

## Supervised learning-aided optimization of expert-driven functional protein sequence annotation

Soinov L<sup>1</sup>, Kanapin A<sup>2</sup>, Kapushesky M<sup>3</sup>

EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

1. Algorithms and methods
2. InterPro and SwissProt data retrieval and encoding
3. Calculations and programming

### Abstract

The aim of this work is to use a supervised learning approach to identify sets of motif-based sequence characteristics, combinations of which can give the most accurate annotation of new proteins. We assess several of InterPro Consortium member databases for their informativeness for the annotation of full-length protein sequences. Thus, our study addresses the problem of integrating biological information from various resources. Decision-rule algorithms are used to cross-map different biological classification systems in order to optimise the process of functional annotation of protein sequences. Various features (e.g., keywords, GO terms, structural complex names) may be assigned to a sequence via its characteristics (e.g., motifs built by various protein sequence analysis methods) with the developed approach. We chose SwissProt keywords as the set of features on which to perform our analysis. From the presented results one can quickly obtain the best combinations of methods appropriate for the description of a given class of proteins.

### Introduction

Availability of a wide variety of effective protein sequence analysis methods calls for an evaluation of their comparative performance and for development of approaches to integrated cross-method consistent annotation. A natural resolution of this problem came with the creation of InterPro [1]. The InterPro database is a single resource collecting sequence pattern data from PROSITE ([2], a repository of regular expressions and profiles), Pfam ([3], based on hidden Markov models), PRINTS ([4], provider of fingerprints) and from several other databases-participants. InterPro is a manually curated database, in which the curation process is supported by various automated procedures.

One of the most straightforward approaches to characterizing a novel sequence is to compare it to the already annotated in InterPro proteins. While this potentially can produce high-quality functional predictions, the motif-focused nature of InterPro complicates the interpretation of such analyses, because in most cases it is impossible to find a single InterPro entry corresponding to the combination of motifs found in a given sequence. On the other hand, there are various systems for direct functional annotation of full-length protein sequences, such as GeneOntology (GO) [5] or SwissProt [6]. The quality of annotation that one can get based on similarity to InterPro entries/motifs can thus be improved by combining these two annotation paradigms. A correspondence between the annotation specific for a group of previously characterised full-length protein sequences and the domain/repeat architecture of a given sequence could help to achieve a more complete functional description of a protein related to the sequence.

The analytical system presented in this work demonstrates an approach to linking keyword-based functional annotation in SwissProt to motif-architecture data from a subset of the oldest InterPro member databases using Machine Learning. In particular, the constructed classification system uses motif information to assign SwissProt keywords to proteins. We chose to use SwissProt keywords as the testing ground for our method. We could have chosen another annotation source, for example, GeneOntology, however, SwissProt keywords present one of the few independent from InterPro manually curated annotation resources. While GO IDs are assigned to only 67% of SwissProt proteins (to 3% of these they are assigned

manually and to 97% via an automated system through InterPro), 83% of SwissProt proteins are covered by SwissProt keywords, which are therefore a better candidate (data from SwissProt release 43.1, UniProt 1.7). The approach taken here can, however, be easily extended to GO in the future.

In this study we use supervised algorithms that are well suited to the situation where source data of diverse types need to be analysed together. A number of studies related to automated annotation of novel proteins have been previously reported [7,8,9,10], however, while our system can predict SwissProt keyword assignments to proteins, our main purpose is to identify those motif-based characteristics whose combinations can be used to achieve the most accurate annotation of new sequences. The problem we aim to address is the inference of such combinations – something that is nearly impossible to do manually. The idea of using features selected by an induction algorithm as inputs to the subsequent classification procedure has proven to be effective in various classification schemes. We show in this work that selected feature combinations, together with a formal index of their informativeness (information content of these combinations useful for the purposes of protein functional annotation), may serve as an aid in expert-driven curation of protein data.

### **Data**

When using supervised classification algorithms it is necessary to present the source data (known as *the training set* in machine learning terms) in the form of examples of correct classifications. These examples consist of sets of characteristics (*features*) that are used to form predictor functions (*classifiers*) that assign a certain property (*class label*) to a previously unseen example correctly.

