue, however, that this kind of limitation in the effectiveness of IBL methods has to do not with the methods per se but rather with the presence of "exceptions" in the data that do not match well with the core characteristics of the methods.

Second, IBL methods do not directly take into account the protein function when computing the similarity between protein sequences. Hence, they can consider two protein sequences as highly similar even when the regions of high similarity are not determinants of protein function [56], [18], in which case the high similarity should not be used to transfer a function from the most similar sequence to the target sequence. A recent study of GO term annotation errors [32] illustrates this point. Jones and colleagues found that the curated annotations with GO evidence code "Inferred by Sequence Similarity (ISS)" had an estimated annotation error of 49 percent, much larger than the annotation error rate for other curated annotations—13 percent to 18 percent. The authors recommend that curators should use the ISS annotation only after carefully examining the annotated similar sequence from which functions are being transferred to the target sequence. This involves checking, for instance, if the similar sequence contains protein domains different

from the ones found in the target sequence, in which case the transfer of GO terms from the former to the latter might not be appropriate [18], [32], [54]. It is important to note, though, that this limitation of current IBL methods is due mainly to the use of unsuitable distance measures, based upon whole-sequence comparisons. As understanding of critical residues improves, new distance measures (e.g., combinations of three-dimensional structures and properties of selected residues) may improve the performance of IBL methods.

Third, there are many cases where no protein function can be assigned to a given protein based on the use of an IBL system, due to the lack of a sufficiently similar and known sequence in the database. If no homologues match the E-value cutoff, no prediction will be made for that target sequence. For instance, at the time the *Arabidopsis thaliana* genome was sequenced, about 30 percent of its genes could not be assigned a function using BLASTP, due to the lack of similar sequences of known function [3]. This example is particularly relevant because *A. thaliana* is an important "model organism" for identifying genes and determining their functions.

Finally, IBL systems miss the opportunity to discover explicit relationships between biochemical properties and protein functions, which could significantly advance the understanding of protein functions by biologists. For example, when the prediction of function is based only on sequence similarity, many other potentially relevant biochemical properties of proteins are ignored [33], [61].

The above issues are a motivation to investigate another approach to predict protein functions, based on the induction of a classification model from data, as discussed in the next section.

2.2 The Model Induction Approach

The IBL systems discussed in the previous section are in contrast with other machine learning paradigms where an algorithm first learns an explicit and abstract classification model from the training data and then uses that model to classify a new test instance. Such explicit predictive classification models usually represent a generalization of the data, and systems following this approach will be hereafter called *model induction systems*. Note that such systems follow an "eager learning" approach, by contrast with the "lazy learning" approach of IBL systems. The term eager learning is a broad term used to refer to any learning technique that learns a classification model from the training set of known-function proteins before any new protein to be classified (in the test set) is observed.

The basic principle of model induction systems is illustrated in Fig. 2. The unknown-function protein has its function predicted by the model—which is supposed to capture the main patterns in the data relevant for function prediction—and so, there is no need to compare the target protein sequence with all sequences in the database like in the case of IBL systems illustrated in Fig. 1.

Concerning the interpretability of the induced classification models, broadly speaking, there are two kinds of models, namely, black-box models and white-box models. Black-box models are typically hard to interpret, whereas white-box models are usually interpretable. Examples of black-box models are artificial neural networks, support

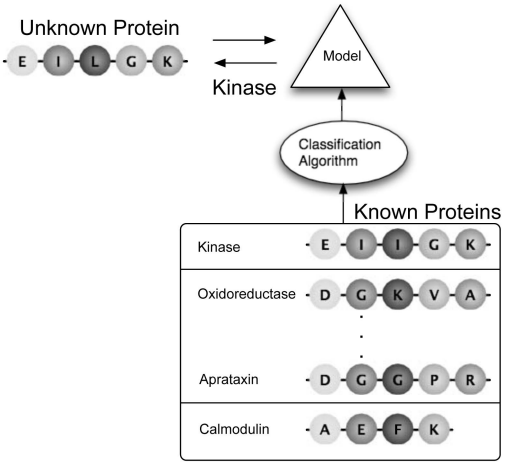

Fig. 2. Basic principle of model induction systems.

vector machines, and hidden Markov models. For instance, an artificial neural network's structure and the corresponding interconnection weights do not provide any insight into the biological nature of the genomic process [44].

Examples of white-box models are decision trees and rule sets—as long as the size of the decision tree or rule set is not too large, which would prevent their interpretation by a biologist. An example of a protein function prediction model expressed by a comprehensible decision tree can be found in [59]. Several examples of function prediction models expressed by a set of comprehensible rules can be found in [11], [24], and [47].

Given the need for interpreting the induced model in many applications, the data mining community developed several methods for transforming a black-box model into a white-box one—in particular methods for extracting comprehensible rules from artificial neural networks [28], [65] or from support vector machines [19], [46]. Only recently, such methods have started to be used by the bioinformatics community [24].

A more detailed discussion of the motivation for inducing white-box models will be presented in Section 3, while a number of examples of comprehensible function prediction models will be discussed in Section 4.2.

The model induction approach can be considered as more general than the IBL approach, because the former can create a model to predict protein functions even in the absence of similarity between the target sequence and sequences in the database [4], [15], [35], [68].

2.3 Hybrid Instance-Based Learning/Model Induction Systems

IBL and model induction methodologies for predicting protein functions are not necessarily mutually exclusive. They have different pros and cons and can be combined into a hybrid system, aiming to get the best of both worlds.

There are at least two different approaches for achieving such a hybrid system. First, we can use the results of a model induction algorithm to improve the results of an IBL method. An example of this approach is found in [33], where first, an IBL method is used to try to retrieve sequences that are homologous to the target sequence. The

problem is that some sequences are considered to be in the "twilight zone," i.e., their homology is uncertain, by the IBL method. Then, a rule induction algorithm is used to induce rules that classify the twilight-zone sequences into homologous or not, reducing the uncertainty of the results produced by the IBL method.

The second approach to hybridize IBL and model induction methods consists of using the result of IBL to improve model induction. An example of this approach is found in [63], where the sequences identified as hits (considered homologous to the target sequence) by an IBL system are used to extract a relevant set of predictor attributes for protein function prediction. As another example of this approach, in [10], the results of PSI-BLAST searches are used to produce homology-based predictor attributes for a model induction system.

3 The Case for Comprehensible Protein Function Prediction Models

This section discusses the motivation for producing a comprehensible classification model mainly in the context of protein function prediction and also, to some extent, in biomedical informatics.

