Motif-based Protein Sequence Classification Using Neural Networks

Konstantinos Blekas*, Dimitrios I. Fotiadis and Aristidis Likas

Department of Computer Science and
Biomedical Research Institute - FORTH
University of Ioannina
GR-45110 Ioannina, Greece

* E-mail: kblekas@cs.uoi.gr
Tel.: (+30) 26510 98816
Fax: (+30) 26510 98890

Abstract

We present a system for multi-class protein classification based on neural networks. The basic issue concerning the construction of neural network systems for protein classification is the sequence encoding scheme that must be used in order to feed the neural network. To deal with this problem we propose a method that maps a protein sequence into a numerical feature space using the matching scores of the sequence to groups of conserved patterns (called motifs) into protein families. We consider two alternative ways for identifying the motifs to be used for feature generation and provide a comparative evaluation of the two schemes. We also evaluate the impact of the incorporation of background features (2-grams) on the performance of the neural system. Experimental results on real datasets indicate that the proposed method is highly efficient and is superior to other well-known methods for protein classification.

Keywords: protein sequence classification, neural networks, probabilistic motifs, MEME algorithm, motif-based features.

1 Introduction

Protein sequence classification constitutes an important problem in biological sciences for annotating new protein sequences and detecting close evolutionary relationships among sequences. It deals with the assignment of sequences to known categories based on homology detection properties (sequence similarity). In several studies, protein classification has been examined at various levels, according to a top-down hierarchy in molecular taxonomy, consisting of superfamilies, families and subfamilies (Dayhoff et al., 1978). Throughout this paper we will use the terms family (or subfamily) and class interchangeably to denote any collection of sequences that are presumed to share common characteristics and belong to the same category.

Various approaches have been developed for solving the protein classification problem. Most of them are based on appropriately modeling protein families, either directly or indirectly. Direct modeling techniques use a training a set of sequences to build a model that characterizes the family of interest. Hidden Markov models (HMMs) are a widely used probabilistic modeling method for protein families (Durbin et al., 1998) that provides a probabilistic measurement (score) of how well an unknown sequence fits to a family. Indirect techniques use direct models as a preprocessing tool in order to extract useful sequence features. In this way, sequences of variable length are transformed into fixed-length input vectors that are subsequently used for training discriminative models, such as neural networks.

In protein sequences, motifs or patterns enclose significant homologous attributes, since they correspond to conserved regions in protein families holding useful structural and functional biological properties. They can be considered as islands of aminoacids conserved in the same order of a given family. Therefore they can be seen as local features characterizing the sequences. Motifs can be either deterministic or probabilistic (Brāzma et al., 1998; Rigoutsos et al., 2000). Deterministic motifs follow grammatical inference properties in order to syntactically describe conserved regions of homologous sequences. The PROSITE database (Hofmann et al., 1999) represents a large collection of such motifs used to identify protein families. On the other hand, probabilistic motifs are more flexible models and they provide a probabilistic matching score of a sequence to a motif. The BLOCKS database (Henikoff and Henikoff, 1994) is an example of ungapped probabilistic motifs. In any case, motif-models are suitable for constructing efficient similarity score functions that can be subsequently used as local features for protein classification. An example is presented in (Ma and Wang, 2000; Wang et al., 2001) where motif-based local features are produced based on the minimum description length (MDL) principle for the case of deterministic motif models.

The background information also constitutes another source for extracting features from

sequence data. The determination of the background features, also defined as global features, is usually made by using the 2-gram encoding scheme that counts the occurrences of two consecutive aminoacids in protein sequences (Wang et al., 2001). In the case of protein sequences (generated from the alphabet of the 20 aminoacids), there are 400 possible 2-grams, that produce a large feature space. A recent approach (Almeida and Vinga, 2002) proposes a scheme for globally encoding sequences, where each aminoacid character is initially represented as a unique binary number with n bits ($n = 5$ for the 20 aminoacids) and then each sequence is mapped into a position inside the n-dimensional hypercube.

In this paper, we focus on building efficient neural classifiers for discriminating multiple protein families by using appropriate local features that have been extracted by efficient probabilistic motif models. As motifs constitute family diagnostic signatures, our aim is to exploit this information by constructing a neural network scheme that exploits motif-based (local) features.

The proposed method can be considered as combining an unsupervised with a supervised learning technique. Starting by applying a motif discovery (unsupervised) algorithm, we identify probabilistic motifs in a training set of multi-class sequences. This can be achieved in two alternative ways: applying the algorithm for motif discovery either to each family training set separately (class-dependent motifs), or to the whole dataset of training sequences (class-independent motifs). The discovered motifs are then used to convert each sequence to a numerical input vector that subsequently can be applied to a typical feedforward neural network. Using a Bayesian regularization training technique, the neural network parameters are adjusted and therefore a classifier is obtained suitable for predicting the family of an unlabeled sequence.

The next section provides a brief overview of statistical and neural techniques proposed for classifying biological sequences, while Section 3 describes the proposed method. Experimental results obtained using several sets of protein families are presented in Section 4, along with a comparison with other known protein classification approaches. Finally, Section 5 summarizes the proposed classification scheme and specifies directions for future research.

2 Protein Classification Methods

One class of methods for protein sequence classification work directly with sequence information and establish classification criteria based on sequence homology properties. In the general scheme, a representative set of sequences is selected for each protein family and used to build an appropriate model for each family. The classification of an unknown sequence is made by selecting the family that best matches according to the model homology mechanism. This can be considered as a simple nearest neighbor scheme that ranks sequence similarities and selects the best ranking.

The popular BLAST tool (Altshul et al., 1990) represents the simplest nearest neighbor approach and exploits pairwise local alignments to measure sequence similarity. The BLAST technique compares protein queries with a protein database of labeled sequences and produces normalized alignment scores for each comparison by calculating the corresponding expectation values (E-values). The classification procedure is based on the selection of the label of the sequence that produces the best pairwise alignment score (i.e. minimum E-value).

Another type of direct modeling methods is based on hidden Markov models (HMMs) (Durbin et al., 1998; Karplus et al., 1998). After constructing an HMM for each family, protein queries can be easily scored against all established HMMs by calculating the log-likelihood of each model for the unknown sequence and then selecting the class label of the most likely model.

The Motif Alignment and Search Tool (MAST) (Bailey and Gribskov, 1998) is based on the combination of multiple motif-based statistical score values. According to this scheme, groups of probabilistic motifs discovered by the MEME algorithm (Bailey and Elkan, 1994), are used to construct protein profiles for the families of interest. The MAST algorithm successively estimates the significance of the match of a query sequence to a family model as the product of the p-values of each motif match score. This measure (called E-value) can then be used to select the family of the unknown sequence.

Neural network schemes for protein classification consist of two stages: the encoding stage, where discriminative numerical features are computed for each training sequence and the decision stage where the created feature vectors are used as input vectors to a neural network classifier. Various encoding schemes have been proposed in the literature that produce numerical features in the encoding stage based on the calculation of background features (global sequence homology) and local features (locally conserved family information) embedded in motifs. In the decision stage feedforward neural networks have been used trained either through backpropagation (Wu et al., 1996) or using Bayesian regularization (Ma and Wang, 2000; Wang et al., 2001). These approaches are characterized by the enormous size of the

extracted input vectors, the imbalance between global and local features (more emphasis on global features) and the need for large training sets (since the number of network inputs is very large). For example in (Ma and Wang, 2000; Wang et al., 2001) only one feature was responsible for carrying local information, while all the others were produced by the 2-grams encoding scheme (background features).

Support vector machines (SVMs) (Vapnik, 1979) have been also applied to protein homology detection problems. Such an approach, which has been introduced in (Logan et al., 2001), feeds probabilistic score values from all motifs available (nearly 10000) in the BLOCKS database (Henikoff and Henikoff, 1994) into an SVM classifier. Obviously, this scheme uses only local features but the dimensionality of the input space is extremely high. Another method has been proposed in (Jaakkola et al., 2000; Karchin et al., 2002) that combines hidden Markov models (HMMs) and SVMs for detecting remote protein homologies. In particular, an HMM is first trained to model a protein family, and then the observed probabilities (in the log space) of each sequence with respect to each parameter of the HMM are calculated. The obtained gradient-log-probability vectors are applied to an SVM to identify the decision boundary between the family and the rest of the protein universe.

3 The proposed method

This paper studies the problem of classifying a set of N protein sequences $\mathbf{S} = \{S_i, i = 1, \ldots, N\}$ into K classes. The set $\mathbf{S}$ is a union of positive example datasets $\mathcal{S}_k$ from K different classes, i.e. $\mathbf{S} = \{ \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_K \}$, and can be seen as a subset of the complete set of all possible sequences over the aminoacid alphabet ($\mathbf{S} \subseteq \Sigma^*$).

Figure 1 near here

Figure 1 illustrates the architecture of the proposed protein classification scheme. It consists of a search tool (unsupervised learning) for discovering probabilistic motifs in a set of K protein families, a feature vector generator that converts protein sequences into feature vectors, and a decision module (neural network) for assigning a protein family to each input sequence. The following subsections describe in detail the major building blocks of the proposed architecture.

3.1 Using motifs for feature generation

Consider a finite alphabet consisting of set of characters $\Sigma = \{\alpha_1, \ldots, \alpha_\Omega\}$ ($\Omega = 20$ for protein sequences). We can probabilistically model a contiguous (ungapped) motif M_j of length W_j using a position weight matrix (PWM_j) that follows a multinomial character distribution. Each column (l) of the matrix corresponds to a position l in the motif sequence ($l = 1, \ldots, W_j$), where the column elements provide the probability of each character of the alphabet $p_{\alpha_\xi, l}$ ($\xi = 1, \ldots, \Omega$) to appear in that position.

Let $s_p = a_{p,1} \ldots a_{p,W_j}$ denote a segment of a sequence S beginning at position p and ending at position $p + W_j - 1$. This represents a subsequence of length W_j. Totally, there are $L - W_j + 1$ such subsequences for a sequence S of length L. Then, we can define the probability that s_p matches the motif M_j, or alternatively, has been generated by the model PWM_j corresponding to that motif, using the following equation:

$$P(s_p|M_j) = \prod_{l=1}^{W_j} p_{a_{p,l},l} \ . \tag{1}$$

A major advantage of using the probabilistic matrix PWM_j is the ability to compute the corresponding position-specific score matrix ($PSSM_j$) in order to score a sequence. The $PSSM_j$ is a log-odds matrix calculating the logarithmic ratio $r_{\alpha_\xi, l}$ of the probabilities $p_{\alpha_\xi, l}$ suggested by the PWM_j and the corresponding general relative frequencies of aminoacids ρ_{α_ξ}

in the family[1]. According to the definition of $PSSM_j$, the score value $f_j(s_p)$ of a subsequence s_p of a sequence S can be defined as:

$$f_j(s_p) = \sum_{l=1}^{W_j} \log(\frac{p_{a_{p,l},l}}{\rho_{a_{p,l}}}) = \sum_{l=1}^{W_j} r_{a_{p,l},l} \ . \tag{2}$$

At the sequence level, the score value of a protein sequence S against a motif M_j can be determined as the maximum value among all scores of the possible subsequences of S, i.e.:

$$f_j(S) = \max_{1 \leq p \leq L - W_j + 1} f_j(s_p) \ . \tag{3}$$

It must be noted that it is possible to adopt other definitions for scoring a sequence, such as setting scores below a certain threshold equal to zero (Bailey and Gribskov, 1998).

If we assume that we have discovered a group of D motifs in the set of sequences $\mathbf{S}$, we can generate a D-dimensional numerical feature space and map each sequence S_i into a vector $\mathbf{x}_i$ in the D-dimensional feature space by calculating the score values $x_{ij} = f_j(S_i)$ $(j = 1, \ldots, D)$ for each of the D motif models.

3.2 Finding probabilistic motifs in protein sequences

Several approaches have been proposed for discovering probabilistic motifs in a set of unaligned biological sequences. The CONSENSUS (Hertz and Stormo, 1999), Gibbs sampler (Lawrence et al., 1993) and MEME (Bailey and Elkan, 1994) are examples of such methods that identify multiple shared motifs in protein families. We have selected the MEME approach for the motif identification component of our strategy, since it has been widely used in biological applications and directly extracts position specific score matrices. Below we briefly describe this algorithm and propose two ways to integrate it in our classification system.

The MEME algorithm follows an iterative procedure, which applies at each iteration a two-component mixture model to discover one motif of length W. In the two-component model, one component describes the motif (ungapped common subsequences of length W) while the other component models the background information. Multiple motifs can be found by sequentially fitting the two-component model to the set of sequences that remain after removing the sequences containing occurrences of the already identified motifs.

In particular, MEME (Bailey and Elkan, 1994) uses the Expectation Maximization (EM) algorithm (Dempster et al., 1977) to maximize the log-likelihood function of the two-component mixture model, i.e. to estimate the elements of the corresponding position weight matrix[2].

[1]The general relative frequencies of aminoacids indicate the background information in a protein family and can be presented as a probabilistic vector ρ of size $\Omega = 20$.

[2]The model used in our experiments assumes that there are zero or more non-overlapping occurrences of the motif in each sequence of the dataset. Alternative models that can be used are the exactly one occurrence per sequence and the zero or one occurrence per sequence.

Furthermore, MEME provides with a strategy for locating efficient initial parameter values in order to prevent the EM algorithm from getting stuck in local optima (Bailey and Elkan, 1994). The D motif models PWM_j ($j = 1, \ldots, D$) discovered by MEME can be of either fixed or variable length W_j. In our experimental studies both types of motifs will be examined to evaluate the impact of this decision on the performance of the neural classifier.

In order to discover a group of motifs from a multi-class training set of sequences (containing sequences of K classes) two alternative approaches can be followed. The first approach is to apply the MEME algorithm K times, separately to the training sequences of each protein family. Then, putting all the discovered K family profiles together we can form the final group of D motifs. An alternative approach is to apply the motif discovery algorithm only once to the total training set $\mathbf{S}$ ignoring class labels. In this way, we do not allow the algorithm to directly create K protein family profiles, but rather to discover D class-independent motifs.

The advantage of the second approach is the ability of taking into account local similarity measurements in the whole training set, without restricting the search procedure to a single class. Therefore, possible partial homologies among sequences from different families can be defined that may prove helpful for the classification task. On the other hand, a disadvantage of the class-independent approach is that the D discovered motifs may not be equally distributed among the K families. This may result in insufficient modeling of some families, thus leading to performance deterioration. During experiments both motif discovery strategies will be considered and evaluated.

3.3 Construction of a neural classifier

After discovering D motifs and constructing the D-dimensional feature space, the last stage in our methodology is to implement and train a feed-forward neural network that will be able to map the input vectors into the protein classes of interest. A typical network architecture is illustrated in Figure 1. To construct the neural classifier we use the training set $\mathbf{X} = \{\mathbf{x}_i, \mathbf{t}_i\}$, $i = 1, \ldots, N$ consisting of positive examples $\mathbf{x}_i$ from the set of K protein families. The target vector $\mathbf{t}_i$ is a binary vector of size K indicating the class label of input $\mathbf{x}_i$, i.e. $t_{ik} = 1$ if $\mathbf{x}_i$ corresponds to a sequence S_i belonging to class k, and 0 otherwise. The output of the classifier is represented by the K-dimensional vector $\mathbf{y}_i$ where component y_{ik} corresponds to class k. Based on this scheme, the predicted class $h(\mathbf{x}_i)$ of an unlabeled feature vector $\mathbf{x}_i$ corresponding to a query sequence S_i is given by the index of the output node with the largest value y_{ic}, i.e.

$$h(\mathbf{x}_i) = c : \;\; y_{ic} = \max_{1 \le k \le K} y_{ik} \; . \tag{4}$$

Setting a threshold value θ ($\in [0,1]$), we can restrict the classifiers' decision only to those input vectors whose maximum output value surpasses this threshold. In this case we can write:

$$h(\mathbf{x}_i, \theta) = c : \;\; y_{ic} = \max_{1 \le k \le K} y_{ik} \;\; \wedge \;\; y_{ic} \ge \theta \; . \tag{5}$$

Parameter θ can be used to specify the sensitivity of the classifier.

In order to train the neural network we used the Gauss-Newton Bayesian Regularization (GNBR) learning algorithm (Foresse and Hagan, 1997). This algorithm applies Bayesian regularization and implements a Gauss-Newton approximation to the Hessian matrix of the objective function.

In the Bayesian regularization framework the objective function is formulated as the weighted sum of two terms: the sum of the squared errors (E_X) and the sum of squares of the network weights (E_W). Using Bayes' rule, the posterior probability distribution for the weights $\mathbf{w}$ of the network given a training set $\mathbf{X}$ can be written as follows:

$$P(\mathbf{w}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{w})P(\mathbf{w})}{P(\mathbf{X})} \; . \tag{6}$$

By properly choosing the prior distribution $P(\mathbf{w})$ and the likelihood function $P(\mathbf{X}|\mathbf{w})$, we can obtain the following expression (Bishop, 1995; Foresse and Hagan, 1997) for the posterior distribution:

$$P(\mathbf{w}|\mathbf{X}) = \frac{1}{Z_F} \exp(-\beta E_X - \alpha E_W) = \frac{1}{Z_F} \exp(-F(\mathbf{w})), \tag{7}$$

where the Z_F corresponds to the normalizing factor that is independent of the weights.

Maximizing the above posterior distribution is equivalent to minimizing the regularized objective function $F(\mathbf{w})$:

$$F(\mathbf{w}) = \frac{\beta}{2} \sum_{i=1}^{N_X} \{\mathbf{y}_i - \mathbf{t}_i\}^2 + \frac{\alpha}{2} \sum_{j=1}^{N_W} w_j^2 \; , \tag{8}$$

where N_X and N_W represent the number of input vectors and network parameters, respectively. In order to estimate the normalizing factor Z_F a Gaussian approximation can be used for the posterior distribution (MacKay, 1992) as obtained by the Taylor expansion of function $F(\mathbf{w})$ around the minimum value of the posterior, $\mathbf{w}_{MP}$. This gives the following estimation: (Bishop, 1995):

$$Z_F^*(\alpha, \beta) = \exp(-F(\mathbf{w}_{MP}))(2\pi)^{N_W/2} |\mathbf{H}|^{-1/2} \; , \tag{9}$$

where $\mathbf{H}$ corresponds to the Hessian matrix of the regularized objective function and, therefore, optimal values for parameters α and β at the minimum point $\mathbf{w}_{MP}$ can be computed as follows:

$$\hat{\alpha} = \frac{\gamma}{2E_W(\mathbf{w}_{MP})} \text{ and } \; \hat{\beta} = \frac{\gamma N_X}{2E_X(\mathbf{w}_{MP})} \; . \tag{10}$$

The quantity γ represents the effective number of network parameters $\mathbf{w}$ and can be defined using the eigenvalues of H^{-1} as $\gamma = N_W - 2\alpha \mathrm{Tr}\mathbf{H}^{-1}$. In cases where the number of effective parameters is equal to the actual ones ($\gamma \approx N_W$), more hidden units must be added to the network. Furthermore, the GNBR algorithm follows a Gauss-Newton approximation method (Foresse and Hagan, 1997) for calculating the Hessian matrix of $F(\mathbf{w})$ at the minimum point $\mathbf{w}_{MP}$, using the Levenberg-Marquardt optimization algorithm (Bishop, 1995). It must be noted that in our experiments, the best results for the GNBR algorithm were obtained by scaling the network inputs in the range $[-1, 1]$.

