Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 10-014

Prediction of Protein Subcellular Localization: A Machine Learning Approach

Kyong Jin Shim

June 24, 2010

CHAPTER 1

PREDICTION OF PROTEIN SUBCELLULAR LOCALIZATION: A MACHINE LEARNING APPROACH

1.1 INTRODUCTION

Over the years, large-scale genomic and proteomic efforts have produced large amounts of sequence data. One of the key challenges in the post-genomic era is to predict functions and roles of gene products. Proteins are essential to the structure and function of all living cells, and many of the them are enzymes or subunits of enzymes that catalyze chemical reactions. Other types of proteins play structural roles as well as engage in mechanical integrity and tissue signalling functions. Eukaryotic cells are comprised of various compartments that are functionally and morphologically distinct. Once different kinds of proteins are synthesized by a cell, they are targeted to one or more appropriate organelles in the cell. It is crucial that a protein is transported to its final destination in order to perform its function optimally. Subcellular localization is a key functional characteristic of proteins [1]. Therefore, the use of resources concerning subcellular localization would be useful for the assignment of functions to uncharacterized proteins.

However, the growth rate of the amount of sequence data far exceeds experimental determination of protein structure and function. Hence, much of the efforts in computational biology today in protein structure and function determination has focused on assigning the protein's putative function from sequences alone, and it is one of the most challenging problems in functional genomics [2]. Over the years, there have been many advances in annotating protein sequences. Continuing efforts in the development of sequence-based protein structure and function prediction can be expected to bring significant improvements in gene annotations.

Sequence-driven protein function prediction techniques can be broadly classified into two categories. The first major category relies on the comparison of a query sequence to other sequences of known functions. Often times, pairwise sequence similarity searches are performed using programs such as FASTA [3] and BLAST [4]. However, sequence similarity-based prediction of putative function suffers some major drawbacks [5]. First, finding the best hit among multiple hits returned by such sequence search programs requires careful interpretation of results often involving human expertise. The best hit that such programs present is only hypothetical, and poor annotation of target sequences (sequences against which similarity search is performed for a query protein) may lead to erroneous results.

The second major category of sequence-driven protein function prediction techniques relies on data mining and machine learning techniques to identify conserved patterns in sequences. The sequence-motif method is one such example. Such methods rely on characteristic signatures extracted from conserved regions in multiple sequence alignments. One of the earliest motif databases is known as PROSITE [6], and it catalogues an extensive list of many known protein families by using regular expressions or generalized profiles. Other well-known and widely used motif databases are eMOTIF [7], BLOCKS [8], SMART [9], Pfam [10], PRINTS [11], and InterPro [12]. Given a query protein sequence, searching a motif database will return a list of conserved sequence patterns and their associated functions. The search results can then be used for determining the putative function of the query protein sequence. Protein motif databases containing a substantial number of protein sequences that are well characterized serve as a good platform for applying data mining and machine learning techniques [13]. Specifically, these techniques are used to construct classifiers for predicting and assigning putative functions to query protein sequences whose functions are unknown. The construction of such classifiers takes a set of training sequences whose functions are known. The classifiers learn certain information that is encoded in the training sequences. Having learned the encoded information, the classifiers are validated using a set of testing sequences. The classifiers can then assign a query protein sequence to one of the functional categories as represented in the training sequences.

1.2 LITERATURE REVIEW

There have been many previous efforts in predicting protein subcellular localization in eukaryotic organisms. ESLPred [14] assigns eukaryotic proteins to nucleus, mitochondrion, cytoplasm or extracellular space by using Support Vector Machine and PSI-BLAST. HSLpred [15] utilizes Support Vector Machine and PSI-BLAST to generate predictions for four localization sites for human proteins. iPSORT [16] is a localization prediction program that classifies eukaryotic N-terminal sorting signals. LOCSVMPSI [17] incorporates evolutionary information into its predictions for eukaryotic localization prediction. It utilizes Support Vector Machine and PSI-BLAST to generate predictions for up to 12 localization sites. LOCtree [18] is a eukaryotic and prokaryotic localization prediction tool. NucPred [19] takes into consideration the presence of nuclear localization signals identified through a genetic programming algorithm for its classification method. predictNLS [20] is a nucleus localization prediction program that considers nuclear localization signal motifs. Predotar [21] is a localization prediction program for identifying the presence of mitochondrial and plastid targeting peptides in plant sequences. Protein Prowler [22] is a localization prediction program for classifying eukaryotic targeting signals as secretory, mitochondrion, chloroplast, or other. Proteome Analyst's Subcellular Localization Server

