Selecting different protein representations and classification algorithms in hierarchical protein function prediction

Carlos N. Silla Jr.* Alex A. Freitas

School of Computing and Centre for Biomedical Informatics
University of Kent, Canterbury, Kent, CT2 7NF, UK
{cns2,A.A.Freitas}@kent.ac.uk

Abstract

Automatically inferring the function of unknown proteins is a challenging task in proteomics. There are two major problems in the task of computational protein function prediction, which are the choice of the protein representation and the choice of the classification algorithm. There are several ways of extracting features from a protein, and the choice of the feature representation might be as important as the choice of the classification algorithm. These problems are aggravated in the case of hierarchical protein function prediction, where a hierarchy of classifiers is built and each of those classifiers' construction has to consider the aforementioned selection problems. In this paper we address these problem by employing three alternative selective hierarchical classification approaches: (a) selecting the best classifier given a fixed representation; (b) selecting the best representation given a fixed classifier; and (c) selecting the best classifier and representation simultaneously, in a synergistic fashion. The analysis of the results have shown that the selective representation approach is almost always ranked number 1 when compared against the different fixed representations and that the use of the selective classifier approach is not able to surpass using only the best classifier for the target problem.

Keywords: Hierarchical Classification, Protein Function Prediction.

1 Introduction

The task of computational prediction of protein function based on the protein's amino acid sequence is an active area of research in the field of proteomics [18, 47]. One approach that can be used to infer proteins functions is supervised

[*] Corresponding author.Tel: +44 (0)1227 823192, Fax: +44 (0)1227 762811, Email: cns2@kent.ac.uk

machine learning – more precisely, the classification task of machine learning or data mining. The goal is to use a set of proteins whose functions are known to build a classification model that can be used to predict the functions of proteins whose functions are unknown. The use of supervised machine learning (classification) algorithms is common practice in the field [43, 20, 1, 37].

There are two major problems in the task of computational protein function prediction with classification algorithms, which are the choice of the protein representation and the choice of the classification algorithm. Those are open problems, even in the conventional scenario of "flat" classification (where there are no hierarchical relationships among classes), as there are many choices and it is not clear which representation and classification algorithm are the best. In the hierarchical classification scenario addressed in this paper, where protein functional classes are organised into a hierarchy, these problems are aggravated, due to the large number of classes and classification sub-problems (where different algorithms and different representations might be best for different class levels).

There are several ways of extracting features from a protein, and the choice of the feature representation might be as important as the choice of the classification algorithm. Apart from a few works, such as [28], the issue of which feature representation to use is often overlooked as the authors are usually more focused on which classification algorithm to use or related issues. One particular challenge is that not all feature sets are available for every experiment, as some biological databases are highly specialized in one particular organism and the same information might not be available for other organisms.

According to [13] there are two broad types of representations that can be derived for proteins: alignment-independent, which are features computed from the sequence by using some computational method without performing sequence alignment, and alignment-dependent, which are features obtained from biological databases of motifs or domains that were typically discovered by performing sequence alignment on a large-scale, in order to identify conserved regions in the sequences of homologous proteins.

In this paper we address the problems of both protein representation selection and algorithm selection in a synergistic way by using selective hierarchical classification approaches. More precisely, we dynamically select the best protein representation and the best classification algorithm for each class in the hierarchy. Both selections are done in a data-driven way.

The remainder of this paper is organized as follows: Section 2 presents a brief introduction to the task of protein function prediction and the 8 different protein representations employed in this work. Section 3 briefly introduces the task of hierarchical classification and the 3 hierarchical selective approaches: (a) selecting the best classifier for each class node given a fixed representation for all class nodes; (b) selecting the best representation for each class node given a fixed classifier for all class nodes; and (c) selecting the best classifier and representation for each class node. Section 4 presents the experimental setup for the experiments. The computational results and their discussion are presented in Section 5 and finally, in Section 6 we state our conclusions and future research directions.

2 Protein Function Prediction

2.1 Overview of the protein function prediction problem

Proteins are large molecules that execute the vast majority of the functions performed by a living cell [2]. Proteins are produced from genes, by transforming the genes' DNA material into amino acids, the building blocks of proteins. Hence, a protein essentially consists of a long sequence of amino acids, which folds into a specific 3D structure, where different protein structures are suitable for performing different functions. In the last few years there has been an enormous progress in genome sequencing technology (giving us the knowledge of the full DNA contents of many organisms, including humans), and as a result the number of proteins with known sequence of amino acids has been growing very fast. Unfortunately, however, the number of proteins with known function is growing at a much lower rate, because it is much more time-consuming and expensive to determine the structure and function of a protein than just determining its sequence of amino acids. Knowledge of protein functions is very important in biomedical sciences, not only for a better understanding of cell biology in general, but also because many diseases are caused by or at least associated with defects in protein functions. Hence, an effective method for the prediction of protein functions can potentially contribute to generate new biological knowledge that can lead to a better treatment and diagnosis of diseases, design of more effective medical drugs, etc.

Therefore, there is a clear motivation to develop data mining methods (specifically classification methods, based on supervised machine learning) that can predict the function of a protein based on its sequence of amino acids. Although such computational predictions are not so reliable as the results of biological experiments that directly determine a protein's function, computational methods are much cheaper and faster, and so they can give researchers valuable clues for the design of future biological experiments.

In this paper we predict protein function from the protein's amino acid sequence by using classification methods. Hence, each example (data instance) corresponds to a protein, the values of the predictor attributes for an example represent properties of the corresponding protein and the classes represent function(s) associated with the protein. The problem is challenging for at least two reasons. First, there are many different types of protein properties that could be used as predictor attributes, and it is not clear which type of property(ies) has(ve) greater predictive power. (To cope with this, our system automatically selects the best type of protein representation in a data-driven manner, as will be explained later.) Secondly, there are a large number of protein functions, which are typically organized into a hierarchy of functions, naturally leading to a hierarchical classification problem, where the classes to be predicted are hierarchically-structured in the form of a tree of class nodes, in the case of our datasets. Hierarchical classification is, by comparison, a much less investigated research area than standard (flat) classification, and the former tends to be a considerably more difficult type of problem due mainly to the large number of classes to be predicted. The definition of predictor attributes for protein function prediction is discussed in the next sub-section (2.2), whilst methods of hierarchical classification in the context of protein function prediction are discussed in Section 3.

2.2 Protein Feature Types

In this section we describe the protein representations used and evaluated in this work. The protein representations in Sub-sections 2.2.1, 2.2.2, 2.2.3, 2.2.4 are alignment-independent representations. The protein representations in Sub-section 2.2.5 are alignment-dependent.

2.2.1 Sequence Length and Molecular Weight

The sequence length is a numerical value which is simply the count of amino acids of a protein. The molecular weight is the sum of the molecular weights of all amino acids in the protein.

These features have been used (with other attributes) in [28, 1, 23]. Since these features are believed to be important for protein functional prediction and they are easily available, we always use them in conjunction with the other protein representation studied in this work.

2.2.2 Z-Values

The z-values [37, 12], also known as Sandberg Descriptors [36, 32], are the principal components of 26 different physicochemical measured and calculated properties of amino acids, and essentially represent hydrophobicity/hydrophilicity (z1), steric/ bulk properties and polarizability (z2), polarity (z3), and electronic effects (z4 and z5) of the amino acids [32].

