Modelling Proteolytic Enzymes With Support Vector Machines

Lionel Morgado[1*], Carlos Pereira[1,2], Paula Veríssimo[3], António Dourado[1]

[1]Center for Informatics and Systems of the University of Coimbra Polo II - University of Coimbra, 3030-290 Coimbra, Portugal

[2]Instituto Superior de Engenharia de Coimbra Quinta da Nora, 3030-199 Coimbra, Portugal

[3]Department of Biochemistry and Center for Neuroscience and Cell Biology of the University of Coimbra, 3004-517 Coimbra, Portugal

Summary

The strong activity felt in proteomics during the last decade created huge amounts of data, for which the knowledge is limited. Retrieving information from these proteins is the next step. For that, computational techniques are indispensable. Although there is not yet a silver bullet approach to solve the problem of enzyme detection and classification, machine learning formulations such as the state-of-the-art Support Vector Machine (SVM) appear among the most reliable options. A SVM based framework for peptidase analysis, that recognizes the hierarchies demarked in the MEROPS database is presented. Feature selection with SVM-RFE is used to improve the discriminative models and build classifiers computationally more efficient than alignment based techniques.

1 Introduction

During the last decade massive amounts of protein data have been collected, making the proteomics field attractive to the data mining and the machine learning communities. The automated classification of proteins has classically been done by means of sequence alignment methods like BLAST and PSI-BLAST [1], searching for similar homologues in a database. These approaches employ considerably extensive computation, considering that the time taken to get a single prediction for a real world sample using ordinary computers can reach several minutes when large databases are utilized. In these conditions, analyzing an average size proteome with few hundred thousands of samples can take a month. It is therefore important to find other means for disclosing answers in a more acceptable period. The Support Vector Machine (SVM) [18] makes part of the most successful methods applied to protein classification and appears as a good candidate to solve the problem of peptidase categorization. Since protein classification is a fundamental task in biology, there is a vast work concerning discriminative classifiers dedicated to subjects such as homology detection [3, 4, 5, 6, 7, 8], structure recognition [9, 10, 11], and protein localization [12, 13], among others. Other important problems in molecular biology include peptidase detection and classification. Peptidases (also known as proteases or proteolytic enzymes) are proteins that can catalyze biochemical reactions like digestion, signal transduction or cell regulation, and represent around 2% of the proteins from

*To whom correspondence should be adressed. E-mail: lionel@dei.uc.pt

organisms. They are attractive drug targets since they are involved in many virus and parasite activity. Peptidase identification and characterization is crucial to understand how they work and their role in a biological system. Considering that no perfect and universal solution has yet been reached, and that the number of new proteomes is still growing, new algorithms, computationally more efficient and more accurate, are needed to extract information embedded in these data within an acceptable period. This paper presents a SVM framework specially developed for peptidase detection and classification according to the hierarchical levels of the MEROPS peptidase database [26]. In the next section, the details about the SVM models developed are exposed. Section 3 brings some concluding remarks, actual limitations and proposes improvements for future framework versions.