3.1 Improving the Biologist's Confidence in the Prediction

First of all, understanding the predictions made by a classification model helps the biologist to get more confidence in the prediction [63], [15]. This is important because the model's predictions are, by definition, computational predictions, rather than evidence from biological experiments.

If the biologist has a high confidence in the model's predictions, she/he is more likely to believe in those predictions and, in principle, more willing to invest the time and money that are required to perform the very time-consuming and expensive biological experiments necessary for a definite confirmation of the model's prediction. Indeed, it can be argued that the ultimate value of any computational prediction method is determined by the cumulative success of the experiments inspired by the method's results [31].

The importance of a high confidence in a computational prediction is well illustrated by a case outside the area of bioinformatics but valid to the point being made here. When there was a major accident in the Three-Mile Island nuclear power plant, the automated system recommended a shutdown, but the human operator did not implement the shutdown because she/he did not believe in the system's recommendation [25].

3.2 Giving the Biologist New Insights about the Data and Ideas for New Hypothesis Creation

Another reason for discovering comprehensible protein function prediction models is that the model can be used not only for predicting functions of individual proteins but also for giving the biologist new insight about the data and associated biological problem, advancing the understanding of protein functions [24], [44], [63]. In particular, a comprehensible model, duly interpreted by the biologist, can provide new evidence confirming or rejecting some previous hypothesis or even lead the biologist to formulate new biological hypotheses.

Several examples of biological hypotheses whose formulation was guided by the interpretation of a comprehensible protein function prediction model can be found in [33]. Such new hypotheses can then be validated by new biological experiments, in order to try to confirm those hypotheses. There are several cases where the comprehensible models produced by a data mining method have been confirmed by further biological experiments [36].

The above point also influences how we can evaluate the protein function prediction model, depending on whether the model is a black-box or a white-box model. When evaluating a black-box predictive model the only criterion is predictive accuracy—or a combination of accuracy and coverage. However, when evaluating a white-box predictive model, we can also evaluate the comprehensibility and "interestingness" of the model to its users.

For instance, Wong and Leung [72] found classification rules with a relatively low or moderate degree of accuracy (around 40 percent-60 percent) that were considered, by senior medical doctors, novel and more accurate than the knowledge of some junior doctors. In addition, as pointed out by Clare and King [9], when measuring predictive accuracy, a lack of statistical significance does not necessarily mean that the rule is not biologically interesting. It is possible that a rule is correct and interesting, but the number of examples used by the classification algorithm was not large enough to produce rules whose accuracy is deemed statistically significant. A biologically interesting rule can still advance our understanding of protein functions, despite the lack of a large statistical confidence in the rule.

Finally, it should be noted that in some cases, human experts find it difficulty to express their knowledge in a formal way, so that the automated induction of a comprehensible model seems a good approach for getting that kind of knowledge. An example of such case is the use of automated induction to discover classification rules predicting the subclass of a tuberculosis agent belonging to the *M. tuberculosis* complex [57].

Several examples of new insights about the data provided by the induction of comprehensible classification models will be mentioned in Section 4.2.

3.3 Detecting Errors in the Model or in the Data

Yet another reason for discovering comprehensible protein function prediction models is to interpret the model in order to potentially detect errors in it—possibly caused by errors in the data. Sources of error in the predictions of a protein functional classification model [63] include limited training data quality and quantity and the use of an algorithm unsuitable for the underlying data. These sources are discussed in more detail below.

First, the quality of the classification model is proportional to the *quality* of the training data. This quality is limited at least by the predictive power of the available attributes—the question of which kind of attribute is most relevant for predicting protein functions is an open question [45]—and by the amount of noise in the data. A simple example of noise in the data, introduced by the use of a

nonstrict methodology for data set creation, can be found in [73]. In this work, the data set included proteins with any GO annotation, regardless of the evidence code, which indicates whether the protein function was inferred by not very reliable means (e.g., electronic annotation) or more reliable means (e.g., manual curation).

In this particular case, the noise associated with not very reliable GO annotations could be avoided by a stricter data set creation methodology, using only proteins whose GO evidence code means that its annotated function is very reliable. In other cases, however, noise is more subtle, and it is associated with a fundamental problem in the nature of the data—at least given current high-throughput technologies.

As an example of noise in the data in the latter case, Zhu et al. [74] points out that high-throughput protein-protein interaction data usually has huge errors, because the data does not contain any information about the condition(s) under which interactions take place, and so, the neighbors of a given protein may be involved in several different pathways, lacking functional consistency. As yet another example of intrinsic noise in the data, in a recent work, Jones et al. [32] found that the *curated* annotations in the GoSeqLite database had an estimated annotation error rate between 28 percent and 30 percent.

In order to select the most relevant attributes in the data, discarding irrelevant or noisy attributes, feature selection techniques can be used. The use of this kind of technique in protein function prediction can be found, for instance, in [30], [2], [13], and [63].

Another potential source of error in the data is that function annotations in biological databases are often incomplete, and so, the lack of a functional annotation (a class) for a protein does not necessarily mean that the protein does not have that class. It may be the case that the protein was simply not annotated yet with that class. An example of such problem is provided in [31], where the system predicted, for a given protein in the test set, the GO term "chromatin binding." From the point of view of the system, this was a wrong prediction—a "false positive"—because this term is not a descendant—in the GO direct acyclic graph—of any GO term annotated for that protein. However, the protein in question catalyzes acetylation of chromatin substrates, so the seemingly "wrong" prediction, from the system's point of view, is biologically relevant anyway. This is just one specific example of the general problem that given the incomplete functional annotations in biological databases, the lack of a given functional annotation for a protein may be due to the lack of experiments confirming that function, rather than to the true absence of that function in the protein [7].

Hence, many apparent "false positive" functional predictions from the system's point of view may not be prediction errors at all. This problem reinforces the need for evaluating a protein function prediction model by not just its predictive accuracy, but also its biological relevance, which can be achieved when a comprehensible classification model is produced.

Second, the predictive accuracy of a model is limited by the *quantity* of training data. In particular, as mentioned earlier, many protein functional classification problems involve a large number of classes, with a correspondingly small number of examples (proteins) per class, which seriously hinders the reliable prediction of the rare classes. In the specific context of hierarchical protein functional classification, classes at the deepest level of the hierarchy (with fewer examples) are often predicted with an accuracy significantly lower than the predictive accuracy of classes at shallower levels of the hierarchy, where there are many more examples per class [23], [35], [58].