All the motif-characteristics (called “signatures” in InterPro terminology) coming from the InterPro member databases form the set of features, while the corresponding SwissProt keywords make up the labels for the protein sequences. In order for our algorithms to make use of the diverse data described, the data need to be represented in a suitable, consistent format – we chose to use the bit vector representation, mapping the collection of InterPro signatures to a vector of 0’s and 1’s: each feature is set to 1 if the corresponding InterPro signature (i.e., motif) is found in the protein sequence and to 0 otherwise.

Two different protein classifications, InterPro and SwissProt keywords, were mapped. Only proteins annotated simultaneously via InterPro signatures and SwissProt keywords were used as examples of correct classification. Judging from the distribution of the number of keywords annotating the proteins, those keywords that are very general (the distribution tails off around keywords matching  $\geq 200$  proteins) were not used, as they are irrelevant for specific functional annotation. This threshold is based on the characteristic distribution of protein numbers in InterPro entries and, although its choice is somewhat arbitrary, it takes into account the majority of SwissProt keywords, corresponding to reasonably small sets of proteins (it would not make sense to take into account a keyword that is assigned to nearly all proteins). Please, note that the decision on assigning a keyword to a protein is taken individually for any given keyword. This implies a separate classification problem for every keyword. As a result, 592 SwissProt keywords were selected and, respectively, 592 separate training sets were created (the chosen keywords are available as supplemental data).

### **Methods**

To form the training sets, we used filter methods for feature subset selection and ROC curves analysis for finding optimal misclassification costs in our cost-sensitive classification scheme [11]. Classifiers were created in the form of decision rules and were used further to identify InterPro signatures most useful for the keyword-based functional annotation of protein sequences. Finally, the relative informativeness of InterPro methods to the purposes of the annotation process was estimated.

### **Training set construction**

The classification problem here is the two-class problem: given a SwissProt keyword we must decide whether or not to assign it to a given protein. Therefore, the training set should consist of proteins to which this keyword is assigned (*positive examples*) and those to which it isn't (*negative examples*). Such a training set would, however, be highly imbalanced: the number of positive examples would often be much smaller than the number of negative ones because the number of proteins that are not labelled by any one keyword is much greater than of those that are. Previous works on this subject have largely ignored the imbalance problem, allowing the majority class (negative examples) to dominate over the information present in the minority class, leading to the creation of trivial classifiers (e.g., assign/not assign a keyword regardless of sequence characteristics of a given protein) [8]. The issue of the training set imbalance is particularly important in our scenario, since it is the positive examples that are in the minority and they are the ones of primary interest. Therefore, to account equally for properties of both positive and negative classes, the disproportion of the example types comprising the training set should be reduced.

### **Feature subset selection**

It would be natural to form the training set by selecting proteins matched by the given keyword as positive examples and all others as negative. However, such a selection would saturate the training set with a lot of negative examples that are irrelevant for classification purposes. The learning procedure should take into account the cases possessing similar sets of features, where the non-trivial problem of discrimination between the two classes exists. Indeed, cases where protein sequences contain motifs that occur in positive examples, but in combinations that result in negative decisions, are the ones that are hard to identify. At the same time, proteins associated with signatures coming only from negative examples are classified as negatives automatically. Selecting as negative examples those proteins that match any of the same motifs, as do the positive example proteins reduced the imbalance of the training set significantly, discarding from 70% to 99% of negative instances per set as irrelevant. We also ranked all features according to the amount of mutual information shared with the label within each training set, keeping only 100 top-ranked features with mutual information more than zero as the most informative [12]. These data preparation steps are described schematically in **Diagram 1**.

### **ROC curves for cost-sensitive classification**

Despite the preliminary filtering and feature subset selection steps, the positive/negative imbalance turns out to still be significant for most of the training sets. Therefore, we used a cost-sensitive learning scheme in order to prevent the occurrence of classifiers that are either trivial or highly biased towards the majority class. Positive and negative examples were given different weights, and ROC curves were constructed for each of the training sets by varying the weight ratio. While ROC curves are usually used to compare the performance of two classification methods against each other [11], we use this method here for the selection of optimal parameters for our classification scheme. In the absence of any *a priori* assumptions about misclassification costs we used the most north-western point of each ROC curve to choose the weights' ratio, thus maximizing the sum of TP (true positive) and TN (true negative) rates.