2 SVM Framework for Peptidase Study

The design of efficient kernels is fundamental for the SVM to generate accurate and fast classifiers able to carry out a prediction task correctly and in the shortest amount of time. Numerous features with reduced computational cost can be created. Nevertheless, only the most informative must be used, since employing a very large feature set to build a discriminator brings some drawbacks. First, the classifier becomes slower when getting a prediction as the number of features increases, and second the decision model is more susceptible to overfitting losing effectiveness to recognize new unseen instances. Feature reduction techniques are for these reasons imperative. The number of features can be decreased either by choosing a subset of features to describe the data or by projecting the original attributes to a new reduced representation, like it is done in popular projection techniques such as Multidimensional Scaling and Principal Component Analysis. The major disadvantages of projection approaches are the loss of the original meaning of the features that compromises the interpretability of the solutions, and the unavoidable need to have always the initial features before projecting them to a lower dimension space. Feature selection approaches don't suffer from these weaknesses. Recursive Feature Elimination (RFE) belongs to this group. It is an iterative procedure that at each step eliminates the least informative features, according to an evaluation criterion, stopping when a given condition is met. Ultimately, the dataset is used to create a discriminative model to distinguish between different membership classes. Inspired by RFE and the state-of-the-art SVM learning algorithm the possibility of using information from a learned decision frontier to weight the features was investigated, emerging a new technique called SVM-RFE [16]. The procedure was here applied to the problem of peptidase detection, and used to build a classifier from a large dataset initially portrayed by thousands of features extracted from the protein primary structure. Then, the feature sets found in this phase as being the ones with higher contribution for peptidase detection, were further explored to create discriminative models for peptidase categorization. Peptidase classifiers were built to recognize the classes from the MEROPS repository defined among hierarchical tiers. Catalytic types, clans and families were targeted.

2.1 Experiments and Results

The construction of the SVM framework included two stages: the creation of a SVM peptidase detector using an optimized feature set, and then the extension of the SVM framework to models capable of performing a classification according to the membership groups defined in the

MEROPS peptidase database.

2.1.1 Peptidase Detection

The SVM-RFE algorithm was applied to a dataset with a large number of features, constructed to simulate peptidase detection. For that purpose 3003 peptidases from the MEROPS database release 8.5 and 3003 non-peptidases from SCOP [17] version 1.75 were randomly collected. Initially, all proteins were subjected to a preprocessing step in order to extract features from their primary structure to be used by the SVM. The list of features computed can be checked in table 1.

The package LIBSVM version 2.9 [2] was adapted to the SVM-RFE scheme, and was after that employed with a gaussian kernel. To promote learning, the SVM cost and the width of the gaussian were tuned using an algorithm that combines a grid search with a hill-climbing approach to discover the best values for the former and the latter parameter, respectively. SVM-RFE was executed until no features remained to describe the instances, following a mixed elimination heuristic: while data had more than 30 attributes the square root of the remaining set was removed and after that a single feature per iteration.

Initially, SVM training was performed with 2/3 of the samples arbitrarily selected and the remaining 1/3 was used in the test phase.

Preliminary studies about the effect of training with normalized features, normalized instances and both normalized features and instances at the same time were made. Because no benefits were noticed from this procedure all the following steps were performed without normalization.

The discriminative capacity of the SVM classifiers was compared with the most used algorithm by the scientific community for searching sequence homologues: PSI-BLAST [1]. PSI-BLAST is a similarity based algorithm that starts by executing a string alignment between a query protein and a search database. After that, it looks for homologues among the aligned sequences with a score higher than a given threshold. This algorithm builds a probabilistic matrix called a profile that is improved by rounds. Here, PSI-BLAST was executed running 2 cycles with the test instances as queries against a database composed by the same examples utilized for SVM training. For each method TP, TN, FP, and FN were recorded (where TP is the number of true positives, TN is the number of true negatives, FP is the number of positive and FN is the number of false negatives) to compute the following performance metrics: accuracy, sensitivity, specificity, precision and the F-measure. Accuracy is defined as

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

, sensitivity is expressed by

$$sensitivity = \frac{TP}{TP + FN}$$

, specificity comes

$$specificity = \frac{TN}{TN + FP}$$

, precision is given by

$$precision = \frac{TP}{TP + FP}$$

, and finally the F-measure is computed combining precision and recall (also known as sensitivity in the binary case) according to

$$F - measure = 2 * \frac{precision * recall}{precision + recall}.$$

A 5-fold cross-validation scheme was implemented to estimate the generalization error of PSI-BLAST and the best decision hyperplane found by SVM-RFE.