In [37] 5 z-values are used to represent each amino acid of the protein sequence. For example, the Alanine (A) amino acid has 5z values: z1 = 0.24, z2 = -2.32, z3 = 0.60, z4 = -0.14, z5 = 1.30. Therefore a protein sequence of length n would be represented by n*5 features. In [12, 37] the authors suggested that the z-values for all amino acids of each protein are averaged so that a protein is represented by just 5 z-values, instead of 5*n. This is needed because most machine learning methods cannot cope with instances (in this case proteins) which have varying number of features (in this case the z-values). It should be noted that they tried more complicated ways of aggregating z-values, but they had better results with this simpler method.

Originally in [37] the authors used the averaged z-values from the whole amino acid sequence. After some experimental research they found out that in order to classify GPCR (G-Protein Coupled Receptor) proteins, it would be better to use 15 z-values [12]. These z-values are then computed as follows:

5-values are computed and averaged over the whole protein sequence. Another 5 z-values are computed from the N-terminus (the first 150 amino acids of the protein sequence) of the protein and further 5 z-values are computed from the C-terminus (the last 150 amino acids of the protein sequence). The number of 150 amino acids was found, in previous experiments, to give the largest improvement in accuracy [12].

In this work we use both 5 z-values and 15 z-values.

2.2.3 Amino Acid Composition (AA)

Another feature which is very simple to compute based on the protein sequence is the percentage occurrence of each amino acid within a protein sequence. This will create a feature set of 20 features, each of them with the percentage of how many times a particular amino acid occurs within the protein's amino acid sequence.

This type of feature has been used in [22, 28, 43, 1].

2.2.4 Local Descriptors (LD)

The local descriptors, also known as global protein sequence descriptors [16], were used in [7, 9, 11, 44].

There are three types of local descriptors used in the aforementioned works (and also used in our own experiments): Composition, Transition and Distribution, which are computed based on the variation of occurrence of functional groups of amino acids within the primary sequence of the protein. The functional groups used were: hydrophobic (amino acids CVLIMFW), neutral (amino acids GASTPHY), and polar (amino acids RKEDQN).

Composition accounts for the percentage composition (relative frequency) of a particular functional group within the amino acid sequence. Therefore, there are three composition features, one for each functional group of amino acids.

Transition features represent the relative frequency in which an amino acid from a particular functional group is followed by an amino acid from another functional group. More precisely, the following transitions are considered: Polar → Neutral or Neutral → Polar; Polar → Hydrophobic or Hydrophobic → Polar; and Neutral → Hydrophobic or Hydrophobic → Neutral.

Distribution features are computed based on the percentage of how many amino acids of a particular functional group are present on the first, 25%, 50%, 75% and 100% of the amino acid sequence.

In total there would be 21 features (3 composition, 3 transition, 15 distribution) if they were computed from the whole amino acid sequence. However, in [11, 44] the authors divided the protein sequence into 10 descriptor regions (A-J) as follows: Regions A,B,C and D are obtained by dividing the entire protein sequence into four equal-length regions. Regions E and F are obtained by diving the protein sequence in two equal-length regions. Region G represents the middle with 50% of the sequence. Region H represents the first 75% of the sequence, Region I the final 75% of the sequence and Region J the middle with

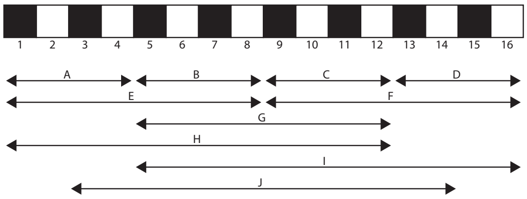

Figure 1: The 10 regions used by the Local Descriptor technique as used in [44, 11]

75% of the sequence. For each region the 21 local descriptors are extracted, resulting in a 210 feature vector. These regions are illustrated in Figure 1.

2.2.5 Motif-Based Features

Instead of computing features directly from the protein sequence, like in the previously described protein representations, it is possible to use features obtained from biological databases. In [15, 5, 6, 20, 46, 23, 31, 35, 25] the authors use the absence/presence of a particular type of protein signatures ("motifs") as binary features.

In this work we use protein signatures from four different databases as features. The employed signatures are PROSITE patterns [26], which use regular expressions to encode the motifs; Fingerprints from the PRINTS [3] database, which are created by considering several motifs to be present in the same protein; motifs from the PFAM [4] database, which are created by using hidden Markov models; and entries from the InterPro [34, 27] database.

The PROSITE patterns are encoded as regular expressions, and the rationale behind its development is that a protein family could be characterized by a single most conserved motif within a multiple alignment of its members sequences, as this would likely encode a key biological feature [21]. However, as pointed out in [21], most protein families are characterized not by one, but by several conserved motifs. This is the rationale behind the development of the fingerprints motifs used in the PRINTS database. Another approach to characterize protein families adopts the principle that the variable regions between conserved motifs also contain valuable information. In the PFAM database, the profiles are encoded using hidden Markov models. Although there is some overlap between these three databases, their content is significantly different. Also, as pointed out in [21], these motifs have different areas of application, e.g.: PROSITE patterns are unreliable in the identification of members of highly divergent superfamilies (where HMMs excel); fingerprints perform relatively poorly in the characterization of very short motifs (where PROSITE patterns do well); and HMMs are less likely to give specific subfamily diagnoses (where fingerprints excel). For these reasons, the curators of these databases (among others) decided to combine efforts in the creation of the INTERPRO database, which combines

Table 1: Summary of number of features per type.

Protein Feature Type	# of features
5 Z-Values (5z)	5
15 Z-Values (15z)	15
Amino Acid Composition (AA)	20
Local Descriptors (LD)	210
Prosite Patterns	582 for EC, 127 for GPCR
Prints Fingerprints	380 for EC, 281 for GPCR
Pfam Profiles	706 for EC, 73 for GPCR
Interpro Entries	1,214 for EC, 448 for GPCR

the information from all these and other databases.

2.2.6 Summary of Protein Features Used in this work

Table 1 presents a summary of the feature types and the respective number of features used in this work. As explained earlier, the top 4 features in Table 1 are alignment-independent features, whilst the bottom 4 features are alignment-dependent. In this table EC and GPCR refer to the Enzyme and GPCR datasets whose creation is explained in detail in section 4.2.

3 Hierarchical Protein Function Prediction

Protein functions are often specified in a functional class hierarchy, with more generic functions at higher levels and more specific functions at deeper levels. For instance, Figure 2 illustrates a small part of the Enzyme Commission hierarchy. On the first level of the hierarchy, there are 6 classes. The meaning of each class is as follows: EC 1 = Oxidoreductases, EC 2 = Transferases, EC 3 = Hydrolases, EC 4 = Lyases, EC 5 = Isomerases, EC 6 = Ligases. The remaining classes shown on Figure 2 have the following functions: EC 1.1 = Acting on the CH-OH group of donors, EC 1.1.1 = With NAD or NADP as acceptor, EC 1.1.1.1 = alcohol dehydrogenase, EC 1.1.1.2 = alcohol dehydrogenase (NADP+), EC 1.1.1.3 = homoserine dehydrogenase.

The existing hierarchical classification methods can be analyzed under different aspects [42, 17, 41], as follows:

- The type of hierarchical structure of the classes, which can be either a tree structure of a DAG (Direct Acyclic Graph) structure. DAG-structured classes are used for instance in the well-known Gene Ontology, but this type of class structure is out of the scope of this paper. In this work, the protein functions are organized into a tree-structured class hierarchy. (This is a consequence of the fact that we are using datasets originally developed in [25], as discussed later.)