### **Classification procedure**

Rule	If	Then
1	PR00499 is not found PF04382 is not found	and Keyword with ID 1031305 is <b>not assigned</b>
2	Not covered by Rule 1 PS50001 is not found PF04382 is found SM00150 is not found	and and and Keyword with ID 1031305 is <b>assigned</b>
3	Not covered by Rules 1-2 SM00150 is found	and Keyword with ID 1031305 is <b>assigned</b>
4	Not covered by Rules 1-3	Keyword with ID 1031305 is <b>not assigned</b>

**Table 1.** Decision rules concerning keyword with ID 1031305. For instance, the first rule here says that if the PRINTS motif with ID PR00499 is not found in the sequence of a given protein and if the Pfam motif with ID PF04382 is not found either, then this keyword should **not** be assigned to that protein. Further, if this rule does not cover the given sequence, following rules can be tried sequentially. The last rule covers all the remaining cases.

The mapping between InterPro motif architectures and SwissProt keywords was defined as a set of *decision rules* that specify which combinations of motifs describe the same properties of protein sequences as a given keyword. See **Table 1** for an example of one such rule-set, consisting of four rules. To construct a list of accurate decision rules, decision tree building algorithms may be used as the first step, and the trees would then be transformed into lists of rules. This approach is not always the best one to follow, since different branches of a decision tree have different accuracy rates and cover different subsets of the training set, thus requiring additional analysis for each derived rule. Alternatively, the decision tree learning procedure can be repeated iteratively, selecting the best branch (i.e., the best decision rule) at each step and discarding those examples that are covered by this branch. Doing so, a hierarchical system of decision rules can be constructed, forming the desired classifier. The advantage of this method is that, although each rule is a part of a whole classifier, it can be considered and evaluated separately just by presenting those rules that are higher in the hierarchy as additional clauses to the one considered. The method used here was the one implemented in the open source machine-learning package WEKA. We used WEKA's J48.PART as the core algorithm within our cost-sensitive learning scheme [13].

#### **Informativeness index**

Clearly, features that repeatedly appear in classification rules of high accuracy should be more informative and useful for annotation than others. Therefore, these are the features we seek to identify as optimal for accurate full-sequence annotation and for which we assessed their informativeness, as described below. The decision rules' accuracy was measured by applying them directly to the training set.

To measure the informativeness of the found feature combinations we identified for each SwissProt keyword the InterPro signatures that were selected by the classification algorithm for constructing the corresponding decision rules. An index chosen as the measure of relative informativeness was calculated as follows: since the generated decision rules are organised into a hierarchical structure (the more examples a rule covers and the more accurate the rule is, the higher it is in the hierarchy), we assign higher values to signatures of higher rules. Each signature's index is defined as the number of times it participates in all positive (ones saying that the keyword should be assigned) rules for a given keyword. Obviously, when annotating a new protein, only positive rules play a useful role. Thus, in the example in **Table 1**, all signatures have index 2, because they all participate in all positive rules (note that Rule 3 includes Rules 1-2, and, hence, all signatures composing those rules). These index counts were then converted into relative percentage contributions. These percentage contributions should not be confused with the actual information content of the features. Informativeness indices provide a good approximation of information load per feature and, more importantly, the means for ranking features according to the information carried by them. However, these indices have a different meaning than information bits per feature and can be used only as relative estimates. As sets of signatures most relevant for keyword-based functional annotation were identified, we proceeded to estimate relative contributions of InterPro methods to the annotation of protein sequences in InterPro. We calculated the average informativeness indices (per method per entry) associated with a given InterPro method by, firstly, considering those SwissProt keywords that were assigned to proteins

covered by the given entry and, secondly, calculating the proportion of signatures provided by each InterPro member database among the signatures within the rules generated for these keywords (**Diagram 2** presents a graphical depiction of these procedures). These indices are a relative measure of how much useful information in the annotation context is delivered by the different methods. We performed this analysis for entries that contain signatures of more than one InterPro method and whose proteins are characterised by the considered set of 592 keywords: 1035 entries. The results of our analysis are presented in **Table 3**.