4 Experimental results

Several experiments were conducted to evaluate the proposed method. The classification accuracy was measured by counting the sensitivity and specificity rates. In all K-class classification problems, each protein family $\mathcal{S}_k$ $(k = 1, \ldots, K)$ was randomly partitioned into training and test sequences, with the training set being only a small percentage (5 - 10%) of the family dataset. Using the training datasets experiments have been carried out using the MEME algorithm to discover groups of motifs. Two cases were considered: in the first case, the MEME algorithm has been applied separately to each training set providing a group of $D_k = 5$ class-dependent motifs for each family $\mathcal{S}_k$[3]. In the second case the MEME algorithm was applied only once to the total training dataset (ignoring the class labels) to provide a group of $D = 5 \times K$ class-independent motifs.

In any case, the obtained final group of D motifs were used to transform each sequence of the dataset into a dataset with numerical D-dimensional feature vectors, denoted $\mathbf{X}_s$ for the class-dependent case and $\mathbf{X}_g$ for the class-independent case. Furthermore, we also experimented with the effect of the length W of the discovered motifs to the performance of the proposed classifier, by applying the MEME algorithm with either fixed or variable motif length. We selected $W = 20$ for the first case and the range $[10, 30]$ for the second case. In summary, we have considered four distinct cases considering the application of MEME: discovering either class-dependent or class-independent motifs with either fixed or variable motif length. Therefore, for each classification problem four distinct neural classifiers will be constructed and tested.

To evaluate classification performance ROC (Receiver Operating Characteristic) analysis was used. More specifically, we used the ROC_{50} curve which is a plot of the sensitivity as a function of false positives for various decision threshold values until 50 false positives are found.

Table 1 near here

Table 2 near here

For our experimental study three real datasets were selected. In particular we have used

[3]Experiments with greater number of motifs did not yield better classification performance.

protein families from the PROSITE database (Hofmann et al., 1999), which is a large collection of protein families together with their characteristic (deterministic) motifs. Two datasets with $K = 6$ (PROSITE 1) and $K = 7$ (PROSITE 2) classes from the PROSITE database (Hofmann et al., 1999) were selected, summarized in Table 1. Moreover, experiments have also been conducted on a dataset of G-protein coupled receptors (GPCR) (Horn et al., 1998), that is a superfamily of cell membrane proteins. The GPCR database is hierarchically classified into five major classes and their subfamilies (Horn et al., 1998). We studied the problem of classifying subfamilies within the class A, since it dominates the whole GPCR database. As indicated in (Karchin et al., 2002), the difficulty of recognizing GPCR subfamilies arises from the fact that the classification of the subfamilies has been made based on chemical properties rather than sequence homology. Therefore, members from different subfamilies may share strong homology thus making their discrimination hard. Among 15 subfamilies consisting class A, seven of them have been selected in our experimental study described in Table 2. The remaining eight subfamilies are of very small size and it is difficult to construct an effective system for their discrimination. Details of the three datasets (family/subfamily names and their protein ID's) used in our experiments are given in the Appendix.

4.1 Local versus global features

In this series of experiments we assessed the impact of using 2-grams (background features) on the performance of the proposed classification scheme. For a sequence S_i with length L_i we define the feature value g_{iq} for each 2-gram q with respect to this sequence as:

$$g_{iq} = \frac{\mathcal{N}(q|S_i)}{L_i - 1} \ , \tag{11}$$

where $\mathcal{N}(q|S_i)$ denotes the number of occurrence of the 2-gram feature q in the sequence S_i. As it is obvious, the above equation gives the relative frequency of a 2-gram feature in a sequence. In a training set $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ of N sequences we can ignore redundant 2-grams and consider only the N_g features g_{iq} that correspond to the most frequently occurring 2-grams. We select the N_g 2-grams occurring in at least half of the training sequences and by computing the corresponding g_{iq} $(q = 1, \ldots, N_g)$ values for each sequence S_i, we construct the corresponding feature vectors to be fed in the neural classifier.

Table 3 near here

Table 3 presents the dimensionality of the feature spaces obtained using 2-grams and motifs for each dataset used in the experiments. It must be noted that we can further reduce the

dimensionality of the 2-gram feature vectors using standard dimension reduction techniques, such as principal component analysis (PCA).

To assess the impact of the several feature types on the performance of the classification system we have considered five different datasets:

- $\mathbf{X_s}$: D motif-based features separately identified for each family (class-dependent),
- $\mathbf{X_g}$: D motif-based class-independent features,
- $\mathbf{X_s} \cup \mathbf{G}$: D motif-based class-dependent features along with N_g 2-gram features,
- $\mathbf{X_g} \cup \mathbf{G}$: D motif-based class-independent features, along with N_g 2-gram features
- $\mathbf{G}$: N_g 2-gram features.

The neural network architecture had one hidden layer of either 10 (for the cases $\mathbf{X}_s$ and $\mathbf{X}_g$) or 20 nodes for the other three cases.

Figure 2 near here

Figure 2 displays the ROC_{50} curves obtained after training the five neural classifiers in each of the three classification problems respectively. For each problem two different graphs are presented concerning motifs of fixed length ($W = 20$) and of variable length $W \in [10, 30]$. As it is obvious, motif-based features itself constitute an excellent source of information able to generate significant features and lead to the construction of efficient classifiers. In all cases, the neural networks trained by mixed features (e.g. $\text{NN}(\mathbf{X}_s \cup \mathbf{G})$) exhibit lower classification accuracy compared to the corresponding classifier trained with only motif-based features (e.g. $\text{NN}(\mathbf{X}_s)$). Furthermore, the 2-grams features alone (case $\text{NN}(\mathbf{G})$) do not seem to contain significant discriminant information.

Another observation that can be made from the ROC_{50} curves in Figure 2 is related to the performance of the neural classifier with class-dependent motifs (network $\text{NN}(\mathbf{X}_s)$) compared to that obtained with class-independent motifs (network $\text{NN}(\mathbf{X}_g)$). In almost all cases we obtained better classification results with the network $\text{NN}(\mathbf{X}_s)$. One explanation for this behaviour is that, when searching for a specific number D of motifs in the whole training set (ignoring class labels) the algorithm may focus on some of the of families and leave the other families explored only partially. This possibly affects the satisfactory modeling of some families, since the discovered class-independent motifs may not be sufficient for describing

them (only a few individual motifs dedicated to this family). Experiments in the $\mathbf{X}_g$ datasets with MEME have shown that the allocation of motifs in most cases was not equal for all the K families.

Figure 3 near here

An example is shown in Figure 3 that illustrates the constructed feature space of the $\mathbf{X}_s$ and $\mathbf{X}_g$ datasets in the case of the GPCR problem (seven classes), after projecting the 35-dimensional numerical to a two-dimensional space using PCA. It can be observed that in the case of class-dependent motifs the protein classes exhibit less overlap while in the reduced feature space of class-independent motifs there is a significant overlapping among class regions, thus making the discrimination harder. A selection of higher values of D probably would lead to better results for the class-independent case, but would simultaneously result in larger feature spaces or to the overestimation of some families.

4.2 Comparison with other approaches

We have also compared the neural classifier (with class-dependent motif-based features) with two other protein classification methods, namely the MAST homology detection algorithm (Bailey and Gribskov, 1998) and the profile HMMs built using SAM (Hughey and Krogh, 1996). In both MAST and SAM each protein family (or subfamily) is transformed (indirectly or directly) into a probabilistic model-profile and the test sequences are classified using the class of the profile with the best score value.

More specifically, the MAST procedure (Bailey and Gribskov, 1998) initially uses the MEME algorithm to discover groups of motifs separately for each one of the K protein families. For each sequence in the testing set, the MAST algorithm combines the calculated p-values and estimates the significance of the observed match (called E-value) of the sequence to each of the K groups of motifs[4]. Then the query sequence is assigned to the class with the minimum E-value. The SAM method (Hughey and Krogh, 1996) works in a similar way by building an HMM for each one of the K protein families (or subfamilies) instead of discovering groups of motifs[5].

Figure 4 provides comparative results from the application of the proposed neural classifier, MAST and SAM to the three datasets. We have created five ROC curves for each method (number of false positives versus sensitivity for several threshold values) until 25 false positives

[4]We use the meme and mast commands from the available MEME package v.3.0.4.

[5]We used the buildmodel and hmmscore commands from the available SAM package v.3.3.1.

were found (ROC_{25}). The performance of the neural classifier and MAST was given by two curves respectively[6] concerning motifs of fixed ($W = 20$) and variable length ($W = [10, 30]$), while the last one corresponds to SAM performance. In the case of MAST and SAM methods, ROC curves were obtained by setting several E-value thresholds. When the lowest estimated E-value for a query sequence was greater than the threshold then the test sequence was considered unclassified.

Figure 4 near here

The superior classification of the proposed neural approach is obvious from the plotted curves in all problems, offering greater sensitivity rates with perfect specificity (zero false positives). For the GPCR dataset which is more difficult to discriminate, the classification improvement is more clear: a sensitivity rate of 99.30% was measured with only 11 false positives, while the corresponding results for MAST and SAM are $(95.76\%, 25)$ and $(95.38\%, 25)$, respectively. It is also important to stress the higher accuracy that the neural scheme achieves compared with the MAST (dot lines). Although these two methods use the same groups of motifs, our method seems to offer a more efficient scheme for combining the motif match scores compared to the combination of their p-values as suggested by MAST. In addition, the neural classifier achieves less false positives with higher sensitivity rates in all datasets concerning either fixed or variable motif length. Again the improvement is more clear in the plots corresponding to the GPCR dataset.

Regarding more carefully the three selected datasets, they can be considered as three different types of protein sequence classification problems. In particular, the PROSITE 1 dataset consists of diverse protein families in the sense that their corresponding PROSITE motifs are not very specific (such as in the case of PS00030 and PS00198) and they can be found in sequences from a large number of protein families. Hence, this application can be seen as a diverse protein family recognition problem. On the other hand, the PROSITE 2 dataset consists of protein families with more specific PROSITE motifs that can be distinguished more easily. Finally, the third dataset, GPCR, is related to the recognition of protein subfamilies within a broader protein family domain sharing strong homology.

In all the above three types of protein sequences classification problems our approach has shown a superior classification performance providing better results in comparison with the two other approaches. As illustrated in Figure 4, the SAM method seems to be unsuccessful

[6]The curves for the neural classifier performance were the best plots from the corresponding ROC_{50} diagrams in Figure 2.

in recognizing diverse protein families (PROSITE 1 case) and the obtained classification rate was low (the individual classification error for each diverse family was about 50%). On the other hand, the performance of the MAST method was lower in the case of the GPCR subfamily recognition problem where sequences from different subfamilies share strong homology. Finally, in the case of recognizing simple protein families (PROSITE 2 dataset) all the three approaches provide similar classification rates, with the proposed neural scheme offering slightly better results.

5 Conclusions

In this paper we have presented a neural network approach for the classification of protein sequences. The proposed methodology is motivated by the principle that in biological sequence analysis motifs can provide major diagnostic features for determining the class label of the unknown sequences. The method is implemented in two steps, where a pre-processing step (based on the MEME algorithm) is initially applied for discovering a group of probabilistic motifs appearing in the sequences. We have suggested and evaluated two alternative ways for motif discovery in a set of K-class sequences depending on whether the class labels are taken into account or not. Using the discovered motifs a numerical feature vector is generated for each sequence by computing the matching score of the sequence to each motif. At the second stage of the proposed method, the extracted feature vectors are used as inputs to a feed-forward neural network trained using the Gauss-Newton Bayesian Regularization algorithm that provides the class label of a sequence.

Experiments were conducted on real datasets (using very small training sets) and comparisons were made with the MAST and SAM probabilistic methods. ROC curves were used as a performance indicator and the experimental results clearly illustrate the superiority of the proposed neural system. In addition we have shown that background features do not constitute a useful source of information for the classification task since they do not lead to performance improvement.

In what concerns our future work, more extensive experiments could be conducted to assess the performance of the method on specific protein superfamilies of important biological functions, as was the case with the GPCR dataset. Also, alternative methods could be implemented and tested, both in the classification stage (mixture models, SVMs etc) and in the motif discovery stage.

6 Acknowledgments

The authors wish to thank the anonymous referees for the valuable comments that have improved the quality and the presentation of this work.

References

Almeida, J. S. and Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, 3(6).

Altshul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment tool. *Journal of Molecular Biology*, 215:403–410.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, Menlo Park, California. AAAI Press.

Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14:48–54.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Univ. Press Inc., New York.

Brāzma, A., Jonasses, I., Eidhammer, I., and Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):277–303.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). *Atlas of protein sequence and structure*. Natl. Biomed. Res. Found., Washington, Dc., Vol. 5.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38.

Durbin, R., Eddy, S., Krough, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acid*. Cambridge University Press, New York, NY.

Foresse, F. D. and Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian regularization. In *Proceedings of the 1997 International Joint Conference on Neural Network*, pages 1930–1935.

Henikoff, S. S. and Henikoff, J. G. (1994). Protein family classification based on searching a database of blocks. *Genomics*, 19:97–107.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7/8):563–577.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucleic Acids Research*, 27:215–219.

Horn, F., Weare, J., Beukers, M. W., Hörsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F., and Vriend, G. (1998). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Research*, 21(1):227–281.

Hughey, R. and Krogh, A. (1996). Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12(2):95–107.

Jaakkola, T., Diekhans, M., and Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1,2):95–114.

Karchin, R., Karplus, K., and Haussler, D. (2002). Classifying G-protein coupled receptors with support vector mahcines. *Bioinformatics*, 18(1):147–159.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwland, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 226:208–214.

Logan, B., Moreno, P., Suzek, B., Weng, Z., and Kasif, S. (2001). A study of remote homology detection. Technical Report CRL 2001/05, Cambridge Research Laboratory.

Ma, Q. and Wang, J. T. L. (2000). Application of Bayesian neural networks to protein sequence classification. In *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 305–309, Boston, MA, USA.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4:415–447.

Rigoutsos, I., Floratos, A., Parida, L., Gao, Y., and Platt, D. (2000). The Emergency of Pattern Discovery Techniques in Computational Biology. *Metabolic Engineering*, 2:159–177.

Vapnik, V. N. (1979). *Estimation of Dependencies Based on Empirical Data*. Nauka, Birmingham, AL.

Wang, J. T. L., Ma, Q., Shasha, D., and Wu, C. H. (2001). New techniques for extracting features from protein sequences. *IBM: Systems Journal*, 40(2):426–441.

Wu, C. H., Zhap, S., Chen, H. L., Lo, C. J., and McLarty, J. (1996). Motif identification neural design for rapid and sensitive protein family search. *CABIOS*, 12(2):109–118.

Problem: PROSITE 1 ($K = 6$)			*Problem: PROSITE 2 ($K = 7$)*		
PROSITE family	*Positive data*	*Training set (avg length of seqs)*	*PROSITE family*	*Positive data*	*Training set (avg length of seqs)*
PS00030	302	20 (370)	PS00070	129	15 (558)
PS00038	289	20 (359)	PS00077	155	15 (502)
PS00061	317	20 (299)	PS00118	168	15 (127)
PS00198	300	20 (284)	PS00180	123	15 (408)
PS00211	574	30 (478)	PS00215	123	15 (321)
PS00301	386	20 (517)	PS00217	148	15 (490)
			PS00338	173	15 (212)

Table 1: The two PROSITE families used in the experimental study.

Problem: GPCR ($K = 7$)		
GPCR Class A subfamily	*Positive data*	*Training set (avg length of seqs)*
Amine	306	20 (485)
Peptide	654	30 (383)
Hormone	43	10 (378)
Rhodopsin	270	20 (358)
Olfactory	325	20 (317)
Prostanoid	43	10 (721)
Nucleotide-like	58	10 (348)

Table 2: Seven families from the GPCR class A used in the experimental study.

Problem	N_g *2-gram features*	D *motif-based features*
PROSITE 1	174	$5 \times 6 = 30$
PROSITE 2	285	$5 \times 7 = 35$
GPCR	152	$5 \times 7 = 35$

Table 3: The number of the extracted motif-based (D) and 2-gram (N_g) features that corresponds to each dataset.

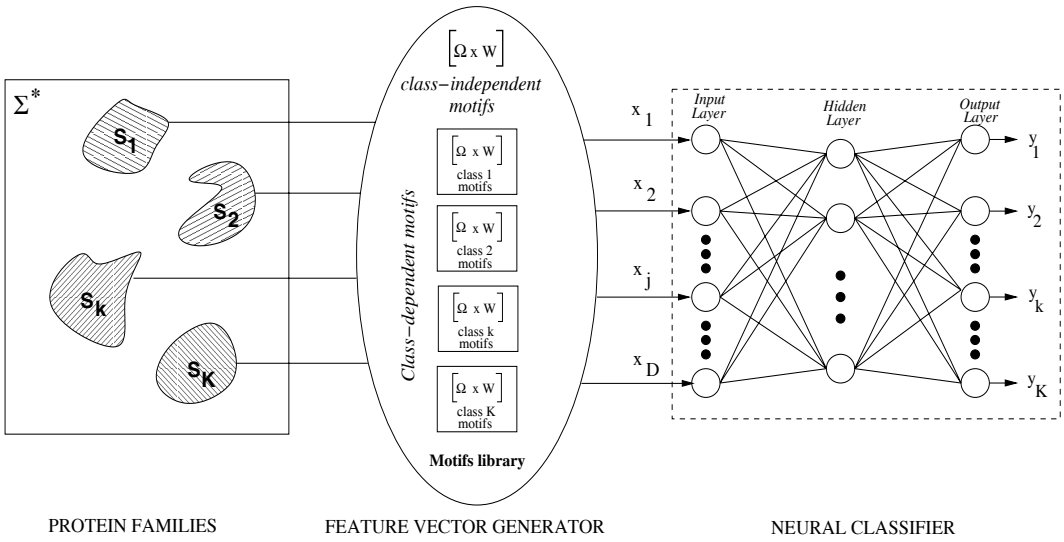

Figure 1: The architecture of the proposed classification scheme.