With SVM-RFE was possible to create discriminative models with less features and a reduced number of support vectors than the best model attained by simply training a SVM with all features, without losing discriminative capacity. The reduced number of support vectors can be seen as a positive aspect for generalization, once the rate of samples kept as support vectors is a direct expression of the training set memorization. There is however a state after which the feature reduction significantly damages the performance of the classifier even despite the number of support vectors increases drastically. To our knowledge there is no formal metric or rule that combines complexity and recognition ability to measure how better a SVM model is than another one so, the classifier that kept the most balanced trade between a reduced complexity and a high accuracy was considered the most suitable. This happened for 148 features that belong to the following sets: amino acid composition, sequence length, isoelectric point and composition of the collocated amino acid pairs. Moreover, the average rate of training examples used by the model to define the decision hyperplane was reduced from an initial value of 42.87% to 19.36%.

Another very important remark is that the SVM model trained with the best features recognizes more accurately the membership of the test examples than PSI-BLAST in this task (see table 2). To assert with confidence that one is better than the other, we used the statistical test defined in [24]. This test permits to calculate the confidence (1- η) by applying the formula:

$$(1 - \eta) = 0.5 + 0.5erf(z_\eta / sqrt(2))$$

, with $z_\eta = \varepsilon t / sqrt(\nu)$, and where t is the number of test examples, ν is the total number of errors (or rejections) that only one of the two classifiers makes, ε is the difference in error rate (or in rejection rate), and $erf(x) = \int_0^x exp(-t^2)\,\mathrm{d}t$ is the error function. This assumes independent identically distributed errors, one-sided risk and the approximation of the Binomial law by the Normal law.

The confidence obtain was nearly 1 (something expected once SVM-RFE outperformed PSI-BLAST in all cross-validation experiments), confirming the SVM as a good alternative to alignment based techniques. Moreover, considering that the MEROPS data bank was built using alignment based approaches, the higher sensitivity (correct recognition of peptidases) and lower specificity (correct classification of proteins as not being peptidases) of PSI-BLAST judged against the discriminative classifiers, suggests that in this kind of tests it may have some advantage over SVMs that is not directly related with the recognition of biological patterns but rather the way how the membership groups inside the repository were formed. Anyway, this was not enough for PSI-BLAST to outperform the SVM models in terms of recognition ability.

No less significant are the results for processing time needed to get a prediction (see table 3), calculated for the test set proteins: the optimized SVM classifier was on average 18.66 times faster than PSI-BLAST.

Table 1: Set of features computed from protein primary structure.

Set	Designation	#Feats	Description	Reference
1	AA Composition	20	Account of each aa available in the protein.	[27,28,29]
2	Sequence Length	1	Total number of aas that compose a protein.	[27,28,30]
3	Molecular Weight	1	Protein molecular weight considering contribution of each unit.	[27]
4	Sequence Isoelectric Point	1	Sequence isoelectric point calculation considering its N aas.	[27,28,30]
5	Sequence Average Charge	1	Estimated charge for typical intracellular pH 7.2.	-
6	Composition of Collocated AA Pairs	2000	Account of the dipeptides with gaps between each unit. For each gap size 400 pairs can be defined. Gaps between 0 and 4 were considered.	[20]
7	2-D Structure Probabilities	15	Mean, variance, standard deviation, skewness and kurtosis for the propensity of each aa to as- sume a given 2-D structure (alpha-strand, beta-sheet or turn) according to Chou-Fasman.	[29]
8	Composition Statistics	100	Mean, variance, standard deviation, skewness and kurtosis for each of the 20 amino acids that may compose a protein.	-
9	Physicochemical Properties	80	Autocorrelation coefficients derived from 8 physicochemical properties: aliphatic, tiny, small, aromatic, non-polar, charged, polar and positive. This characterization is non-exclusive (each aa can be associated with more than one group). Lags between 0 and 9 were used.	[29]
10	Radical Group	10	Autocorrelation coefficients derived from 5 mutually exclusive groups encoding aas according to radical groups (non-polar aliphatic, polar uncharged, positively charged, negatively charged and aromatic). Autocorrelation was applyed with lags ranging from 0 to 9.	[20,23]
11	Electronic Groups	10	Autocorrelation coefficients derived from a mutually exclusive 5 groups aa encoding based on electric properties (electron donor, weak electron donor, electron acceptor, negatively charged and neutral). The autocorrelation function with lags from 0 to 9.	[23,27,28]
12	Hydropathy	20	The hydropathy index is a number representing the hydrophobic or hydrophilic properties of aa side-chain. Hydropathy indexes are derived from Kyte and Doolittle charts and Eisenberg consensus scale (ECS). They were used to compute autocorrelation coefficients considering 0 to 9 lags.	[27]

