Automated Structural Classification of Proteins by Using Decision Trees and Structural Protein Features

Slobodan Kalajdziski, Bojan Pepik, Ilinka Ivanovska, Georgina Mirceva, Kire Trivodaliev, and Danco Davcev

Ss. Cyril and Methodious University, Faculty of Electrical Engineering and Information Technologies, Karpos 2 bb, 1000 Skopje, Macedonia
{skalaj,georgina,kiret,etfdav}@feit.ukim.edu.mk,
bojan.pepik@gmail.com, ilinka.ivanovska@yahoo.com

Abstract. The protein function is tightly related to classification of proteins in hierarchical levels where proteins share same or similar functions. One of the most relevant protein classification schemes is the structural classification of proteins (SCOP). The SCOP scheme has one negative drawback; due to its manual classification methods, the dynamic of classification of new proteins is much slower than the dynamic of discovering novel protein structures in the protein data bank (PDB). In this work, we propose two approaches for automated protein classification. We extract protein descriptors from the structural coordinates stored in the PDB files. Then we apply C4.5 algorithm to select the most appropriate descriptor features for protein classification based on the SCOP hierarchy. We propose novel classification approach by introducing a bottom-up classification flow, and a multi-level classification approach. The results show that these approaches are much faster than other similar algorithms with comparable accuracy.

Keywords: Structural Classification of Proteins (SCOP), C4.5 Classification, Protein function prediction.

1 Introduction

Proteins play vital structural and functional role in every cell in living organisms. They are constructed by long chains of amino acid residues folding into complex three-dimensional polypeptide chain structures. This three-dimensional representation of a residue sequence and the way this sequence folds in the 3D space are very important to understand the logic in which a function of a protein is based on. In fact, the concept of function typically acts as an umbrella term for all types of activities that a protein is involved in, be it cellular, molecular or physiological. Also, evolutionary evidence could potentially be derived from conserved protein structures existed in multiple spices. The knowledge of protein function is a crucial link in the development of new drugs, better crops, and even development of synthetic biochemicals.

Since the determining of the first 3D structure of the protein *myoglobin*, up to now, the complexity and the variety of the protein structures has increased as the number of the new determined macromolecules has. Therefore, a need for a classification of proteins is obvious, which may result in a better understanding of these complicated three-dimensional structures, their functions, and the deeper evolutionary procedures that led to their creation. In molecular biology, many classification schemes and databases (CATH [16], FSSP [15] and SCOP [2]) have been developed in order to describe the different kinds of similarity between proteins.

The Structural Classification of Proteins - SCOP database [2] describes the evolutionary relationships between proteins of known structure. It has been accepted as the most relevant and the most reliable classification hierarchy [3]. This is due to the fact that SCOP strictly builds its classification decisions based on visual observations of the structural elements of the proteins made by human experts. Therefore, this manual approach during the classification process of new structures clarifies that SCOP is completely biased towards reliable and precise protein classification. The main levels of the SCOP hierarchy are Family, Superfamily, Fold, and Class. Using the terminology of the SCOP database, two proteins that belong to the same fold share a common three-dimensional pattern with the same major secondary structure elements (SSEs) in the same arrangement with the same topological connections. In the SCOP hierarchy, folds are grouped into different classes, where a class is defined by the topographical arrangement of the secondary structures of its member proteins. Although SCOP is highly reliable and precise system, it has one negative drawback. Namely, due to its manual classification methods, the dynamic of classification of new proteins in SCOP can't follow the dynamic of discovering novel protein structures stored in PDB (38.200 proteins classified in SCOP vs. 59.800 protein entries in PDB in August 2009). This clearly brings in front the necessity of a system which will classify proteins in a precise and reliable manner as SCOP does, but in an automated fashion.

There are various approaches for protein classification which are trying to offer efficient and completely automated protein classification. These approaches have different characteristics in terms of algorithm for determining protein similarity. Basically, the protein similarity metric used defines the complexity and the efficiency of the classification approach.

One way to determine protein similarity is to use sequence alignment algorithms like Needleman–Wunch [19], BLAST [18], PSI-BLAST [17] etc. They offer fast and efficient recognition of overlapping subsequences in two protein structures which leads to detection of closely related protein structures, but these methods cannot recognize proteins with remote homology.