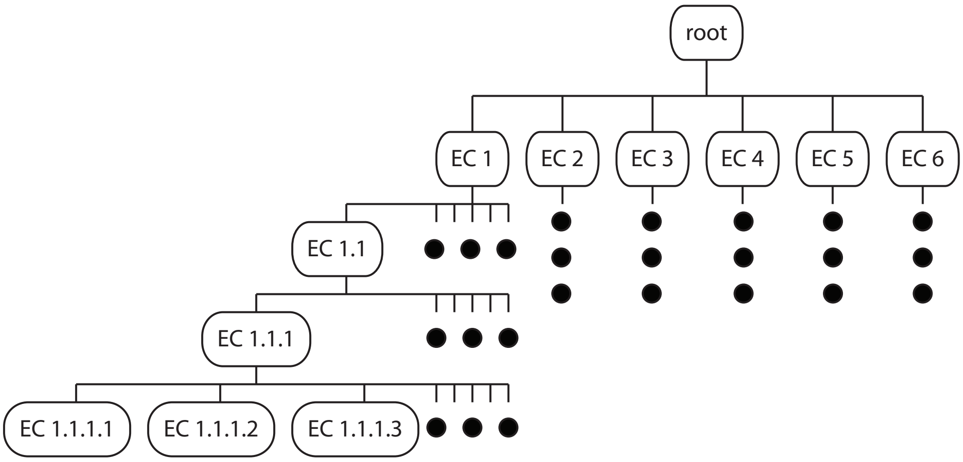

Figure 2: An excerpt of the Enzyme Commission class hierarchy

- How deep the classification in the hierarchy is performed. I.e., if the output of the classifier is always a leaf class node (which [17] refers to as Mandatory Leaf-Node Prediction and [42] refers to as Virtual Category Tree) or if the most specific ("deepest") class predicted by the classifier for a given example could be a node at any level of the class hierarchy (which [17] refers to as Non-Mandatory Leaf Node Prediction and [42] refers to as Category Tree). In this work, we are dealing with a mandatory leaf node prediction problem.

- Whether an example (protein) is assigned to exactly one leaf node in the class hierarchy or potentially assigned to two or more leaf nodes. In this work we use datasets where each protein is assigned to just one leaf node.

- How the hierarchical class structure is explored by the algorithm. The existing hierarchical classification approaches can be broadly classified into: local and global approaches. In this work we use a set of local approaches.

In the global–model approach, a single (relatively complex) classification model is built from the training set, taking into account the class hierarchy as a whole during a single run of the classification algorithm. When used during the test phase, each test example (unseen during training) is classified by the induced model, a process that can assign classes at potentially every level of the hierarchy to the test example [42, 17, 39].

The local–model approach consists of creating a local classifier for every parent node (i.e., any non-leaf node) [30] in the class hierarchy (assuming a multi-class classifier is available) or a local binary classifier for each class node (parent or leaf node, except for the root node) [10]. In the former case the classifier's goal is to discriminate among the child classes of the classifier's corresponding node. In the latter case, each binary classifier predicts whether or not an example belongs to its corresponding class. In both cases, these approaches are creating classifiers with a local view of the problem.

Despite the differences on creating and training the classifiers, these approaches are often used with the same "top-down" class prediction strategy in the testing phase. The top-down class-prediction approach works in the testing phase as follows. For each level of the hierarchy (except the top level), the decision about which class is predicted at the current level is based on the class predicted at the previous (parent) level. More precisely, once an instance (protein) is assigned a class at a certain level, only the subclasses (child nodes) of that class are considered as candidate classes to be assigned at the next lower level. The main disadvantage of the local approach with the top-down class-prediction approach is that a classification mistake at a high level of the hierarchy is propagated through all the descendant nodes of the wrongly assigned class.

3.1 Selective Classifier and Representation Approaches

In [37, 12] the authors hypothesise that it would be possible to improve the predictive accuracy of the local, top-down approach by using different classification algorithms at different nodes of the class hierarchy. The choice of which classifier to use at a given class node is made on a data-driven manner using the training set. More precisely, in order to determine which classifier should be used at each node of the class hierarchy, during the training phase, the training set is randomly split into mutually-exclusive sub-training and validation sets. Different classifiers are then trained using this sub-training set and are then evaluated on the validation set. The classifier chosen for the current class node is the one with the highest classification accuracy on the validation set. In this approach the protein representation is fixed, i.e. all classifiers are trained with the same feature set. This approach is referred to as the Selective Classifier (Sel. C.) approach.

In this work as components of the Sel. C. approach we have employed the k nearest neighbor (k-NN) with k = 3, Naive Bayes (NB) and Support Vector Machines (SVM) classifiers. All these classifiers were used with the WEKA Data mining Tool [45] with default parameters. The rationale behind the choice of these particular classifiers is that they are well-known classifiers which have been successfully used in flat (non-hierarchical) protein function prediction problems and also they have very different inductive biases, meaning that they will construct different classification models, therefore insuring a diversity of predictions to be exploited by the Sel. C. approach.

Inspired by the selective classifier approach, in [40] the authors proposed that instead of selecting the best classifier, it might be better to select the best representation (feature set) at each node of the class hierarchy. In this approach the classifier is fixed, i.e. at all class nodes the same type of classifier is trained with each of the different types of feature set, and the best type of feature (on the validation set) is chosen at each class node. This approach is referred to as the Selective Representation (Sel. R.) approach.

Another alternative [40] to selecting only the best classifier or only the best representation is to try to select the best combination of both for each node of the class hierarchy. In this approach, all the classifiers are trained with all the available representations, and the best joint combination of classifier and representation is selected. This approach is, therefore, the combination of Sel. C. and Sel. R.. This approach has the advantage of having a greater flexibility as it considers the interactions between classifiers and representations. However, it has the drawback that it is computationally expensive, meaning that in practice, we need to limit the number of classifiers and representations to a small number.

It should be noted that the Sel. R. and Sel. C. + Sel. R. approaches proposed in [40] have originally been evaluated in a music genre classification dataset, while in this work, we perform many more experiments on a very different application domain, namely protein function prediction.

Note that, at a very high level of abstraction, the idea of representation selection seems similar to the well-known idea of feature selection in data mining [33]. However, the motivation for representation selection rather than feature selection in a hierarchical classification scenario is explained by the following reasons: (a) it is much more efficient (faster) to select a representation at each class node than to perform feature selection at each class node; (b) Representation selection produces results at a coarser grain of information, possibly providing new insights to biologists, that is, it might reveal that some broad type of representations (sets of features of the same type, rather than single features) are particularly more effective to classify protein functions at particular levels. It also differs from feature selection as different representations in a dataset are actually just "candidate representations", because just one will be chosen, unlike in feature selection where any subset of features could be chosen.

4 Experimental Setup

4.1 Evaluation Metrics for Hierarchical Predictive Accuracy

Unfortunately, in the task of hierarchical classification there are no standard measures to evaluate the results. Comprehensive reviews of hierarchical classification measures can be found in [42, 8]. An aspect that can be criticized in the field is that most researchers still use flat classification measures to evaluate their hierarchical classification algorithms. Therefore, the question that naturally arises, since there is no consensus in the literature, is "What evaluation metric to use?". In order to evaluate the algorithms we have used the metrics of hierarchical precision (hP), hierarchical recall (hR) and hierarchical f-measure (hF) proposed in [29]. These measures are extended versions of the well known metrics of precision, recall and f-measure but tailored to the hierarchical classification scenario. They are defined as follows:

$hP = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{P}_i|}$, $hR = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{T}_i|}$, $hF = \frac{2 * hP * hR}{hP + hR}$, where $\hat{P}_i$ is the set consisting of the most specific class predicted for test example i and all its ancestor classes and $\hat{T}_i$ is the set consisting of the most specific true class of test example i and all its ancestor classes. The main advantage of using this particular metric is that it can be applied to any hierarchical classification scenario (i.e. single label, multi-label, tree-structured, dag-structured, mandatory-leaf node or non-mandatory leaf node problems).