Table 2: Best results for the algorithms studied during the development of the peptidase detector: SVM (without feature selection), SVM-RFE (148 features) and PSI-BLAST. The parameter 'SVs rate' refers to the rate of train examples kept as support vectors by a trained model, which is not applicable (NA) to PSI-BLAST. Mean, Max, Min and Std, stand respectively for the mean, the maximum, the minimum and the standard deviation of the quality metric determined from the 5-fold cross-validation procedure accomplished.

Quality measure[%]		SVM	SVM-RFE	PSI-BLAST
Accuracy	Mean	93.76	95.77	92.55
	Max	94.75	96.09	93.44
	Min	92.84	95.34	91.95
	Std	0.70	0.32	0.55
Sensitivity	Mean	93.83	95.66	99.74
	Max	95.32	96.61	100.00
	Min	92.79	94.46	99.52
	Std	0.98	0.78	0.18
Specificity	Mean	93.68	95.86	86.76
	Max	94.15	96.27	97.89
	Min	92.89	95.52	85.62
	Std	0.47	0.35	1.00
Precision	Mean	93.67	95.86	88.34
	Max	94.56	96.08	89.81
	Min	92.79	95.55	86.90
	Std	0.70	0.22	1.07
F-measure	Mean	93.75	95.76	93.69
	Max	94.94	96.22	94.41
	Min	92.79	95.26	92.92
	Std	0.82	0.37	0.54
SVs rate	Mean	42.87	19.36	NA
	Max	43.79	20.67	NA
	Min	41.66	18.53	NA
	Std	0.90	0.85	NA

Table 3: Processing time (in seconds) to get a prediction for a single protein sequence: optimized SVM classifier versus PSI-BLAST. Mean, Max, Min and Std, are by this order the mean, the maximum, the minimum and the standard deviation for the times collected for 2002 test sequences.

Algorithm	Mean	Max	Min	Std
SVM (148 features)	0.124	0.483	0.050	0.047
PSI-BLAST	2.314	4.711	0.112	0.898

2.2 Peptidase Categorization

Unfortunately, SVM-RFE is associated to a heavy processing time and is unfeasible for the large scale and huge multiclass problem posed by the MEROPS repository (hundreds of thousands of proteins belonging to hundreds of membership groups). The technique was avoided and instead the set of features the algorithm revealed in the previous stage as being the most relevant in the peptidase detection problem was computed for this extended assignment. The multiclass system was erected to recognize a total of 7 catalytic types, 51 clans and 209 families, by training SVM classifiers according to an all-versus-all strategy.

Approximately 20% of all sequences stored in the database were used. They were randomly selected but respecting the proportion of each group in the repository. Training used 2/3 of the samples and testing the remaining 1/3. The discriminative ability was measured using accuracy as a quality metric.

Once again, the performance of the classifiers was compared to the one from PSI-BLAST. PSI-BLAST executed 2 search cycles using the test proteins as queries against the train set utilized for SVM training.