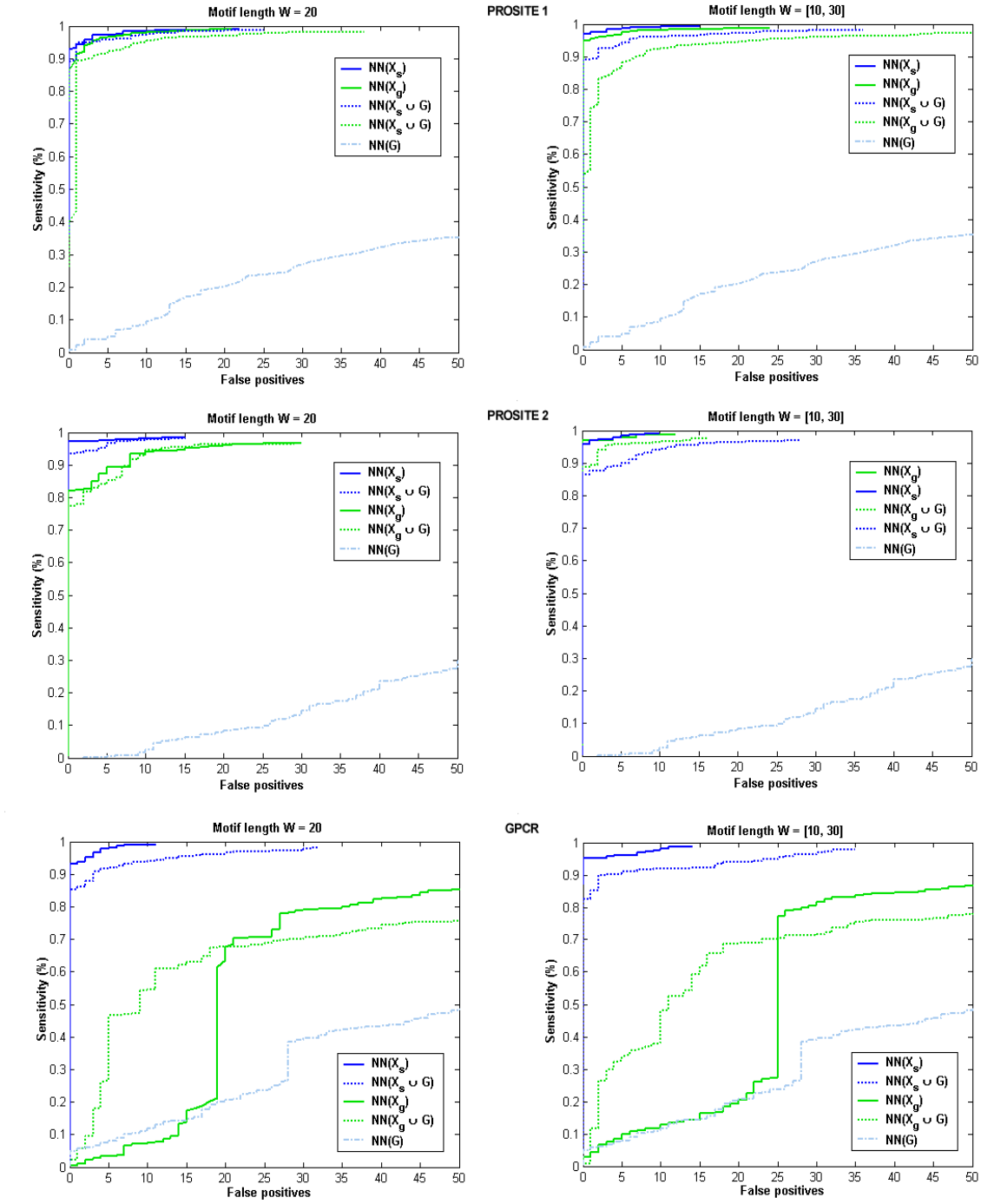

Figure 2: ROC_{50} curves illustrating the performance of the neural classifier on the three datasets using the five different feature vectors.

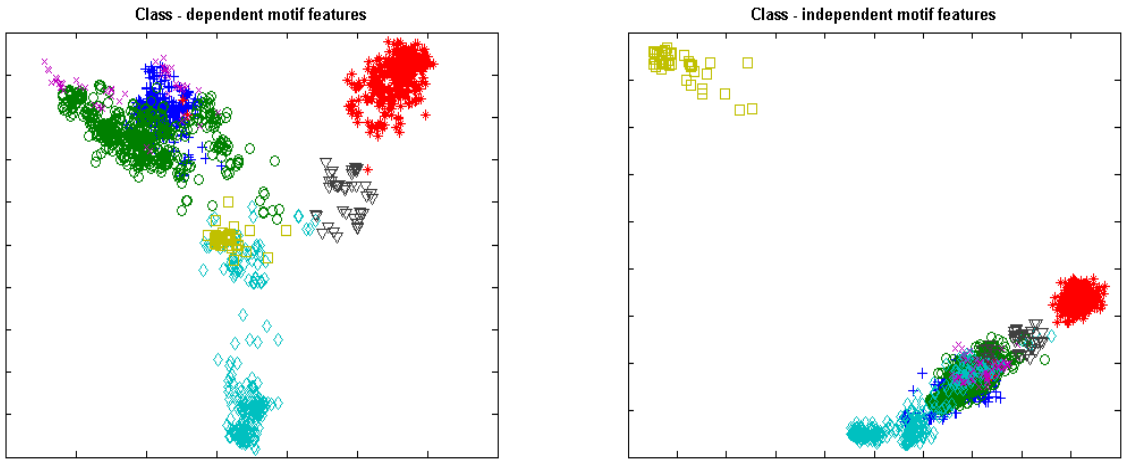

Figure 3: The seven class regions in the GPCR dataset in the case of class-dependent and class-independent features. The data have been projected in two dimensions using PCA.

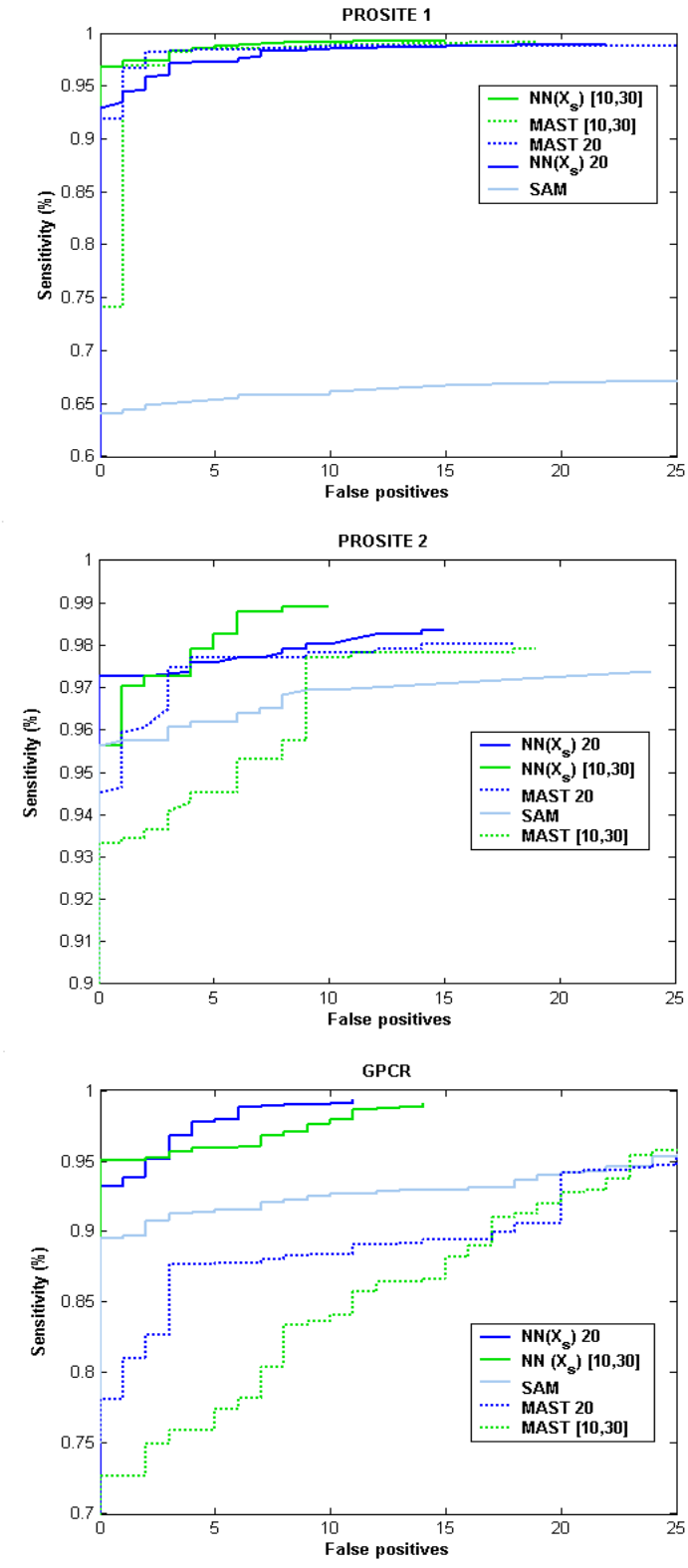

Figure 4: ROC_{25} curves for the three methods (neural (NN), MAST and SAM) on the three datasets.

7 Appendix: Datasets

In the next tables proteins with bold ID's correspond to the training examples and the rest of them to the test set.

Table 4: Description of the PROSITE 1 dataset.

Family	*Protein ID's*
PS00030	**CB20-HUMAN GAR2-SCHPO HRB1-YEAST HS49-YEAST IF34-MOUSE NAB4-YEAST PAB3-ARATH RB27-DROME RN15-YEAST ROA1-MOUSE ROC3-NICSY RU17-HUMAN RU17-YEAST RU1A-DROME RU2B-HUMAN SFPQ-HUMAN U2AF-CAEEL U2AF-HUMAN U2AG-HUMAN K682-HUMAN** A2BP-HUMAN A2BP-MOUSE ARP2-PLAFA ROAA-MOUSE CAZ-DROME CB20-XENLA CG79-HUMAN PM14-MOUSE CIRP-HUMAN CIRP-MOUSE CIRP-XENLA CPO-DROME CST2-HUMAN CSX1-SCHPO CTF1-SCHPO CUG1-HUMAN CUG1-MOUSE CWF5-SCHPO CYPE-DROME CYPE-HUMAN CYPE-MOUSE D111-ARATH ELAV-DROME ELAV-DROVI ELV1-HUMAN ELV1-MOUSE ELV2-HUMAN ELV2-MOUSE ELV3-HUMAN ELV3-MOUSE ELV4-HUMAN ELV4-RAT ELV4-RAT EWS-HUMAN EWS-MOUSE FCA-ARATH FUS-BOVIN FUS-HUMAN FUS-MOUSE G3B2-HUMAN G3B2-MOUSE G3BP-HUMAN G3BP-MOUSE G3BP-SCHPO GBP2-YEAST GR10-BRANA GRF1-HUMAN GRP1-SINAL GRP1-SORBI GRP2-SINAL GRP2-SORBI GRP7-ARATH GRP8-ARATH GRPA-MAIZE GRP-DAUCA IF34-CAEEL IF34-HUMAN IF34-SCHPO IF34-YEAST IF39-HUMAN IF39-SCHPO IF39-TOBAC IF39-YEAST IF4B-HUMAN IF4B-YEAST IF4H-HUMAN IF4H-MOUSE JSN1-YEAST LAA-XENLA LAB-XENLA LAH1-SCHPO LAH1-YEAST LA-AEDAL LA-BOVIN LA-DROME LA-HUMAN LA-MOUSE LA-RABIT LA-RAT MAT3-HUMAN MAT3-RAT MEI2-SCHPO MLO3-SCHPO MODU-DROME MSSP-HUMAN NAB3-YEAST NAM8-YEAST NGR1-YEAST NONA-DROME NOP3-YEAST NOP4-YEAST NOP8-YEAST NOT4-YEAST NR54-HUMAN NRD1-SCHPO NRD1-YEAST NRP1-YEAST NSR1-YEAST NUCL-CHICK NUCL-HUMAN NUCL-MESAU NUCL-MOUSE NUCL-RAT NUCL-XENLA PAB1-HUMAN PAB1-MOUSE PAB2-ARATH PAB2-HUMAN PAB4-HUMAN PAB5-ARATH PABP-DROME PABP-SCHPO PAB1-XENLA PABP-YEAST PABX-ARATH PES4-YEAST PRT1-PICAN PTB-HUMAN PTB-MOUSE PTB-PIG PTB-RAT PUB1-YEAST RB56-HUMAN RB87-DROME RB97-DROME RBM3-HUMAN RBM3-MOUSE RBM5-HUMAN RBM6-HUMAN RBM7-HUMAN RB8A-HUMAN RBM9-HUMAN RBMA-HUMAN RBMA-RAT RBMB-HUMAN RBMS-CHICK RBMS-HUMAN RBMS-MOUSE RBMS-XENLA RBP1-DROME RBPA-SYNY3 RDP-HUMAN RDP-MOUSE RN24-SCHPO RNP1-YEAST RO21-XENLA RO22-XENLA RO31-XENLA RO32-XENLA ROA0-HUMAN ROA1-BOVIN ROA1-DROME ROA1-HUMAN ROA1-MACMU ROA1-RAT ROA1-SCHAM ROA1-XENLA ROA2-HUMAN ROA2-MOUSE ROA3-HUMAN ROAB-ARTSA ROC1-ARATH ROC1-NICPL ROC1-NICSY ROC1-SPIOL ROC2-ARATH ROC2-NICPL ROC2-NICSY ROC3-ARATH ROC4-NICSY ROC5-NICSY ROC-HUMAN ROC-RAT ROC-XENLA ROD-HUMAN ROD-RAT ROF-HUMAN ROG-HUMAN ROG-MOUSE ROH1-HUMAN ROH2-HUMAN ROL-HUMAN ROM-HUMAN ROR-HUMAN ROU2-HUMAN RS31-ARATH RS40-ARATH RS41-ARATH RT19-ARATH RU17-ARATH RU17-DROME RU17-MOUSE RU17-XENLA RU1A-HUMAN RU1A-XENLA RU1A-YEAST RX21-DROME S3B4-HUMAN SCE3-SCHPO SFR1-ARATH SFR1-HUMAN SFR2-CAEEL SFR2-CHICK SFR2-HUMAN SFR2-MOUSE SFR3-HUMAN SFR4-HUMAN SFR5-HUMAN SFR5-MOUSE SFR5-RAT SFR6-HUMAN SFR6-RABIT SFR7-HUMAN SFR9-HUMAN SFRB-HUMAN SQD-DROME SR55-DROME SRA4-HUMAN SRA4-RAT SRP1-SCHPO SSB1-YEAST SXL-CERCA SXL-CHRRU SXL-DROME SXL-DROSU SXL-MEGSC SXL-MUSDO TIA1-HUMAN TIA1-MOUSE TIAR-HUMAN TIAR-MOUSE TR2A-HUMAN TRA2-DROME TRA2-DROVI U2AF-CAEBR U2AF-DROME U2AF-MOUSE U2AF-SCHPO U2AG-DROME U2AG-MOUSE U2AG-SCHPO U2R1-HUMAN U2R1-MOUSE U2R2-HUMAN U2R2-MOUSE WHI3-YEAST XMS2-DROME Y117-HUMAN YAC4-SCHPO YAG3-SCHPO YAS9-SCHPO YBF1-YEAST YBLC-SCHPO YCQ9-SCHPO YD3D-SCHPO YDB2-SCHPO YDC1-SCHPO CWF2-SCHPO YDR6-SCHPO YFK2-YEAST YG5B-YEAST YHC4-YEAST YHH5-YEAST YIS1-YEAST YIS5-YEAST YIS9-YEAST YKV4-YEAST YLF1-CAEEL YML7-YEAST YN26-YEAST YN8T-YEAST YNL0-YEAST YNR5-YEAST YP68-YEAST YP85-CAEEL YQO1-CAEEL YQO4-CAEEL YQOA-CAEEL YQOC-CAEEL YRA1-YEAST YSO7-CAEEL YSX2-CAEEL

Table 4: Description of the PROSITE 1 dataset.