4.2 Data Preparation

The protein datasets used in this work were originally developed by [25]. These datasets were originally created from the information about two types of proteins (Enzymes and GPCRs – G-Protein Coupled Receptors) obtained from different protein databases. For both datasets, the classes (protein functions) form a tree where each node represents a class. An excerpt of the class tree associated with the Enzymes dataset was shown in Figure 2, where classes at different levels are separated by a "." E.g., as shown in that figure, there are 6 classes at the first level, each of them sub-divided into sub-classes, and so on, until the fourth class level. Each class essentially refers to the type of chemical reaction catalyzed by an enzyme. In the case of the GPCR dataset, each class essentially denotes the type of ligand that binds to the GPCR (GPCRs essentially transmit signals received from ligands outside the cells to other molecules inside the cell). For further details of the meaning of the functional classes in these two datasets, see [25].

Originally there were 8 datasets (4 for Enzymes and 4 for GPCR) created based on protein data available in the Uniprot database and motif information obtained from the Interpro, Pfam, Prints and Prosite databases. Each of those datasets contained only one type of motif representation. For example, the EC-Interpro dataset had as predictive attributes only the Interpro entries motifs.

It should be noted that proteins obtained from biological databases contain non-standard amino acids and in such cases we have made the following substitutions, as it has been done in [12]: B (either an asparagine or aspartic acid) → N (asparagine); Z (either a glutamine or a glutamic acid) → Q (glutamine); X (unknown residues) → A (alanine); U (selenocysteine) → C (cysteine).

One of the objectives of this work is to evaluate the impact of the many different types of representations discussed in Section 2.2. Therefore, we expanded the number of representations used in each of the original eight datasets by extracting the alignment-independent attributes described in Section 2.2. This means that each of these 8 datasets now has 5 representations (5z,15z,AA,LD,one type of motif). These datasets are hereafter referred to as single-motif datasets.

Although these datasets allow us to verify the impact of each of the alignment-independent features against each of the motif representations, they do not allow us to verify if there is any difference in the predictive power of the different motifs representations. For this reason, we have also created two new datasets, which we refer to as "multiple-motif EC" and "multiple-motif

GPCR" which were created from the common proteins that appeared in all four corresponding specific datasets, i.e. the four datasets about EC or the four datasets about GPCR. Therefore, each multiple-motif dataset has 8 representations (5z,15z,AA,LD,Interpro motifs, Pfam motifs, Prints motifs and Prosite motifs).

Table 2 presents a summarized description of the datasets. The last column of Table 2 presents the number of classes at each level of the hierarchy (1st/2nd/3rd/ 4th levels). Note that concerning the number of protein representations, the multiple-motif datasets are more comprehensive than their single-motif counterpart datasets, because the 5 candidate representations used in a single-motif dataset are a proper subset of the 8 candidate representations used in the corresponding multiple-motif dataset. However, the motivation for performing the experiments on both single-motif and multiple-motif datasets is that the latter datasets have a reduced number of examples (specially in the case of Enzymes), since a protein is included in a multiple-motif dataset only if it appears in all the four single-motif datasets for the protein in question (Enzymes or GPCRs). Hence, the single-motif datasets have considerably more examples, offering a better statistical support to some experiments. The datasets used in the experiments are available at: http://sites.google.com/site/carlossillajr/resources.

Table 2: Dataset Details.

Dataset	# of Examples (Proteins)	# Classes per Level
Multiple-motif EC	5,221	6/35/47/70
EC-Interpro	14,027	6/41/96/187
EC-Pfam	13,987	6/41/96/190
EC-Prints	14,025	6/45/92/208
EC-Prosite	14,041	6/42/89/187
Multiple-motif GPCR	5,156	7/42/74/49
GPCR-Interpro	7,444	12/54/82/50
GPCR-Pfam	7,053	12/52/79/49
GPCR-Prints	5,404	8/46/76/49
GPCR-Prosite	6,246	9/50/79/49

5 Computational Results and Discussion

In this section, we will first discuss the impact of the different protein representations (which is less investigated in the literature) on the task of hierarchical protein function prediction and we will also discuss the impact of the different classifiers. Also, all the experiments were performed using 10-fold cross-validation.

Table 3: Hierarchical F-Measure (hF) for the single-motif datasets with 5 protein representations

Dataset	Type of Feature	Knn hF	SVM hF	NB hF	Sel.C. hF
EC-Interpro	5z	42.36	18.56	21.68	40.85
	15z	50.44	20.98	24.39	47.80
	AA	51.86	23.92	25.29	49.14
	LD	57.76	31.52	15.71	55.79
	Interpro Motif	84.28	83.02	77.01	83.01
	Sel. R.	84.26	83.00	79.65	82.97
EC-Pfam	5z	40.39	18.74	21.98	39.40
	15z	50.30	21.15	24.64	47.60
	AA	53.13	23.86	25.76	50.29
	LD	58.94	33.45	17.81	57.28
	Pfam Motif	83.94	82.36	76.30	82.73
	Sel. R.	83.95	82.49	78.77	82.60
EC-Prints	5z	39.17	19.32	21.93	39.25
	15z	49.84	22.48	21.61	50.35
	AA	53.04	25.63	25.50	53.29
	LD	59.83	41.10	24.50	60.24
	Prints Motif	83.10	80.63	79.96	82.04
	Sel. R.	83.19	81.17	81.39	83.08
EC-Prosite	5z	42.92	16.60	19.73	43.65
	15z	51.82	21.21	22.91	52.52
	AA	53.23	23.42	24.34	54.50
	LD	59.26	32.60	14.14	58.52
	Prosite Motif	85.19	83.57	81.96	85.26
	Sel. R.	85.25	83.85	83.16	85.48
GPCR-Interpro	5z	60.80	45.06	46.98	60.58
	15z	73.07	57.11	51.35	72.93
	AA	78.03	63.56	53.19	77.95
	LD	82.12	77.51	60.35	82.27
	Interpro Motif	79.44	74.36	65.80	79.52
	Sel. R.	86.16	81.66	74.72	86.39
GPCR-Pfam	5z	62.24	46.40	48.29	62.10
	15z	74.82	59.43	52.85	74.78
	AA	79.68	65.72	55.55	79.80
	LD	83.54	78.79	62.06	83.57
	Pfam Motif	68.06	59.07	57.44	67.27
	Sel. R.	85.19	84.00	74.70	85.23
GPCR-Prints	5z	67.91	50.56	52.09	67.67
	15z	77.25	60.35	56.01	77.29
	AA	80.97	66.21	55.93	81.08
	LD	83.30	79.02	61.30	83.86
	Prints Motif	76.64	72.02	64.54	76.64
	Sel. R.	83.33	81.09	74.31	83.90
GPCR-Prosite	5z	67.05	49.45	50.95	66.87
	15z	76.27	58.14	54.65	76.21
	AA	80.79	64.09	53.97	80.83
	LD	82.69	78.47	61.73	82.92
	Prosite Motif	64.54	53.56	49.80	64.54
	Sel. R.	82.69	78.52	63.67	82.97