A related problem involves the issue of imbalanced class distributions, where some class(es) are much more frequent than others, making it particularly difficult to predict the minority class(es). A practical approach to try to solve this problem involves undersampling the majority class, in order to avoid a bias toward predicting the majority class in general. This approach has produced good results in some cases—e.g., [2] and [12].

Third, some prediction errors of the model will be caused by the use of a classification algorithm that is not the ideal algorithm for the underlying data. Actually, it is well known in the machine learning and data mining literature that no classification algorithm is superior to all others in all application domains—this point has been shown both theoretically [55], [52] and empirically [43], [42]. This is because every classification algorithm has an inductive bias, and the adequacy of a bias depends on the nature of the data being mined. Despite some significant progress in the area of "metalearning" [5], [43], [49], [50]—where the goal is essentially to try to automatically select the best classification algorithm for the data being mined—the selection of the best classification algorithm for a given data set is still an open problem.

Given all the previously discussed problems, it is important to analyze an induced protein function prediction model in order to detect errors in the model or in the data. One example of a system that detects errors in automated protein functions annotations by analyzing a comprehensible classification model is the Xanthippe system [71]. Xanthippe is a collection of exclusion rules used to postprocess the output of other prediction systems or to detect erroneous annotation in protein annotation databases. For example, if a system predicts the "SUBCELLULAR LOCATION: Mitochondrion" Swiss-Prot comment for a bacterial protein, Xanthippe prevents the attachment of this prediction to the protein entry. Xanthippe collects exclusion rules by analyzing mutually exclusive annotation items in Swiss-Prot entries and by collecting input from curators. It also extracts less obvious exclusion rules using a decision tree algorithm. The attributes for the latter are primarily sequence patterns, such as PFAM or PROSITE. The focus of the decision tree algorithm lies on predicting the absence of a protein annotation rather than the presence.

Another approach to detect errors in protein function predictions consists of analyzing the examples (proteins) wrongly predicted by a classification rule. Since a rule represents an abstract generalized representation of a correlation between the conditions in the rule antecedent and the functional class predicted by the rule consequent,

examining the examples that represent exceptions to such a generalized pattern can give new insights to a biologist. By exceptions, we mean proteins that satisfy all the conditions of the rule but do not have the functional class predicted by the rule.

For instance, Pappa et al. [47] examined the exceptions of rules that predict postsynaptic activity based on conditions referring to the presence or absence of PROSITE patterns and found some of these exceptions quite revealing about the relationship between some PROSITE patterns and postsynaptic activity.

As another example, in [11], an analysis of exceptions of classification rules predicting protein synthesis has revealed that the used functional classification scheme did not capture well the common nature of some proteins—an important conclusion that could not have been drawn by simply measuring the predictive accuracy of a black-box classification model.

4 Discovering Comprehensible Functional Classification Models

The previous section discussed why it is desirable to induce comprehensible ("white-box") protein function prediction models from data—at least in cases where the discovered knowledge will be interpreted and validated by biologists. The main goal of this section is to present a critical review of the pros and cons of different knowledge representations that have been used to produce such comprehensible models. This review is presented specifically in Section 4.1. This is then followed by several examples of discovered knowledge represented by comprehensible protein function prediction models in Section 4.2.

4.1 The Pros and Cons of Different Knowledge Representations for Comprehensible Models

Although there is no consensus in the data mining literature about which knowledge representation is "the most" comprehensible one [48], there is a reasonable consensus that representations such as decision trees and rule sets are more comprehensible than black-box representations such as neural networks and support vector models.

Decision trees have the advantage of being a graphical representation of discovered knowledge, and the hierarchical structure of the tree provides information about the relative importance of the attributes used for prediction: the closer an attribute is to the root of the tree, the more relevant it was deemed to be by the decision tree building algorithm. By contrast, rule sets are not a graphical nor a hierarchical representation, they rather consist of a set of modular IF-THEN rules.

As shown in Fig. 3, a decision tree can be easily converted into a set of IF-THEN rules by creating one rule for each path in the tree from the root to a leaf node, and doing this conversion often helps to simplify the discovered knowledge. The reason for this simplification is that the rule set representation can be considered somewhat more flexible than the decision tree one as follows.

Each rule can be easily interpreted in a modular "local" fashion, independent of other rules, without the need to maintain a "global" decision tree structure. Note, for

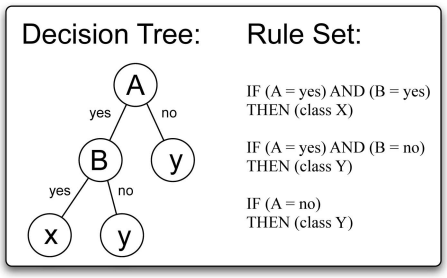

Fig. 3. Example of a decision tree with three leaf nodes converted into a set of three classification rules.

instance, that the attribute at the root node of the decision tree has to be used to predict the classes of all examples, because that node is in every path leading to a leaf node—where a class prediction is made. This seems counterintuitive in some cases, because it is possible that different proteins should have their functional classes predicted by different sets of attributes, and the use of the same attribute in all predictions can be unnatural—and possibly be overfitting the data. Rule sets are not obliged to have the rigid structure of using a given attribute for all predictions, because different rules can naturally refer to different sets of attributes. Of course, it is possible that all rules in a rule set refer to the same attribute, but this will happen only if the algorithm producing the rules decides that this is the best way of maximizing predictive accuracy, rather than being a structural constraint imposed in the model regardless of its predictive accuracy as it is the case with decision trees.

For this and related reasons, a rule set is often considered to be simpler to interpret than its equivalent decision tree counterpart [24], [51], and in practice, several protein function prediction works use the approach of converting a decision tree into a set of rules [38], [11].

It is also appropriate to discuss here the recent trend—both in data mining and bioinformatics—of using an ensemble of classifiers to improve predictive accuracy [6], [30], [61], [67], [66]. It is well known that in general, an ensemble of classifiers improves predictive accuracy by comparison with the use of a single classifier [64]. On the other hand, the use of ensembles also tends to reduce the comprehensibility of the predictive model, in comparison with the use of a single classifier, as follows.