[23] is designed to classify Gram-negative, Gram-positive, fungi, plant, and animal proteins to several localization sites. pSLIP [24] uses Support Vector Machine and a variety of physiochemical properties of amino acid residues to assign a eukaryotic protein to one of the six localization sites. PSLT [25] is a Bayesian network-based method designed to predict human protein localization based on motif or domain co-occurrence. PSORT [26] is a localization prediction program designed for plant sequences. PSORT II [27] is a localization prediction program designed for eukaryotic sequences. pTARGET [28] takes into consideration localization-specific Pfam domains and amino acid composition to assign a eukaryotic protein sequence to one of the nine localization sites. SecretomeP [29] is a prediction program for eukaryotic proteins that are secreted via a non-traditional secretory mechanism. SignalP [30] is a prediction program for prokaryotic and eukaryotic proteins that considers traditional N-terminal signal peptides. SubLoc [31] is a localization prediction program that uses Support Vector Machine. It is designed to assign a prokaryotic protein to the cytoplasmic, periplasmic, or extracellular sites. Also, it assigns a eukaryotic protein to the cytoplasmic, mitochondrial, nuclear, or extracellular sites. TargetP [32] is designed to predict the presence of signal peptides, chloroplast transit peptides, and mitochondrial targeting peptides in plant proteins. It is also designed to predict the presence of signal peptides and mitochondrial targeting peptides in eukaryotic proteins.

1.3 MOTIVATION

1.3.1 Gene Ontology

The Gene Ontology (GO) is developed by the Gene Ontology Consortium [33]. The main goal of the GO consortium is to annotate gene products with a consistent, controlled, and structured vocabulary. The GO is independent from any biological species. Using a controlled vocabulary across a variety of species enables dynamic maintenance and interoperability among many different types of biological databases. It represents terms in a Directed Acyclic Graph (DAG), providing a vocabulary of genetic annotation terms in three categories, and they are molecular function, biological process, and cellular component. The GO graph consists of over 18,000 terms where each term is represented as a node within the DAG. The terms are connected by relationships that are represented as edges in the DAG. Terms can have multiple parents as well as multiple children. There are two different types of relationship between terms. The first relationship is the "is-a" relationship. The second type of relationship is the "part of" relationship that describes, for instance, that regulation of cell differentiation is part of cell differentiation. Providing a standard vocabulary across any biological resources, the GO enables researchers to use this information for automatic data analysis done by computers and not by humans.

Some of the recent efforts in protein subcellular localization have focused on incorporating gene ontology into their prediction systems. Chou and Cai [34,35,36,37] devised hybrid approaches for protein subcellular localization prediction by combining functional domain composition, pseudo-amino acid composition, and gene ontology. A recent study by Lu and Hunter [38] examined the relationship between GO molecular function annotations and localization information, and found highly predictive GO molecular function terms with respect to subcellular location: endoplasmic reticulum, extracellular, membrane, mitochondrion, Golgi, and nucleus. More specifically, the contribution that molecular function and its existing annotation in GO make to the prediction of subcellular localization was explored in their work. Information gain was used as a measure of the amount of knowledge that

each of the selected highly discriminating GO terms had in terms of providing information regarding protein subcellular localization. Table 1 lists six subcellular localization sites and corresponding GO molecular function terms. In this paper, we explore machine learning approaches to the construction of classifiers for assigning protein sequences to appropriate GO subcellular localizations as defined in [38] using a kernel representation of amino acid sequences and secondary structures. Specifically, we represent protein sequences and secondary structures using the spectrum kernel [39].