Instead of sequence alignment methods, structure alignment methods like CE [5], MAMMOTH [6], DALI [7], etc. are used to detect and highlight distant homology relations between protein structures. In general these methods are very precise and efficient and they have high degree of successful mapping of existing structures in new proteins. Structure alignment methods perform one-against-all proteins comparison in order to find the most similar existing protein to a novel protein structure. Having in mind that the number of classified proteins, for example in SCOP, is ever increasing and that structure alignment methods are quite cost expensive, the speed of classification with these methods is always questioned. For

example CE takes 209 days [5] to classify 11.000 novel protein structures. The bottom-up classification approaches proposed in this paper took around 6 hours to classify 9.994 proteins.

There are numerous research approaches that combine sequence and structure alignment of the proteins. SCOPmap [8] is a system that uses a pipelined architecture for the classification. SCOPmap uses four sequence alignment methods: BLAST [18], PSI-BLAST [17], RPS-BLAST and COMPASS [11] and two structure alignment methods: VAST [20] and DaliLite [10]. This pipelined approach brings to front high complexity of the classification process. FastSCOP [9] is another, more efficient system than SCOPmap which is based on 3D-BLAST [21] and MAMMOTH [6]. 3D-BLAST is structure alignment method that is used as a preprocessing filter to produce the top 10 scores. Afterwards, these top 10 results are used by MAMMOTH in order to find the most similar protein structures to the query structure. Although fastSCOP possesses high precision, the used combination of methods eventually in future, considering the ever increasing number of novel proteins, will produce increasing classification complexity.

Instead of using the alignment methods, the classification based on the mapping of the protein structure in the high-dimensional uniform descriptor space can be found as very promising. In [14] protein descriptor is formed by first producing distance matrix, which is treated as image, and local and global protein features are extracted from the image histograms. The whole descriptor dimension is 33, consisted from 24 local features and 9 global features. This protein descriptor afterwards is used for protein classification into the SCOP hierarchy based on the E-predict algorithm [14]. In [1] protein descriptor is generated solely from the protein sequence information in order to avoid complex structure comparison. The protein descriptor gives information for the number of different amino acids, the hydrophobicity, the polarity, the Van der Waals volume, the polarizability and for the secondary structures in the protein structure. With this protein descriptor, proteins are classified hierarchically into the SCOP hierarchy with Naive Bayes and boosted C4.5 methods [1].

In this work, we propose two classification processes based on generated protein descriptors in combination with C4.5 decision tree classification algorithm. As a classification scheme, the SCOP classification hierarchy is used. First we introduce the classification flow that is based on a bottom-up classification according to the SCOP hierarchy. The implemented classification logic is original and new due to the fact that according to the related work there were no protein classification approaches which use similar classification architecture. Second we adjust the multi-level modification of the C.4.5 decision tree algorithm to solve the SCOP classification. The implemented approaches introduce tremendous speed up compared with the structure alignment algorithms. They are ~816 times faster than CE [5] and ~68 times faster than MAMMOTH [6]. They are less correct than fastSCOP [9] which has accuracy of 98% for the SUPERFAMILY level compared with our 84% for bottom-up approach and 80% for multi-level approach. However, our algorithms are ~70 times faster than MAMMOTH.

In section 2 we present the classification process architecture and the used classification methods. Section 3 presents the experimental results, while the section 4 concludes the paper.

2 Classification Process Architecture and Methods

The classification process architecture has three main features. First, this architecture is based, and it uses the SCOP classification scheme. Second, it uses 3D protein descriptor [13] which transforms the protein tertiary structure into N-dimensional feature vector, and additionally gives some other protein structural features. And finally, as a classification algorithm decision trees trained with C4.5 are used.