Table 4: Hierarchical F-measures (hF) for the multiple-motif datasets with 8 protein representations

Dataset	Type of	Knn	SVM	NB	Sel.C.
	Feature	hF	hF	hF	hF
Multiple-motif EC	5z	54.92	33.39	40.99	63.56
	15z	62.55	36.97	45.00	65.66
	AA	63.70	40.11	47.34	73.32
	LD	65.67	56.44	39.94	74.42
	Interpro	79.91	79.59	79.07	81.93
	Pfam	79.43	79.07	77.39	79.37
	Prints	79.59	78.56	81.18	82.34
	Prosite	79.44	78.51	79.64	79.09
	Sel. R.	79.82	79.56	82.33	81.58
Multiple-motif GPCR	5z	68.63	51.85	53.39	68.56
	15z	77.97	61.87	57.39	77.96
	AA	81.58	67.66	57.08	81.65
	LD	83.58	79.60	62.49	84.30
	Interpro	79.79	75.47	64.54	79.89
	Pfam	63.32	51.70	51.79	62.66
	Prints	76.83	71.90	65.13	76.46
	Prosite	65.82	54.91	52.29	65.78
	Sel. R.	86.61	83.53	74.91	86.94

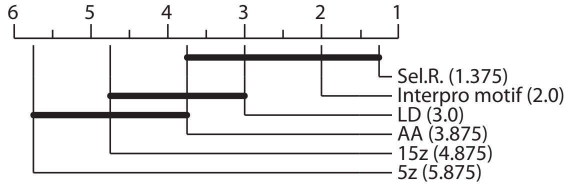

Figure 3: Analysis of relative protein representation importance on the Interpro-motif based datasets.

5.1 Impact of the different protein representations

5.1.1 Results for the single-motif datasets

One of the main contributions of this paper is to asses the impact of the choice of a type of protein representation on the hierarchical protein function prediction problem. Table 3 presents the results obtained by each representation on each single-motif dataset. However, verifying the particular importance of each protein representation is not straightforward, since as seen in section 2.2.5 different motif representations have very different rationale behind their development. For this reason, in the analysis of the different protein representations based on the single-motif datasets, we break down the analysis by the type of motif. That is, for each of the 4 types of motif, we analyse the result for both EC and GPCR datasets with that motif as a candidate representation to be selected. E.g., taking into account the results on both EC-Interpro and GPCR-Interpro, as they have the same type of motif-based protein representation.

For the Interpro-motif based datasets, considering all the representations (including the selective representation method), the average ranking of the protein representations (computed by the Friedman statistical test, considering the hierarchical f-measure values) is: Sel. R. (1.375), Interpro motifs (2.0), LD (3.0), AA (3.875), 15z (4.875) and 5z (5.875) (the smaller the rank number, the better the method). This ranking provides an overall order of the effectiveness of each protein representation across all datasets without going into the merit of wins/loses in individual datasets [14]. In order to identify on which pairwise comparisons there is a statistical difference between the results, we conduct a post-hoc test. As strongly recommended by [19] we use the Shaffer static procedure for $\alpha = 0.05$. This combination of Friedman statistical test and Shaffer post-hoc test was used to produce all results shown in Figures 3 to 8. Figure 3 shows the result of this test in a graphical way as suggested by [14]. In Figure 3 the bold horizontal lines connect the representations whose results are not found to be statistically significantly different. (This graphical representation is also used in Figures 4 through 8.) The analysis of the results in Figure 3 shows that there is no statistical difference, when comparing the Sel. R., Interpro Motifs, LD and AA. There is a statistical difference when comparing the Sel. R., Interpro Motifs and LDs with 5z and 15z.

For the Pfam-motif based datasets, the average ranking of the protein rep-

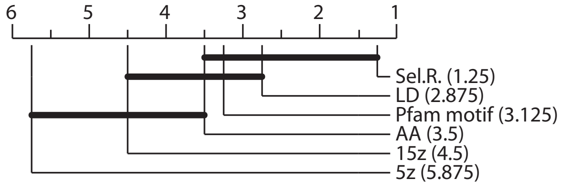

Figure 4: Analysis of relative protein representation importance on the Pfam-motif based datasets.

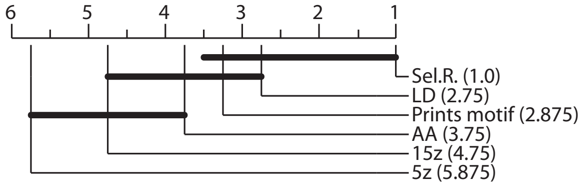

Figure 5: Analysis of relative protein representation importance on the Prints-motif based datasets.

resentations is: Sel. R. (1.25), LD (2.875), Pfam motifs (3.125), AA (3.5), 15z (4.5) and 5z (5.875). Figure 4 shows the graphical result of the Shaffer static post-doc test. The analysis of the results in Figure 4 shows that there is no statistical difference, when comparing the Sel. R., LD, Pfam Motifs and AA. There is a statistical difference when comparing the Sel. R., Pfam Motifs and LDs with 5z and 15z.

For the Prints-motif based datasets, the average ranking of the protein representations is: Sel. R. (1.0), LD (2.75), Prints motifs (2.875), AA (3.75), 15z (4.75) and 5z (5.875). The analysis of the results in Figure 5 shows that there is no statistical difference, when comparing the Sel. R., LD, Pfam Motifs and AA. There is a statistical difference when comparing the Sel. R., Prints Motifs and LDs with 5z and 15z.

For the Prosite-motif based datasets, the average ranking of the protein representations is: Sel. R. (1.0625), LD (2.8125), AA (3.5), Prosite motifs (3.875), 15z (4.25) and 5z (5.5). The analysis of the results in Figure 6 shows that there is no statistical difference, when comparing the Sel. R., LD, and AA. There is a statistical difference when comparing the Sel. R. with Prosite Motifs, 5z and 15z.

5.1.2 Results for the Multiple-Motif datasets

Recall that apart from the single-motif datasets, we have also created two multiple-motif datasets in order to evaluate the performance of each particular type of motif against the others as well as against the alignment-independent features and the Sel. R. approach. Table 4 presents the predictive accuracy (measured by hierarchical precision, recall and f-measure values) by each rep-

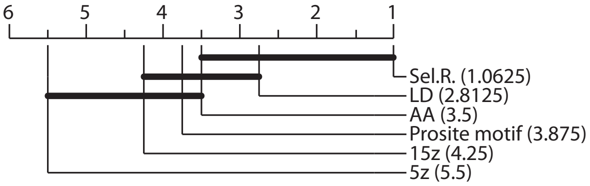

Figure 6: Analysis of Relative protein representation importance on the Prosite-motif based datasets.