1.3.2 Protein Secondary Structure

In living cells, proteins perform a variety of biological tasks. Each protein has a particular three-dimensional structure that determines its function. Structure is more conserved than sequence, and protein structure is central for understanding protein functions. Protein secondary structure prediction is one of the key problems in computational biology today. The prediction of a protein's secondary structure is an important problem because it serves as an essential step to predicting the full three-dimensional structure of a protein. If the secondary structure of a protein is known, it is possible to derive a comparatively small number of possible tertiary structures using knowledge about the ways that secondary structural elements pack.

Many recent efforts in protein secondary structure prediction have led to a sustained three-state prediction accuracy in the range of $77\% \sim 78\%$. Furthermore, in some cases, combinations of secondary structure prediction programs may lead to higher prediction accuracy by one to two percentage points [40]. These improvements are due to an increasing number of experimentally determined tertiary structures and also due to the use of evolutionary information coupled with advances in algorithms. With much emphasis on and recent improvements in protein secondary structure prediction programs, we explore in this paper machine learning approaches that use the resources concerning protein secondary structure for the purposes of predicting protein subcellular localization sites.

1.4 MATERIALS AND METHODS

1.4.1 Support Vector Machines

Support vector machine (SVM) is a learning algorithm proposed by Vapnik [41,42]. Given a training set in a vector space, a support vector machine finds the best decision hyperplane that separates two classes. The distance between two hyperplanes (also called margin) parallel to the decision hyperplane and touching the closest data points of each class determines the quality of a decision hyperplane. The decision hyperplane with the maximum margin is the best one. Support vector machines are only applicable for binary classification tasks. Thus, discriminating among more than two classes must be treated as a series of dichotomous classification problems. Typical kernels include linear kernels, polynomial kernels, and radial kernels.

Let $\tilde{\mathbf{x}}$ be the feature vector to be classified. The SVM classifies $\tilde{\mathbf{x}}$ to either -1 or 1 using

$$y(\mathbf{x}) = \left\{ \begin{array}{ccc} 1 & if & L(\tilde{\mathbf{x}}) > 0 \\ -1 & & otherwise \end{array} \right\}$$

where the discriminant function is given by

$$L(\tilde{\mathbf{x}}) = \sum_{i=1}^{T} \alpha_i y_i K(\tilde{\mathbf{x}}, \mathbf{x}_i)$$

Table 1.1. Selected highly discriminating GO molecular function terms [38]

Location	Predictive GO Molecular Function terms
Nucleus	GO:0003676 Nucleic acid binding GO:0008134 Transcription factor binding GO:0030528 Transcription regulator activity
Membrane	GO:0004872 Receptor activity GO:0015267 hannel/pore class transporter activity GO:0008528 Peptide receptor activity, G-protein coupled
Extracellular	GO:0005125 Cytokine activity GO:0030414 Protease inhibitor activity GO:0005201 Extracellular matrix structural constituent
Mitochondrion	GO:0015078 Hydrogen ion transporter activity GO:0004738 Pyruvate dehydrogenase activity GO:0003995 Acyl-CoA dehydrogenase activity GO:0015290 Electrochemical potential-driven transporter activity
Endoplasmic reticulum	GO:0004497 Monooxygenase activity GO:0016747 Transferase activity, transferring groups other than amino-acyl groups
Golgi	GO:0016757 Transferase activity, transferring glycosyl groups GO:0015923 Mannosidase activity GO:0005384 Manganese ion transporter activity

where $\{\mathbf{x}_i\}_{i=1}^T$ is a set of training vectors and $\{y_i\}_{i=1}^T$ are the corresponding classes ($y_i \in$ -1, 1). $K(\mathbf{x}_i, \mathbf{x}_j)$ is denoted a kernel and is often chosen as a polynom of degree d, i.e.

$$K(\tilde{\mathbf{x}}, \mathbf{x}_i) = (\mathbf{x}^T \mathbf{x}_i + 1)^d$$

Finally, α_i is the weight of training sample $\mathbf{x}_i$. It expresses the strength with which that sample is embedded in the final decision function. Only a subset of the training vectors will be associated with a non-zero α_i. These vectors are called *support vectors*.