## Results

### Decision rules

All the decision rules obtained for the considered set of 592 keywords and their accuracy rates on the training sets are available as supplemental data. 3065 rules were generated with average accuracy of 71% on the training data; for 1782 rules it was  $\geq 90\%$  and 1497 had 100% accuracy. While the primary aim of this work has been to discover optimal (for keyword-based functional annotation of proteins) combinations of features comprising the constructed rules, it is worthwhile to remark here that these rules can themselves be used for annotation with fairly high accuracy. For 24 keywords no positive rules were obtained (all rules were negative or trivial, *i.e.*, “never assign the keyword”) – rules generated for the 568 remaining keywords were used in the subsequent analysis.

### Method informativeness for keyword-based annotation

For each of the 568 keywords, the relative informativeness index of each InterPro method was calculated as an intermediate step of the calculations described in the

Keywords		Relative Informativeness of InterPro Member Databases						
		PROSITE Patterns	PROSITE Profiles	Pfam	PRINTS	ProDom	SMART	TIGRFAMs
ID	Keyword							
100050	Amino-acid biosynthesis	8.93%	7.14%	60.71%	0.00%	0.00%	14.29%	8.93%
1441138	Pentaxin	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
189923	Tumor antigen	33.33%	0.00%	16.67%	16.67%	0.00%	16.67%	16.67%
1140070	Acetylcholine receptor inhibitor	16.67%	0.00%	33.33%	33.33%	16.67%	0.00%	0.00%
17790019	Potassium channel	6.79%	15.43%	19.75%	58.02%	0.00%	0.00%	0.00%

**Table 2** Excerpt from supplemental data, *Method informativeness for keyword-based annotation*.

“Informativeness index” subsection of Methods. The informativeness indices of InterPro methods for the annotation of proteins characterised by a given SwissProt keyword are given in the supplemental data. For example, as **Table 2** shows (an excerpt from the full supplemental data table), proteins characterised by the keyword “Potassium channel” are best annotated by motifs coming from PRINTS, while for those to which “Pentaxin” is assigned, “PROSITE Patterns” turns out to be the most effective source of information.

### General method informativeness

**Table 3** contains the calculated informativeness indices of InterPro methods, per InterPro method per entry. The informativeness index is a relative measure of useful

Member DB	Average informativeness index	Maximum informativeness index
PROSITE patterns	2.52	100.00
PROSITE profile	0.70	50.00
Pfam	4.78	100.00
PRINTS	2.01	100.00
ProDom	1.13	33.33
SMART	1.03	50.00
TIGRFAMs	0.83	100.00

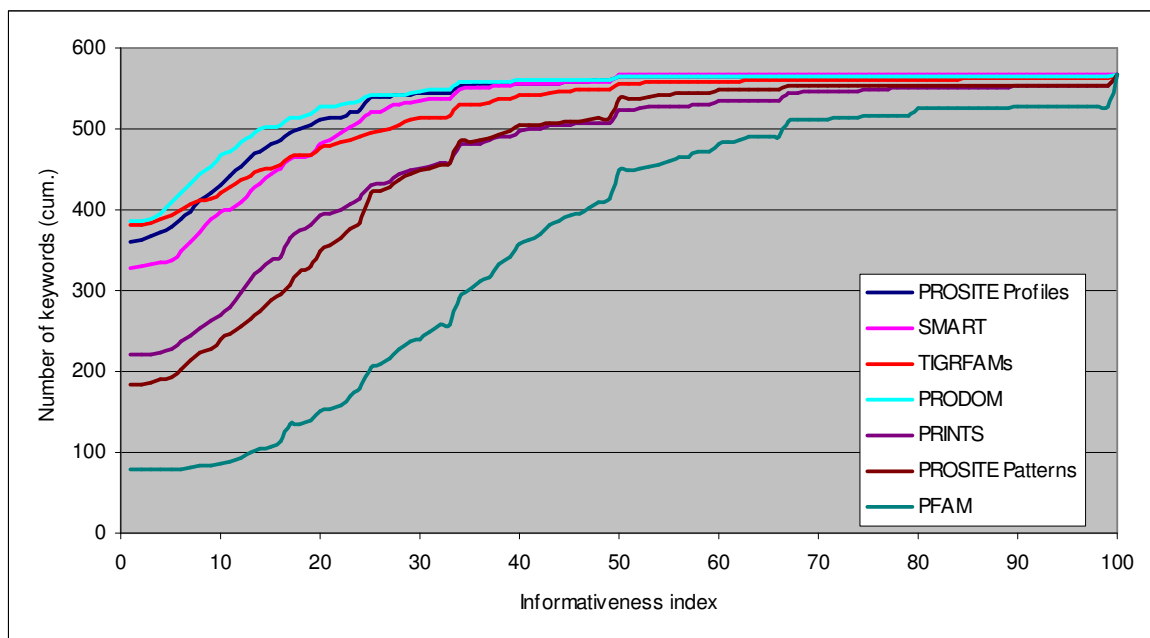
**Table 3** Relative informativeness per method per InterPro entry for different InterPro member databases, average and maximum among the 1035 selected InterPro entries.