Family	*Protein ID's*
PS00038	**AHR-RAT ARRS-MAIZE CBF1-YEAST DA-DROME ESM7-DROME HEN2-MOUSE HES3-MOUSE MAD4-MOUSE MITF-RAT MXI1-BRARE MYC-HYLLA MYF6-HUMAN MYOD-BRARE NDF1-MESAU SIM1-MOUSE TAL2-MOUSE TAL-HUMAN TF21-MOUSE TWS1-HUMAN USF2-RAT** AHR-HUMAN AHR-MOUSE ARLC-MAIZE ARN2-HUMAN ARN2-MOUSE ARNT-DROME ARNT-HUMAN ARNT-MOUSE ARNT-RABIT ARNT-RAT ASC1-HUMAN ASC1-MOUSE ASC1-RAT ASC1-XENLA ASC2-HUMAN ASC2-MOUSE ASC2-RAT AST3-DROME AST4-DROME AST5-DROME AST8-DROME ATH1-HUMAN ATH1-MOUSE NDF6-MOUSE NDF4-MOUSE NDF4-XENLA NGN2-MOUSE NGN3-MOUSE ATO-DROME BET3-MESAU BMAL-HUMAN CBF1-KLULA CLOC-DROME CLOC-HUMAN CLOC-MOUSE CYCL-DROME DEI-DROME DPN-DROME EMC-DROME ESC1-SCHPO ESM3-DROME ESM5-DROME ESM8-DROHY ESM8-DROME ESMB-DROME ESMC-DROME ESMD-DROME HAIR-DROME HAIR-DROVI HAN1-CHICK HAN1-HUMAN HAN1-MOUSE HAN1-RAT HAN1-SHEEP HAN1-XENLA HAN2-BRARE HAN2-CHICK HAN2-HUMAN HAN2-MOUSE HAN2-XENLA HEN1-HUMAN HEN1-MOUSE HEN2-HUMAN HES1-CHICK HES1-HUMAN HES1-MOUSE HES1-RAT HES2-HUMAN HES2-MOUSE HES2-RAT HES3-RAT HES5-MOUSE HES5-RAT HEY1-CANFA HEY1-HUMAN HEY1-MOUSE HIFA-HUMAN HIFA-MOUSE HTF4-CHICK HTF4-HUMAN HTF4-MESAU HTF4-MOUSE HTF4-PAPHA HTF4-RAT HTF4-XENLA ID1-HUMAN ID1-MOUSE ID1-RAT ID2-HUMAN ID2-MOUSE ID2-RAT ID3-HUMAN ID3-MOUSE ID3-RAT ID4-HUMAN ID4-MOUSE INO2-YEAST INO4-YEAST ITF2-CHICK ITF2-CHICK ITF2-HUMAN ITF2-MOUSE ITF2-RAT LYL1-HUMAN LYL1-MOUSE MAD4-HUMAN MAD-HUMAN MAD-MOUSE MAX-BRARE MAX-CHICK MAX-HUMAN MAX-MOUSE MAX-RAT MAX-XENLA MF25-XENLA MITF-HUMAN MITF-RAT MLX-HUMAN MLX-MOUSE MNT-HUMAN MNT-MOUSE MUSC-HUMAN MUSC-MOUSE MXI1-HUMAN MXI1-MOUSE MXI1-RAT MYC1-CYPCA MYC1-XENLA MYC2-CYPCA MYC2-MARMO MYC2-SPEBE MYC2-XENLA MYCB-RAT MYCL-HUMAN MYCL-MOUSE MYCL-XENLA MYCM-HUMAN MYCM-XENLA MYCN-CHICK MYCN-HUMAN MYCN-MARMO MYCN-MOUSE MYCN-RAT MYCN-SERCA MYCN-XENLA MYCS-RAT MYC-ASTVU MYC-AVIM2 MYC-AVIMC MYC-AVIMD MYC-AVIME MYC-AVIOK MYC-BRARE MYC-CALJA MYC-CANFA MYC-CARAU MYC-CHICK MYC-FELCA MYC-FLVTT MYC-HUMAN MYC-MARMO MYC-MOUSE MYC-ONCMY MYC-PANTR MYC-PIG MYC-RAT MYC-SHEEP MYF5-BOVIN MYF5-CHICK MYF5-COTJA MYF5-HUMAN MYF5-MOUSE MYF5-NOTVI MYF5-XENLA MYF6-CHICK MYF6-MOUSE MYF6-RAT MYF6-XENLA MYO1-ONCMY MYO2-ONCMY MYOD-CAEBR MYOD-CAEEL MYOD-CHICK MYOD-COTJA MYOD-DROME MYOD-HUMAN MYOD-MOUSE MYOD-PIG MYOD-RAT MYOD-SHEEP MYOD-XENLA MYOG-CHICK MYOG-COTJA MYOG-HUMAN MYOG-MOUSE MYOG-PIG MYOG-RAT NDF1-CHICK NDF1-HUMAN NDF1-MOUSE NDF1-RAT NDF1-XENLA NDF2-HUMAN NDF2-MOUSE NDF2-RAT NGN1-HUMAN NGN1-MOUSE NGN1-RAT NDFM-CHICK NPA1-HUMAN NPA1-MOUSE NPA2-HUMAN NPA2-MOUSE NUC1-NEUCR PAS1-HUMAN PAS1-MOUSE PHO4-YEAST OLG2-HUMAN RTG1-YEAST RTG3-YEAST SCL-CHICK SIM1-HUMAN SIM2-HUMAN SIM2-MOUSE SIMA-DROME SIM-DROME SRE1-CRIGR SRE1-HUMAN SRE1-MOUSE SRE1-PIG SRE1-RAT SRE2-CRIGR SRE2-HUMAN SUM1-LYTVA SYFA-DROME TAL2-HUMAN TAL-MOUSE TAP4-HUMAN TAP-DROME TF21-HUMAN TFE2-HUMAN TFE2-MESAU TFE2-MOUSE TFE2-RAT TFE2-XENLA TFE3-HUMAN TFE3-MOUSE TFEB-HUMAN TFEB-MOUSE TRH-DROME TWST-DROME TWS1-MOUSE TWST-XENLA TYE7-YEAST USF1-HUMAN USF1-MOUSE USF1-RABIT USF1-XENBO USF2-HUMAN USF2-MOUSE USF-STRPU WS14-HUMAN WS14-MOUSE YAWC-SCHPO YLB7-CAEEL YMH7-CAEEL YNP2-CAEEL YRY3-CAEEL TWST-CAEEL
PS00061	**2BHD-STREX ADH2-DROMN ADH-DROMA ADH-DROMM ADH-DROSL DECR-RAT DHB7-RAT DHGA-BACME DHG-BACSU DHI2-RABIT DHI2-SHEEP DHK2-STRVN ENTA-ECOLI MAS1-AGRT9 PGDH-HUMAN Y019-THEMA YAEB-SCHPO YF43-MYCTU YWC4-CAEEL OXIR-STRAT** 25KD-SARPE 3BHD-COMTE ACT3-STRCO ADH1-CERCA ADH1-DROHY ADH1-DROMN ADH1-DROMO ADH1-DROMT ADH1-DROMU ADH1-DRONA ADH2-CERCA ADH2-DROAR ADH2-DROBU ADH2-DROHY ADH2-DROMO ADH2-DROMU ADH-DROMY ADH2-DROWH ADHR-DROAM ADHR-DROER ADHR-DROGU ADHR-DROIM ADHR-DROLE ADHR-DROMA ADHR-DROMD ADHR-DROME ADHR-DROPE ADHR-DROPS ADHR-DROSU ADHR-DROTE ADHR-DROYA ADH-BACOL ADH-DROAD ADH-DROAF ADH-DROAM ADH-DROBO ADH-DROCR ADH-DRODI ADH-DROER ADH-DROFL ADH-DROGR ADH-DROGU ADH-DROHA ADH-DROHE ADH-DROIM ADH-DROLA ADH-DROLE ADH-DROMD ADH-DROME ADH-DRONI ADH-DROOR ADH-DROPE ADH-DROPI ADH-DROPL ADH-DROPS ADH-DROSI ADH-DROSU ADH-DROTE ADH-DROTS ADH-DROWI ADH-DROYA ADH-SCAAL ADH-ZAPTU ARDH-CANAL ARDH-CANTR ARDH-PICST BA71-EUBSP BA72-EUBSP BDHA-ALCEU BDHA-RHIME BDH-BOVIN BDH-HUMAN BDH-RAT BEND-ACICA BLI4-NEUCR TODD-PSEPU BPHB-BURCE BPHB-COMTE BPHB-PSEPS BPHB-PSEPU BPHB-PSES1 BPHB-RHOGO BUDC-KLEPN BUDC-KLETE CBR2-CAEEL CBR2-MOUSE CBR2-PIG CMTB-PSEPU DECR-HUMAN DHB1-HUMAN DHB1-MOUSE DHB1-RAT DHB2-HUMAN DHB2-MOUSE DHB2-RAT DHB3-HUMAN DHB3-MOUSE DHB3-RAT DHB4-HUMAN DHB4-MOUSE DHB4-RAT DHB7-HUMAN DHB7-MOUSE DHB8-CALJA DHB8-HUMAN DHB8-MOUSE DHBA-BACSU DHBK-MOUSE DHBV-CAEEL DHBW-CAEEL DHBX-ANAPL DHBX-CAEEL DHBY-CAEEL DHC3-HUMAN DHCA-HUMAN DHCA-MOUSE DHCA-RABIT DHCA-RAT DHG1-BACME DHG2-BACME DHG2-BACSU DHG3-BACME DHG4-BACME DHGB-BACME DHG-BACME DHI1-HUMAN DHI1-MOUSE DHI1-RAT DHI1-SAISC DHI1-SHEEP DHI2-BOVIN DHI2-HUMAN DHI2-MOUSE DHI2-RAT DHK1-STRVN DHKR-STRCM DHMA-FLAS1 DHPR-HUMAN DHPR-RAT DHSO-RHOSH DIDH-COMTE DIDH-PSESP DLTE-BACSU EPHD-MYCTU FABG-ACTAC FABG-AQUAE FABG-ARATH FABG-BACSU FABG-BRANA FABG-BUCAI FABG-CHLMU FABG-CHLPN FABG-CHLTR FABG-CUPLA FABG-ECOLI FABG-HAEIN FABG-MYCAV FABG-MYCSM FABG-MYCTU FABG-PERAE FABG-PSEAE FABG-RICPR FABG-SALTY FABG-VIBCH FABG-VIBHA FAG1-SYNY3 FAG2-SYNY3 FBP2-DROME FIXR-BRAJA FOX2-CANTR FOX2-NEUCR FOX2-YEAST FVT1-HUMAN GNO-GLUOX GS39-BACSU HCAB-ECOLI HCD2-BOVIN HCD2-DROME HCD2-HUMAN HCD2-MOUSE HCD2-RAT HDHA-CLOSO HDHA-ECOLI DHS2-HUMAN HETN-ANASP IDNO-ECOLI KDUD-BACSU KDUD-ECOLI KDUD-ERWCH L767-CAEEL LIGD-PSEPA LINC-PSEPA LINX-PSEPA MOAE-KLEAE MS11-AGRRH MS12-AGRRH MTDH-UROFA NAHB-PSEPU NODG-AZOBR NODG-RHIME NODG-RHIS3 NOG4-RHIME NOR1-ASPPA OXIR-STRLI PHAB-ACISP PHAB-PARDE PHBB-ALCEU PHBB-CHRVI PHBB-RHIME PHBB-ZOORA PTMA-CAMCO PTR1-LEIMA PTR1-LEITA RDH1-BOVIN RDH1-HUMAN RHLG-PSEAE RIDH-KLEAE ROH1-RAT ROH2-RAT ROH3-RAT SDR1-PICAB SORD-KLEPN SOU1-CANAL SOU2-CANAL SP19-YEAST SRLD-ECOLI STCE-EMENI STCU-EMENI T4HR-MAGGR TODD-PSEPU TRN1-DATST TRN2-DATST TRN2-HYONI TRNH-DATST TS2-MAIZE UCPA-ECOLI UCPA-SALTY VDLC-HELPJ VDLC-HELPY VER1-ASPPA VIBA-VIBCH XYLL-PSEPU Y048-HAEIN Y1J1-MYCTU Y325-THEMA Y432-LISMO Y452-LISIN Y484-MYCTU Y4EK-RHISN Y4EL-RHISN Y4LA-RHISN Y4MP-RHISN Y4VI-RHISN Y945-MYCTU YAY8-SCHPO YB09-YEAST YB45-SCHPO YBBO-ECOLI YCIK-ECOLI YL46-BRAJA YD1F-SCHPO YD50-MYCTU YDFG-BACNO YDFG-ECOLI YDFG-HAEIN YDFG-SALTY YDGB-ECOLI YGCW-ECOLI YGFF-ECOLI YGHA-ECOLI YHDF-BACSU YHXC-BACSU YHXD-BACSU YIM4-YEAST YIV5-YEAST YIV6-YEAST YJGI-ECOLI YK02-MYCTU YK73-MYCTU YKF5-YEAST YKUF-BACSU YM71-YEAST YMEC-METEX YMP3-STRCO YOHF-ECOLI YOXD-BACSU YQJQ-BACSU YURA-MYXXA YUSZ-BACSU YUXG-BACSU YV06-PSEAE YVX3-CAEEL YWFD-BACSU YWFH-BACSU YXBG-BACSU YXEK-CAEEL YXJF-BACSU

Table 4: Description of the PROSITE 1 dataset.

Family	*Protein ID's*
PS00198	**DHSB-CYACA DHSB-PARDE DHSB-RICCN FER3-PLEBO FER-ALIAC FER-CLOST FIXG-RHIME FIXX-BRAJA HMC6-DESVH MAUM-METEX NIFJ-ECOLI NUIC-ARATH NUIM-NEUCR PORD-METJA PSAC-ORYSA RNFB-PASMU RNFC-ECO57 Y208-METJA YD49-METJA YFHL-ECOLI** AEGA-ECOLI ASRA-SALTY ASRC-SALTY COOF-RHORU DCA1-METMA DCA2-METMA DCMA-METJA DCMA-METSO DCMA-METTE DCMA-METTH DCMG-METTE DHSB-SCHPO DHSB-BACSU DHSB-CAEEL DHSB-CHOCR DHSB-COXBU DHSB-DROME DHSB-ECOLI DHSB-HUMAN DHSB-MYCGR DHSB-PORPU DHSB-RAT DHSB-RECAM DHSB-RICPR DHSB-SCHPO DHSB-USTMA DHSB-YEAST DMSB-ECOLI DMSB-HAEIN DPYD-BOVIN DPYD-CAEEL DPYD-HUMAN DPYD-PIG DSRB-ARCFU DSVB-DESGI DSVB-DESVH FDHB-METFO FDHB-METJA FDHB-METTF FDHB-WOLSU FDNH-ECOLI FDOH-ECOLI FDXH-HAEIN FDXN-ANASP FDXN-ANAVA FDXN-AZOCH FDXN-BRAJA FDXN-RHILT FDXN-RHIME FDXN-RHISN FDXN-RHOCA FER1-AZOVI FER1-CAUCR FER1-CHLLI FER1-DESAF FER1-DESDN FER1-DESVM FER1-METJA FER1-RHOPA FER1-RHORU FER1-SULTO FER2-CHLLI FER2-DESDN FER2-DESVM FER2-METJA FER2-RHOCA FER2-RHORU FER2-SULTO FER2-THEAC FER3-ANASP FER3-ANAVA FER3-DESAF FER3-METJA FER3-RHISN FER3-RHOCA FER4-METJA FER5-METJA FER6-METJA FER7-METJA FER8-METJA FERN-AZOVI FERV-AZOVI FER-ACIAM FER-BACSC FER-BACST FER-BACSU FER-BACTH FER-BUTME FER-CHLLT FER-CHRVI FER-CLOAC FER-CLOBU FER-CLOPA FER-CLOPE FER-CLOSP FER-CLOTM FER-CLOTS FER-DESGI FER-ENTHI FER-MEGEL FER-METBA FER-METTE FER-METTL FER-MOOTH FER-MYCSM FER-MYCTU FER-PEPAS FER-PSEPK FER-PSEST FER-PYRAB FER-PYRFU FER-PYRIS FER-RICPR FER-SACER FER-STRGR FER-SULAC FER-THEAC FER-THELI FER-THEMA FER-THETH FIXX-AZOCA FIXX-AZOVI FIXX-ECOLI FIXX-RHILE FIXX-RHILP FIXX-RHILT FIXX-RHIME FIXX-RHISN FPRB-MYCLE FPRB-MYCTU FRD1-AQUAE FRD2-AQUAE FRDB-ECOLI FRDB-HAEIN FRDB-HELPJ FRDB-HELPY FRDB-MYCTU FRDB-PROVU FRDB-WOLSU FRHG-METJA FRHG-METTH FRHG-METVO GLCF-ECOLI GLPC-ECOLI GLPC-HAEIN HMC2-DESVH HYBA-ECOLI HYCB-ECOLI HYCF-ECOLI HYDN-ECOLI HYFA-ECOLI HYFH-ECOLI IORA-ARCFU IORA-METTH IORA-PYRAB IORA-PYRHO IORA-PYRKO MAUM-METFL MAUM-METME MAUM-PARDE MAUN-METEX MAUN-METFL MAUN-PARDE NAPF-ECOLI NAPF-HAEIN NAPG-ECOLI NAPG-HAEIN NAPH-ECOLI NAPH-HAEIN NIFJ-ANASP NIFJ-ENTAG NIFJ-KLEPN NIFJ-RHORU NIFJ-SYNY3 NQO9-PARDE NQO9-THETH NRFC-ECOLI NRFC-HAEIN NUG2-RHIME NUIC-MAIZE NUIC-MARPO NUIC-MESVI NUIC-ORYSA NUIC-PLEBO NUIC-SPIOL NUIC-SYNY3 NUIC-TOBAC NUIC-WHEAT NUIM-ARATH NUIM-BOVIN NUIM-CAEEL NUIM-HUMAN NUIM-RECAM NUIM-SOLTU NUIM-TOBAC NUIM-TRYBB NUOI-BUCAI NUOI-ECOLI NUOI-MYCTU NUOI-RHOCA NUOI-RICCN NUOI-RICPR PHF1-CLOPA PHFL-DESVH PHFL-DESVO PHSB-SALTY PORD-METTH PORD-PYRAB PORD-PYRFU PORD-PYRHO PORD-THEMA PSAC-ANASP PSAC-ANTSP PSAC-ARATH PSAC-CHLRE PSAC-CHLVU PSAC-CYACA PSAC-CYAPA PSAC-EUGGR PSAC-FREDI PSAC-GUITH PSAC-MAIZE PSAC-MARPO PSAC-MASLA PSAC-MESVI PSAC-ODOSI PSAC-PEA PSAC-PINTH PSAC-PORPU PSAC-SKECO PSAC-SPIOL PSAC-SYNEL PSAC-SYNP2 PSAC-SYNP6 PSAC-SYNY3 PSRB-WOLSU RDXA-RHOSH RDXB-RHOSH RNFB-BUCAI RNFB-ECO57 RNFB-ECOLI RNFB-HAEIN RNFB-PSEAE RNFB-RHOCA RNFB-VIBCH RNFC-BUCAI RNFC-ECOLI RNFC-HAEIN RNFC-PASMU RNFC-PSEAE RNFC-RHOCA RNFC-VIBCH VORC-METTH VORD-PYRAB VORD-PYRFU VORD-PYRHO Y092-METJA Y264-METJA Y492-MYCTU Y578-METJA Y726-METJA Y870-METJA YA43-HAEIN YCCM-ECOLI YCXI-PORPU YDIJ-ECOLI YDIJ-HAEIN YDIT-ECOLI YEIA-ECOLI YFHL-HAEIN YFRA-PROVU YG84-METTH YGFS-ECOLI YGFT-ECOLI YGL5-BACST YJES-ECOLI YJJW-ECOLI YKGF-ECOLI YNFG-ECOLI YSAA-ECOLI

Table 4: Description of the PROSITE 1 dataset.