The training of the SVM consists of determining the weights α_i that maximizes the distance between the training samples from the two classes. The goal of the training process is to learn a set of weights that maximizes the following objective function:

$$J(\alpha) = \sum_{i=1}^{T} \alpha_i (2 - y_i L(\mathbf{x}_i))$$

$$= \quad 2\sum_{i=1}^{T}\alpha_i - \sum_{i=1}^{T}\sum_{j=1}^{T}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to the following constraints

$$\alpha_i \geq 0 \quad \sum_i \alpha_i y_i = 0 \quad i = 1, ..., T$$

The output of the SVM learning algorithm is the optimized set of weights $\alpha_1, \alpha_2, ..., \alpha_T$.

1.4.2 Feature Representations and SVM Kernel Selection

In this paper, we explore an SVM-based learning method to predict a protein's subcellular localization. We classify six subcellular localization sites: endoplasmic reticulum, extracellular, Golgi, membrane, mitochondrion, and nucleus). The two kinds of information we use for feature representation are amino acid sequences and secondary structures. As for secondary structures, they are known for training data (secondary structures are derived from DSSP[43]) and predicted using PSI-PRED [44] for testing data. We use the publicly available SVM^{light} package to learn the binary classifiers [45].

The spectrum kernel is used to generate SVM feature vectors. The feature space of this kernel is a set of sequence models, and each component of the feature space representation measures the extent to which a given sequence fits the model. The spectrum kernel models a sequence in the space of all k-mers (subsequences of k-length), and its features count the number of times each k-mer appears in the sequence. $\mathcal{X}$ is the input space of all finite length sequences of characters from an alphabet $\mathcal{A}$, $|\mathcal{A}| = l$. A feature map from $\mathcal{X}$ to $\Re^{l^k}$ is defined as

$$\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$$

where $\phi_a(x)$ is the number of times a occurs in x.

In our experiment, we move a k-length sliding window across a protein sequence, look up the current k-length subsequence in the look-up table, and increment the classier value by the associated coefficient. The same procedure is used for secondary structures. Using a window size of three, we can map each amino acid residue to an 8000-dimension vector by the feature map Φ_{AA} and its secondary structure to a 27-dimension vector by the feature map Φ_{SS}. Different feature representation can be combined by concatenation of feature vectors. Φ_{AA} x Φ_{SS} represents the direct product of amino acid residue and secondary structure feature maps.

1.4.3 Datasets

A total of 2,897 protein sequences derived from SwissProt [46] database release 48.7 and secondary structures were used in the experiment. Each protein sequence has a corresponding PDB [47] ID and GO ID. The GO ID of each protein sequence indicates its subcellular localization as defined in Table 1.1. For example, if a sequence's GO ID is GO:0004497, we assign the sequence to the endoplasmic reticulum localization. The secondary structure of training sequences is either derived from the DSSP program or predicted using the PSI-PRED program. Eight states from DSSP are converted to three secondary structure states:

alpha = (H, G, I), beta = (E, B), and coil = (T, S, ' '). For testing sequences, secondary structures are predicted using the PSI-PRED program and we use the three states: alpha, beta, and coil.

Table 1.2. SwissProt Dataset - Six Subcellular Localization

Localization	Total Number of Sequences
Endoplasmic reticulum	507
Extracellular	588
Golgi	463
Membrane	462
Mitochondrion	182
Nucleus	695
Total	2,897

1.4.4 Evaluation Method

The training and testing were carried out with two-fold jackknife cross-validation where 50% of the protein sequences were used as training cases while the remaining 50% of the protein sequences were used in testing, and the processes repeat two times each with a different 50% of the protein sequences. A total of six SVM binary classifiers, one SVM classifier for each of the six protein subcellular localizations, are built in this experiment. We compute ROC_{50} scores and ROC_{ALL} scores for each binary classifier. The ROC score represents the area under the receiver operating characteristic curve. ROC_{50} is a plot of true positives as a function of false positives up to the first 50 false positives [48]. ROC_{ALL} is a plot of true positives as a function of false positives in all of the samples. On a scale of zero to one, a score of one means that there is perfect separation between positives and negatives. A score of zero means that among the samples selected by the method (top 50 for ROC_{50} and all for ROC_{ALL}), none is positive.