information contribution towards keyword-based functional annotation and may be used to assess only the comparative performance of different InterPro methods. The average and maximum informativeness indices are given for each InterPro member database. The meaning of the values in this table is as follows, taking PRINTS as an example: its average informativeness index of 2.01 means that every PRINTS- signature, contributing to an InterPro entry, on average carries 2.01% of information (the meaning of “information” here is explained in the “Informativeness index” subsection of Methods) towards annotation of proteins of this entry. At the same time, for some entries, a PRINTS signature can carry up to 100.00% of information useful for annotation of the respective proteins.

**Figure 1** contains plots of the cumulative distributions of numbers of keywords across the entire range of method informativeness for all the InterPro methods surveyed. In general, the lower the graph for a given method, the higher the number of keywords about which the signatures derived by that method carry a significant share of information. This plot complements the data in **Table 3** and clearly shows that Pfam dominates over the other member databases in terms of average informativeness towards keyword-based functional annotation of proteins.

### Discussion

In the majority of cases accurate functional annotation of uncharacterised protein sequences is possible only by integrating different methods and types of information (cf. **Figure 1**). Often a human expert may find it difficult or impossible to identify the optimal combinations of data sources for the most efficient annotation. Supervised techniques are



**Figure 1** Plots of the cumulative distributions of numbers of keywords across the range of informativeness indices (from 0 to 100%). Pfam is clearly the most informative method on average.

frequently considered to be the most effective approach when diverse types of data are to be analysed together [12]. In this study we considered the methods used by the member databases involved in the InterPro Consortium [1] and employed supervised learning algorithms as a means of assessing the informativeness of different sequence analysis methods that are used for protein classification. We also identified those combinations of methods that are optimal for keyword-based functional annotation of proteins.

The features (InterPro signatures) selected by the used classification algorithm and the combinations in which they appear in the constructed classifiers represent subsets of InterPro

methods that are most efficient in protein annotation. InterPro signatures that are not included in the rules could require additional consideration or could be seen as carriers of information about proteins' functional properties that are not described by the considered keywords.

Our analysis underscores the significance of all the methods involved in InterPro. However, it shows that they should be treated differently when classifying different groups of proteins. The results indicate that several subsets of proteins annotated with SwissProt keywords can be characterised using combinations of only a few of the available methods. These combinations are not trivial and can be easily extracted manually by searching through the lists of signatures. In fact, our approach complements expert annotation and does not contradict it. The process of integration of different sequence analysis methods would benefit from the fusion of expert opinion and automated assessment of the results. As for the properties of the considered methods, some of them are important for the isolation of vast groups of functionally related proteins, while others are critical for providing highly specific classifications. It is also interesting that the best coverage and specificity of classification can be achieved for different groups of proteins by using different methods (see **Table 3** and classification results for the keywords, excerpted in **Table 2**, full data in the supplement). For instance, informativeness indices of signatures comprising the decision rules for the keyword "Pentaxin" indicate the effectiveness of PROSITE patterns [14] in identifying the group of proteins characterised by this keyword. As a family with a distinct function (lipoprotein ligand-binding), according to InterPro it can be described by a number of other methods, such as PRINTS, Pfam, ProDom and SMART. Therefore, our study confirms that PROSITE patterns are particularly strong in this case [14], because the functionality of these proteins is defined by the presence of a binding site.