A single comprehensible predictive model can be interpreted by a biologist, but it is not practical to ask a biologist to interpret an ensemble consisting of a large number of comprehensible models. In addition to the obvious problem that such an interpretation would be time consuming and tedious to the biologist, there is also a more fundamental conceptual problem. This is the fact that the classification models being combined in an ensemble are often to some extent inconsistent with each other—this inconsistency is necessary to achieve diversity in the ensemble, which in turn is necessary to increase the predictive accuracy of the ensemble. Considering that each model can be regarded as a "hypothesis" to explain predictive patterns hidden in the data, this means that an ensemble does not represent a single coherent hypothesis

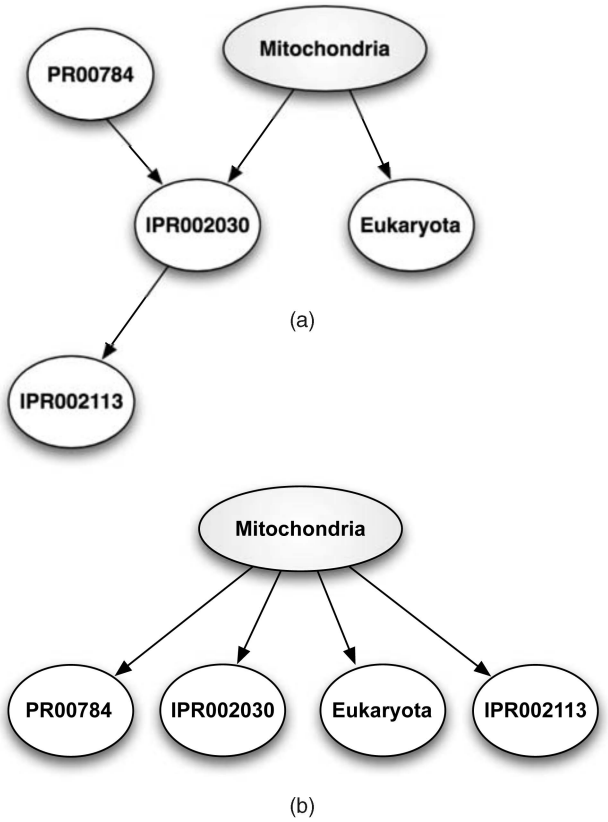

Fig. 4. Examples of a Bayesian network and a naive Bayes classifier's network structure. (a) Bayesian network example. (b) Naive Bayes example.

about the data but rather a large set of mutually inconsistent hypotheses, which in general would be too confusing for a biologist.

In the specific context of protein function prediction, an example of the use of an ensemble of decision trees can be found in [61]. Although this type of ensemble clearly reduces the advantages of interpretability associated with decision trees, for the reasons explained above, it might be argued that an ensemble of decision trees still offers some opportunities for a limited form of interpretation (being at least more interpretable than a black-box model). For instance, one can examine the attributes at the top node(s) in the trees, in order to try to detect attributes that have been consistently chosen to label the top node(s) across the majority of the trees in the ensemble. Such attributes would then be considered the most relevant ones for function prediction.

In addition to decision trees and rule sets, another type of white-box model that is often used in bioinformatics are Bayesian networks—including the naïve Bayes classifier, which can be considered as the simplest kind of Bayesian network for the classification task of data mining. Fig. 4a shows an example of a Bayesian network, where PR00784 is a PRINTS signature, and IPR002030 and IPR002113 are InterPro signatures. The edges in the network represent dependences between the attributes (nodes in the graph). As shown in Fig. 4b, a naïve Bayes classifier is represented by a simple network structure where every attribute directly depends on the class attribute (Mitochondria in this example network) only, so that the attributes are assumed to be independent from each other given the class.

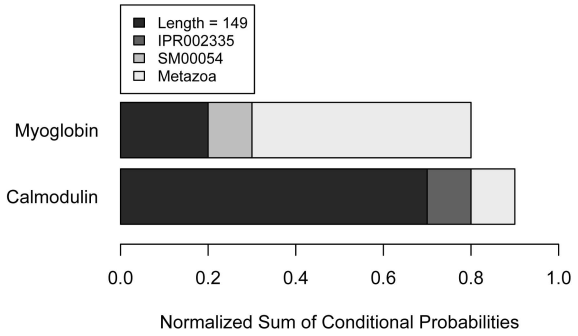

Fig. 5. Example of a bar-graph-based visualization of a naive Bayes classifier.

Despite the potential of Bayesian networks for representing comprehensible knowledge due to the network's graphical structure [37], in practice, this potential seems largely unexplored in the bioinformatics literature. In general, works using Bayesian networks report only a measure of predictive accuracy of the network, without showing the actual network (not even a small part of it). This is regrettable, and we encourage authors working with Bayesian networks to report (at least the main part(s) of) the constructed networks. Even the vast majority of works using the naïve Bayes classifier focus only on reporting its predictive accuracy, rather than trying to interpret the relevance of the computed conditional probabilities.

It is not difficult to interpret the probabilities of a naïve Bayes classifier, assuming that the user is familiar with conditional probability concepts (which is the case for most scientists), and this interpretation can be facilitated by the use of visualization techniques. For instance, Szafron et al. [62], [63] report classification models using a graphical representation, based on bar graphs. In essence, each functional class is associated with a stacked bar, consisting of several subbars. Each subbar is associated with a predictor attribute, and the length of a subbar within the bar of a given class is proportional to the probability of observing that attribute value in the proteins of that class. An example of such a bar-graph-based visualization of a naïve Bayes classifier is illustrated in Fig. 5.

In any case, it should be noted that black-box and white-box models are not mutually exclusive approaches. Both types of models can be used in a hybrid system, trying to combine the potentially somewhat greater predictive power of black-box models with the advantage of comprehensibility of white-box models. An example of such a hybrid approach is found in [24], where in essence, a support vector machine is used as a preprocessing method for a decision tree induction algorithm.

It is interesting to note that concerning the explanation of the predictions, IBL protein function prediction methods are at a kind of intermediate position between the two extremes of a completely black-box or completely white-box model. On one hand, they do provide an "explanation"—in the sense of corroborating evidence—for predicted functions [26]. The explanation in question consists of showing

to the biologist the sequences and functional annotations of the nearest hit(s) to the target sequence. On the other hand, this explanation is specific to each target sequence, and it is an explanation at a low level of abstraction—i.e., at the level of individual protein sequences.

Note that a white-box classification model can explain its predictions at a higher level of abstraction, consisting of rules or another type of comprehensible model that represents a generic relationship in the data, referring to many proteins at a time—e.g., referring to all the proteins that satisfy the antecedent of the rule in the case of rule set representations.