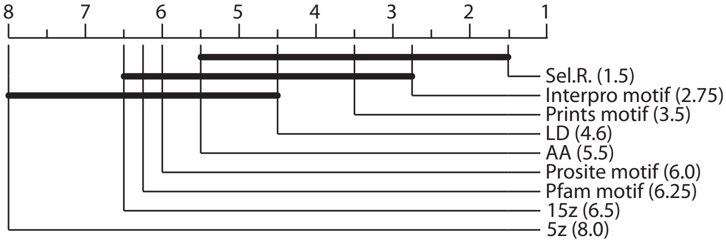

Figure 7: Analysis of relative protein representation importance on the multiple-motif datasets.

resentation on the multiple-motif datasets. The average ranking of the protein representations (computed by the Friedman statistical test, considering the hierarchical f-measure values) is: Sel. R. (1.5), Interpro motifs (2.75), Prints motifs (3.5), LD (4.625), AA (5.5), Prosite motifs (6.0), Pfam motifs (6.625), 15z (6.5), 5z (8.0). Again, this ranking provides an overall order of the effectiveness of each protein representation across all datasets and classification algorithms without going into the merits of individual wins/loses. Figure 7 shows the graphical result of the Shaffer static post-doc test. The analysis of the results in Figure 7 shows that there is no statistical difference, when comparing the Sel. R., Intepro motifs, Prints motifs, LD, and AA. There is a statistical difference when comparing the Sel. R. with Pfam and Prosite Motifs, 5z and 15z.

5.1.3 Discussion of Results for Different Protein Representations

The overall analysis of the results shows some interesting points. First, although not statistically significantly different from some representations, the Sel. R. has ranked 1st in all experiments, meaning that it is an interesting approach to deal with the problem of hierarchical protein function prediction.

Second, the result that 15z is better than 5z (although not statistically significant) corroborates with the experiments of [11, 38] where the authors came to the same conclusion. Note, however, that in their experiments they used only one GPCR dataset, while in this study we have employed 4 GPRC and 4 Enzyme datasets. Our work therefore, validates their initial proposal in a larger number of datasets. According to [37], the z-values representation provides a numerical description of the proteins' physiochemical properties that potentially results in a higher predictive accuracy than the use of amino acid sequence composition. However, in our experiments we have empirically verified that this is not the case, since the AA features are always ranked above both 5z and 15z features (although this difference is not statistically significant).

Third, the best performing alignment-independent feature is the LD. Its results are better than all the other alignment-independent features (although only statistically significantly different from 5z on some motif-based datasets).

Fourth, the use of the alignment-dependent features (motifs) on the single-motif datasets have ranked 2nd for Intepro motifs, 3rd for Pfam and Prints motifs and 4th for the Prosite motifs. On the multiple-motif datasets the alignment-dependent features (motifs) have ranked 2nd (Intepro), 3rd (Prints), 6th (Prosite) and 7th (Pfam). Considering the rankings it is clear that the use of Interpro motifs lead to higher predictive accuracies than the use of other types of motifs. This is an expected result, since (as previously discussed) Interpro is a joint effort from curators of all its members databases which includes Prosite, Prints and Pfam among others.

Note that the Sel. R. was the best protein representation on both experiments (single-motif datasets and multiple-motif datasets). Hence, it is interesting to analyse which features were selected the most by the selective representation approach at each level of the class hierarchy. Tables 5 and 6 present the percentage of how many times a particular protein representation was selected in each dataset at each level of the class hierarchy for the datasets with 5 and 8 representations, respectively, corresponding to single-motif and multiple-motif datasets, respectively.

The analysis of Table 5 shows that for the single-motif datasets, the motif features are highly predictive for the classes at the first level of the class hierarchy being selected on average in 90.6% of the time. In fact, the only dataset where other type of protein representation is selected at this level is the GPCR-Prosite dataset, where the LD representation is selected 75% of the time. For the other three class levels, it seems that the motifs are often selected for the EC datasets, while a combination of alignment-independent features are selected for GPCRs. An explanation for this was presented in [13] were the authors claim that there are several instances where the application of alignment-free techniques have been proven to be more effective than alignment-based techniques. And the GPCRs are an example of this, because they have a great structural and/or functional homology but a low degree of sequence similarity.

For the multiple-motif datasets presented in Table 6 the same conclusions can be drawn. That is, the motif features are highly discriminative at the 1st level of the class hierarchy, specially the Interpro motifs. There is a significant difference in the number of times that motif-based and alignment-independent features are selected for the Enzyme and GPCRs datasets. These results confirm that the Sel. R. approach effectively determines which protein representation is the best to be used with each classifier across different levels in the class hierarchy structure.

Table 5: Percentage of times each representation is selected by the Sel. R. method per class level per dataset for 5 protein representations

Rep.	Dataset	Class Level			
		1	2	3	4
5z	EC-Interpro	0	0	8	22
	EC-Pfam	0	0	8	21
	EC-Prints	0	0	7	14
	EC-Prosite	0	0	4	13
	GPCR-Interpro	0	7	10	7
	GPCR-Pfam	0	8	10	8
	GPCR-Prints	0	8	11	8
	GPCR-Prosite	0	17	11	8
	Average	0	5.625	10.875	14.375
15z	EC-Interpro	0	4	6	11
	EC-Pfam	0	0	5	11
	EC-Prints	0	0	9	17
	EC-Prosite	0	0	5	9
	GPCR-Interpro	0	8	15	17
	GPCR-Pfam	0	11	16	19
	GPCR-Prints	0	17	13	15
	GPCR-Prosite	0	5	18	16
	Average	0	5.625	10.875	14.375
AA	EC-Interpro	0	3	12	14
	EC-Pfam	0	2	11	19
	EC-Prints	0	0	14	19
	EC-Prosite	0	0	10	18
	GPCR-Interpro	0	32	15	29
	GPCR-Pfam	0	31	20	30
	GPCR-Prints	0	29	18	33
	GPCR-Prosite	0	23	22	32
	Average	0	15	15.25	24.25
LD	EC-Interpro	0	1	7	7
	EC-Pfam	0	6	12	8
	EC-Prints	0	12	14	16
	EC-Prosite	0	1	8	13
	GPCR-Interpro	0	17	44	36
	GPCR-Pfam	0	28	54	41
	GPCR-Prints	0	22	40	31
	GPCR-Prosite	75	43	48	42
	Average	9.375	16.25	28.375	24.25
Motifs	EC-Interpro	100	92	67	46
	EC-Pfam	100	92	64	41
	EC-Prints	100	88	56	34
	EC-Prosite	100	99	73	47
	GPCR-Interpro	100	36	16	11
	GPCR-Pfam	100	22	0	2
	GPCR-Prints	100	24	18	13
	GPCR-Prosite	25	12	1	2
	Average	90.625	58.125	36.875	24.5

Table 6: Percentage of times each representation is selected by the Sel. R. method per class level per dataset for 8 protein representations

Rep.	Dataset	Class Level			
		1st	2nd	3rd	4th
5z	Multiple-motif EC	0	0	5	23
	Multiple-motif GPCR	0	15	11	9
	Average	0	7.5	8	16
15z	Multiple-motif EC	0	7	5	4
	Multiple-motif GPCR	0	7	13	14
	Average	0	7	9	9
AA	Multiple-motif EC	0	5	11	33
	Multiple-motif GPCR	0	17	14	31
	Average	0	11	12.5	32
LD	Multiple-motif EC	0	5	8	12
	Multiple-motif GPCR	0	28	32	28
	Average	0	16.5	20	20
Interpro	Multiple-motif EC	95	37	49	23
	Multiple-motif GPCR	73	26	22	15
	Average	84	31.5	35.5	19
Pfam	Multiple-motif EC	0	12	2	0
	Multiple-motif GPCR	25	7	0	0
	Average	12.5	9.5	1	0
Prints	Multiple-motif EC	0	13	18	1
	Multiple-motif GPCR	2	0	8	3
	Average	1	6.5	13	2
Prosite	Multiple-motif EC	5	21	2	4
	Multiple-motif GPCR	0	0	0	0
	Average	2.5	10.5	1	2

5.2 Impact of the different classifiers

It is a well-known fact in machine learning that there is “no free-lunch”, i.e. a classifier which is the best for all applications do not exist. Recall that in this work we are employing the selective classifier approach with three classification algorithms: k-NN, SVM and NB. To measure the performance of the classifiers we consider their average ranking over all datasets and over all representations (computed by the Friedman statistical test, considering the hierarchical f-measure values). The resulting ranking is: Sel. C. (1.5454), Knn (1.560606), SVM (3.3181), NB (3.5757).