Our results show that, on average, Pfam, PROSITE patterns and PRINTS are the most useful sources of information towards keyword-based functional annotation. Pfam has a broad area of application, while at the same time SMART, TIGRFAMs and PROSITE profiles are highly informative only within certain groups. Indeed, (cf. **Figure 1**) for the majority of keywords the informativeness of the latter methods is comparatively low, while they, in combinations with other methods, can characterise some other keywords fully. It is interesting to note that we did not find, among the selected 1035 entries, any, whose keywords can be fully characterised by SMART, ProDom or PROSITE profile signatures (**Table 3**). Also we did not find any SwissProt keywords assigned to proteins that were fully described (via generated decision rules) solely by these methods. We would suggest, then, that methods with small average informativeness indices should not be used for protein functional annotation on their own, however, in combinations with the other methods could augment the annotation with specific details.

### **Conclusions**

Supervised classification methods have proven to be effective for the prediction of properties of uncharacterized objects on the whole and also for feature subset selection problems specifically. Classification results may be approached as providers of optimal feature sets for subsequent analysis. Our study demonstrates the effectiveness of this approach and provides computationally supported guidance in selecting best methods for expert driven annotation of proteins and protein families.

### **Acknowledgements**

We are thankful to Dr. Maria Krestyaninova of InterPro and Dr. Alvis Brazma for useful discussions and comments. Lev Soinov and Misha Kapushesky would like to acknowledge that they are funded by the Wellcome Trust support of the BioMap project.

### **Supplemental data**

Please see <http://www.ebi.ac.uk/~ostolop/skk04> for additional data for this work.