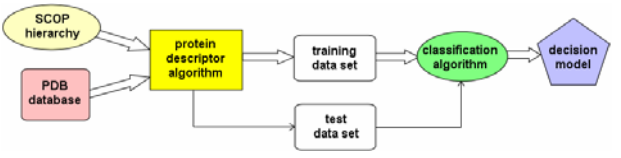

Fig. 1. Classification process architecture

Chronologically (as can be seen from the Fig. 1) the system is consisted of two phases: training phase and testing or classification phase. The training or offline data flow takes into consideration the knowledge given in the SCOP hierarchical database to build predicative classification flow for each SCOP hierarchy level. The training procedure can be divided in two general processes. The first one is the descriptor extraction (shown on Fig. 2) and data set generation process. Descriptors consisting of 450 features (416 of them describe the protein's geometry, while 34 of them give information for the primary and secondary protein structure) are generated for each protein forming a training set for the C4.5 decision tree algorithm. This descriptor relies on the geometric 3D structure of the proteins. After triangulation, normalization and voxelization of the 3D protein structures, the Spherical Trace Transform is applied to them to produce geometry - based descriptors, which are completely rotation invariant.

The second process is the process of forming and training of the decision trees for the protein classifiers. We propose two approaches in solving this classification task.

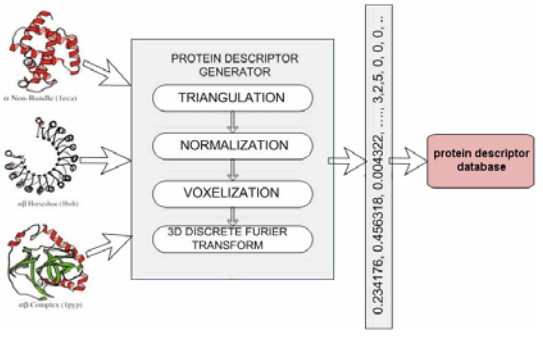

Fig. 2. Protein descriptor generation process

First one is the bottom-up classification approach, while the second one is the multi-level modification of the C4.5 algorithm. The classification logic of both classifiers is based on the fact that the SCOP classification hierarchy is tree–like hierarchy, providing only one parent node for every child node in the hierarchy. According to this fact, if we know the *domain* of the protein, then we know the upper SCOP levels (the whole hierarchy) for that protein. These two classification approaches are explained in the following subsections.

2.1 Bottom-Up Classification Approach

Basically this classification flow is by all means very much similar to the classical top – down approach, except that the starting point is changed. Instead of starting the classification from the root, it is started from the leafs.

Decision trees are built for each level of the SCOP hierarchy, providing separate trees for classification in *class*, *fold*, *superfamily*, *family* and *domain*. Also we provide additional level-specific decision trees that are trained for classification in specific