4.2 Examples of Discovered Knowledge Represented by Comprehensible Protein Function Prediction Models

To illustrate the potential of classification rules, several example works are worth mentioning here. We emphasize that the goal of this section is not to discuss in detail the biological meaning and relevance of the discovered rules mentioned here, since most of the works mentioned below already include such a detailed discussion. Rather, the goal of this section is mainly to show the diversity of the kinds of predictor attributes that can be used to discover comprehensible protein function prediction models to illustrate that several different types of biological insights can be obtained by inducing comprehensible models from data. Example works include the following:

- Clare and King [8] mined data about mutant phenotype growth experiments with *S. cerivisiae*. The predictor attributes represented growth media for mutant phenotypes—attribute values denoted the observed sensitivity or resistance of the mutant compared with the wildtype. The predicted classes were defined by the MIPS protein functional classification scheme. The rule induction algorithm discovered many comprehensible rules that had just one or two conditions and had a good predictive accuracy. These rules were simple to interpret and clearly identified the most relevant attributes for protein function prediction out of all the attributes, giving experimenters knowledge about which growth media are more informative for identifying different functional classes of disruption mutants.
- Clare and King [10] did further experiments with a rule induction algorithm applied to *S. cerivisiae* data, using several different kinds of predictor attributes and again predicting MIPS functional classes. They reported the discovery of several comprehensible rules, in particular 1) a rule having conditions referring to the predicted secondary structure (lengths and relative positions of alpha, beta, and coil parts of the structure) and 2) a rule having conditions referring to the degree of homology to other kinds of proteins. These rules were shown to be consistent with biological knowledge.
- Kretschmann et al. [38] described how they use sequence patterns and organism information to build decision trees and rules. Due to many user requests, the rules predicting annotations for UniProtKB/TrEMBL are displayed on the UniProt web page (www.uniprot.org) so that users can easily trace back the origin of the annotation.

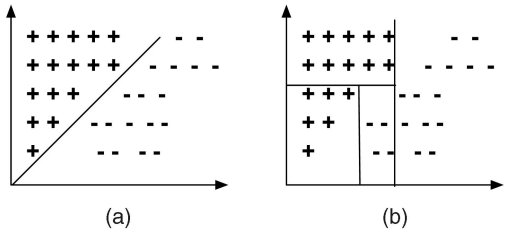

Fig. 6. Example of the flexibility of oblique splits. (a) Oblique split. (b) Axis-parallel splits.

It should be noted that the previously quoted example works in general report "high-level" rules, whose conditions involved predictor attributes at a high level of abstraction, gave new insight about the data, and led to the explicit formulation of new biological hypotheses. The next example works differ with respect to the level of abstraction of the discovered rules:

- He et al. [24] obtained relatively "low-level" rules predicting transmembrane segments, where the rule conditions refer to the presence of specific amino acids in specific positions of the sequence.
- This kind of low-level rule—with conditions referring to specific amino acids in specific positions—was also discovered by Huang et al. [27], with the difference that in this case, the rules predict protein stability changes upon mutations.

Although the rules discovered in the previous two works do not give as much insight about the data as rules having higher level attributes, they were still considered interpretable by the authors, and they were found to be useful to guide "wetlab" experiments because the explicit sequence features that caused the prediction to be made are identified, and therefore, a specific mutation can be made to validate the prediction. Therefore, biologists can narrow the experiment scope by focusing only on certain changes in amino acid sequence.

The discussion up to now assumed the use of a conventional decision tree representation (from which rules were derived), using axis-parallel splits in the attribute space. Hayete and Bienkowska [23] compared axis-parallel and oblique splits in the prediction of GO terms. Oblique splits are more flexible than axis-parallel splits. This flexibility is illustrated in the abstract two-dimensional data space shown in Fig. 6, where only one oblique split is needed to separate positive and negative classes in Fig. 6a, while three axis-parallel splits are needed in Fig. 6b. However, oblique splits are considerably more difficult to be interpreted by biologists (because they are typically produced by mathematical equations involving several variables), as well as requiring much more processing time than axis-parallel splits. Interestingly, in the previously mentioned work, oblique splits did not improve predictive performance with respect to simpler axis-parallel splits, so axis-parallel splits were preferred due to their better comprehensibility and shorter processing time.

5 Conclusions

This paper presented a critical evaluation of the conventional approach for measuring the performance of a protein function prediction system, which is typically based on a single quality criterion—namely, predictive accuracy. This approach has very often led to the induction of "black-box" predictive models that in general, although very accurate, cannot be interpreted by biologists and so do not offer any new insights to the latter.

As a refreshing alternative to this conventional approach, the paper reviewed the importance of discovering comprehensible ("white-box") predictive models, which can be interpreted by biologists and be used as a source of ideas for the creation of new hypotheses and insights about the data and the target biological research problem. The paper also discussed different knowledge representations that lend themselves more naturally to the expression of comprehensible models than the representations typically used in black-box systems, therefore facilitating the goal of discovering models that are easily interpretable by biologists.

Hence, it is hoped that this paper will motivate other researchers and practitioners to pay more attention to the important issues of discovering and interpreting—in the context of current biological knowledge—comprehensible protein function prediction models.

We emphasize that the induction of comprehensible models with data mining techniques is by no means the only useful approach for protein function prediction. Such prediction is of course a very challenging problem in general, and so, there are plenty of opportunities for using alternative techniques. In particular, the induction of comprehensible white-box models should be used as a *complementary* approach to (rather than replacing) the more conventional approaches of IBL systems and the induction of black-box models aimed at maximizing predictive accuracy only.

Concerning future research directions, although there are several case studies reported in the literature where many discovered rules were analyzed with respect to their biological meaning and relevance, there are still many problems to be solved. One major problem is that the number of discovered rules is often very large, seriously hindering rule interpretation by the biologist user. As examples of this problem, He et al. [24], Kretschmann et al. [38], and Laegreid et al. [39] report the discovery of about 20,000, 11,306, and 18,064 rules, respectively.

To cope with the interpretation of such a large number of rules, one possible direction could be to apply a clustering algorithm to the discovered rule set, in order to group similar rules into the same cluster and then summarize the contents of each cluster with a "typical prototype" rule or a higher level rule that summarizes the original rules for that cluster.

An alternative approach for future research could be to borrow, from the data mining literature, methods that try to select the most "interesting" (novel and unexpected) rules out of a large set of discovered rules [17], [42], [53], [60]. Such methods seem still unexplored in the context of bioinformatics.

Acknowledgments

This work was partly funded by the UniProt NIH Grant 2 U01HG02712-04.