The general accuracy for the experts can be checked in table 4. It shows that the SVM was not so effective in this last task as PSI-BLAST. Still, the classifiers remain as a low computational cost complement to alignment algorithms, or even a sustainable alternative for high confidence predictions. More detailed information about the models performance is registered in tables 5 to 13. There is observable that for some classes the detection capacity was very low or even zero. A miticulous analysis revealed that more than 90% of the classes without detections used less than 6 samples for training. On the other hand, many classes with 100% accuracy used an equally reduced number of examples for training. Consequently, although some other memberships used few hundred samples for training, we cannot say that the bad results for the minor size classes are due to the presence of strongly unbalanced groups but rather that the distribution of the examples influenced learning. Typically, in this kind of studies the classes with few units are excluded. However, because a complete expert system must include all of them, we decided not to make such excision. Despite not being conventional, improving the accuracy described in this preliminary work, may demand that future models use all instances during the learning of the injured classes. This methodology won't have a significant impact in the general accuracy and is expected to promote learning by providing potentially missing patterns.

3 Conclusions

To our knowledge, this was the first work presenting a SVM based system for peptidase detection and classification in agreement with the MEROPS taxonomy. The SVM classifiers showed ability to detect subtle patterns when dealing with examples not considered by the MEROPS data bank. The benefit of using SVMs for protease examination is emphasized by its superior capacity to distinguish between peptidases and non-peptidases, where the approach gets results that outperform PSI-BLAST in terms of recognition. The possibility that SVM classifiers offer to get a prediction in a very short time against the time spent by alignment techniques that can take several seconds or even some minutes is an important functional aspect (a speedup

of nearly 20 times). Our contribution opens the possibility to decrease the overall processing time needed to analyze very large collections of proteins like entire proteomes, by combining SVM classifiers for peptidase detection with PSI-BLAST for an extended analysis of those cases which show a higher potential to be of major interest. A rough estimation points to a time reduction from several days or weeks to few hours for proteomes with few hundred thousand samples. Another key topic for future work is the adaptation of the framework to the paradigms of high concurrency and processing parallelization to decrease the considerable computation time needed for very large jobs which are common in proteomics. In this stage, the use of graphics processing units and standards such as MPI and OpenMP, for local and distributed computation parallelization, may come into play to aid solving this issue.

Acknowledgements

This work was supported by Fundação para a Ciência e Tecnologia and FEDER through Program COMPETE (QREN) under the project FCOMP-01-0124-FEDER-010160 (PTDC/EIA/71770/2006), designated BIOINK – Incremental Kernel Learning for Biological Data Analysis.

References

[1] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. Lipman: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res 25:3389-3402, 1997.

[2] C. Chang and C. Lin: LIBSVM: a Library for Support Vector Machines, 2004.

[3] T. Jaakkola, M. Diekhans and D. Haussler: Using the Fisher Kernel Method to Detect Remote Protein Homologies. Proc Int Conf Intell Syst Mol Biol, 1999.

[4] A. Krogh, M. Brown, I. Mian, K. Sjolander and D. Haussler: Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J Mol Biol 235:1501-1531, 1994.

[5] R. Kuang, E. Ie, K. Wang, M. Siddiqi, Y. Freund and C. Leslie: Profile-Based String Kernels for Remote Homology Detection and Motif Extraction. J Bioinform Comput Biol 3:527-550, 2005.

[6] C. Leslie, E. Eskin and W. Noble: The Spectrum Kernel: a String Kernel for SVM Protein Classification. Proc Pac Symp Biocomput 7:564-575, 2002.

[7] C. Leslie, E. Eskin, A. Cohen, J. Weston and W. Noble: Mismatch String Kernels for Discriminative Protein Classification. Bioinform 20:467-476, 2004.

[8] I. Melvin, E. Ie, R. Kuang, J. Weston, W. Noble and C. Leslie: Svm-fold: a Tool for Discriminative Multi-class Protein Fold and Superfamily Recognition. BMC Bioinform 8(4), 2007.

[9] Z. Aydin, Y. Altunbasak, I. Pakatci and H. Erdogan: Training Set Reduction Methods for Protein Secondary Structure Prediction in Single-Sequence Condition. Proc 29th Annual Int Conf IEEE EMBS, 2007.

[10] L. Kurgan and K. Chen: Prediction of Protein Structural Class for the Twilight Zone Sequences. Biochem Biophys Res Commun 357(2):453-60, 2007.