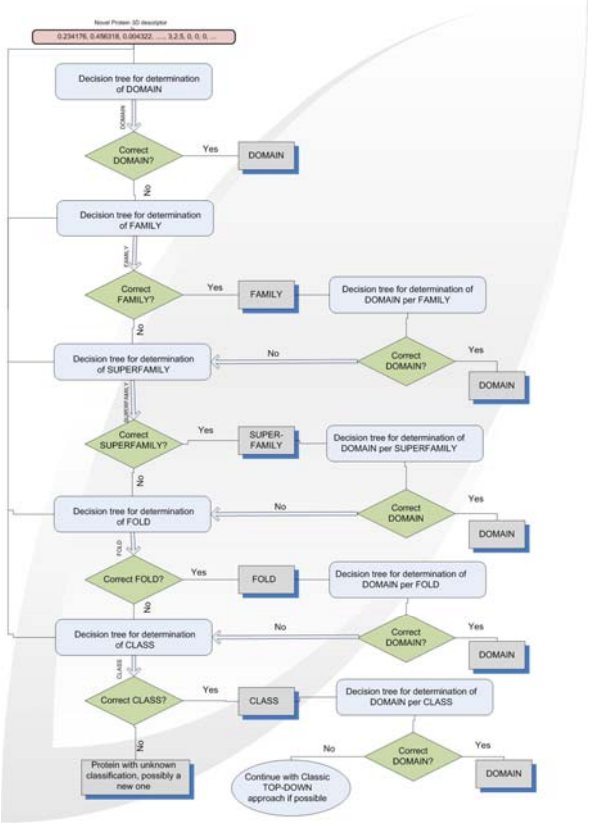

Fig. 3. Bottom-up classification approach

cases. Namely if we use *domain* specific decision tree trained for all protein instances it is obvious that this classifier will be low accurate. In cases when we know the upper level of the SCOP hierarchy of a given protein, than we can use additional decision trees trained on the subset of the protein data taking into account only the descendant proteins grouped in the lower levels of the SCOP hierarchy. In this way, there are *class* specific, *fold* specific, *superfamily* specific and *family* specific decision trees for determination of *domain*. Also, additional decision trees are built for determination of one-level up SCOP levels (*class* specific decision trees for *fold* determination etc.). These decision trees are afterwards used in the classification process of novel proteins.

As can be seen from the Fig. 3, the unknown protein is passed through the *domain* specific classifier. If the *domain* classifier correctly classifies unknown protein, then there is no need to classify the protein in the upper levels. The classification process can associate the upper hierarchy labels from the background knowledge extracted from the SCOP hierarchy. If the *domain* classifier incorrectly classifies the protein, than the protein is being preceded to the classifier into the higher level of the hierarchy, in this case the decision tree for *family* determination. If the classified *family* is correct, than the rest of the levels of the hierarchy for the protein are known, except the *domain*, which remains unknown. To correct this, we precede the protein to the level-specific decision tree for determination of *domain*, trained only with instances from the predicted *family*. In this way we lower the false positive hits in the classification flow if we use separate decision trees for separate SCOP levels. If the classified *family* is mistaken, the classifier continues one level up into the hierarchy, in this case it continues with predicting the *superfamily* level of the new protein. This step is also the next step if the *family* specific decision tree mistakes the protein *domain*. Otherwise the classification is finished. The process of protein classification continues with the backward recursion explained in the previous paragraph until it reaches the top level of the SCOP hierarchy, the *class* level. If there is no success in recognizing the correct protein *domain*, then the protein is announced as a protein with unknown SCOP label, possibly a candidate for a new label in the SCOP hierarchy.

2.2 Multi-level Modification of the C4.5 Algorithm

Traditional *single-label* classification is concerned with learning from a set of examples that are associated with a single label l from a set of disjoint labels L, $|L|>1$. If $|L|=2$, then the learning problem is called a *binary* classification problem, while if $|L|>2$, then it is called a *multi-class* classification problem. The problem of classifying proteins in SCOP hierarchy is a standard *multi-class* or *multi-level* classification problem.

For the purposes of the protein classification, we have used and readapted the modification of the C4.5 algorithm [22] for multi-label data. In order to automate the SCOP classification we need to: (1) have information about the hierarchy of classes, (2) calculate the entropy, and (3) check the membership of the new protein in the existing hierarchy. We have provided flat text file with the SCOP hierarchy labels

organized in tree-like manner. The modification of the C4.5 algorithm for multi-label data is made by changing the entropy calculation:

$$entropy(S) = -\sum_{i=1}^{N} \left(p(c_i) \log p(c_i) + q(c_i) \log q(c_i) \right) \quad (1)$$

where $p(c_i)$ = relative frequency of class c_i and $q(c_i) = 1 - p(c_i)$. Also there is allowed multiple labels in the leaves of the tree.

By applying the multi-label C4.5 classification the end result is one decision tree that can classify new protein structures in branch of the SCOP hierarchy at once.

3 Experimental Results

All of the experiments were conducted on a PC with a 2.2 GHz Intel Core 2 CPU and 2GB RAM. The SCOP hierarchy from the last two versions SCOP1.71 and SCOP 1.73 were downloaded and integrated into Oracle 10g database. The protein descriptor generation was made in C++. The bottom-up classification approach was implemented in C#.NET by using the C4.5 decision trees generated by WEKA data mining toolkit. The multi-level modification of the C4.5 algorithm was implemented in C#.NET.

In the training phase, 73.642 out of 75.930 classified protein chains from SCOP v1.71 were used for training the decision trees for both classification strategies. For the bottom-up approach, instead of having one decision tree for each level of the SCOP hierarchy, ensemble of decision trees is used for each level of the SCOP hierarchy. The idea for ensemble of trees came as a result of the huge memory requirements of the approach with one tree per level. The number of trees per ensemble in one level and number of output classes per tree in each level are presented in Table 1.