Family	*Protein ID's*
PS00211	**ABC2-HUMAN APPD-BACSU FTSE-HAEIN HISP-SALTY KST1-ECOLI LCCL-LACLA LMRA-LACLC LOLD-BUCAI MDLB-BUCAI MKL-MYCTU MODC-HAEIN MRP2-RABIT NIKD-ECOLI NODI-AZOCA NODI-RHISN NOSF-PSEST NRTD-SYNY3 OPPF-LACLA OPPF-MYCPN POTA-MYCGE RFBB-MYXXA SUFC-ECOLI UVRA-BRUAB UVRA-STRMU VEXC-SALTI WHIT-ANOAL Y348-CHLPN YF08-METJA YJJK-HAEIN YXDL-BACSU** AAPP-RHILV AB11-HUMAN AB11-MOUSE AB11-RABIT AB11-RAT ABC1-HUMAN ABC1-MOUSE ABC1-SCHPO ABC2-MOUSE ABC3-HUMAN ABC6-HUMAN ABC7-HUMAN ABC7-MOUSE ABC8-HUMAN ABCA-AERSA ABCR-HUMAN ABCX-ANTSP ABCX-CYACA ABCX-CYAPA ABCX-GALSU ABCX-GUITH ABCX-ODOSI ABCX-PORPU ABCX-STRMU METN-ECOLI METN-HAEIN ABD2-HUMAN ABD3-HUMAN ABD3-MOUSE ABD3-RAT ABD4-HUMAN ABD4-MOUSE ABF2-HUMAN ABG1-HUMAN ABG1-MOUSE ABG2-HUMAN ABG3-MOUSE ABG4-HUMAN ABG5-HUMAN ABG5-MOUSE ABG5-RAT ABG8-HUMAN ABG8-MOUSE ABG8-RAT ACC8-CRICR ACC8-HUMAN ACC8-RAT ACC9-HUMAN ACC9-MOUSE ACC9-RABIT ACC9-RAT ADCC-STRPN ADP1-YEAST FBPC-ACTPL FBPC-ECOLI FBC1-HAEIN AGLK-RHIME ALD-HUMAN ALD-MOUSE ALSA-ECOLI AMIE-STRPN AMIF-STRPN AOTP-PSEAE APPF-BACSU APRD-PSEAE ARAG-ECOLI ARTP-ECOLI ARTP-HAEIN ATM1-YEAST BCRA-BACLI BEXA-HAEIN BFR1-SCHPO BPT1-YEAST BRAF-PSEAE BRAG-PSEAE BROW-DROME BROW-DROVI BTUD-ECOLI BZTD-RHOCA CBIO-SALTY CBRD-ERWCH CCMA-BRAJA CCMA-ECOLI CCMA-HAEIN CCMA-PARDE CCMA-RHOCA CDR1-CANAL CDR2-CANAL CDR3-CANAL CDR4-CANAL CFTR-BOVIN CFTR-CAVPO CFTR-HUMAN CFTR-MACMU CFTR-MOUSE CFTR-RABIT CFTR-RAT CFTR-SHEEP CFTR-SQUAC CFTR-XENLA CHVA-AGRT5 CHVD-AGRTU COMA-STRPN CTRD-NEIMA CTRD-NEIMB CVAB-ECOLI CYAB-BORPE CYDC-BACSU CYDC-ECOLI CYDC-HAEIN CYDD-BACSU CYDD-ECOLI CYDD-HAEIN CYSA-CHLVU CYSA-ECOLI CYSA-MARPO CYSA-MESVI CYSA-SALTY CYSA-SYNP7 CYSA-SYNY3 DPPD-BACSU DPPD-ECOLI DPPD-HAEIN DPPF-ECOLI DPPF-HAEIN DRRA-STRPE ECSA-BACSU EF3A-YEAST EF3B-YEAST EF3-CANAL EF3-PNECA EF3-SCHPO EGO-ECOLI EXP8-STRPN FECE-ECOLI FEPC-ECOLI FHUC-BACSU FHUC-ECOLI FTSE-ECOLI GC20-YEAST GLNQ-BACST GLNQ-ECOLI GLTL-ECOLI CLUA-CORGL HEPA-ANASP HFAC-CAUCR HISP-ECOLI FBC2-HAEIN HLY2-ECOLI HLYB-ACTAC HLYB-ECOLI HLYB-PASHA HLYB-PASSP HLYB-PROVU HMT1-SCHPO HMUV-YERPE HST6-CANAL KST5-ECOLI LACK-AGRRD LCN3-LACLA LCNC-LACLA LIVF-ARCFU LIVF-ECOLI LIVF-METJA LIVF-SALTY LIVG-ARCFU LIVG-ECOLI LIVG-METJA LIVG-SALTY LOLD-BUCAP LOLD-ECOLI LOLD-HAEIN LOLD-NEIMA LOLD-NEIMB LOLD-VIBCH LOLD-XYLFA MACB-ECOLI MALK-ECOLI MALK-ENTAE MALK-PHOLU MALK-SALTY MAM1-SCHPO MCHF-ECOLI MDL1-CANAL MDL1-YEAST MDL2-YEAST MDLA-BUCAI MDLA-ECOLI MDLB-ECOLI MDR1-CAEEL MDR1-CRIGR MDR1-ENTHI MDR1-HUMAN MDR1-LEIEN MDR1-MOUSE MDR1-RAT MDR2-CRIGR MDR2-MOUSE MDR2-RAT MDR3-CAEEL MDR3-CRIGR MDR3-ENTHI MDR3-HUMAN MDR3-MOUSE MDR4-DROME MDR4-ENTHI MDR5-DROME MDR-LEITA MDR-PLAFF MESD-LEUME MGLA-ECOLI MGLA-HAEIN MGLA-MYCGE MGLA-MYCPN MGLA-SALTY MGLA-TREPA MKL-MYCLE MNTA-SYNY3 MODC-AZOVI MODC-ECOLI MODC-MYCTU MODC-RHOCA MODF-ECOLI MRP1-HUMAN MRP2-HUMAN MRP2-RAT MRP3-HUMAN MRP3-RAT MRP4-HUMAN MRP5-HUMAN MRP5-MOUSE MRP5-RAT MRP6-HUMAN MRP6-RAT MSBA-ECOLI MSBA-HAEIN MSMK-STRMU MSMX-BACSU MSRA-STAEP NASD-KLEPN NATA-BACSU NDVA-RHIME NIKE-ECOLI NIST-LACLA NOCP-AGRT5 NODI-BRAJA NODI-RHIGA NODI-RHILO NODI-RHILT NODI-RHILV NDI1-RHIME NODI-RHIS3 NRTC-SYNP7 NRTC-SYNY3 NRTD-SYNP7 OCCP-AGRTU OCCP-RHIME OPAA-BASCU OPBA-BACSU OPCA-BACSU OPPD-BACSU OPPD-ECOLI OPPD-HAEIN OPPD-LACLA OPPD-LACLC OPPD-MYCGE OPPD-MYCPN OPPD-SALTY OPPF-BACSU OPPF-ECOLI OPPF-HAEIN OPPF-MYCGE OPPF-SALTY OPPF-STRMU OPPF-STRPY P29-MYCGE P29-MYCHR P29-MYCPN PDR5-YEAST PDRA-YEAST PDRB-YEAST PDRC-YEAST PDRF-YEAST PEBC-CAMJE PEDD-PEDAC PHNC-ECOLI PHNK-ECOLI PHNL-ECOLI PMD1-SCHPO POTA-ECOLI POTA-HAEIN POTA-MYCPN POTA-SALTY POTG-ECOLI PROV-ECOLI PROV-SALTY PRTD-ERWCH PSTB-ECOLI PSTB-EDWTA PSTB-ENTCL PSTB-METJA PSTB-MYCGE PSTB-MYCIT PSTB-MYCPN PSTB-MYCTU PSTB-PASMU PSTB-RHILO PSTB-SALTY PSTB-XYLFA PXA1-YEAST PXA2-YEAST RBSA-BACSU RBSA-ECOLI RBSA-HAEIN RFB1-KLEPN RFB2-KLEPN RFBE-YEREN RT1B-ACTPL RT3B-ACTPL SAPD-ECOLI SAPD-HAEIN SAPD-SALTY SAPF-ECOLI SAPF-HAEIN SAPF-SALTY SCRT-DROME FBPC-SERMA SMOK-RHOSH SNQ2-YEAST SPAT-BACSU SRTF-STRPY SSUB-BASCU SSUB-ECOLI STE6-YEAST SYRD-PSESY TAGB-DICDI TAGC-DICDI TAGH-BACSU TAP1-HUMAN TAP1-MOUSE TAP1-RAT TAP2-HUMAN TAP2-MOUSE TAP2-RAT TAUB-ECOLI THIQ-ECOLI THIQ-HAEIN TLRC-STRFR TROB-TREPA UGPC-ECOLI UUP1-HAEIN UUP2-HAEIN UUP-BUCAI UUP-ECOLI UVRA-AQUAE UVRA-BACHD UVRA-BACSU UVRA-BORBU UVRA-CHLMU UVRA-CHLPN UVRA-CHLTR UVRA-DEIRA UVRA-ECOLI UVRA-HAEIN UVRA-HELPJ UVRA-HELPY UVRA-LACLA UVRA-METTH UVRA-MICLU UVRA-MYCGE UVRA-MYCPN UVRA-MYCTU UVRA-NEIGO UVRA-PARDE UVRA-PASMU UVRA-PROMI UVRA-PSELE UVRA-RHIME UVRA-RICPR UVRA-SALTY UVRA-SERMA UVRA-STRCO UVRA-SYNY3 UVRA-THEMA UVRA-THETH UVRA-TREPA UVRA-VITST UVRA-ZYMMO V296-BACSU WHIT-ANOGA WHIT-CERCA WHIT-DROME WHIT-LUCCU XYLG-ECOLI XYLG-HAEIN Y014-MYCGE Y014-MYCPN Y015-MYCGE Y015-MYCPN Y035-METJA Y035-TREPA Y036-HAEIN Y065-MYCGE Y065-MYCPN Y068-CHLTR Y075-SYNY3 Y089-METJA Y121-METJA Y124-THEMA Y179-MYCGE Y179-MYCPN Y180-MYCGE Y180-MYCPN Y182-SYNY3 Y187-MYCGE Y187-MYCPN Y303-MYCGE Y303-MYCPN Y304-MYCGE Y304-MYCPN Y318-BORBU Y339-CHLMU Y352-THEMA Y354-HAEIN Y361-HAEIN Y382-RHIME Y412-METJA Y415-SYNY3 Y416-CHLTR Y467-MYCGE Y467-MYCPN Y46B-MYCGE Y46B-MYCPN Y4FO-RHISN Y4GM-RHISN Y4MK-RHISN Y4OS-RHISN Y4TH-RHISN Y4TR-RHISN Y4TS-RHISN Y542-CHLPN Y663-HAEIN Y664-HAEIN Y697-CHLMU Y700-RICPR Y719-METJA Y796-METJA Y7I9-ANASP Y873-METJA Y888-HELPJ Y888-HELPY Y986-MYCTU YA23-METJA YA51-HAEIN YA78-HAEIN YADG-ECOLI YATR-BACFI YAWB-SCHPO YBBA-ECOLI YBBL-ECOLI YBHF-ECOLI YBIT-ECOLI YBT1-YEAST YBXA-BACSU YC72-HAEIN YC72-MYCTU YC73-MYCTU YC81-MYCTU YCBN-BACSU YCFI-YEAST YCJV-ECOLI YCKI-BACSU YCXD-CYAPA YD34-MYCPN YD48-MYCTU YD49-MYCTU YD67-METJA YDCT-ECOLI YDDA-ECOLI YDDO-ECOLI YDDP-ECOLI YDIF-BACSU YE67-HAEIN YE70-HAEIN YE74-HAEIN YECC-ECOLI YEHX-ECOLI YEJF-ECOLI YEM6-YEAST YD01-SCHPO YFC8-YEAST YFEB-YERPE YFIB-BACSU YFIC-BACSU YNT9-SCHPO YG18-HAEIN YHBG-AZOCA YHBG-ECOLI YHBG-HAEIN YHBG-KLEPN YHBG-PSEPU YHBG-THIFE YHCG-BACSU YHCH-BACSU YHD5-YEAST YHDZ-ECOLI YHES-ECOLI YHES-HAEIN YHIH-ECOLI YI19-MYCTU YJJK-ECOLI YK83-YEAST CED7-CAEEL YLIA-ECOLI YMEB-LACLA YN26-MYCTU YN99-YEAST YNJD-ECOLI YOH5-YEAST YOJI-ECOLI YOR1-YEAST YP64-MYCTU YPC3-CAEEL YPHE-ECOLI YQ5C-CAEEL YQGJ-BACSU YQGK-BACSU YQIZ-BACSU YRBF-ECOLI YRBF-HAEIN YSC1-STRGC YTFR-ECOLI MNTB-BACSU YTMN-BACSU YTRE-BACSU YWJA-BACSU YXEO-BACSU YYBJ-BACSU ZNUC-BUCAI ZNUC-ECOLI ZNUC-HAEIN ZURA-LISIN ZURA-LISMO

Table 4: Description of the PROSITE 1 dataset.

Family	*Protein ID's*
PS00301	**CYSN-RHITR CYSN-XYLFA EF1A-ARCFU EF1A-DICDI EF1A-SULSO EF1S-PORPU EF2-CHICK EF2-MESAU EFTU-CHLTR EFTU-FERIS EFTU-GRALE EFTU-MYCPN EFTU-NEPOL EFTU-TOBAC EFTU-XYLFA LEPA-MYCHY LEPA-MYCLE LEPA-MYCPN TETQ-PREIN TYPA-SYNY3** CYSN-BUCAI CYSN-ECOLI CYSN-MYCTU CYSN-PSEAE CYSN-RHIME EF10-XENLA EF11-CRIGR EF11-DAUCA EF11-DROME EF11-EUPCR EF11-HORVU EF11-HUMAN EF11-MOUSE EF11-RHIRA EF11-SCHPO EF11-XENLA EF12-DAUCA EF12-DROME EF12-EUPCR EF12-HORVU EF12-HUMAN EF12-MOUSE EF12-RHIRA EF12-SCHPO EF12-XENLA EF13-RHIRA EF12-SCHPO EF13-XENLA EF11-SCHPO EF1A-ABSGL EF1A-AERPE EF1A-AJECA EF1A-APIME EF1A-ARATH EF1A-ARTSA EF1A-ARXAD EF1A-ASHGO EF1A-AURPU EF1A-BLAHO EF1A-BOMMO EF1A-BRARE EF1A-CAEEL EF1A-CANAL EF1A-CHICK EF1A-CRYNE EF1A-CRYPV EF1A-DESMO EF1A-EIMBO EF1A-ENTHI EF1A-EUGGR EF1A-GIALA EF1A-HALHA EF1A-HALMA EF1A-HELVI EF1A-HYDAT EF1A-LYCES EF1A-MAIZE EF1A-MANES EF1A-METJA EF1A-METTH EF1A-METVA EF1A-NEUCR EF1A-ONCVO EF1A-ORYSA EF1A-PEA EF1A-PLAFK EF1A-PODAN EF1A-PODCU EF1A-PUCGR EF1A-PYRAB EF1A-PYRAE EF1A-PYRHO EF1A-PYRWO EF1A-RHYAM EF1A-SCHCO EF1A-SORMA EF1A-SOYBN EF1A-STYLE EF1A-SULAC EF1A-TETPY EF1A-THEAC EF1A-THECE EF1A-TOBAC EF1A-TRIRE EF1A-TRYBB EF1A-VICFA EF1A-WHEAT EF1A-YARLI EF1A-YEAST EF1C-PORPU EF2-AERPE EF2-ARCFU EF2-BETVU EF2-BLAHO EF2-CAEEL EF2-CANAL EF2-CHLKE EF2-CRIGR EF2-CRYPV EF2-DESMO EF2-DICDI EF2-DROME EF2-ENTHI EF2-HALHA EF2-HUMAN EF2-METBU EF2-METJA EF2-METMT EF2-METTE EF2-METTH EF2-METVA EF2-MOUSE EF2-PYRAB EF2-PYRHO EF2-PYRFU EF2-RABIT EF2-RAT EF2-SCHPO EF2-SULAC EF2-SULSO EF2-THEAC EF2-YEAST EFG1-BORBU EFG1-STRCO EFG1-SYNY3 EFG1-TREPA EFG1-YEAST EFG2-BORBU EFG2-STRCO EFG2-SYNY3 EFG2-TREPA EFG2-YEAST EFGC-PEA EFGC-SOYBN EFGL-MYCTU EFGL-SYNY3 EFGL-THEMA EFG1-RAT EFG-AGRTU EFG-APPPP EFG-AQUAE EFG-AQUPY EFG-BACHD EFG-BACST EFG-BACSU EFG-BUCAI EFG-CHLMU EFG-CHLPN EFG-CHLTR EFG-ECOLI EFG-HAEIN EFG-HELPJ EFG-HELPY EFG-MICLU EFG-MYCGE EFG-MYCLE EFG-MYCPN EFG-MYCTU EFG-NEIGO EFG-PASMU EFG-PLARO EFG-RICCN EFG-RICPR EFG-SALTY EFG-SPIPL EFG-STAAM EFG-STRPY EFG-STRRA EFG-SYNP6 EFG-THEMA EFG-THETH EFG-THICU EFG-UREPA EFT1-SOYBN EFT1-STRCO EFT1-STRCU EFT1-STRRA EFT2-SOYBN EFT2-STRRA EFT3-STRCO EFT3-STRRA EFT1-PASMU EFT2-PASMU EFTU-AGRTU EFTU-APPPP EFTU-AQUAE EFTU-AQUPY EFTU-ARATH EFTU-ASTLO EFTU-BACFR EFTU-BACHD EFTU-BACST EFTU-BACSU EFTU-BORBU EFTU-BOVIN EFTU-BRELN EFTU-BRYPL EFTU-BUCAI EFTU-BUCAP EFTU-BUCMH EFTU-BUCSC EFTU-BURCE EFTU-CAMJE EFTU-CHACO EFTU-CHLAU EFTU-CHLMU EFTU-CHLPN EFTU-CHLRE EFTU-CHLVI EFTU-CHLVU EFTU-CODFR EFTU-COLOB EFTU-CORGL EFTU-COSCS EFTU-CYAPA EFTU-CYCME EFTU-CYTLY EFTU-DEIRA EFTU-DEISP EFTU-DERMA EFTU-ECOLI EFTU-EIKCO EFTU-EUGGR EFTU-FIBSU EFTU-FLAFE EFTU-FLESI EFTU-GLOS1 EFTU-GLOVI EFTU-GONPE EFTU-GUITH EFTU-GYMST EFTU-HAEIN EFTU-HELPJ EFTU-HELPY EFTU-HERAU EFTU-HUMAN EFTU-MANSQ EFTU-MESVI EFTU-MICLU EFTU-MYCGA EFTU-MYCGE EFTU-MYCHO EFTU-MYCLE EFTU-MYCTU EFTU-NEIGO EFTU-ODOSI EFTU-PANMO EFTU-PEA EFTU-PHOEC EFTU-PLARO EFTU-PLEBO EFTU-PORPU EFTU-PROHO EFTU-PSEAE EFTU-RECAM EFTU-RHILO EFTU-RICPR EFTU-SALTY EFTU-SCHPO EFTU-SHEPU EFTU-SPIAU EFTU-SPIPL EFTU-STIAU EFTU-STRAU EFTU-STRCJ EFTU-STRLU EFTU-STRMU EFTU-STROR EFTU-STRPY EFTU-SYNP6 EFTU-SYNP7 EFTU-SYNY3 EFTU-TAXOC EFTU-THEAQ EFTU-THEMA EFTU-THETH EFTU-THICU EFTU-TREHY EFTU-TREPA EFTU-UREPA EFTU-WOLSU EFTU-YEAST ERF2-CANAL ERF2-PICPI ERF2-SCHPO ERF2-YEAST GSP1-HUMAN GUF1-YEAST HBS1-YEAST LEPA-AQUAE LEPA-BACHD LEPA-BACSU LEPA-BORBU LEPA-BORPE LEPA-BUCAI LEPA-CHLMU LEPA-CHLPN LEPA-CHLTR LEPA-ECOLI LEPA-HAEIN LEPA-HELPJ LEPA-HELPY LEPA-LACLA LEPA-MYCGE LEPA-MYCTU LEPA-PASMU LEPA-PSEFL LEPA-RICPR LEPA-SALTY LEPA-STRCO LEPA-SYNY3 LEPA-THEMA LEPA-TREPA NODQ-AZOBR NODQ-RHIME NODQ-RHIS3 NODQ-RHISB NODQ-RHITR OTRA-STRRM RF3-BACNO RF3-BUCAI RF3-ECOLI RF3-HAEIN RF3-LACLA RF3-PASMU RF3-SALTY RF3-STAAU RF3-SYNY3 SELB-DESBA SELB-ECOLI SELB-HAEIN SELB-HUMAN SELB-METJA SELB-MOOTH SELB-MOUSE SN14-YEAST TET1-ENTFA TET5-ENTFA TET9-ENTFA TETM-NEIME TETM-STAAU TETM-STRLI TETM-STRPN TETM-UREUR TETO-CAMCO TETO-CAMJE TETO-STRMU TETO-STRPN TETP-CLOPE TETQ-BACFR TETQ-BACTN TETQ-PRERU TETS-LACLA TETS-LISMO TETW-BUTFI TYPA-BACSU TYPA-BUCAI TYPA-ECOLI TYPA-HAEIN TYPA-HELPJ TYPA-HELPY U5S1-HUMAN U5S1-MOUSE YE14-SCHPO YNQ3-YEAST YO81-CAEEL