[11] J. Cheng and P. Baldi: A Machine Learning Information Retrieval Approach to Protein Fold Recognition. Bioinform 22(12):1456-1463, 2006.

[12] S. Mei andW. Fei: Amino Acid Classification Based Spectrum Kernel Fusion for Protein Subnuclear Localization. BMC Bioinform 11(1):S17, 2010

[13] P. Du and Y. Li: Prediction of Protein Submitochondria Locations by Hybridizing Pseudo-amino acid Composition with Various Physicochemical Features of Segmented Sequence. BMC Bioinform 7:518, 2006.

[14] G. Lanckriet, M. Deng, N. Cristianini, M. Jordan and W. Noble: Kernel-based Data Fusion and Its Application to Protein Function Prediction in Yeast. Pac Symp Biocomput: 300-311, 2004.

[15] R. Kuang, J. Gu, H. Cai and Y. Wang: Improved Prediction of Malaria Degradomes by Supervised Learning with SVM and Profile Kernel,Genetica 36(1):189-209, 2009.

[16] I. Guyon, J. Weston, S. Barnhill and V. Vapnik: Gene Selection for Cancer Classification Using Support Vector Machines. Mach Learn 46:389-422, 2002.

[17] A. Murzin, S. Brenner, T. Hubbard and C. Chothia: SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structure. J Mol Biol 247:536-540, 1995.

[18] V. Vapnik: Statistical Learning Theory. Wiley, New York, 1998.

[19] S. Niijima and S. Kuhara: Recursive Gene Selection Based on Maximum Margin Criterion: a Comparison with SVM-RFE. BMC Bioinform 7, 2006.

[20] K. Chen, L. Kurgan and J. Ruan: Optimization of the Sliding Window Size for Protein Structure Prediction. Int Conf Comput Intell Bioinfo Comput Biol: 366-372, 2006.

[21] Y. Tang, Y. Zhang and Z. Huang: Development of two-stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis. IEEE/ACM TransacComputBiol-Bioinform 4:365-381, 2007.

[22] R. Varshavsky, M. Fromer, A. Man and M. Linial: When Less is More: Improving Classification of Protein Families with a Minimal Set of Global Features. Lect Notes in Computer Science 4645:12-24, 2007.

[23] Website of the Laboratory of Mass Spectrometry and Gaseous Ion Chemistry of the University of Rockefeller: http://prowl.rockefeller.edu Accessed 1 October, 2009.

[24] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik: What size test set gives good error rate estimates?. PAMI 20 (1): 52-64, 1998.

[25] X. Yang and B. Wang: Weave Amino Acid Sequences for Protein Secondary Structure Prediction. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery:80-8, 2003.

[26] N. Rawlings, A. Barrett and A. Bateman: MEROPS: the Peptidase Database.Nucleic Acids Res 38, 2010.

[27] L. Kurgan and L. Homaeian: Prediction of Structural Classes for Protein Sequences and Domains: Impact Prediction Algorithms, Sequence Representation and Homology, and Test Procedures on Accuracy. Pattern Recognit 39(12):2323-2343, 2006.

[28] K. Kedarisetti, L. Kurgan and S. Dick: Classifier Ensembles for Protein Structural Class Prediction with Varying Homology. Biochem. and Biophys. Res Communications 384:981-988, 2006.

[29] S. Muskal and S. Kim: Predicting Protein Secondary Structure Content: a Tandem Neural Network Approach. J Mol Biol 225:713-727, 1992.

[30] U. Hobohm and C. Sander: A Sequence Property Approach to Searching Protein Databases. J Mol Biol 251:390-399, 1995.

Table 4: Accuracy values for the SVM system and PSI-BLAST.