1.5 RESULTS AND DISCUSSION

We report results that are computed using two-fold jackknife cross-validation. All the results shown in Table 1.3 and Table 1.4 are produced using the polynomial kernel with the default parameters. We also performed experiments using the linear kernel and the radial (RBF) kernel, but no significant improvements in terms of ROC scores were obtained.

In our experiment, we compare the ROC scores between three different types of feature maps for SVM classification: Φ_{AA}, Φ_{SS}, and Φ_{AA} x Φ_{SS}. The method involving Φ_{SS} is divided into two cases. In one case, secondary structures of training sequences are derived from the DSSP program, and such feature map is denoted as Φ_{SS-D}. In the other case, secondary structures are predicted using the PSI-PRED program, and such feature map is denoted as Φ_{SS-P}.

Table 1.3 and 1.4 compares the ROC_{50} and ROC_{ALL} scores between three different types of feature maps for SVM classification. The second column, Φ_{AA}, shows the ROC

Table 1.3. Comparative Performance of Different Feature Maps (ROC_{50} scores)

Localization	Φ_{AA}	Φ_{SS-D}	Φ_{SS-P}	Φ_{AA} **x** Φ_{SS-D}	Φ_{AA}**x** Φ_{SS-P}
E.R.	0.824	0.312	0.327	0.820	0.824
Extracellular	0.777	0.066	0.056	0.734	0.721
Golgi	0.693	0.058	0.106	0.692	0.703
Membrane	0.777	0.020	0.148	0.754	0.752
Mitochondrion	0.717	0.035	0.182	0.797	0.840
Nucleus	0.556	0.081	0.110	0.543	0.546
Average	0.724	0.095	0.155	0.724	0.731

scores achieved by SVM classifiers when only amino acid residue information was used for feature representation. The third and fourth columns, Φ_{SS-D} and Φ_{SS-P}, show the ROC scores when only secondary structure information was used for feature representation. The last two columns, Φ_{AA} xΦ_{SS-D} and Φ_{AA} xΦ_{SS-P}, show the ROC scores when both amino acid residue information and secondary structure information were used for feature representation.

In Table 1.3, it is shown that when using the amino acid residue information alone, we obtain an average ROC_{50} score of 0.724, which is substantially higher than the ROC_{50} scores obtained when using secondary structure information alone. It is clear that using secondary structure information alone does not serve as an accurate predictor of protein subcellular localization. Figure 1.1 shows the results, ROC_{ALL} curves, produced by endoplasmic reticulum binary classifiers developed using the polynomial kernel. Clearly, using secondary structure information alone, leads to inferior ROC scores.

We now explore the effect of combining amino acid residue information with secondary structure information. An average ROC_{50} score of 0.724 is achieved when combining amino acid residue information with secondary structure information (derived from DSSP for training data and predicted using PSI-PRED for testing data). When using PSI-PRED predicted secondary structure for both training and testing, an average ROC_{50} score of 0.731 is achieved. The score is not substantially higher than that achieved by using amino acid residue information alone. From these results, it is clear that the addition of secondary structure information to amino acid residue information does not lead to better performance in protein subcellular localization prediction. However, it is noticeable that the addition of secondary structure information brings a slight improvement in mitochondrion localization prediction.

1.6 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we explored machine learning approaches to construction of classifiers for assigning protein sequences to appropriate GO subcellular localizations as defined in [38] using a kernel representation of amino acid sequences and secondary structures. Optimally combining available information is one of the key challenges in knowledge-based subcellular localization prediction approaches.

In this paper, the two kinds of information we use for features are amino acid sequences and secondary structures. In using the spectrum kernel for feature vector generation, a fixed

Table 1.4. Comparative Performance of Different Feature Maps (ROC_{ALL} scores)

Localization	Φ_{AA}	Φ_{SS-D}	Φ_{SS-P}	Φ_{AA} **x** Φ_{SS-D}	Φ_{AA}**x** Φ_{SS-P}
E.R.	0.970	0.804	0.804	0.967	0.973
Extracellular	0.975	0.762	0.694	0.966	0.966
Golgi	0.968	0.708	0.806	0.962	0.964
Membrane	0.969	0.705	0.728	0.969	0.969
Mitochondrion	0.975	0.693	0.746	0.985	0.988
Nucleus	0.944	0.526	0.635	0.941	0.942
Average	0.967	0.700	0.735	0.965	0.967