Again we employ the Shaffer static post-hoc test and the graphical representation of the result of the test is shown in Figure 8. The analysis of the results in Figure 8 shows that the Sel. C. and Knn are both (statistically significant) better than SVM and NB, but there is no statistically significant difference between the results of Sel. C. and KNN. Also, there is no statistically significant difference between the results of SVM and NB.

Considering we are using the Sel. C. approach and it gives results just slightly better than the Knn classifier, a question that naturally arises is if in the internal classifier selection procedure of the Sel. C. method the Knn classifier is almost always chosen. Table 7 presents the relative classifier importance for each dataset, i.e. the number of times a particular classifier is selected at each class level. The analysis of Table 7 reveals that at the first level the Knn classifier is selected in about 90% o the experiments. This result corroborates with the experiments reported in [38] where for one GPCR dataset the Knn classifier was always selected at the first class level. For the other class levels it seems that, although the Sel. C. approach actually selects different classifiers, this does not impact significantly on the results. Other studies on hierarchical protein function prediction that employed the Sel. C. approach achieved similar conclusions [37, 24], i.e. the Sel. C. is better than most classifiers but is not statistically significantly different from a Knn classifier, even though the former employs several classifiers, which has the disadvantage of considerably increasing the training time of the hierarchical classification system. Therefore, it seems that the use of the Sel. C. approach does not bring the same benefits as the Sel. R. approach.

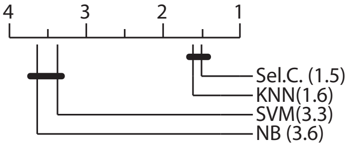

Figure 8: Analysis of relative classifier importance over all datasets.

Moreover, the negative impact of using a bad representation even with a good classifier (e.g. 5z with Knn on EC-Intepro has an hierarchical f-measure of 42.36%) seems to be greater than the impact of using a bad classifier with a good representation (e.g. Motif with NB on EC-Interpro has an hierarchical

Table 7: Percentage of times each classifier is selected by the Sel. C. method per class level per dataset

Classifier	Dataset	Class Level			
		1	2	3	4
SVM	EC-Interpro	20	35	41	59
	EC-Pfam	20	36	41	59
	EC-Prints	0	30	39	47
	EC-Prosite	8	23	37	50
	GPCR-Interpro	0	51	50	56
	GPCR-Pfam	18	49	42	44
	GPCR-Prints	0	49	47	60
	GPCR-Prosite	0	44	43	44
	Multiple-motif EC	12	57	52	66
	Multiple-motif GPCR	14	36	36	47
	Average	9.2	41	42.8	53.2
KNN	EC-Interpro	80	65	49	20
	EC-Pfam	80	64	50	15
	EC-Prints	100	70	44	27
	EC-Prosite	92	77	61	23
	GPCR-Interpro	100	36	42	30
	GPCR-Pfam	82	40	48	38
	GPCR-Prints	100	40	44	28
	GPCR-Prosite	100	46	45	39
	Multiple-motif EC	88	26	37	15
	Multiple-motif GPCR	86	55	56	43
	Average	90.8	51.9	47.6	27.8
NB	EC-Interpro	0	0	10	21
	EC-Pfam	0	0	9	26
	EC-Prints	0	0	17	26
	EC-Prosite	0	0	2	27
	GPCR-Interpro	0	13	8	14
	GPCR-Pfam	0	11	10	18
	GPCR-Prints	0	11	9	12
	GPCR-Prosite	0	10	12	17
	Multiple-motif EC	0	17	11	19
	Multiple-motif GPCR	0	9	8	10
	Average	0	7.1	9.6	19

f-measure of 77.01%). Interestingly, most papers in protein function prediction are more concerned with trying different classification algorithms than different features and their impact on predictive accuracy.

6 Conclusions

In this work we presented an empirical study analyzing the impact of different protein representations and different types of classification algorithms for the task of hierarchical protein function prediction. We have employed 8 types of protein representation, 4 of which are alignment-independent representations computed from the protein sequence: 5 z-values (5z), 15z-values (15z), Amino Acid Composition (AA) and Local Descriptors (LD); and 4 alignment-dependent protein signatures (motifs) from the biological databases Interpro, Pfam, Prints and Prosite. To perform the classification we have used 3 classifiers: k-NN, SVM and Naive Bayes and 3 selective approaches: one fixing the classifier for all nodes in the class hierarchy and selecting the best representation at each class node, one fixing the representation and selecting the best classifier at each class node, and one that selects the best match of representation and classifier at each node of the class hierarchy.

We have carried out the experiments on 10 datasets, being 5 datasets with G-Protein Coupled Receptors (GPCR) proteins and 5 with Enzymes. Our experimental results show that in general, regardless of the type of protein: 15 z-values are better than 5 z-values; AA is a very good descriptor with k-NN since it is simple and provides better results than z-values; LD is the best alignment-independent representation that can be computed directly from the protein sequence.

Considering the results specifically for GPCRs, the LD provides very good results (except for the GPCR-Interpro, in which its results are similar to the results of the motif representation) considering they can be computed directly from the sequence. Concerning the results specific to the EC datasets the motif representation performs better than the alignment independent features computed from the sequence. The fact that these better results of alignment-independent features was observed for GPCRs but not for enzymes is possibly explained by the fact that GPCRs have a great structural and/or functional homology but a low degree of sequence similarity, which does not seem to be the case for enzymes.

Therefore, our recommendation (based on our experimental results) is that when using alignment-independent features derived from the sequence, we suggest the use of Local Descriptors. When motif features are available, we recommend the use of the Interpro entries as they provide in general better results than the other types of motifs for GPCRs and all motif features are roughly equally effective for Enzyme classification.

Future research would include performing experiments with other types of protein representations, more classifiers and with other types of proteins. Another direction for future research is to perform more controlled experiments to see if the number of features has a significant influence on the effectiveness of a particular type of feature: e.g. the z-values representation has a very small number of features, what if z-values were computed for the same 10 regions as the LD approach (considerably increasing the number of z-value features)?

Acknowledgment

The first author is financially supported by CAPES – a Brazilian research-support agency (process number 4871-06-5).

References

[1] Al-Shahib, A., Breitling, R., and Gilbert, D. Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics 4*, 3 (2005), 195–203.

[2] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. *Molecular Biology of the Cell.* Garland Publishing, 2002.

[3] Attwood, T. K. The prints database: A resource for identification of protein families. *Briefings in Bioinformatics 3*, 3 (2002), 252–263.

[4] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. The pfam protein families database. *Nucleic Acids Research 32* (2004), D138–D141. Database issue.

[5] Ben-Hur, A., and Brutlag, D. Remote homology detection: a motif based approach. *Bioinformatics 19*, Suppl. 1 (2003), i26–i33.