Table 5: Description of the PROSITE 2 dataset

Family	*Protein ID's*
PS00070	**DHAE-MACPR DHAX-HUMAN DHA1-BOVIN HPCC-ECOLI YHJ9-YEAST GABD-ECOLI MAOC-ECOLI DHA4-YEAST DHA3-BACSU DHA5-YEAST YLQ6-CAEEL DHAS-CHICK DHAM-BOVIN PUT2-HUMAN MMSA-CAEEL** ALDA-ECOLI ALDB-ECOLI ASTD-ECOLI ASTD-PSEAE CALB-CAUCR CALB-PSEAE CALB-PSESP CROM-OCTDO CROM-OMMSL DHA1-BACSU DHA1-CHICK DHA1-ENTHI DHA1-HORSE DHA1-HUMAN DHA1-MOUSE DHA1-RAT DHA1-SHEEP DHA2-ALCEU DHA2-BACST DHA2-BACSU DHA2-HUMAN DHA2-MOUSE DHA2-RAT DHA2-YEAST DHA3-YEAST DHA4-HUMAN DHA4-MOUSE DHA4-RAT DHA5-BOVIN DHA5-HUMAN DHA6-HUMAN DHA6-YEAST DHA7-HUMAN DHA8-HUMAN DHA9-POLMI DHAB-AMAHP DHAB-ATRHO DHAB-BACSU DHAB-BETVU DHAB-ECOLI DHAB-GADCA DHAB-HORVU DHAB-ORYSA DHAB-RHIME DHAB-SPIOL DHAC-RAT DHAE-ELEED DHAF-VIBHA DHAG-HUMAN DHAG-PIG DHAL-AGABI DHAL-ALTAL DHAL-ASPNG DHAL-BACST DHAL-CLAHE DHAL-DEIRA DHAL-ECOLI DHAL-EMENI DHAL-ENCBU DHAL-MYCTU DHAL-PSEOL DHAL-PSESP DHAL-RHORU DHAL-STRCO DHAL-VIBCH DHAM-HORSE DHAM-HUMAN DHAM-LEITA DHAM-MESAU DHAM-MOUSE DHAM-RAT DHAN-MACPR DHAP-BOVIN DHAP-HUMAN DHAP-MOUSE DHAP-RAT DHAX-PEA DHAX-YEAST DHAY-YEAST DMPC-PSESP FEAB-ECOLI FTDH-HUMAN FTDH-RAT GABD-DEIRA GABD-RHISN GABD-SYNY3 GAPN-MAIZE GAPN-NICPL GAPN-PEA GAPN-STRMU MMSA-BACSU MMSA-BOVIN MMSA-HUMAN MMSA-PSEAE MMSA-RAT NAHF-PSESP PUT2-AGABI PUT2-YEAST PUTA-ECOLI PUTA-KLEAE PUTA-RHIME PUTA-SALTY ROCA-BACSU SSDH-HUMAN SSDH-RAT THCA-RHOER UGA5-YEAST XYC2-ACIGB XYLC-PSEPU XYLG-PSEPU Y4UC-RHISN YDCW-ECOLI YM00-YEAST YNEI-ECOLI

Table 5: Description of the PROSITE 2 dataset

Family	*Protein ID's*
PS00077	**COX1-THETH COX1-BACFI COX1-DIDMA COX1-ASCSU COX1-HORSE COX1-EPHEQ FIXN-AZOCA COX1-SYNVU COX1-CRION COX1-ALLMA AOX1-AERPE COX1-PEA COX1-RHOSH COX1-SOYBN COX1-PLABE** CO13-THETH CO14-BRAJA COX1-ACACA COX1-ALBCO COX1-ALBTU COX1-AMICA COX1-ANAPL COX1-ANOGA COX1-ANOQU COX1-APILI COX1-APTAU COX1-ARATH COX1-ARTSF COX1-ASTPE COX1-BACP3 COX1-BACSU COX1-BALMU COX1-BALPH COX1-BETVU COX1-BOVIN COX1-BRAJA COX1-CAEEL COX1-CANFA COX1-CANSI COX1-CAPHI COX1-CARAU COX1-CASBE COX1-CERSI COX1-CHICK COX1-CHLRE COX1-CHOBI COX1-CHOCR COX1-CHOFU COX1-CHOOC COX1-CHORO COX1-COTJA COX1-CROLA COX1-CYACA COX1-CYPCA COX1-DASNO COX1-DINSE COX1-DROME COX1-DRONO COX1-DROYA COX1-EMENI COX1-EQUAS COX1-FELCA COX1-GADMO COX1-GEOSD COX1-GOMVA COX1-HALGR COX1-HALHA COX1-HANWI COX1-HIPAM COX1-HUMAN COX1-KLULA COX1-LATCH COX1-LEITA COX1-LEPOC COX1-LEPSP COX1-LOCMI COX1-LUMTE COX1-MACRO COX1-MAIZE COX1-MARPO COX1-MEGAT COX1-METSE COX1-MOUSE COX1-MYCTU COX1-MYTED COX1-MYXGL COX1-NEUCR COX1-NOTPE COX1-OENBE COX1-ONCMY COX1-ORNAN COX1-ORYSA COX1-PANBU COX1-PAPHA COX1-PARLI COX1-PARTE COX1-PECMA COX1-PELSU COX1-PETMA COX1-PHOVI COX1-PHYME COX1-PHYPO COX1-PIG COX1-PISOC COX1-PLACH COX1-PLAFA COX1-PODAN COX1-POLOR COX1-POLSP COX1-POLSX COX1-POMNI COX1-PONPA COX1-PROWI COX1-RABIT COX1-RAT COX1-RHEAM COX1-RHILE COX1-RHISA COX1-RHIUN COX1-RHOCA COX1-RICPR COX1-SACDO COX1-SALSA COX1-SALTR COX1-SCAPL COX1-SCHPO COX1-SCYCA COX1-SHEEP COX1-SORBI COX1-SQUAC COX1-STRCA COX1-STRPU COX1-SYNY3 COX1-TETPY COX1-TINMA COX1-TRIRU COX1-TRYBB COX1-WHEAT COX1-XENLA COX1-YEAST COXN-BRAJA CX1A-PARDE CX1B-PARDE CYOB-BUCAI CYOB-ECOLI CYOB-PSEPU FIXN-AGRT7 FIXN-BRAJA FIXN-RHIME NORB-PSEAE NORB-PSEST QOX1-ACEAC QOX1-BACSU QOX1-SULAC QOXM-SULAC
PS00118	**PA21-NAJMO PA21-HORSE PA2H-BUNFA PA2E-PSEAU PA2C-CRODU PA2H-BOTJR PA2C-PSEAU PA2Z-HUMAN PA22-BUNMU PA23-NAJNG PA21-TRIGA PA21-ACAAN PA21-BOTPI PA2X-RAT PA22-PIG** OC90-CAVPO OC90-HUMAN OC90-MOUSE PA20-BUNMU PA20-NOTSC PA20-PSEAU PA21-AGKHA PA21-AGKHP PA21-AGKPI PA21-BOTAS PA21-BOTJA PA21-BOTJR PA21-BOTMO PA21-BOVIN PA21-BUNMU PA21-CANFA PA21-CAVPO PA21-ERIMA PA21-HEMHA PA21-HUMAN PA21-LATSE PA21-MATBI PA21-MOUSE PA21-NAJME PA21-NAJOX PA21-NOTSC PA21-OXYSC PA21-PIG PA21-PSEAU PA21-RAT PA21-SHEEP PA21-TRIFL PA21-VIPAA PA21-VIPAZ PA22-ACAAN PA22-AGKHA PA22-AGKHP PA22-ASPSC PA22-BITNA PA22-BOTAS PA22-BOTMO PA22-BOTPI PA22-CERGO PA22-ERIMA PA22-HELSU PA22-LATCO PA22-MATBI PA22-NAJKA PA22-NAJME PA22-NAJMO PA22-NOTSC PA22-OXYSC PA22-TRIGA PA22-TRIST PA22-VIPAZ PA23-AGKHP PA23-BOTAS PA23-BOTPI PA23-BUNMU PA23-HELSU PA23-HUMAN PA23-LATSE PA23-NAJKA PA23-NAJME PA23-NAJMO PA23-NOTSC PA23-OXYSC PA23-PSEAU PA23-TRIGA PA24-BUNMU PA24-DABRU PA24-LATSE PA24-TRIGA PA25-HUMAN PA25-MOUSE PA25-PSEAU PA25-RAT PA25-TRIGA PA25-TRIST PA26-BUNFA PA26-TRIGA PA27-DABRU PA27-TRIGA PA29-PSEAU PA2A-BUNFA PA2A-CRODU PA2A-HUMAN PA2A-MICNI PA2A-MOUSE PA2A-PSEAU PA2A-PSEPO PA2A-PSETE PA2A-RABIT PA2A-RAT PA2A-VIPAA PA2A-VIPPA PA2B-BUNFA PA2B-CRODU PA2B-MICNI PA2B-PSEPO PA2B-PSETE PA2B-TRIFL PA2B-TRIMU PA2B-VIPAA PA2C-MOUSE PA2C-PSETE PA2C-RAT PA2C-VIPAA PA2D-HUMAN PA2D-MOUSE PA2D-PSEAU PA2D-PSETE PA2E-HUMAN PA2E-MOUSE PA2F-HUMAN PA2F-MOUSE PA2G-PSEAU PA2H-AGKPI PA2H-ATRNM PA2H-LATCO PA2H-XENLA PA2I-VIPAA PA2L-VIPAA PA2M-AGKCL PA2M-CAVPO PA2M-CROSS PA2N-BUNFA PA2N-CROSS PA2N-ECHCA PA2N-VIPAA PA2X-BUNFA PA2X-HUMAN PA2X-MOUSE PA2X-NOTSC PA2X-TRIFL PA2Y-HUMAN PA2Y-MOUSE PA2Y-TRIFL PA2Z-MOUSE PA2-AIPLA PA2-APIME PA2-BITCA PA2-BITGA PA2-BOMTE PA2-CERCE PA2-CROAD PA2-CROAT PA2-DABRR PA2-ENHSC PA2-HELHO PA2-LATLA PA2-NAJAT PA2-NAJNA PA2-NAJPA PA2-OPHHA PA2-RHONO PA2-TRIOK PA2-VIPBB
PS00180	**GLNA-COLGL GLN4-PEA GLN2-DROME GLNA-HELPY GLNA-PANAR GLN3-RHILP GLN1-ARATH GLN5-MAIZE GLNA-PIG GLNA-PYRHO GLNA-THIFE GLNA-SALTY GLN3-PHAVU GLNA-NICPL GLN2-DAUCA** GLN1-ALNGL GLN1-BRAJA GLN1-CHLRE GLN1-DAUCA GLN1-DROME GLN1-FRAAL GLN1-LOTJA GLN1-MAIZE GLN1-MEDSA GLN1-MYCTU GLN1-ORYSA GLN1-PEA GLN1-PHAVU GLN1-RHILV GLN1-RHIME GLN1-SOYBN GLN1-STRRP GLN1-STRVR GLN1-VITVI GLN2-ARATH GLN2-BRAJA GLN2-CHLRE GLN2-FRAAL GLN2-HORVU GLN2-MAIZE GLN2-MEDSA GLN2-MYCTU GLN2-ORYSA GLN2-PEA GLN2-PHAVU GLN2-RHILP GLN2-RHIME GLN2-SOYBN GLN2-STRHY GLN2-STRVR GLN2-VITVI GLN3-HORVU GLN3-LUPAN GLN3-MAIZE GLN3-MEDSA GLN3-ORYSA GLN3-PEA GLN3-RHIME GLN4-MAIZE GLN4-PHAVU GLNA-AGABI GLNA-ANASP GLNA-AQUAE GLNA-ARCFU GLNA-AZOBR GLNA-AZOCA GLNA-AZOVI GLNA-BACCE GLNA-BACFR GLNA-BACSU GLNA-BOVIN GLNA-BUTFI GLNA-CAEEL GLNA-CHICK GLNA-CLOSA GLNA-CRILO GLNA-DUNSA GLNA-ECOLI GLNA-FREDI GLNA-HAEIN GLNA-HALN1 GLNA-HALVO GLNA-HELPJ GLNA-HUMAN GLNA-LACDE GLNA-LACLA GLNA-LACSA GLNA-LUPLU GLNA-METCA GLNA-METJA GLNA-METMP GLNA-METTH GLNA-METVO GLNA-MOUSE GLNA-NEIGO GLNA-PASMU GLNA-PINSY GLNA-PROVU GLNA-PYRAB GLNA-PYRFU GLNA-PYRKO GLNA-PYRWO GLNA-RAT GLNA-RHOCA GLNA-RHOSH GLNA-SCHPO GLNA-SQUAC GLNA-STAAU GLNA-STRCO GLNA-SULAC GLNA-SULSO GLNA-SYNP2 GLNA-SYNY3 GLNA-THEMA GLNA-TRITH GLNA-VIBAL GLNA-VIBCH GLNA-VIGAC GLNA-XENLA GLNA-YEAST GLNC-BRANA GLNC-MAIZE YCJK-ECOLI
PS00215	**UCP5-HUMAN AR13-NEUCR SA18-MOUSE YIA6-YEAST SHM1-YEAST ADT1-BOVIN ADT2-WHEAT UCP3-BOVIN M2OM-RAT YAD8-SCHPO UCP1-MOUSE TXTP-HUMAN DNC-HUMAN ADT3-YEAST ADT3-HUMAN** ADT1-ARATH ADT1-GOSHI ADT1-HUMAN ADT1-MAIZE ADT1-MOUSE ADT1-RAT ADT1-SOLTU ADT1-WHEAT ADT1-YEAST ADT2-ARATH ADT2-HUMAN ADT2-MAIZE ADT2-MOUSE ADT2-RAT ADT2-SOLTU ADT2-YEAST ADT3-BOVIN ADT-ANOGA ADT-CHLKE ADT-CHLRE ADT-DROME ADT-KLULA ADT-NEUCR ADT-ORYSA ADT-SCHPO BT1-MAIZE CG69-HUMAN CMC1-CAEEL CMC1-DROME CMC1-HUMAN CMC1-YEAST CMC2-CAEEL CMC2-HUMAN CMC2-MOUSE CMC3-CAEEL DIC-HUMAN DIC-MOUSE ECHP-MOUSE FLX1-YEAST GDC-BOVIN GDC-HUMAN GDC-RAT LEU5-YEAST M2OM-BOVIN M2OM-HUMAN M2OM-MOUSE MCAT-HUMAN MCAT-RAT MFT-HUMAN MPCP-BOVIN MPCP-CAEEL MPCP-CHOFU MPCP-HUMAN MPCP-RAT MPCP-YEAST MRS3-YEAST MRS4-YEAST ODC1-YEAST ODC2-YEAST ODC-HUMAN ORT1-HUMAN ORT1-MOUSE ORT1-YEAST ORT2-HUMAN P47A-CANBO P47B-CANBO PET8-YEAST PM34-HUMAN PM34-MOUSE PMT-YEAST RIM2-YEAST SA18-HUMAN SFC1-YEAST TXTP-BOVIN TXTP-CAEEL TXTP-RAT TXTP-YEAST UCP1-BOVIN UCP1-HUMAN UCP1-MESAU UCP1-RABIT UCP1-RAT UCP2-BRARE UCP2-CANFA UCP2-CYPCA UCP2-HUMAN UCP2-MOUSE UCP2-PIG UCP2-RAT UCP3-CANFA UCP3-HUMAN UCP3-MOUSE UCP3-PIG UCP3-RAT UCP4-HUMAN UCP5-MOUSE YD1K-SCHPO YDE9-SCHPO YE08-SCHPO YEA6-YEAST YEO3-YEAST YFL5-YEAST YG2O-YEAST YG5F-YEAST YM39-YEAST YMC1-YEAST YMC2-YEAST YQ51-CAEEL

Table 5: Description of the PROSITE 2 dataset

Family	*Protein ID's*
PS00217	**GTR1-RAT IOLF-BACSU CSBC-BACSU GTR5-HUMAN KHT2-KLULA PH84-YEAST NANT-ECOLI GHT3-SCHPO HUP1-CHLKE HGT1-CANAL GTR4-RAT GTR1-CHICK MMLH-ALCEU OUSA-ERWCH PHDK-NOCSK** AGT1-YEAST ARAE-BACSU ARAE-ECOLI ARAE-KLEOX BENK-ACICA CIT1-ECOLI CIT1-KLEPN CIT1-SALTY GAL2-YEAST GALP-ECOLI GHT2-SCHPO GHT4-SCHPO GHT5-SCHPO GHT6-SCHPO GIT1-YEAST GLCP-SYNY3 GLF-ZYMMO GT10-HUMAN GT11-HUMAN GTR1-BOVIN GTR1-HUMAN GTR1-LEIDO GTR1-MOUSE GTR1-PIG GTR1-RABIT GTR1-SHEEP GTR2-BOVIN GTR2-CHICK GTR2-HUMAN GTR2-LEIDO GTR2-MOUSE GTR2-PIG GTR2-RAT GTR3-BOVIN GTR3-CANFA GTR3-CHICK GTR3-DROME GTR3-HUMAN GTR3-MOUSE GTR3-PIG GTR3-RABIT GTR3-RAT GTR3-SHEEP GTR4-BOVIN GTR4-CANFA GTR4-HUMAN GTR4-MOUSE GTR4-PIG GTR5-BOVIN GTR5-MOUSE GTR5-RABIT GTR5-RAT GTR6-HUMAN GTR8-BOVIN GTR8-HUMAN GTR8-MOUSE GTR8-RAT GTR9-HUMAN HEX6-RICCO HGT1-KLULA HUP2-CHLKE HUP3-CHLKE HXT0-YEAST HXT1-YEAST HXT2-YEAST HXT3-YEAST HXT4-YEAST HXT5-YEAST HXT6-YEAST HXT7-YEAST HXT8-YEAST HXT9-YEAST HXTA-YEAST HXTC-YEAST HXTD-YEAST HXTE-YEAST HXTF-YEAST HXTG-YEAST ITR1-SCHPO ITR1-YEAST ITR2-SCHPO ITR2-YEAST JEN1-YEAST KGTP-ECOLI LACP-KLULA MA3T-YEAST MA6T-YEAST MAXT-YEAST MHPT-ECOLI MUCK-ACICA MYCT-HUMAN PCAK-ACICA PCAK-PSEPU PRO1-LEIEN PROP-ECOLI PROP-SALTY QAY-NEUCR QUTD-EMENI RAG1-KLULA RCO3-NEUCR RGT2-YEAST SHIA-ECOLI SNF3-YEAST STA-RICCO STC-RICCO STL1-YEAST STP1-ARATH STP-SPIOL TH11-TRYBB TH12-TRYBB TH23-TRYBB TH2A-TRYBB XYLE-ECOLI XYLT-LACBR Y281-HAEIN Y418-HAEIN YAAU-ECOLI YAEC-SCHPO YB04-HAEIN YB91-YEAST YCEI-BACSU YDFJ-ECOLI YDJE-ECOLI YDJK-ECOLI YFE0-YEAST YFIG-BACSU YGCS-ECOLI YGK4-YEAST YHJE-ECOLI YIR0-YEAST YJHB-ECOLI YOU1-CAEEL YYAJ-BACSU
PS00338	**SOMA-TRIVU PRL-CHICK PRL-PAROL SOMA-MACMU PRL-MOUSE SOMA-ACALA PLL2-MESAU SOML-SIGGU SOMA-ESOLU SOM2-CARAU SOMA-CANFA PRL-SHEEP SOM2-HUMAN PRL-HORSE SOMA-PANTR** GHR1-RAT GHR3-RAT GHR4-RAT PLF1-MOUSE PLF2-MOUSE PLF3-MOUSE PLFR-MOUSE PLL1-BOVIN PLL1-MOUSE PLL1-RAT PLL2-BOVIN PLL2-MOUSE PLL2-RAT PLLV-RAT PLL-HUMAN PLL-SHEEP PRL1-ALLMI PRL1-CRONO PRL1-ONCKE PRL1-OREMO PRL2-ALLMI PRL2-CRONO PRL2-ONCKE PRL2-ONCTS PRL2-OREMO PRL-ANGAN PRL-BALBO PRL-BOVIN PRL-BUFJA PRL-CAMDR PRL-CAPHI PRL-CARAU PRL-CHEMY PRL-CORAU PRL-CYPCA PRL-DICLA PRL-FELCA PRL-HUMAN PRL-HYPMO PRL-HYPNO PRL-ICTPU PRL-LOXAF PRL-MACMU PRL-MELGA PRL-MESAU PRL-MONDO PRL-MUSVI PRL-ONCMY PRL-PIG PRL-PROAT PRL-RABIT PRL-RAT PRL-SALSA PRL-SPAAU PRL-TRIVU PRR1-BOVIN PRR2-BOVIN PRR3-BOVIN PRR4-BOVIN PRRA-RAT PRRB-RAT PRRC-RAT SOM1-ACIGU SOM1-CARAU SOMA-ONCKE SOM1-ONCNE SOM1-SPAAU SOM2-ACIGU SOM2-MACMU SOM2-ONCMY SOM2-ONCNE SOM2-PANTR SOM2-SPAAU SOMA-ACABU SOMA-ANAPL SOMA-ANGJA SOMA-BALBO SOMA-BOVIN SOMA-BUBBU SOMA-BUFMA SOMA-CALJA SOMA-CARDE SOMA-CEREL SOMA-CHEMY SOMA-CHICK SOMA-CORAU SOMA-CORLV SOMA-CRONO SOMA-CTEID SOMA-CYPCA SOMA-DICLA SOMA-FELCA SOMA-FUGRU SOMA-GALSE SOMA-HETFO SOMA-HORSE SOMA-HUMAN SOMA-ICTPU SOMA-KATPE SOMA-LABRO SOMA-LAMPA SOMA-LATCA SOMA-LEPOS SOMA-LOXAF SOMA-MELGA SOMA-MESAU SOMA-MISMI SOMA-MONDO SOMA-MORSA SOMA-MOUSE SOMA-MUSVI SOMA-NYCPY SOMA-ODOAR SOMA-ONCKE SOMA-ONCKI SOMA-ONCMA SOMA-ONCTS SOMA-OREMO SOMA-ORENI SOMA-PAGMA SOMA-PANPG SOMA-PAROL SOMA-PERFV SOMA-PIG SOMA-PRIGL SOMA-PROAN SOMA-PSECR SOMA-RABIT SOMA-RANCA SOMA-RAT SOMA-SAIBB SOMA-SALSA SOMA-SCIOC SOMA-SEBSC SOMA-SERQU SOMA-SHEEP SOMA-SIGGU SOMA-SOLSE SOMA-SPAAU SOMA-STRCA SOMA-THUAL SOMA-THUTH SOMA-TRITC SOMA-VERVA SOMA-VULVU SOMA-XENLA SOMB-XENLA SOML-ACITR SOML-ANGAN SOML-CARAU SOML-CYCLU SOML-GADMO SOML-HIPHI SOML-ICTPU SOML-ONCKE SOML-PAROL SOML-PROAN SOML-SCIOC SOML-SOLSE SOML-TETMU