References

[1] D.W. Aha, ed., *Artificial Intelligence Rev.*, special issue on lazy learning, vol. 11, 1997.
[2] A. Al-Shahib, R. Breitling, and D. Gilbert, "Feature Selection and the Class Imbalance Problem in Predicting Protein Function from Sequence," *Applied Bioinformatics*, vol. 4, no. 3, pp. 195-203, 2005.
[3] The Arabidopsis Genome Initiative, "Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis Thaliana*," *Nature*, vol. 408, pp. 796-815, 2000.
[4] J.R. Bock and D.A. Gough, "In Silico Biological Function Attribution: A Different Perspective," *Biosilico*, vol. 2, no. 1, pp. 30-37, Jan. 2004.
[5] P.B. Brazdil, C. Soares, and J.P. Costa, "Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results," *Machine Learning*, vol. 50, no. 3, pp. 251-277, Mar. 2003.
[6] A. Bulashevska and R. Eils, "Predicting Protein Subcellular Locations Using Hierarchical Ensemble of Bayesian Classifiers Based on Markov Chains," *BMC Bioinformatics*, vol. 7, p. 298, 2006.
[7] Y. Chen and D. Xu, "Genome-Scale Protein Function Prediction in Yeast *Saccharomyces cerevisiae* through Integrating Multiple Sources of High-Throughput Data," *Proc. Pacific Symp. Biocomputing (PSB '05)*, vol. 10, pp. 471-482, 2005.
[8] A. Clare and R.D. King, "Knowledge Discovery in Multi-Label Phenotype Data," *Proc. Fifth European Conf. Principles of Data Mining and Knowledge Discovery (PKDD '01)*, pp. 42-53, 2001.
[9] A. Clare and R.D. King, "Machine Learning of Functional Class from Phenotype Data," *Bioinformatics*, vol. 18, no. 1, pp. 160-166, 2002.
[10] A. Clare and R.D. King, "Predicting Gene Function in *Saccharomyces Cerevisiae*," *Bioinformatics*, vol. 19, no. Suppl. 2, pp. ii42-ii49, 2003.
[11] A. Clare, A. Karwath, H. Ougham, and R.D. King, "Functional Bioinformatics for *Arabidopsis thaliana*," *Bioinformatics*, vol. 22, no. 9, pp. 1130-1136, 2006.
[12] R.J. Dobson, P.B. Munroe, M.J. Caufield, and M.A.S. Saqi, "Predicting Deleterious nsSNPs: An Analysis of Sequence and Structural Attributes," *BMC Bioinformatics*, vol. 7, p. 217, 2006.
[13] E.S. Correa, A.A. Freitas, and C.G. Johnson, "A New Discrete Particle Swarm Algorithm Applied to Attribute Selection in a Bioinformatics Data Set," *Proc. Genetic and Evolutionary Computation Conf. (GECCO '06)*, J. Keijzer et al., eds., pp. 35-42, 2006.
[14] M.N. Davies, D.E. Gloriam, A. Secker, A.A. Freitas, M. Mendao, J. Timmis, and D.R. Flower, "Proteomic Applications of Automated GPCR Classification," *Proteomics*, vol. 7, no. 16, pp. 2800-2814, Aug. 2007.
[15] M. Doderer, K. Yoon, J. Salinas, and S. Kwek, "Protein Subcellular Localization Prediction Using a Hybrid of Similarity Search and Error-Correcting Output Code Techniques That Produces Interpretable Results," *In Silico Biology*, vol. 6, 2006.
[16] D. Filmore, "It's a GPCR World," *Modern Drug Discovery*, pp. 24-27, Nov. 2004.
[17] A.A. Freitas, "Are We Really Discovering "Interesting" Knowledge from Data," *Expert Update (the BCS-SGAI Magazine)*, vol. 9, no. 1, pp. 41-47, Autumn 2006.
[18] I. Friedberg, "Automated Protein Function Prediction—The Genomic Challenge," *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 225-242, 2006.
[19] G. Fung, S. Sandilya, and R.B. Rao, "Rule Extraction from Linear Support Vector Machines," *Proc. ACM SIGKDD '05*, pp. 32-40, 2005.
[20] J.A. Gerlt and P.C. Babbitt, "Can Sequence Determine Function," *Genome Biology*, vol. 1, no. 5, 2000.

[21] GO Consortium, "The Gene Ontology (GO) Database and Informatics Resource," *Nucleic Acids Research,* vol. 32, pp. D258-D261, 2004.

[22] GO Consortium, "The Gene Ontology (GO) Project in 2006," *Nucleic Acids Research,* vol. 34, pp. D322-D326, 2006.

[23] B. Hayete and J.R. Bienkowska, "GOTrees: Predicting GO Associations from Protein Domain Composition Using Decision Trees," *Proc. Pacific Symp. Biocomputing (PSB '05),* vol. 10, pp. 127-138, 2005.

[24] J. He, H.-J. Hu, R. Harrison, P.C. Tai, and Y. Pan, "Transmembrane Segments Prediction and Understanding Using Support Vector Machine and Decision Tree," *Expert Systems with Applications,* vol. 30, pp. 64-72, 2006.

[25] R.J. Henery, "Classification," *Machine Learning, Neural and Statistical Classification,* D. Michie, D.J. Spiegelhalter, and C.C. Taylor, eds., pp. 6-16, Ellis Horwood, 1994.

[26] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, and K. Nakai, "WoLF PSORT: Protein Localization Predictor," *Nucleic Acids Research Advance Access,* May 2007.

[27] L.-T. Huang, M.M. Gromiha, and S.-Y. Ho, "iPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations," *Bioinformatics,* vol. 23, no. 10, pp. 1292-1293, 2007.

[28] H. Jacobson, "Rule Extraction from Recurrent Neural Networks: A Taxonomy and Review," *Neural Computation,* vol. 17, pp. 1223-1263, 2005.

[29] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H.H. Staerfeldt, K. Rapacki, C. Workman, C.A.F. Andersen, S. Snudsen, A. Krogh, A. Valencia, and S. Brunak, "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features," *J. Molecular Biology,* vol. 319, pp. 1257-1265, 2002.

[30] L.J. Jensen, R. Gupta, H.-H. Staerfeldt, and S. Brunak, "Prediction of Human Protein Function According to Gene Ontology Categories," *Bioinformatics,* vol. 19, no. 5, pp. 635-642, 2003.

[31] T. Jiang and A.E. Keating, "AVID: An Integrative Framework for Discovering Functional Relationships among Proteins," *BMC Bioinformatics,* vol. 6, no. 136, 2005.

[32] C.E. Jones, A.L. Brown, and U. Baumann, "Estimating the Annotation Error Rate of Curated GO Database Sequence Annotations," *BMC Bioinformatics,* vol. 8, no. 170, 2007.

[33] A. Karwath and R.D. King, "Homology Induction: The Use of Machine Learning to Improve Sequence Similarity Searches," *BMC Bioinformatics,* vol. 3, no. 11, 2002.