[6] Ben-Hur, A., and Brutlag, D. *Feature extraction, foundations and applications.* Springer, 2006, ch. Protein sequence motifs: Highly predictive features of protein function, pp. 625–645.

[7] Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research 31*, 13 (2003), 3962–3697.

[8] Costa, E., Lorena, A., Carvalho, A., and Freitas., A. A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for Machine Learning II: papers from the 2007 AAAI Workshop* (2007), AAAI Press, pp. 1–6.

[9] Cui, J., Han, L. Y., Li, H., Ung, C. Y., Tang, Z. Q., Zheng, C. J., Cao, Z. W., and Chen, Y. Z. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Molecular Immunology 44* (2007), 514–540.

[10] D´ Alessio, S., Murray, K., Schiaffino, R., and Kershenbaum, A. The effect of using hierarchical classifiers in text categorization. In *Proc. of the 6th Int. Conf. Recherche d´ Information Assistee par Ordinateur* (2000), pp. 302–313.

[11] Davies, M., Secker, A., Freitas, A., Clark, E., Timmis, J., and Flower, D. Optimizing amino acid groupings for GPCR classification. *Bioinformatics 24*, 18 (2008), 1980–1986.

[12] Davies, M., Secker, A., Freitas, A., Mendao, M., Timmis, J., and Flower, D. On the hierarchical classification of G protein-coupled-receptors. *Bioinformatics 23*, 23 (2007), 3113–3118.

[13] Davies, M., Secker, A., Freitas, A., Timmis, J., Clark, E., and Flower, D. Alignment-independent techniques for protein classification. *Current Proteomics 5*, 4 (2008), 217–223.

[14] Demsar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7* (2006), 1–30.

[15] Drawid, A., and Gerstein, M. A bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *Journal of Molecular Biology 301* (2000), 1059–1075.

[16] Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S.-H. Prediction of protein folding class using global description of amino acid sequence. In *Proceedings of the National Academy of Sciences of the United States of America* (1995), vol. 92, pp. 8700–8704.

[17] Freitas, A. A., and de Carvalho, A. C. P. L. F. *Research and Trends in Data Mining Technologies and Applications.* Idea Group, 2007, ch. A Tutorial on Hierarchical Classification with Applications in Bioinformatics, pp. 175–208.

[18] Friedberg, I. Automated protein function prediction – the genomic challenge. *Briefings in Bioinformatics 7*, 3 (2006), 225–242.

[19] García, S., and Herrera, F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research 9* (2008), 2677–2694.

[20] Hayete, B., and Bienkowska, J. Gotrees: Predicting go associations from protein domain composition using decision trees. In *Proc. of the Pacific Symp. on Biocomputing* (2005), pp. 127–138.

[21] Higgs, P. G., and Attwood, T. K. *Bioinformatics and Molecular Evolution.* Blackwell Publishing, 2005.

[22] Hobohm, U., and Sander, C. A sequence property approach to searching protein databases. *Journal of Molecular Biology 251* (1995), 390–399.

[23] Holden, N., and Freitas, A. A. Hierarchical classification of g-protein-coupled receptors with a pso/aco algorithm. In *Proc. of the 3rd IEEE Swarm Intelligence Symposium* (2006), pp. 77–84.

[24] Holden, N., and Freitas, A. A. Improving the performance of hierarchical classification with swarm intelligence. In *Proc. 6th European Conf. on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio)* (2008), vol. 4973 of *Lecture Notes in Computer Science*, Springer, pp. 48–60.

[25] Holden, N., and Freitas, A. A. Hierarchical classification of protein function with ensembles of rules and particle swarm optimisation. *Soft Computing Journal 13* (2009), 259–272.

[26] Hulo, N., A., A. B., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P., M., P., and Sigrist, C. The prosite database. *Nucleic Acids Research 34* (2006), D227–D230.

[27] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. Interpro: the integrative protein signature database. *Nucleic Acids Research 37* (2009), D211–D215. Database issue.

[28] King, R. D., Karwath, A., Clare, A., and Dehaspe, L. The utility of different representations of protein sequence for predicting functional class. *Bioinformatics 17*, 5 (2001), 445–454.

[29] Kiritchenko, S., Matwin, S., and Famili, A. F. Functional annotation of genes using hierarchical text categorization. In *Proc. of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics* (2005).

[30] Koller, D., and Sahami, M. Hierarchically classifying documents using very few words. In *Proc. of the 14th Int. Conf. on Machine Learning* (1997), pp. 170–178.

[31] Kong, W., Tan, T. S., Tham, L., and Choo, K. W. Improved prediction of allergenicity by combination of multiple sequence motifs. *Silico Biology 7*, 1 (2007), 77–86.

[32] Lapinsh, M., Prusis, P., Lundstedt, T., and Wikberg, J. E. S. Proteochemometrics modeling of the interaction of amine g-protein coupled receptors with a diverse set of ligands. *Molecular Pharmacology 61*, 6 (2002), 1465–1475.

[33] Liu, H., and Motoda, H., Eds. *Computational Methods of Feature Selection.* Chapman & Hall, 2007.

[34] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, E., Kejariwal, A., Labarga, A., Langendijk-genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., Mcanulla, C., Mcdowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, R., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J. A., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. New developments in the interpro database. *Nucleic Acids Research 35* (2007), D224–D228. Database Issue.

[35] Nariai, N., Kolaczyk, E. D., and Kasif, S. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE 2*, 3 (2007), e377.

[36] Sandberg, M., Eriksson, L., Jonsson, J., Sjostrom, M., and Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medical Chemistry 41* (1998), 2481–2491.

[37] Secker, A., Davies, M., Freitas, A., Timmis, J., Mendao, M., and Flower, D. An experimental comparison of classification algorithms for the hierarchical prediction of protein function. *Expert Update (the BCS-SGAI Magazine) 9*, 3 (2007), 17–22.

[38] Secker, A., Davies, M., Freitas, A. A., Clark, E., Timmis, J., and Flower, D. R. Hierarchical classification of g-protein-coupled-receptors with data-driven selection of attributes and classifiers. *International Journal of Data Mining and Bioinformatics* (2010). To appear.

[39] Silla Jr., C. N., and Freitas, A. A. A global-model naive bayes approach to the hierarchical prediction of protein functions. In *Proc. of the IEEE Int. Conf. on Data Mining* (2009), pp. 992–997.

[40] Silla Jr., C. N., and Freitas, A. A. Novel top-down approaches for hierarchical classification and their application to automatic music genre classification. In *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics* (2009), pp. 3599–3604.

[41] Silla Jr., C. N., and Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* (2010), To appear.

[42] Sun, A., and Lim, E.-P. Hierarchical text classification and evaluation. In *Proc. of the IEEE Int. Conf. on Data Mining* (2001), pp. 521–528.

[43] Syed, U., and Yona, G. Using a mixture of probabilistic decision trees for direct prediction of protein function. In *Proceedings of the seventh annual international conference on Research in computational molecular biology* (2003), pp. 289–300.

[44] Tong, J. C., and Tammi, M. T. Prediction of protein allergenicity using local description of amino acid sequence. *Frontiers in Bioscience 13* (2008), 6072–6078.

[45] Witten, I. H., and Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, 2005.

[46] Zhang, Z., Kochhar, S., and Grigorov, M. G. Descriptor-based protein remote homology identification. *Protein Science 14* (2005), 431–444.

[47] Zhao, X. M., Chen, L., and Aihara, K. Protein function prediction with high-throughput data. *Amino Acids 35*, 3 (2008), 517–530.