Table 6: Description of the GPCR dataset

Subfamily	*Protein ID's*
Amine	**5H1A-RAT 5H1B-CAVPO 5H1B-CRIGR 5H1B-HUMAN 5H1B-RABIT 5H1D-MOUSE 5H2A-CRIGR 5H2A-MOUSE 5HTB-DROME ACM1-DROME ACM3-PIG ACM4-MOUSE B2AR-MESAU DBDR-XENLA HH2R-MOUSE O44198 O61232 OAR2-LOCMI** 5H1A-FUGRU 5H1A-HUMAN 5H1A-MOUSE 5H1B-DIDMA 5H1B-FUGRU 5H1B-MOUSE 5H1B-RAT 5H1B-SPAEH 5H1D-CANFA 5H1D-CAVPO 5H1D-FUGRU 5H1D-HUMAN 5H1D-RABIT 5H1D-RAT 5H1E-HUMAN 5H1F-CAVPO 5H1F-HUMAN 5H1F-MOUSE 5H1F-RAT 5H2A-HUMAN 5H2A-MACMU 5H2A-PIG 5H2A-RAT 5H2B-HUMAN 5H2B-MOUSE 5H2B-RAT 5H2C-HUMAN 5H2C-MOUSE 5H2C-RAT 5H4-CAVPO 5H4-HUMAN 5H4-MOUSE 5H4-RAT 5H5A-HUMAN 5H5A-MOUSE 5H5A-RAT 5H5B-MOUSE 5H5B-RAT 5H6-HUMAN 5H6-MOUSE 5H6-RAT 5H7-CAVPO 5H7-HUMAN 5H7-MOUSE 5H7-RAT 5H7-XENLA 5HT1-APLCA 5HT1-DROME 5HT2-APLCA 5HTA-DROME 5HT-BOMMO 5HT-HELVI 5HT-LYMST A1AA-BOVIN A1AA-CAVPO A1AA-HUMAN A1AA-MOUSE A1AA-ORYLA A1AA-RABIT A1AA-RAT A1AB-HUMAN A1AB-MESAU A1AB-MOUSE A1AB-RAT A1AD-HUMAN A1AD-MOUSE A1AD-RABIT A1AD-RAT A2AA-BOVIN A2AA-CAVPO A2AA-HUMAN A2AA-MOUSE A2AA-PIG A2AA-RAT A2AB-CAVPO A2AB-HUMAN A2AB-MOUSE A2AB-ORYAF A2AB-RAT A2AC-CAVPO A2AC-DIDMA A2AC-HUMAN A2AC-MOUSE A2AC-RAT A2AR-CARAU A2AR-LABOS ACM1-HUMAN ACM1-MACMU ACM1-MOUSE ACM1-PIG ACM1-RAT ACM2-CHICK ACM2-HUMAN ACM2-MOUSE ACM2-PIG ACM2-RAT ACM3-BOVIN ACM3-CHICK ACM3-GORGO ACM3-HUMAN ACM3-MOUSE ACM3-PANTR ACM3-PONPY ACM3-RAT ACM4-CHICK ACM4-HUMAN ACM4-RAT ACM4-XENLA ACM5-HUMAN ACM5-MACMU ACM5-RAT B1AR-BOVIN B1AR-CANFA B1AR-FELCA B1AR-HUMAN B1AR-MACMU B1AR-MELGA B1AR-MOUSE B1AR-PIG B1AR-RAT B1AR-SHEEP B1AR-XENLA B2AR-BOVIN B2AR-CANFA B2AR-FELCA B2AR-HUMAN B2AR-MACMU B2AR-MOUSE B2AR-PIG B2AR-RAT B3AR-BOVIN B3AR-CANFA B3AR-CAPHI B3AR-FELCA B3AR-HUMAN B3AR-MACMU B3AR-MOUSE B3AR-RAT B3AR-SHEEP B4AR-MELGA D1DR-CARAU D1DR-FUGRU D1DR-OREMO D2D1-XENLA D2DR-BOVIN D2DR-CERAE D2DR-FUGRU D2DR-HUMAN D2DR-MELGA D2DR-MOUSE D3DR-CERAE D3DR-HUMAN D3DR-MOUSE D3DR-RAT D4DR-HUMAN D4DR-MOUSE D4DR-RAT D5DR-FUGRU DADR-DIDMA DADR-HUMAN DADR-MACMU DADR-PIG DADR-RAT DADR-XENLA DBDR-HUMAN DBDR-RAT DCDR-XENLA DOP1-DROME DOP2-DROME GRE1-BALAM GRE2-BALAM HH1R-BOVIN HH1R-CAVPO HH1R-HUMAN HH1R-MOUSE HH1R-RAT HH2R-CANFA HH2R-CAVPO HH2R-HUMAN HH2R-RAT HH3R-CAVPO HH3R-HUMAN HH3R-RAT HH4R-HUMAN O02146 O15969 O15970 O17470 O17496 O18512 O42315 O42316 O42317 O42322 O60451 O61730 O76267 O77254 O96716 O97171 OAR1-LOCMI OAR1-LYMST OAR2-LYMST OAR-BOMMO OAR-DROME OAR-HELVI P90927 P91096 P97842 Q13167 Q13675 Q13729 Q24038 Q63004 Q923X5 Q923X6 Q923X7 Q923X8 Q923X9 Q923Y0 Q923Y1 Q923Y2 Q923Y3 Q923Y4 Q923Y5 Q923Y6 Q923Y7 Q923Y8 Q923Y9 Q969N4 Q96RI8 Q96RI9 Q96RJ0 Q98841 Q98842 Q98843 Q98844 Q98998 Q99MB0 Q9BMA9 Q9BZK0 Q9BZK1 Q9D282 Q9DBL0 Q9GJS6 Q9GJT0 Q9GJU1 Q9GK99 Q9GKA0 Q9GKI2 Q9GL56 Q9GL57 Q9GLP5 Q9GQ54 Q9MYI8 Q9MZ00 Q9MZU2 Q9MZU3 Q9N263 Q9N296 Q9N297 Q9N298 Q9N2B0 Q9N2B1 Q9N2B2 Q9N2B7 Q9NG02 Q9NZR3 Q9PSA6 Q9PSA7 Q9PTF6 Q9QW44 Q9QW71 Q9QW77 Q9QWS2 Q9QX37 Q9TSW7 Q9TTM9 Q9U5A7 Q9U7D5 Q9UD63 Q9UD67 Q9UPA9 Q9V8Q3 Q9V8Q9 Q9VDJ6 Q9W180 Q9YHA5

Table 6: Description of the GPCR dataset

Subfamily	*Protein ID's*
Peptide	**BRS3-HUMAN CCR3-HUMAN CKR5-MACMU CKR5-TRAFR FML1-MOUSE FML1-PANTR GP37-HUMAN IL8A-PANTR IL8A-RABIT IL8B-GORGO MC5R-RAT MSHR-CEREL NK3R-HUMAN NMBR-MOUSE NTR2-HUMAN NTR2-RAT NY5R-MOUSE O57317 O57585 O88535 O93247 O97505 OX1R-RAT Q98U14 Q9BXA0 Q9D392 Q9DGM2 Q9GK75 Q9TQT0 SSR1-MOUSE** ACTR-BOVIN ACTR-CAVPO ACTR-HUMAN ACTR-MESAU ACTR-MOUSE ACTR-SHEEP ADMR-HUMAN ADMR-MOUSE ADMR-RAT AG22-HUMAN AG22-MERUN AG22-MOUSE AG22-RAT AG2R-BOVIN AG2R-CANFA AG2R-CAVPO AG2R-CHICK AG2R-HUMAN AG2R-MELGA AG2R-MERUN AG2R-MOUSE AG2R-PIG AG2R-RABIT AG2R-RAT AG2R-SHEEP AG2R-XENLA AG2S-HUMAN AG2S-MOUSE AG2S-RAT AG2S-XENLA APJ-HUMAN APJ-MACMU APJ-MOUSE APJ-RAT APJ-XENLA AVT-CATCO BRB1-HUMAN BRB1-MOUSE BRB1-RABIT BRB1-RAT BRB2-CAVPO BRB2-HUMAN BRB2-MOUSE BRB2-RABIT BRB2-RAT BRS3-CAVPO BRS3-MOUSE BRS3-SHEEP BRS4-BOMOR C3AR-CAVPO C3AR-HUMAN C3AR-MOUSE C3AR-RAT C3X1-HUMAN C3X1-MOUSE C3X1-RAT C5AR-CANFA C5AR-CAVPO C5AR-GORGO C5AR-HUMAN C5AR-MACMU C5AR-MOUSE C5AR-PANTR C5AR-PONPY C5AR-RABIT C5AR-RAT CCKR-CAVPO CCKR-HUMAN CCKR-MOUSE CCKR-RABIT CCKR-RAT CCKR-XENLA CCR3-MOUSE CCR4-BOVIN CCR4-CERTO CCR4-FELCA CCR4-HUMAN CCR4-MACFA CCR4-MACMU CCR4-MOUSE CCR4-PAPAN CCR4-RAT CCR5-HUMAN CCR5-MOUSE CCR5-RAT CCR6-CERAE CCR6-HUMAN CCR6-MACMU CCR6-MACNE CKD6-HUMAN CKD6-MOUSE CKD6-RAT CKR1-HUMAN CKR1-MACMU CKR1-MOUSE CKR2-HUMAN CKR2-MACMU CKR2-MOUSE CKR2-RAT CKR3-CAVPO CKR3-CERAE CKR3-HUMAN CKR3-MACMU CKR3-MOUSE CKR3-RAT CKR4-HUMAN CKR4-MOUSE CKR5-CERAE CKR5-CERTO CKR5-GORGO CKR5-HUMAN CKR5-HYLLE CKR5-MOUSE CKR5-PANTR CKR5-PAPHA CKR5-PONPY CKR5-PYGBI CKR5-PYGNE CKR5-RAT CKR5-TRAPH CKR6-HUMAN CKR6-MOUSE CKR7-HUMAN CKR7-MOUSE CKR8-HUMAN CKR8-MACMU CKR8-MOUSE CKR9-HUMAN CKR9-MOUSE CKRA-HUMAN CKRA-MOUSE CKRB-BOVIN CKRB-HUMAN CKRV-MOUSE CML1-HUMAN CML1-MOUSE CML1-RAT CML2-HUMAN CML2-RAT CXC1-HUMAN CXC1-MOUSE EBP2-HUMAN ET1R-BOVIN ET1R-HUMAN ET1R-PIG ET1R-RAT ET3R-XENLA ETBR-BOVIN ETBR-HORSE ETBR-HUMAN ETBR-MOUSE ETBR-PIG ETBR-RAT FML1-GORGO FML1-HUMAN FML1-MACMU FML1-PONPY FML2-HUMAN FMLR-GORGO FMLR-HUMAN FMLR-MACMU FMLR-MOUSE FMLR-PANTR FMLR-PONPY FMLR-RABIT GALR-HUMAN GALR-MOUSE GALR-RAT GALS-HUMAN GALS-MOUSE GALS-RAT GALT-HUMAN GALT-MOUSE GALT-RAT GASR-BOVIN GASR-CANFA GASR-HUMAN GASR-MOUSE GASR-PRANA GASR-RABIT GASR-RAT GP44-HUMAN GP44-MOUSE GP72-CANFA GP72-HUMAN GP72-MOUSE GPRW-HUMAN GPRX-MOUSE GRPR-HUMAN GRPR-MOUSE GRPR-RAT IL8A-GORGO IL8A-HUMAN IL8A-RAT IL8B-BOVIN IL8B-CANFA IL8B-HUMAN IL8B-MACMU IL8B-MOUSE IL8B-PANTR IL8B-RABIT IL8B-RAT ITR-CATCO MC3R-HUMAN MC3R-MOUSE MC3R-RAT MC4R-HUMAN MC4R-PIG MC4R-RAT MC5R-BOVIN MC5R-HUMAN MC5R-MOUSE MC5R-PANTR MC5R-SHEEP MSHR-ALCAA MSHR-BOVIN MSHR-CANFA MSHR-CAPCA MSHR-CAPHI MSHR-CHICK MSHR-DAMDA MSHR-HUMAN MSHR-MOUSE MSHR-OVIMO MSHR-PANTR MSHR-RANTA MSHR-SHEEP MSHR-VULVU MTR-BUFMA NFF1-HUMAN NFF1-RAT NFF2-HUMAN NFF2-RAT NK1R-CAVPO NK1R-HUMAN NK1R-MOUSE NK1R-RANCA NK1R-RAT NK2R-BOVIN NK2R-CAVPO NK2R-HUMAN NK2R-MESAU NK2R-MOUSE NK2R-RABIT NK2R-RAT NK3R-RABIT NK3R-RAT NK4R-HUMAN NMBR-HUMAN NMBR-RAT NTR1-HUMAN NTR1-MOUSE NTR1-RAT NTR2-MOUSE NY1R-CANFA NY1R-CAVPO NY1R-HUMAN NY1R-MOUSE NY1R-PIG NY1R-RAT NY1R-XENLA NY2R-BOVIN NY2R-CAVPO NY2R-CHICK NY2R-HUMAN NY2R-MACMU NY2R-MOUSE NY2R-PIG NY4R-HUMAN NY4R-MOUSE NY4R-RAT NY5R-CANFA NY5R-HUMAN NY5R-PIG NY5R-RAT NY6R-MOUSE NY6R-RABIT NYR-DROME O00421 O35457 O42324 O42402 O42445 O43192 O43664 O55040 O57463 O70171 O73667 O73671 O73733 O73734 O73739 O75307 O76067 O76873 O77488 O77776 O77808 O77833 O88313 O88536 O88537 O88538 O88634 O88721 O93237 O93239 O93259 O97724 O97774 O97962 O97975 OPRD-HUMAN OPRD-MOUSE OPRD-RAT OPRK-CAVPO OPRK-HUMAN OPRK-MOUSE OPRK-RAT OPRM-BOVIN OPRM-HUMAN OPRM-MACMU OPRM-MOUSE OPRM-PIG OPRM-RAT OPRX-CAVPO OPRX-HUMAN OPRX-MOUSE OPRX-PIG OPRX-RAT OX1R-HUMAN OX2R-CANFA OX2R-HUMAN OX2R-RAT OXYR-BOVIN OXYR-HUMAN OXYR-MACMU OXYR-MOUSE OXYR-PIG OXYR-RAT OXYR-SHEEP PAR2-HUMAN PAR2-MOUSE PAR2-RAT PAR3-HUMAN PAR3-MOUSE Q16144 Q16433 Q25396 Q25397 Q62973 Q91548 Q91V45 Q924N0 Q924U1 Q94736 Q969F8 Q96F42 Q96QG0 Q96RV1 Q98UH1 Q99463 Q99647 Q9BDP3 Q9BDQ4 Q9BDQ5 Q9BDS5 Q9BDS6 Q9BDS8 Q9BGN5 Q9BGN6 Q9BUN4 Q9BYX5 Q9DBV6 Q9DEC5 Q9DG05 Q9DG06 Q9DGI1 Q9DGJ9 Q9DGQ6 Q9DHG8 Q9DHV5 Q9EPP3 Q9EQ16 Q9EQM7 Q9EQR9 Q9ERC0 Q9ERH5 Q9ESQ4 Q9GK73 Q9GLN9 Q9GZQ4 Q9H573 Q9HB89 Q9HBV6 Q9HCA5 Q9I7W8 Q9JIB1 Q9JIB2 Q9JII9 Q9JIN4 Q9JIY1 Q9JJI5 Q9JK40 Q9JKN0 Q9JLP2 Q9JLY8 Q9MYJ8 Q9MYJ9 Q9MZ99 Q9MZA0 Q9MZA1 Q9MZA2 Q9MZA3 Q9N0M0 Q9N0U1 Q9N0W7 Q9N0Z0 Q9NBC8 Q9NRA6 Q9NS48 Q9NYK7 Q9PTF7 Q9PUA0 Q9PUG3 Q9PVF9 Q9PVG0 Q9PVY7 Q9QW13 Q9QW17 Q9QW18 Q9QW32 Q9QWG9 Q9QWN6 Q9QY42 Q9QYC5 Q9QYC6 Q9R0D1 Q9R1L9 Q9R1M0 Q9R1V0 Q9TQR2 Q9TQR8 Q9TQT1 Q9TQT2 Q9TQT3 Q9TQU3 Q9TQU4 Q9TQU5 Q9TQU6 Q9TQU7 Q9TQV0 Q9TQV2 Q9TQV3 Q9TQV5 Q9TQV6 Q9TQW0 Q9TQW2 Q9TQW4 Q9TQX0 Q9TQX2 Q9TQX3 Q9TSK1 Q9TSN2 Q9TSN3 Q9TSQ1 Q9TSQ2 Q9TSQ3 Q9TSQ4 Q9TSQ7 Q9TSQ8 Q9TUQ4 Q9TUQ5 Q9TUQ6 Q9TUQ7 Q9TUQ8 Q9TUQ9 Q9TUR0 Q9TUR1 Q9TUR2 Q9TUR3 Q9TUR4 Q9TUR5 Q9TUR6 Q9TUR7 Q9TUR8 Q9TUR9 Q9TUS0 Q9TUS1 Q9TUS2 Q9TUS3 Q9TUS4 Q9TUS5 Q9TUS6 Q9TUS7 Q9TUS8 Q9TUS9 Q9TUT0 Q9TUT1 Q9TUT2 Q9TUT3 Q9TUT4 Q9TUT5 Q9TUT6 Q9TUT7 Q9TUT8 Q9TUT9 Q9TUU0 Q9TUU1 Q9TUU2 Q9TUU3 Q9TUU4 Q9TUU5 Q9TUU6 Q9TUU7 Q9TUU8 Q9TUU9 Q9TUV0 Q9TUV1 Q9TUV2 Q9TUV3 Q9TUV4 Q9TUV5 Q9TUV6 Q9TUV8 Q9TUV9 Q9TUW0 Q9TUW1 Q9TUW2 Q9TUW3 Q9TUW4 Q9TUW5 Q9TUW6 Q9TUW7 Q9TUW8 Q9TUW9 Q9TUX0 Q9TUX1 Q9TV16 Q9TV42 Q9TV43 Q9TV44 Q9TV45 Q9TV46 Q9TV47 Q9TV48 Q9TV49 Q9TV50 Q9TV93 Q9U721 Q9UBF7 Q9UBJ7 Q9UBT9 Q9UD23 Q9UDE6 Q9UN23 Q9UN24 Q9UN25 Q9UN26 Q9UN27 Q9UN28 Q9UPA4 Q9UPG0 Q9UQQ6 Q9VAD2 Q9VAU0 Q9VB87 Q9VGX8 Q9VW75 Q9W4R0 Q9W6I3 Q9WTV8 Q9WTV9 Q9XS35 Q9XS99 Q9XSD7 Q9XT12 Q9XT13 Q9XT14 Q9XT76 Q9YGC3 Q9YHX1 Q9Z0G3 Q9Z0T7 Q9Z2D4 SSR1-HUMAN SSR1-RAT SSR2-BOVIN SSR2-HUMAN SSR2-MOUSE SSR2-PIG SSR2-RAT SSR3-HUMAN SSR3-MOUSE SSR3-RAT SSR4-HUMAN SSR4-MOUSE SSR4-RAT SSR5-HUMAN SSR5-MOUSE SSR5-RAT THRR-CRILO THRR-HUMAN THRR-MOUSE THRR-PAPHA THRR-RAT THRR-XENLA TLR1-DROME TLR2-DROME UR2R-HUMAN UR2R-RAT V1AR-HUMAN V1AR-MOUSE V1AR-RAT V1AR-SHEEP V1BR-HUMAN V1BR-MOUSE V1BR-RAT V2R-BOVIN V2R-HUMAN V2R-PIG V2R-RAT YLD1-CAEEL
Hormone	**FSHR-EQUAS FSHR-SHEEP Q14751 Q98T84 Q9BG55 Q9DGC5 Q9DGC6 Q9I8N7 Q9I948 TSHR-BOVIN** FSHR-BOVIN FSHR-CHICK FSHR-HORSE FSHR-HUMAN FSHR-MACFA FSHR-MOUSE FSHR-PIG FSHR-RAT LSHR-BOVIN LSHR-CALJA LSHR-HUMAN LSHR-MOUSE LSHR-PIG LSHR-RAT LSHR-SHEEP Q15996 Q27986 Q64183 Q98T85 Q98TF4 Q9BG56 Q9BGN4 Q9D697 Q9DGF5 Q9I949 Q9PVN9 Q9PVP0 Q9PW16 TSHR-CANFA TSHR-HUMAN TSHR-MOUSE TSHR-RAT TSHR-SHEEP