[34] T. Kenakin, "New Bull's Eyes for Drugs," *Scientific Am.,* pp. 32-39, Oct. 2005.

[35] R.D. King, A. Karwath, A. Clare, and L. Dehaspe, "The Utility of Different Representations of Protein Sequence for Predicting Functional Class," *Bioinformatics,* vol. 17, no. 5, pp. 445-454, 2001.

[36] R.D. King, P.H. Wise, and A. Clare, "Confirmation of Data Mining Based Predictions of Protein Function," *Bioinformatics,* vol. 20, no. 7, pp. 1110-1118, 2004.

[37] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence.* Chapman & Hall/CRC, 2004.

[38] E. Kretschmann, W. Fleischmann, and R. Apweiler, "Automatic Rule Generation for Protein Annotation with the C4.5 Data Mining Algorithm Applied on SWISS-PROT," *Bioinformatics,* vol. 17, no. 10, pp. 920-926, 2001.

[39] A. Laegreid, T. Hvidsten, H. Midelfart, J. Komorowski, and A.K. Sandvik, "Predicting Gene Ontology Biological Process from Temporal Gene Expression Patterns," *Genome Research,* vol. 13, pp. 965-979, 2003.

[40] T.S. Lim, W.Y. Loh, and Y.S. Shih, "A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-Three Old and New Classification Algorithms," *Machine Learning,* vol. 40, no. 3, pp. 203-228, 2000.

[41] J. McDowall, "InterPro, Exploring a Powerful Protein Diagnostic Tool," *Tutorial at the Fourth European Conf. Computational Biology (ECCB '05),* Sept. 2005.

[42] K. McGarry, "A Survey of Interestingness Measures for Knowledge Discovery," *Knowledge Eng. Rev.,* vol. 20, no. 1, pp. 39-61, 2005.

[43] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, 1994.

[44] B. Mirkin and O. Ritter, "A Feature-Based Approach to Discrimination and Prediction of Protein Folding Groups," *Genomics and Proteomics: Functional and Computational Aspects,* S. Suhai et al., eds., pp. 157-177, Kluwer Academic/Plenum Publishers, 2000.

[45] N. Nariai, E.D. Kolaczyk, and S. Kasif, "Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data," *PLoS One,* vol. 2, no. 3, p. e337, 2007.

[46] H. Nunez, C. Angulo, and A. Catala, "Rule Extraction from Support Vector Machines," *Proc. European Symp. Artificial Neural Networks (ESANN '02),* pp. 107-202, 2002.

[47] G.L. Pappa, A.J. Baines, and A.A. Freitas, "Predicting Post-Synaptic Activity in Proteins with Data Mining," *Bioinformatics,* vol. 21, no. Suppl. 2, pp. ii19-ii25, 2005.

[48] M.J. Pazzani, "Knowledge Discovery from Data," *IEEE Intelligent Systems,* pp. 10-13, Mar./Apr. 2000.

[49] Y. Peng, P.A. Flach, C. Soares, and P. Brazdil, "Improved Dataset Characterisation for Meta-Learning," *Proc. Fifth Int'l Conf. Discovery Science (DS '02),* pp. 141-152, 2002.

[50] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier, "Landmarking Various Learning Algorithms," *Proc. 17th Int'l Conf. Machine Learning (ICML '00),* pp. 743-750, 2000.

[51] J.R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[52] R.B. Rao, D. Gordon, and W. Spears, "For Every Generalization Action, Is There Really an Equal and Opposite Reaction? Analysis of the Conservation Law for Generalization Performance," *Proc. 12th Int'l Conf. Machine Learning (ICML '95),* p. 471, 1995.

[53] W. Romao, A.A. Freitas, and I.M.S. Gimenes, "Discovering Interesting Knowledge from a Science and Technology Database with a Genetic Algorithm," *Applied Soft Computing,* vol. 4, pp. 121-137, 2004.

[54] B. Rost, J. Liu, R. Nair, K.O. Wrzeszczynski, and Y. Ofran, "Automatic Prediction of Protein Function," *CMLS Cellular and Molecular Life Sciences,* vol. 60, pp. 2637-2650, 2003.

[55] C. Schaffer, "A Conservation Law for Generalization Performance," *Proc. 11th Int'l Conf. Machine Learning (ICML '94),* pp. 259-265, 1994.

[56] J. Schug, S. Diskin, J. Mazzarelli, B.P. Brunk, and C.J. Stoeckert Jr., "Predicting Gene Ontology Functions from ProDom and CDD Protein Domains," *Genome Research,* vol. 12, pp. 648-655, 2002.

[57] M. Sebban, I. Mokrousov, N. Rastogi, and C. Sola, "A Data-Mining Approach to Spacer Oligonucleotide Typing of Mycobacterium Tuberculosis," *Bioinformatics,* vol. 18, no. 2, pp. 235-243, 2002.

[58] A. Secker, M.N. Davies, A.A. Freitas, J. Timmis, M. Mendao, and D. Flower, "An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function," *Expert Update (the BCS-SGAI Magazine),* vol. 9, no. 3, pp. 17-22, Autumn 2007.

[59] M. Singh, P.K. Wadhwa, and P.W. Sandhu, "Human Protein Function Prediction Using Decision Tree Induction," *Int'l J. Computer Science and Network Security,* vol. 7, no. 4, pp. 92-98, Apr. 2007.

[60] E. Suzuki, "Discovering Interesting Exception Rules with Rule Pair," *Proc. PKDD Workshop Advances in Inductive Rule Learning,* pp. 163-178, 2004.

[61] U. Syed and G. Yona, "Using a Mixture of Probabilistic Decision Trees for Direct Prediction of Protein Function," *Proc. Seventh Ann. Int'l Conf. Research in Computational Molecular Biology (RECOMB),* 2003.

[62] D. Szafron, P. Lu, R. Greiner, D.S. Wishart, B. Poulin, R. Eisner, Z. Lu, B. Poulin, R. Eisner, J. Anvik, and C. Macdonell, "Proteome Analyst—Transparent High-Throughput Protein Annotation: Function, Localization and Custom Predictors," *Proc. ICML Workshop Bioinformatics,* 2003.

[63] D. Szafron, P. Lu, R. Greiner, D.S. Wishart, B. Poulin, R. Eisner, Z. Lu, J. Anvik, C. Macdonell, A. Fyshe, and D. Meeuwis, "Proteome Analyst: Custom Predictions with Explanations in a Web-Based Tool for High-Throughput Proteome Annotations," *Nucleic Acids Research,* vol. 32, pp. W365-W371, 2004.