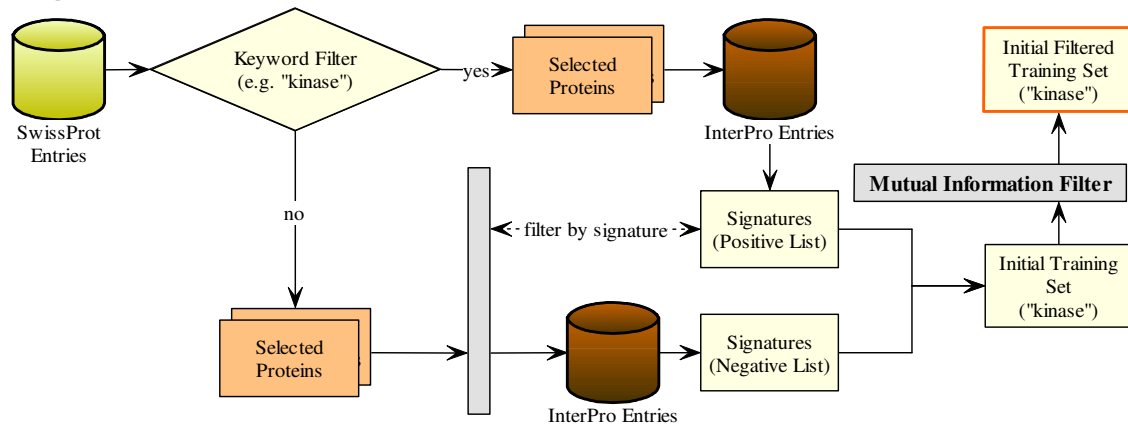
## References

1. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. *Nucl. Acids. Res.* (2003). Jan 1;31(1):315-8.
2. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database, its status in 2002. *Nucleic Acids Res.* (2002) Jan 1;30(1):235-8.
3. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res.* (2000) Jan 1;28(1):263-6.
4. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* (2000) Jan 1;28(1):225-7.
5. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* (2004) Jan 1;32, Database issue:D258-61.
6. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* (2003) Jan 1;31(1):365-70.
7. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* (2003) Mar 22;19(5):635-42.
8. Bazzan AL, Engel PM, Schroeder LF, Da Silva SC. Automated annotation of keywords for proteins related to mycoplasmataceae using machine learning techniques. *Bioinformatics* (2002) Oct;18 Suppl 2:S35-43.
9. Kretschmann E, Fleischmann W, Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* (2001) Oct;17(10):920-6.
10. Pavlidis P, Weston J, Cai J, Noble WS. Learning gene functional classifications from multiple data types. *J Comput Biol.* (2002) 9(2):401-11.
11. Provost F, Fawcett T, Kohavi R. Building the Case Against Accuracy Estimation for Comparing Induction Algorithms. ICML-98.
12. Witten I, Frank E. Data Mining-Practical Machine Learning Tools and Techniques with JAVA Implementations, Morgan Kaufmann, 1999.
13. WEKA. (<http://www.cs.waikato.ac.nz/~ml/weka>).
14. Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A. Recent improvements to the PROSITE database. *Nucleic Acids Res.* (2004) 32, 134-7.
15. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M,

Bradley P, Bork P, Bucher P, Copley R, Courcelle E, Durbin R, Falquet L, Fleischmann W, Gouzy J, Griffith-Jones S, Haft D, Hermjakob H, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Orchard S, Pagni M, Peyruc D, Ponting CP, Servant F, Sigrist CJ; InterPro Consortium. InterPro: An integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.* (2002) Sep;3(3):225-35.

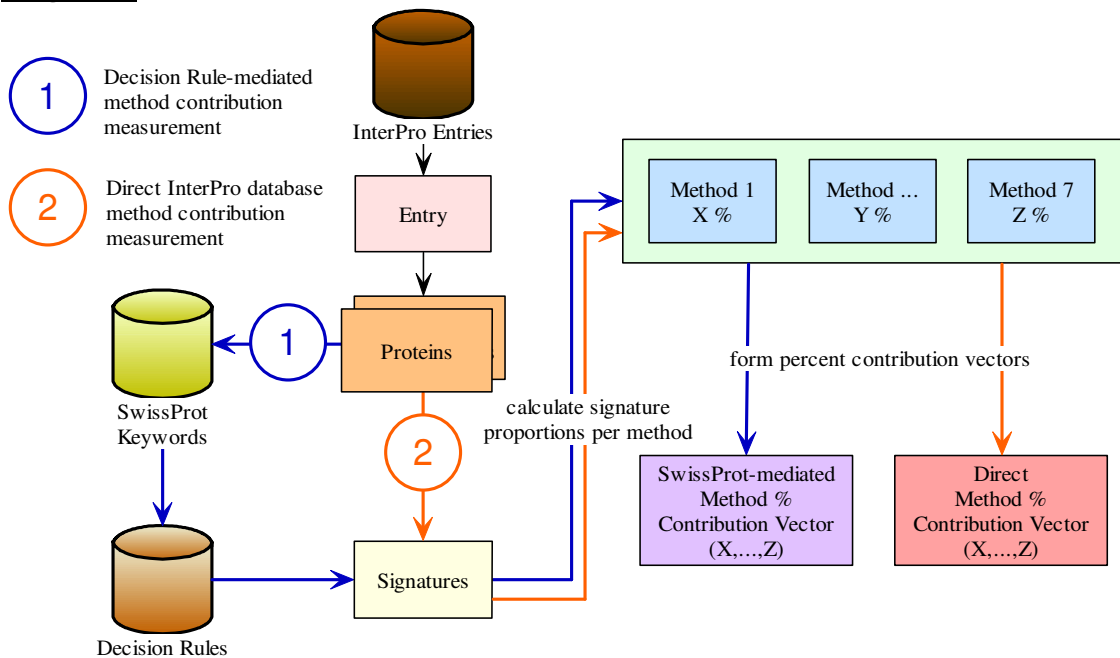
## DIAGRAMS

**Diagram 1**



**Caption:** Data preparation steps for an example keyword (kinase), demonstrating various filters and selection procedures aimed at creating the initial training set to be then subjected to ROC curves analysis for obtaining optimal weight ratios of positive and negative example classes.

**Diagram 2**



**Caption:** Calculation of the informativeness index vectors of the InterPro methods, (1) via the supervised learning classification scheme and (2) via the direct count of method contribution per InterPro entry.