MEROPS hierarchical level	SVM Accuracy [%]	PSI-BLAST Accuracy [%]
Catalytic type	74.62	98.96
Clan	78.82	99.45
Aspartic families	96.75	1000.0
Cysteine families	86.45	100.00
Glutamic families	100.00	100.00
Metallo families	86.03	100.00
Serine families	83.38	100.00
Threonine families	98.32	100.00
Unknown catalytic type families	96.60	100.00

Table 5: Accuracy values for the SVM system by clans.

#	Clans	Accuracy [%]	#	Clans	Accuracy [%]	#	Clans	Accuracy [%]
1	AA	86.40	18	GA	80.00	35	PB	73.06
2	AB	0.00	19	MA	86.58	36	PC	76.42
3	AC	94.44	20	MC	68.27	37	SB	82.56
4	AD	93.18	21	MD	65.33	38	SC	80.07
5	AE	61.11	22	ME	74.62	39	SE	61.94
6	AF	100.00	23	MF	84.34	40	SF	79.61
7	A-	100.00	24	MG	85.02	41	SH	40.00
8	CA	82.35	25	MH	78.01	42	SJ	88.73
9	CD	52.17	26	MJ	71.72	43	SK	79.62
10	CE	76.74	27	MK	65.79	44	SP	55.56
11	CF	55.00	28	MM	86.43	45	SQ	70.59
12	CH	36.36	29	MN	75.00	46	SR	78.13
13	CL	46.88	30	MO	85.86	47	SS	31.03
14	CM	100.00	31	MP	79.69	48	ST	84.88
15	CN	100.00	32	MQ	45.00	49	S-	94.44
16	CO	85.05	33	M-	62.86	50	T-	100.00
17	C-	66.67	34	PA	87.11	51	U-	65.00

Table 6: Accuracy values for the SVM system by catalytic types.

#	C. type	Accuracy [%]	#	C. type	Accuracy [%]	#	C. type	Accuracy [%]
1	A	70.86	4	M	77.10	7	U	44.19
2	C	69.09	5	S	78.40			
3	G	100.00	6	T	67.80			

Table 7: Accuracy values for the SVM system for families from catalytic type A.

#	Catalytic type	Family	Accuracy [%]	#	Catalytic type	Family	Accuracy [%]
1	A	A1	99.22	9	A	A8	100.00
2		A2	82.93	10		A22	96.67
3		A3	33.33	11		A24	93.22
4		A9	100.00	12		A25	100.00
5		A11	97.98	13		A31	96.77
6		A33	100.00	14		A26	100.00
7		A6	100.00	15		A5	100.00
8		A21	100.00				

Table 8: Accuracy values for the SVM system for families from catalytic type C.

#	Catalytic type	Family	Accuracy [%]	#	Catalytic type	Family	Accuracy [%]
1	C	C1	94.83	37	C	C15	65.00
2		C2	92.86	38		C46	63.64
3		C10	50.00	39		C60	50.00
4		C12	84.62	40		C82	76.47
5		C16	33.33	41		C18	100.00
6		C19	95.83	42		C9	100.00
7		C28	0.00	43		C40	88.79
8		C39	78.79	44		C6	100.00
9		C47	100.00	45		C7	0.00
10		C51	85.71	46		C8	0.00
11		C54	76.47	47		C21	100.00
12		C58	50.00	48		C23	100.00
13		C64	100.00	49		C27	100.00
14		C65	87.50	50		C31	0.00
15		C66	100.00	51		C32	0.00
16		C67	100.00	52		C33	0.00
17		C71	0.00	53		C36	0.00
18		C76	50.00	54		C42	0.00
19		C78	85.71	55		C53	100.00
20		C83	50.00	56		C70	0.00
21		C85	66.67	57		C74	0.00
22		C86	85.71	58		C75	100.00
23		C87	0.00	59		C84	0.00
24		C88	85.71	60		C3	90.00
25		C11	80.00	61		C4	100.00
26		C13	71.43	62		C24	100.00
27		C14	76.92	63		C30	100.00
28		C25	50.00	64		C37	100.00
29		C50	44.44	65		C62	100.00
30		C80	0.00	66		C44	92.42
31		C5	100.00	67		C45	37.50
32		C48	87.88	68		C59	88.89
33		C55	100.00	69		C69	85.71
34		C57	50.00	70		C89	80.00
35		C63	100.00	71		C26	82.90
36		C79	100.00	72		C56	93.95

Table 9: Accuracy values for the SVM system and PSI-BLAST.

