In Proc. of the 6th Asia-Pacific Bioinformatics Conference (APBC 2008), Kyoto, Japan, pp. 29-38

Supervised Ensembles of Prediction Methods for Subcellular Localization

Johannes Aßfalg, Jing Gong, Hans-Peter Kriegel,
Alexey Pryakhin, Tiandi Wei, Arthur Zimek

Institute for Informatics, Ludwig-Maximilians-Universität München, Germany
www: `http://www.dbs.ifi.lmu.de`
E-mail: `{assfalg,gongj,kriegel,pryakhin,tiandi,zimek}@dbs.ifi.lmu.de`

In the past decade, many automated prediction methods for the subcellular localization of proteins have been proposed, utilizing a wide range of principles and learning approaches. Based on an experimental evaluation of different methods and on their theoretical properties, we propose to combine a well balanced set of existing approaches to new, ensemble-based prediction methods. The experimental evaluation shows our ensembles to improve substantially over the underlying base methods.

1. Introduction

In cells, different regions have different functionalities. Certain functionalities are performed by specific proteins. To function properly, a protein must be localized in the proper region of a cell. Co-translational or post-translational transport of proteins into specific subcellular localizations is therefore a highly regulated and complex cellular process. Knowing of the subcellular localization of a protein helps to annotate its possible interaction partners and functionalities.

Starting in the mid-nineties of the last century, until now a plethora of automated prediction methods for the subcellular localization of proteins has emerged. These methods are based on different sources of information like the amino acid composition of the protein, specific sorting signals or targeting sequences contained in the protein sequence, or homology search in databases of proteins with known localization. Furthermore, hybrid methods combine the different sources of information often in a very specialized way. Besides different sources of information, prediction methods differ in the employed learning algorithms (like naive Bayes and Bayes networks, k-nearest neighbor methods, support vector machines (SVM), and neural networks). Due to their different sources of information, prediction methods differ widely in their coverage of different localizations. For example, methods based on targeting sequences generally have a low coverage of only a few localizations. Methods based on amino acid composition vary considerably in their coverage. The coverage of a method is also directly related to the available classes in the data sets used for training of the corresponding method. As most prediction methods are trained and evaluated on data sets suitable to their requirements in coverage, it is a hard task to compare different methods w.r.t. their performance.[9]

In this paper, we survey shortly prominent methods for prediction of subcellular local-

ization of proteins, particularly considering their different properties (Section 2). Based on a diverse selection of the best methods, we propose combined methods using a well balanced set of prediction methods as new ensemble-methods (Section 3). Section 4 presents the evaluation of selected localization prediction methods in comparison to our new ensemble methods. Finally, Section 5 concludes the paper.

2. Survey on Prominent Prediction Methods for Subcellular Localization

For our evaluation of localization prediction methods, we confined the selection to those that are available (excluding methods like NNPSL[24] or fuzzy_loc[16]), and that focus on eukaryotic localization prediction (excluding methods like PSORT-B[11] or PSLPred[3]). In the following, we survey prominent examples from these methods, choosing representatives for the different sources of information the methods are based upon.

2.1. *Amino Acid Composition*

Predicting the subcellular localization based on amino acid composition was suggested by Nakashima and Nishikawa.[22] They presented a method to discriminate between intracellular and extracellular proteins using the amino acid composition. In the following years, a number of approaches using the amino acid composition was proposed.

SubLoc[15] uses one-versus-rest support vector machines (SVM) to predict the localization. No additional information aside from the amino acid composition (like, e.g., dipeptide composition) is used for the prediction. In contrast to SubLoc, PLOC[23] additionally considers the dipeptide composition and the gapped amino acid composition aside from the standard amino acid composition. Like SubLoc, this method employs one-versus-rest SVMs. By using pairs of peptides the authors take more sequence-order information than SubLoc into account. The gapped pair composition corresponds to periodic occurrences of certain amino acids in the sequence. Similar to PLOC, CELLO[17] incorporates several kinds of compositions, including single, dipeptide, and partitioned amino acid compositions. Furthermore, compositions based on physicochemical properties of the amino acids were derived. These features are again used as input for one-versus-rest SVMs.

2.2. *Sorting Signals*

One of the earliest works trying to identify a certain location based on protein sorting signals was already presented in 1986.[27] Most of the methods based on sorting signals are very specialized. For example, Mitoprot[5] predicts only mitochondrial proteins, SignalP[2] predicts only proteins of the secretory pathway. More general methods in this category are iPSORT[1] and Predotar.[25] The comparison of these two methods is especially interesting because they use very different computational approaches: iPSORT uses simple and interpretable rules based on protein sequence features. These features are derived from the so-called amino acid index, a categorization of amino acids based on different kinds of properties. iPSORT uses N-terminal sorting signal sequences. Predotar considers N-terminal sorting signals as well and processes the input information with a feed forward neural network. As an out-

put value, this method yields probability values for the presence of a certain localization sequence rather than an abstract score.