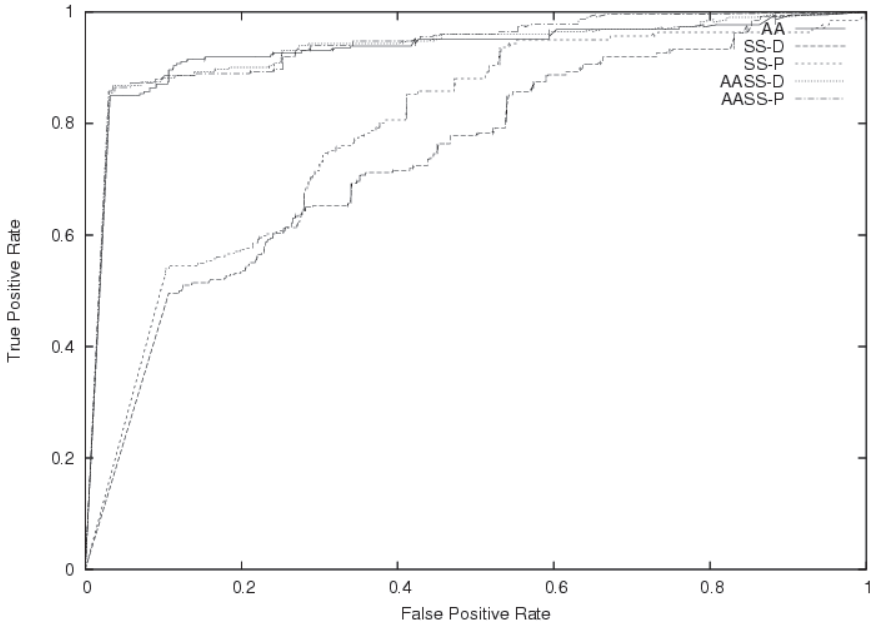

Figure 1.1. ROC_{ALL} curve for different feature maps

length of three was used as the window size. One possible extension to the current work is to use different lengths of k to incorporate more neighboring amino acid residues and secondary structures. Also, the spectrum kernel is not the only method to represent amino acid residue information and secondary structure information as feature vectors for use with SVM. There are other types of kernels such as mismatch kernel [71] and profile-based kernel [72]. Our experiment shows that using amino acid residue information alone can be a good predictor. Better representation of amino acid residue information by using other types of kernels may lead to higher overall prediction accuracy.

Another possible extension is to use a variety of secondary structure programs. Many of today's secondary structure prediction programs can be largely classified into three cat-

egories and they are neighbor-based, model-based, and meta-predictor-based. The prediction approach taken by the neighbor-based programs [49,50,51] identifies a set of similar sequence-fragments whose secondary structures are known. The model-based programs [52,53,54,55] utilize machine learning techniques to build a predictive model learned from training sequences with known secondary structures. Lastly, the meta-predictor-based programs [56,57] combine the predictions obtained from a variety of neighbor and model-based programs. It would be interesting to see how different types of secondary structure prediction approaches contribute to the prediction of protein subcellular localization with or without the use of amino acid residues as part of feature representation.

In our efforts to optimally combine available information for the prediction of protein subcellular localization, one interesting extension is to use local protein structure and solvent accessibility. Local protein structure is known to describe an amino acid residue's environment as well as its relationship to neighboring amino acid residues in a three-dimensional space [58]. Recent efforts in local protein structure include backbone angle prediction [59,60] and novel local structure alphabet [43, 58,60,61,62]. Residue solvent accessibility prediction can help identify the relationship between sequence and structure. Various approaches have been developed over the years, and these include neural network [63,64,65,66], Bayesian statistics [67], multiple linear regression [68], information theory [69], and support vector machine [70].