[64] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining,* Section 5.6, Addison-Wesley, 2006.

[65] A.B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The Truth Will Come to Light: Directions and Challenges in Extracting Knowledge Embedded within Trained Artificial Neural Networks," *IEEE Trans. Neural Networks,* vol. 9, no. 6, pp. 1057-1068, 1998.

[66] K. Tu, H. Yu, Z. Guo, and X. Li, "Learnability-Based Further Prediction of Gene Functions in Gene Ontology," *Genomics*, vol. 84, pp. 922-928, 2004.
[67] A. Vinayagam, R. Konig, J. Moormann, F. Schubert, R. Eils, K.-H. Glatting, and S. Suhai, "Applying Support Vector Machines for Gene Ontology Based Gene Function Prediction," *BMC Bioinformatics*, vol. 5, no. 116, 2004.
[68] A. Vinayagam, C. Del Val, F. Schubert, R. Eils, K.-H. Glatting, S. Suhai, and R. Konig, "GOPET: A Tool for Automated Predictions of Gene Ontology Terms," *BMC Bioinformatics*, vol. 7, no. 161, 2006.
[69] A. Vinayagam, R. Konig, J. Moormann, F. Schubert, R. Eils, K.-H. Glatting, and S. Suhai, "Applying Support Vector Machines for Gene Ontology Based Gene Function Prediction," *BMC Bioinformatics*, vol. 5, no. 116, 2004.
[70] W.R. Weinert and H.S. Lopes, "Neural Networks for Protein Classification," *Applied Bioinformatics*, vol. 3, no. 1, pp. 41-48, 2004.
[71] D. Wieser, E. Kretschmann, and R. Apweiler, "Filtering Erroneous Protein Annotation," *Bioinformatics*, vol. 20, no. Suppl. 1, pp. i342-i347, 2004.
[72] M.L. Wong and K.S. Leung, *Data Mining Using Grammar-Based Genetic Programming and Applications*. Kluwer Academic Publishers, 2000.
[73] J. Xiong, S. Rayner, K. Luo, Y. Li, and S. Chen, "Genome Wide Prediction of Protein Function via a Generic Knowledge Discovery Approach Based on Evidence Integration," *BMC Bioinformatics*, vol. 7, no. 628, 2006.
[74] M. Zhu, L. Gao, Z. Guo, Y. Li, D. Wang, J. Wang, and C. Wang, "Globally Predicting Protein Functions Based on Co-Expressed Protein-Protein Interaction Networks and Ontology Taxonomy Similarities," *Gene*, vol. 391, pp. 113-119, 2007.

Alex A. Freitas received the BSc degree in computer science from the Faculdade de Tecnologia de São Paulo (FATEC-SP), Brazil, in 1989, the MSc degree in computer science from the Federal University of São Carlos (UFSCar), Brazil, in 1993, and the PhD degree in computer science from the University of Essex, United Kingdom, in 1997. He worked as a visiting lecturer at Centro Federal de Educacao Tecnologica do Parana (CEFET-PR), Brazil, in 1998, and as an associate professor at the Pontifícia Universidade Católica do Paraná (PUC-PR), Brazil, from 1999 to June 2002. In July 2002, he moved to the University of Kent, Canterbury, United Kingdom, where he is now a senior lecturer and the head of research in the Computing Laboratory. He is a member of the editorial board of three international journals, namely, *Intelligent Data Analysis*, the *International Journal of Data Warehousing and Mining*, and the *International Journal of Computational Intelligence and Applications*. He has authored two research-oriented books (both in the area of data mining) and has published more than 10 invited book chapters and more than 100 peer-reviewed papers in journals and conferences. His current research interests include data mining and knowledge discovery, biologically inspired computational intelligence algorithms, and bioinformatics. He is a member of the IEEE, the Association for the Advancement of Artificial Intelligence (AAAI), the British Computer Society's Specialist Group on Artificial Intelligence (BCS-SGAI), and the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD).

Daniela C. Wieser studied biotechnology with a specialization in bioinformatics at the University of Applied Sciences, Freising/Munich. Since the end of 2005, she has been enrolled in a part-time PhD program in the Department of Computer Science, University of Sheffield, where she works on machine learning topics applied to the field of computational biology; the funding is fully awarded by the university. She is also employed as a senior software engineer at the European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom, where she works on large-scale automatic annotation of uncharacterized proteins. She also acts as a tutor at the University of Cambridge to assist students with their bioinformatics projects. She has published papers in peer-reviewed journals and conference proceedings and has coauthored a book chapter on the functional annotation of proteins.

Rolf Apweiler received the MSc and PhD degress in biology from the University of Heidelberg, Germany. He has been working on the Swiss-Prot database at the European Molecular Biology Laboratory (EMBL) since 1987, and in 1994, he became the leader of the Swiss-Prot group at the European Bioinformatics Institute (EBI), Hinxton, Cambridge, United Kingdom. He is an EMBL senior scientist and, in 2001, became head of the Sequence Database group at the EBI. Since early 2007, he has been the joint head of the EBI's PANDA group of around 140 life and computer scientists, which develops and maintains main public high-quality data resources and bioinformatics services like the EMBL nucleotide sequence database, the Ensembl Genome Browser, the UniProt protein sequence database, and the InterPro database of protein families, domains, and functional sites. He is the chair of the Human Proteome Organization (HUPO) Proteomics Standards Initiative (HPSI) and a member of several committees, review panels, and advisory boards, including the Nomenclature Committee of IUBMB (the "Enzyme Commission"), the expert committee on "Bibliometric Mapping of Excellence in the Area of Life Sciences" of the European Commission, the FlyBase advisory board, the Scientific Committee of the German Research Centre for Biotechnology (GBF), the Scientific Advisory Board of the Centre of Human Proteomics, Dublin, Ireland, the Scientific Advisory Board of the Human Proteome Resource (HPR) program, Sweden, the Scientific Advisory Board of the Systematic Protein Annotation and Modeling (SPAM) project, San Diego, and the Bioinformatics review panel of the German Ministry of Research. He is also an editor of the *FEBS Journal* and a member of several other editorial or advisory boards of other scientific journals such as the *European Journal of Biochemistry*, *BioSilico*, *Biochimica et Biophysica Acta*, *Journal of Proteome Research*, and *Expert Review of Proteomics*. He has published more than 170 papers and book chapters and is a frequent invited speaker for lectures and tutorials at universities, companies, and conferences. In 2003, he was elected to "Database Doyen" by the All Stars Faculty of "Genome Technology," and he was the elected president of HUPO until 2008.