Table 6: Description of the GPCR dataset

Subfamily	*Protein ID's*
Rhodopsin	**O57422 O57447 OPS1-DROPS OPSB-SAIBB OPSD-ICTPU OPSD-MACFA OPSD-SARMI OPSD-SARSP OPSD-SHEEP OPSG-SCICA OPSR-FELCA OPSV-CHICK Q90226 Q98UJ5 Q9GU63 Q9IB87 Q9PTX9 Q9PUE9 Q9UAM9 Q9W609** O02464 O02465 O46554 O57448 O57605 O61473 O61474 O62860 O70363 O76123 O76124 O76125 O96107 O97901 OPN3-HUMAN OPN3-MOUSE OPN4-HUMAN OPN4-MOUSE OPS1-CALVI OPS1-DROME OPS1-HEMSA OPS1-LIMPO OPS1-PATYE OPS1-SCHGR OPS2-DROME OPS2-DROPS OPS2-HEMSA OPS2-LIMPO OPS2-PATYE OPS2-SCHGR OPS3-DROME OPS3-DROPS OPS4-DROME OPS4-DROPS OPS4-DROVI OPS5-DROME OPS6-DROME OPSB-ANOCA OPSB-APIME OPSB-ASTFA OPSB-BOVIN OPSB-CARAU OPSB-CHICK OPSB-CONCO OPSB-GECGE OPSB-HUMAN OPSB-MOUSE OPSB-ORYLA OPSB-RAT OPSD-ALLMI OPSD-AMBTI OPSD-ANGAN OPSD-ANOCA OPSD-APIME OPSD-ASTFA OPSD-ATHBO OPSD-BOVIN OPSD-BUFBU OPSD-BUFMA OPSD-CAMAB OPSD-CAMHU OPSD-CAMLU OPSD-CAMSC OPSD-CANFA OPSD-CARAU OPSD-CATBO OPSD-CHELB OPSD-CHICK OPSD-CRIGR OPSD-CYPCA OPSD-DELDE OPSD-DICLA OPSD-DIPAN OPSD-DIPVU OPSD-GALML OPSD-GAMAF OPSD-GLOME OPSD-GOBNI OPSD-HUMAN OPSD-LAMJA OPSD-LITMO OPSD-LIZAU OPSD-LIZSA OPSD-LOLFO OPSD-MESBI OPSD-MOUSE OPSD-MUGCE OPSD-MULSU OPSD-MYRBE OPSD-MYRVI OPSD-NEOAR OPSD-NEOAU OPSD-NEOSA OPSD-OCTDO OPSD-ORCAU OPSD-ORCVI OPSD-ORYLA OPSD-PETMA OPSD-PHOGR OPSD-PHOVI OPSD-PIG OPSD-POERE OPSD-POMMI OPSD-PROCL OPSD-PROML OPSD-PROSE OPSD-RABIT OPSD-RAJER OPSD-RANCA OPSD-RANPI OPSD-RANTE OPSD-RAT OPSD-SALPV OPSD-SARDI OPSD-SARPI OPSD-SARPU OPSD-SARSL OPSD-SARTI OPSD-SARXA OPSD-SCYCA OPSD-SEPOF OPSD-SOLSO OPSD-SPAAU OPSD-SPHSP OPSD-TETNG OPSD-TODPA OPSD-TRIMA OPSD-TURTR OPSD-XENLA OPSD-ZEUFA OPSD-ZOSOP OPSF-ANGAN OPSG-ASTFA OPSG-CARAU OPSG-CAVPO OPSG-CHICK OPSG-GECGE OPSG-HUMAN OPSG-MOUSE OPSG-ORYLA OPSG-RABIT OPSG-RAT OPSH-ASTFA OPSH-CARAU OPSI-ASTFA OPSL-CALJA OPSO-SALSA OPSP-CHICK OPSP-COLLI OPSP-ICTPU OPSP-PETMA OPSR-ANOCA OPSR-ASTFA OPSR-CAPHI OPSR-CARAU OPSR-CHICK OPSR-HUMAN OPSR-ORYLA OPSR-XENLA OPSU-BRARE OPSU-CARAU OPSV-APIME OPSV-ORYLA OPSV-XENLA OPSX-HUMAN OPSX-MOUSE Q96FC5 Q98TH3 Q98TS2 Q98UH7 Q98UJ4 Q9BGI7 Q9DEW8 Q9ERF2 Q9GU64 Q9I852 Q9I853 Q9I854 Q9I855 Q9I856 Q9I857 Q9I8P4 Q9I8R6 Q9I960 Q9I975 Q9I9I5 Q9I9I6 Q9I9I7 Q9I9I8 Q9I9I9 Q9I9J0 Q9I9J1 Q9I9R2 Q9IA33 Q9IA34 Q9IA35 Q9IA36 Q9IAH7 Q9IAH8 Q9IAH9 Q9IAI0 Q9IB88 Q9IBH2 Q9JLS7 Q9NJC9 Q9PUA1 Q9PUA2 Q9PUA9 Q9PWN3 Q9PWN4 Q9TU70 Q9TVV5 Q9TX52 Q9TX54 Q9TX55 Q9UAM6 Q9UAM7 Q9UAM8 Q9W684 Q9W685 Q9W6A5 Q9W6A6 Q9W6A7 Q9W6A8 Q9W6A9 Q9W6I4 Q9W6I5 Q9W6J6 Q9W6K3 Q9W6S0 Q9W6S1 Q9W771 Q9W772 Q9W773 Q9W7C1 Q9W7K8 Q9XS34 Q9XSF1 Q9XSX2 Q9XSX3 Q9XSX4 Q9YGY7 Q9YGY8 Q9YI51 Q9YI52 Q9YI53 REIS-TODPA RGR-BOVIN RGR-HUMAN RGR-MOUSE
Olfactory	**O1C1-HUMAN O70266 O70270 O8B8-HUMAN OLF1-CANFA OLF6-CHICK OLF6-RAT Q9EQB2 Q9H340 Q9H341 Q9I8B8 Q9I8C2 Q9I8Z2 Q9I8Z8 Q9IBD9 Q9PSU4 Q9PVU0 Q9QZ19 Q9TU89 Q9UDD9** GU27-RAT O13036 O1A1-HUMAN O1A2-HUMAN O1D2-HUMAN O1D4-HUMAN O1E1-HUMAN O1E2-HUMAN O1F1-HUMAN O1G1-HUMAN O1I1-HUMAN O1Q1-HUMAN O2A4-HUMAN O2B2-HUMAN O2B3-HUMAN O2B6-HUMAN O2C1-HUMAN O2D2-HUMAN O2F1-HUMAN O2F2-HUMAN O2G1-HUMAN O2H1-HUMAN O2H3-HUMAN O2J2-HUMAN O2J3-HUMAN O2S2-HUMAN O2T1-HUMAN O2W1-HUMAN O35184 O35434 O3A1-HUMAN O3A2-HUMAN O3A3-HUMAN O3A4-HUMAN O42165 O42166 O42167 O42168 O42169 O42170 O42171 O42172 O42605 O42609 O4D1-HUMAN O4F3-HUMAN O57597 O5F1-HUMAN O5I1-HUMAN O5U1-HUMAN O5V1-HUMAN O6A1-HUMAN O6B1-HUMAN O70265 O70267 O70268 O70269 O70271 O77756 O77757 O77758 O7A5-HUMAN O7AA-HUMAN O7AH-HUMAN O7C1-HUMAN O7C2-HUMAN O8D2-HUMAN O93549 O93550 O93551 O95499 OAA4-HUMAN OAA5-HUMAN OAH1-HUMAN OAH2-HUMAN OAH3-HUMAN OAJ1-HUMAN OBA1-HUMAN OCD3-HUMAN OL15-MOUSE OLF0-RAT OLF1-CHICK OLF1-RAT OLF2-CANFA OLF2-CHICK OLF2-RAT OLF3-CANFA OLF3-CHICK OLF3-RAT OLF4-CANFA OLF4-CHICK OLF4-RAT OLF5-CHICK OLF5-RAT OLF7-RAT OLF8-RAT OLF9-RAT OLFD-CANFA OXB2-HUMAN OXB4-HUMAN OXE2-HUMAN OXE2-RAT OXI1-HUMAN OXI2-HUMAN OYA1-HUMAN OYD1-HUMAN P70526 Q62007 Q62942 Q62943 Q62944 Q63394 Q63395 Q90426 Q90428 Q98SH0 Q99NH4 Q9D3U9 Q9D4F9 Q9DFC7 Q9DFC8 Q9DGH4 Q9EP55 Q9EP67 Q9EPF5 Q9EPF6 Q9EPF7 Q9EPF8 Q9EPF9 Q9EPG0 Q9EPG1 Q9EPG2 Q9EPG3 Q9EPG4 Q9EPG5 Q9EPG6 Q9EPN8 Q9EPN9 Q9EPV0 Q9EPV1 Q9EQ84 Q9EQ85 Q9EQ86 Q9EQ87 Q9EQ88 Q9EQ89 Q9EQ90 Q9EQ91 Q9EQ92 Q9EQ93 Q9EQ94 Q9EQ95 Q9EQ96 Q9EQ97 Q9EQ98 Q9EQ99 Q9EQA0 Q9EQA1 Q9EQA2 Q9EQA3 Q9EQA4 Q9EQA5 Q9EQA6 Q9EQA7 Q9EQA8 Q9EQA9 Q9EQB0 Q9EQB1 Q9EQB3 Q9EQB4 Q9EQB5 Q9EQB6 Q9EQB7 Q9EQB8 Q9EQG1 Q9EQQ5 Q9EQQ6 Q9EQQ7 Q9ERU6 Q9GKV8 Q9GZK1 Q9GZK6 Q9H206 Q9H2C5 Q9H2C6 Q9H2C8 Q9H339 Q9H342 Q9H345 Q9I835 Q9I8B6 Q9I8B7 Q9I8B9 Q9I8C0 Q9I8C1 Q9I8C3 Q9I8C4 Q9I8Y7 Q9I8Y8 Q9I8Y9 Q9I8Z0 Q9I8Z1 Q9I8Z3 Q9I8Z4 Q9I8Z5 Q9I8Z6 Q9I8Z7 Q9I9B2 Q9IBE0 Q9IBE1 Q9IBE2 Q9IBE3 Q9IBE4 Q9JHB2 Q9JHE2 Q9JHW3 Q9JKA6 Q9JM16 Q9PRJ2 Q9PSJ1 Q9PSJ2 Q9PSJ3 Q9PSJ4 Q9PSJ5 Q9PSJ6 Q9PSJ7 Q9PSJ8 Q9PSJ9 Q9PSU3 Q9PSY2 Q9PVP1 Q9PVP2 Q9PVP3 Q9PVP4 Q9PVP5 Q9PVP6 Q9PVP7 Q9PVU1 Q9PVU2 Q9PVW1 Q9PVW2 Q9QW34 Q9QW35 Q9QW36 Q9QW37 Q9QW38 Q9QWU6 Q9QY00 Q9QZ17 Q9QZ18 Q9QZ20 Q9QZ21 Q9QZ22 Q9R0K1 Q9R0K2 Q9R0K3 Q9R0K4 Q9R0K5 Q9R0Z2 Q9TQX4 Q9TSM7 Q9TSM8 Q9TSN0 Q9TU84 Q9TU86 Q9TU88 Q9TU90 Q9TU92 Q9TU93 Q9TU94 Q9TU95 Q9TU97 Q9TU99 Q9TUA0 Q9TUA1 Q9TUA2 Q9TUA3 Q9TUA4 Q9TUA6 Q9TUA7 Q9TUA8 Q9TUA9 Q9WU86 Q9WU88 Q9WU89 Q9WU90 Q9WU91 Q9WU93 Q9WU94 Q9WV09 Q9WV11 Q9WV13 Q9WV14 Q9WVD7 Q9WVD8 Q9WVD9 Q9WVN4 Q9WVN5 Q9WVN6 Q9YH55 Q9YH79 Q9YH80 Q9YHY2 Q9YHY3 Q9Z1V0
Prostanoid	**O00326 PD2R-MOUSE PE22-MOUSE PE23-BOVIN PE23-HUMAN PE23-RABIT PF2R-MOUSE PI2R-BOVIN Q9R261 TA2R-BOVIN** O00325 O15191 O35932 O46657 O75228 PD2R-HUMAN PE21-HUMAN PE21-MOUSE PE21-RAT PE22-CANFA PE22-HUMAN PE22-RAT PE23-MOUSE PE23-PIG PE23-RAT PE24-HUMAN PE24-MOUSE PE24-RABIT PE24-RAT PF2R-BOVIN PF2R-HUMAN PF2R-RAT PF2R-SHEEP PI2R-HUMAN PI2R-MOUSE PI2R-RAT Q9BGL8 Q9D627 Q9TU16 TA2R-CERAE TA2R-HUMAN TA2R-MOUSE TA2R-RAT
Nucleotide-like	**AA1R-BOVIN AA1R-RAT AA2A-RAT O57466 P2Y3-MELGA P2Y6-HUMAN P2YR-RAT Q99MT6 Q9ERK9 Q9H1C0** AA1R-CANFA AA1R-CAVPO AA1R-CHICK AA1R-HUMAN AA1R-RABIT AA2A-CANFA AA2A-CAVPO AA2A-HUMAN AA2A-MOUSE AA2B-CHICK AA2B-HUMAN AA2B-MOUSE AA2B-RAT AA3R-CANFA AA3R-HUMAN AA3R-RABIT AA3R-RAT AA3R-SHEEP GPRZ-HUMAN GPRZ-MOUSE O00398 O08766 O35811 P2UR-HUMAN P2UR-MOUSE P2UR-RAT P2Y3-CHICK P2Y4-HUMAN P2Y5-CHICK P2Y5-HUMAN P2Y6-RAT P2Y8-XENLA P2Y9-HUMAN P2YR-BOVIN P2YR-CHICK P2YR-HUMAN P2YR-MELGA P2YR-MOUSE Q9BXA5 Q9BXC1 Q9BYU4 Q9CPZ4 Q9DE05 Q9JJS7 Q9N1U0 Q9PU18 Q9R202 Q9W6C4