REFERENCES

1. F. Eisenhaber and P. Bork. Wanted: subcellular localization of proteins based on sequence. Trends Cell Biol, 8:169-70, 1998.
2. D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. Nature, 405:823-6, 2000.
3. W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA, 85:2444-2448, 1988.
4. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. J Mol Biol, 215:403-410, 1990.
5. P. Bork and E. V. Koonin. Predicting functions from protein sequences - where are the bottlenecks? Nat Genet, 18:313-318, 1998.
6. K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. Nucleic Acids Res, 27:215-219, 1999.
7. J. Y. Huang and D. L. Brutlag. The EMOTIF database. Nucleic Acids Res, 29:202-204, 2001.
8. S. Henikoff, J. G. Henikoff, and S. Pietrokovski. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics, 15:471-479, 1999.
9. J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res, 28:231-234, 2000.
10. A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer. The Pfam protein families database. Nucleic Acids Res, 28:263-266, 2000.
11. T. K. Attwood, M. D. Croning, D. R. Flower, A. P. Lewis, J. E. Mabey, P. Scordis, J. N. Selley, W. Wright. PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res, 28:225-227, 2000.
12. R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, and et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res, 29:37-40, 2001.

13. P. Baldi and S. Brunak. Bioinformatics: The Machine Learning Approach. Cambridge, MA: MIT press, 1998.

14. M. Bhasin and G. P. S. Raghava. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST .Nucleic Acids Research, 32(Web Server Issue):W414-W419, 2004.

15. A. Garg, M. Bhasin, and G. P. Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem, Apr 15;280(15):14427-32, 2005. Epub 2005 Jan 12.

16. H. Bannai, Y. Tamada Y, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics, Feb;18(2):298-305, 2002.

17. D. Xie, A. Li, M. Wang, Z. Fan, and H. Feng. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Research, 33(Web Server Issue):W105-W110, 2005.

18. R. Nair and B. Rost. Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol, Apr 22;348(1):85-100, 2005.

19. A. Heddad, M. Brameier, and R. M. MacCallum. Evolving Regular Expression-based Sequence Classifiers for Protein Nuclear Localisation. Accepted paper at the 2nd European Workshop on Evolutionary Bioinformatics (EvoBIO2004, April 2004), 2004.

20. M. Cokol, R. Nair, and B. Rost. Finding nuclear localization signals. EMBO Rep, Nov;1(5):411-5, 2000.

21. I. Small, N. Peeters, F. Legeai, and C. Lurin. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics, Jun;4(6):1581-90, 2004.

22. M. Boden and J. Hawkins. Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics, May 15;21(10):2279-86, 2005. Epub 2005 Mar 3.

23. Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Avnik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics, Mar 1;20(4):547-56, 2004. Epub 2004 Jan 22.

24. D. Sarda, G. H. Chua, K. B. Li, and A. Krishnan. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics, Jun 17;6:152, 2005.

25. M. S. Scott, D. Y. Thomas, and M. T. Hallett. Predicting subcellular localization via protein motif co-occurrence. Genome Res, Oct;14(10A):1957-66, 2004.

26. K. Nakai and M. Kanehisa. Expert system for predicting protein localization sites in gram-negative bacteria. Proteins, 11(2):95-110, 1991.

27. P. Horton and K. Nakai. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. Proc Int Conf Intell Syst Mol Biol, 5:147-52, 1997.

28. C. Guda and S. Subramaniam. pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics, Nov 1;21(21):3963-9, 2005. Epub 2005 Sep 6.

29. J. D. Bendtsen, L. J. Jensen, N. Blom, G. Von Heijne, and S. Brunak. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel, Apr;17(4):349-56, 2004. Epub 2004 Apr 28.

30. J. Bendtsen, H. Nielsen, G. von Heijne and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol., 340:783-795, 2004.

31. S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. Bioinformatics, Aug;17(8):721-8, 2001.

32. O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol, Jul 21;300(4):1005-16, 2000.

33. M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, Jan 1;32(Database issue):D258-61, 2004.

34. K. C. Chou and Y. D. Cai. J. Biol. Chem, 277:45765-45769, 2002.

35. K. C. Chou. PROTEINS: Structure, Function, and Genetics. 43:246-255, 2001 (Erratum: Proteins: Struct. Funct. Genet. 2001, vol. 44, 60).