Table 1. Number of trees per ensemble in one level and number of output classes per tree in each level

Level	Classes per tree	Number of trees per level
CLASS	11	1
FOLD	256	5
SUPERFAMILY	169	12
FAMILY	341	11
DOMAIN	659	14

In the test phase, 3.576 protein chains from the SCOP v1.73 were taken. We have selected only those proteins that were not previously classified in the SCOP v1.71, and the *domain* of the selected protein from the SCOP v1.73 must be present in the SCOP v1.71 hierarchy.

The classification results for our proposed classification approaches, bottom-up classification approach and multi-level modified C4.5 approach are shown on the Table 2 and Table 3 respectively.

Table 2. Results of the classification with the bottom-up approach

Level	Correctly classified	Incorrectly classified	Accuracy
CLASS	3154	422	88,2%
FOLD	3027	549	84,65%
SUPERFAMILY	2993	583	83,69%
FAMILY	2886	690	80,7%
DOMAIN	2839	737	79,39%

Table 3. Results of the classification with the multi-level modification of C4.5

Level	Correctly classified	Incorrectly classified	Accuracy
CLASS	3050	526	85,29%
FOLD	2957	619	82,69%
SUPERFAMILY	2864	712	80,09%
FAMILY	2839	737	79,38%
DOMAIN	2821	755	78,88%

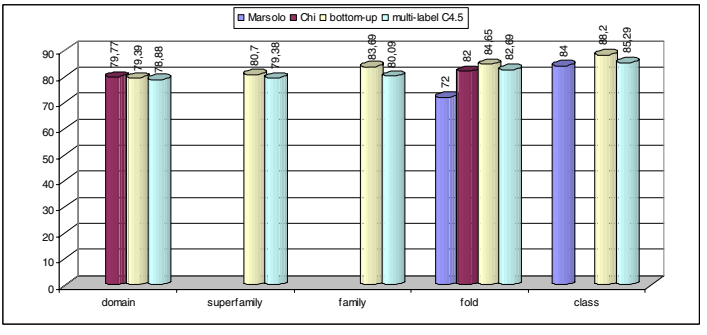

Fig. 4. Classification accuracy comparison of our approaches and the approaches presented in [14] and [1]

We have compared our results with the protein classification approaches given in [1] and [14]. The dataset used in [1] is based on SCOP v1.67 and consists of 311 training proteins taken from 27 most populated SCOP folds with no more than 35% sequence similarity between any two proteins. This approach presents only classifications by *class* and *fold* levels of the SCOP hierarchy. The dataset used in [14] is based on SCOP v1.69, and the test proteins are taken among two consequence SCOP versions. This approach provides only classification results for the *fold* level of the SCOP hierarchy. On Fig.4 we show the comparison results between classification results given in [1], [14], and our bottom-up approach and multi-level modification of C4.5 algorithm. As can be seen from the obtained results, the precision of the classification is satisfying. In classifying *fold* in [14] the precision is 92% for SCOP v1.69 when E-predict is used as a classification algorithm, but if C4.5 decision tree is used as classifier the precision for *fold* prediction is 82%. From this point of view our classifiers have comparable accuracy and produce classification results for the whole SCOP hierarchy, not by partial levels. It should be mentioned here, that fastSCOP [9]

predicts the protein *superfamily* with 98% accuracy compared with 84% accuracy in our bottom-up classifier, but our bottom-up classifier classifies proteins nearly ~70 times faster then fastSCOP thus providing much higher efficiency.

Also we have conducted the speed performance testing (as shown on the Fig. 5) of our classification approaches compared with the CE [5] and MAMMOTH [6] approaches. We have randomly selected around 10.000 proteins and passed them to all four systems. The bottom-up classification lasted 6 hours on a Intel Core 2 Duo machine with 2 GB RAM, while multi-level C4.5 modification lasted around 11,5 hours (~1,95 times slower) on the same machine. The MAMMOTH results are obtained on Intel Pentium 2.8 GHz machine, while the CE results are obtained on Sun Ultra Sparc II machine.