#	Catalytic type	Family	Accuracy [%]
1	G	G1	100.00

Table 10: Accuracy values for the SVM system for families from catalytic type M.

#	Catalytic type	Family	Accuracy [%]	#	Catalytic type	Family	Accuracy [%]
1	M	M1	96.90	29	M	M64	66.67
2		M2	90.00	30		M66	100.00
3		M3	87.50	31		M72	33.33
4		M4	83.33	32		M78	25.00
5		M5	0.00	33		M14	79.81
6		M6	66.67	34		M15	81.82
7		M7	100.00	35		M74	90.00
8		M8	87.50	36		M16	90.42
9		M9	100.00	37		M44	66.67
10		M10	84.93	38		M17	89.16
11		M11	0.00	39		M24	91.50
12		M12	94.53	40		M18	88.89
13		M13	84.85	41		M20	86.94
14		M26	33.33	42		M28	76.53
15		M27	100.00	43		M42	77.78
16		M30	0.00	44		M19	72.09
17		M32	80.95	45		M38	75.74
18		M34	0.00	46		M22	78.95
19		M35	66.67	47		M50	95.71
20		M36	100.00	48		M55	62.50
21		M41	97.27	49		M23	94.76
22		M43	63.64	50		M67	95.31
23		M48	86.15	51		M29	45.00
24		M54	71.43	52		M49	63.64
25		M56	85.00	53		M73	100.00
26		M57	100.00	54		M75	100.00
27		M60	33.33	55		M76	100.00
28		M61	66.67	56		M77	57.14

Table 11: Accuracy values for the SVM system for families from catalytic type S.

#	Catalytic type	Family	Accuracy [%]	#	Catalytic type	Family	Accuracy [%]
1	S	S1	93.24	24	S	S12	66.67
2		S3	100.00	25		S13	100.00
3		S6	80.00	26		S24	33.33
4		S7	100.00	27		S26	25.00
5		S29	100.00	28		S21	79.81
6		S30	60.00	29		S16	81.82
7		S31	100.00	30		S50	90.00
8		S32	0.00	31		S69	90.42
9		S39	33.33	32		S14	66.67
10		S46	77.78	33		S41	89.16
11		S55	83.33	34		S49	91.50
12		S64	0.00	35		S59	88.89
13		S45	76.92	36		S58	86.94
14		S51	33.33	37		S60	76.53
15		S8	91.63	38		S66	77.78
16		S53	70.00	39		S54	72.09
17		S9	78.01	40		S48	75.74
18		S10	77.14	41		S62	78.95
19		S15	73.68	42		S63	95.71
20		S28	73.09	43		S68	62.50
21		S33	78.68	44		S71	94.76
22		S37	50.00	45		S72	95.31
23		S11	88.00	46		S73	45.00

Table 12: Accuracy values for the SVM system and PSI-BLAST.

#	Catalytic type	Family	Accuracy [%]	#	Catalytic type	Family	Accuracy [%]
1	T	T1	99.47	4	T	T4	100.00
2		T2	93.94	5		T5	97.67
3		T3	98.86				

Table 13: Accuracy values for the SVM system and PSI-BLAST.

#	Catalytic type	Family	Accuracy [%]	#	Catalytic type	Family	Accuracy [%]
1	U	U4	60.00	7	U	U49	0.00
2		U9	100.00	8		U56	60.00
3		U32	100.00	9		U57	50.00
4		U35	100.00	10		U62	98.81
5		U40	100.00	11		U68	100.00
6		U48	100.00	12		U69	100.00