36. Y. D. Cai and K. C. Chou. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Commun, May 30;305(2):407-11, 2003.

37. K. C. Chou and Y. D. Cai. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun, Nov 21;311(3):743-7, 2003.

38. Z. Lu and L. Hunter. Go molecular function terms are predictive of subcellular localization. Pac Symp Biocomput 10: 151-161, 2005.

39. C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput, 564-75, 2002.

40. B. Rost. Review: Protein secondary structure prediction continues to rise. Journal of Structural Biology, 134:204.218, 2001.

41. M. A. Shipp, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine, 8:68-74, 2002.

42. M. Bribskov and N. L. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Computers and Chemistry, 20(1):25-33, 1996.

43. W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, Dec;22(12):2577-637, 1983.

44. L. J. McGuffin, K. Bryson, and D. T. Jones. The PSIPRED protein structure prediction server. Bioinformatics, Apr;16(4):404-5, 2000.

45. T. Joachims. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

46. A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res. 26:38-42, 1998.

47. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242, 2000.

48. M. Gribskov and N. Robinson. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. Computational Chemistry, 20:2533, 1996.

49. A. A. Salamov and V. V. Solovyev. Protein secondary structure prediction using local alignments. J. Mol. Biol., 268:31.36, 1997.

50. D. Frishman and P. Argos. Seventy-five percent accuracy in protein secondary structure prediction. PROTEINS: Structure, Function, and Genetics, 27:329.335, 1997.

51. K. Joo, J. Lee, S. Kim, I. Kum, J. ee, and S. Lee. Profile-based nearest neighbor method for pattern recognition. J. of the Korean Physical Society, 54(3):599.604, 2004.

52. B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232:584.599, 1993.

53. D. T. Jones. Protein secondary structure prediction based on position-specific scoring matricies. J. Mol. Biol., 292:195.202, 1999.

54. D. Przybylski and B. Rost. Alignments grow, secondary structure prediction improves. PROTEINS: Structure, Function, and Genetics, 46:197.205, 2002.

55. G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. PROTEINS: Structure, Function, and Genetics, 47:228.235, 2002.

56. J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. PROTEINS: Structure, Function, and Genetics, 34:508.519, 1999.

57. G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics, 21:1719.1720, 2005.

58. R. Karchin, M. Cline, Y. Mandel-Gutfreund, and K. Karplus. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins, Jun 1;51(4):504-14, 2003.

59. R. Kuang, C. Leslie, and A. Yang. Protein backbone angle prediction with machine learning approaches. Bioinformatics 20(10):1612-1621, 2004.

60. C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol, Aug 4;301(1):173-90, 2000.

61. A. G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins, Nov 15;41(3):271-87, 2000.

62. D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. Proteins, Dec;23(4):566-79, 1995.

63. B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein families. Proteins, 20, 216-226, 1994.

64. G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio. Prediction of coordination number and relative solvent accessbility in proteins. Proteins, 47, 142-153, 2002.

65. R. Adamczak, A. Porollo, and J. Meller. Accurate prediction of solvent accessibility using neural networks based regression. Proteins, 56, 753-767, 2004.

66. B. Rost. How to use protein ID structure predicted by PROFphd. In J. E. Walker (ed.), The Proteomics Protocols Handbook. Humana, Totowa, NJ, pp. 875-901, 2004.

67. M. J. Thompson and R. A Goldstein. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins, 25, 38-47, 1996.

68. X. Li and X.-M. Pan. New method for accurate prediction of solvent accessibility from protein sequence. Proteins, 42, 1-5, 2001.

69. H. Naderi-Manesh, M. Sadeghi, S. Arab, and A. A. M. Movahedi. Prediction of protein surface accessibility with information theory. Proteins, 42, 452-459, 2001.

70. Z. Yuan, K. Burrage, and J. Mattick. Prediction of protein solvent accessibility using support vector machines. Proteins, 48, 566-570, 2002.

71. C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. Noble. Mismatch String Kernels for Discriminative Protein Classification. Bioinformatics, 20:4, pp. 467-476, 2004.

72. R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote protein homology detection and motif extraction, Proc. Comput. Systems Bioinfo., 2004.