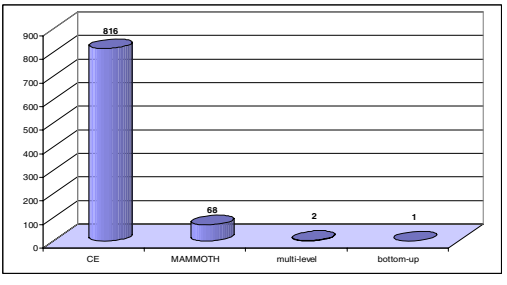

Fig. 5. Speed comparison of protein classification by using different classification approaches

4 Conclusion

The objective of this paper was to make empirical tests of the usefulness and the contribution of the different protein features to the decision tree classification precision and the usefulness of the bottom-up classification approach and multi-level modification of the C4.5 algorithm. It is evident that these approaches which use protein descriptor and decision tree algorithms for classification introduces high level of efficiency which can be concluded from the time taken to classify unknown protein and the percent of correctly classified proteins. The multi-level modification of C4.5 does not find as many rules as would be found by learning all the levels individually. This is to be expected, as the criteria for choosing nodes in the decision tree are slightly different, and a different amount of information is available.

The provided comparison with some other relevant works proves the satisfying results obtained with this work. In the future we plan to extend the classification in order to solve the problem of classification of proteins in novel SCOP branches.

References

1. Marsolo, K., Parthasarathy, S., Ding, C.: A Multi-Level Approach to SCOP Fold Recognition. In: IEEE Symposium on Bioinformatics and Bioeng., pp. 57–64 (2005)
2. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247, 536–540 (1995)

3. Camoğlu, O., Can, T., Singh, A.K., Wang, Y.F.: Decision tree based information integration for automated protein classification. Journal of Bioinformatics and Computational Biology 3(3), 717–724 (2005)
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
5. Shindyalov, H.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 9, 739–747 (1998)
6. Ortiz, A.R., Strauss, C.E., Olmea, O.: Mammoth: An automated method for model comparison. Protein Science 11, 2606–2621 (2002)
7. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology 233, 123–138 (1993)
8. Cheek, S., Qi, Y., Krishna, S.S., Kinch, L.N., Grishin, N.V.: SCOPmap: Automated assignment of protein structures to evolutionary superfamilies. BMC Bioinformatics 5, 197–221 (2004)
9. Tung, C.H., Yang, J.M.: FastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. Nucleic Acids Res. 35, W438–W443 (2007)
10. Holm, L., Sander, C.: Dali: a network tool for protein structure comparison. Trends in Biochemical Science 20, 478–480 (1995)
11. Sadreyev, R., Grishin, N.: COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. 326, 317–336 (2003)
12. Yang, J.M., Tung, C.H.: Protein structure database search and evolutionary classification. Nucleic Acids Research 34, 3646–3659 (2006)
13. Kalajdziski, S., Mirceva, G., Trivodaliev, K., Davcev, D.: Protein Classification by Matching 3D Structures. In: Frontiers in the Convergence of Bioscience and Information Technologies 2007, Jeju Island, Korea, pp. 147–152 (2007)
14. Chi, P.H.: Efficient protein tertiary structure retrievals and classifications using content based comparison algorithms. PhD thesis, University of Missouri-Columbia (2007)
15. Holm, L., Sander, C.: The FSSP Database: Fold Classification Based on Structure-Structure Alignment of Proteins. Nucleic Acids Research 24, 206–210 (1996)
16. Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH - A hierarchic classif. of protein domain structures. Structure 5(8), 1093–1108 (1997)
17. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17), 3389–3402 (1997)
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
19. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Bio. 48(3), 443–453 (1970)
20. Madej, T., Gibrat, J.F., Bryant, S.H.: Threading a database of protein cores. Proteins 23, 356–369 (1995)
21. Tung, C.H., Huang, J.W., Yang, J.M.: Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biology 8(3), 31–46 (2007)
22. Clare, A.: Machine learning and data mining for yeast functional genomics. PhD thesis, University of Wales Aberystwyth (2003)