2.3. *Homology*

Prominent methods based on homology search are PredictNLS[6] and PA_SUB.[19] PredictNLS is also based on sorting signals, as it is trained on a data set of experimentally confirmed nuclear localization signal (NLS) sequences. This data set is extended by homology search. Nevertheless, NLSPred is specialized on recognizing nuclear proteins. PA_SUB is purely based on PSI-BLAST homology search using database annotations from homologous proteins. In many cases, homology search is very accurate. However, the result will be arbitrary if no homologous protein with localization annotation is available. The combination of homology search with other methods is a common way to overcome this shortcoming.

2.4. *Hybrid Methods*

As in PredictNLS, most of the methods using homology search combine this technique with some other sources of information. In this category, great effort was already spent to develop refined combinations of information and methods. One often finds series of related approaches from certain groups like the PSORT series (PSORT,[21] PSORT-II,[20] PSORT-B,[10,11] and WoLFPSORT[14]) or ESLPred,[4] HSLPred,[12] and PSLPred.[3] The PSORT-B approaches and PSLPred are specialized for bacteria. PSORT is one of the earliest methods at all, based on amino acid composition, N-terminal targeting sequence information, and motifs. Like iPSORT, it is based on a set of rules. PSORT-II uses a k-NN approach. WoLFPSORT uses a feature selection procedure and incorporates new features, based on new sequence data, simultaneously increasing the coverage of localizations and organisms. ESLPred uses an SVM approach, combining amino acid composition, dipeptide composition, overall physicochemical properties, and PSI-BLAST scores. The extensions HSLPred and PSLPred focus on human and prokaryotic proteins, respectively. MITOPRED[13] uses Pfam domains and amino acid composition, and is specialized for mitochondrial proteins. MultiLoc[18] traines SVMs based on N-terminal targeting sequences, sequence motifs, and amino acid composition.

3. Ensemble Methods

In preliminary tests on our data set, the accuracy of all compared methods was not as high as reported in their original literature for other data sets, meaning our data set can be considered as not too easy. Furthermore, there were sequences with certain localizations always wrongly predicted by some methods, e.g. there was no protein with localization vacuole within fungi group predicted positively although there were 68 vacuole proteins in this group. Some other methods could predict more accurately for these proteins while they might be incapable of accurate prediction of other localizations. In other words, each method has its own advantages and disadvantages. These findings motivate the idea to combine some of these methods.

3.1. *Theory*

Combining several self-contained predicting algorithms to an ensemble to yield a better performance in terms of accuracy than any of the base predictors, is backed by a sound theoretical background.[7,8,26] In short, a predictive algorithm can suffer from several limitations such as statistical variance, computational variance, and a strong bias. *Statistical variance* describes the phenomenon that different prediction models result in equally good performance on training data. Choosing arbitrarily one of the models can then result in deteriorated performance on new data. Voting among equally good classifiers can reduce this risk. *Computational variance* refers to the fact, that computing the truly optimal model is usually intractable and hence any classifier tries to overcome computational restrictions by some heuristics. These heuristics, in turn, can lead to local optima in the training phase. Obviously, trying several times reduces the risk of choosing the wrong local optimum. A restriction of the space of hypotheses a predictive algorithm may create is refered to as *bias* of the algorithm. Usually, the bias allows for learning an abstraction and is, thus, a necessary condition of learning a hypothesis instead of learning by heart the examples of the training data (the latter resulting in random performance on new data). However, a strong bias may also hinder the representation of a good model of the true laws of nature one would like to learn. A weighted sum of hypotheses may then expand the space of possible models.

To improve over several self-contained classifiers by building an ensemble of those classifiers requires the base algorithms being accurate (i.e., at least better than random) and diverse (i.e., making different errors on new instances). It is easy to understand why these two conditions are necessary and also sufficient. If several individual classifiers are not diverse, then all of them will be wrong whenever one of them is wrong. Thus nothing is gained by voting over wrong predictions. On the other hand, if the errors made by the classifiers were uncorrelated, more individual classifiers may be correct while some individual classifiers are wrong. Therefore, a majority vote by an ensemble of these classifiers may be also correct. More formally, suppose an ensemble consisting of k hypotheses, and the error rate of each hypothesis is equal to a certain $p < 0.5$ (assuming a dichotomous problem), though independently. The ensemble will be wrong, if more than $k/2$ of the ensemble members are wrong. Thus the overall error rate $\bar{p}$ of the ensemble is given by the area under the binomial distribution, where $k \geq \lceil k/2 \rceil$, that is for at least $\lceil k/2 \rceil$ hypotheses being wrong: $\bar{p}\,(k,p) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} p^i (1-p)^{k-i}$. The overall error-rate is rapidly decreasing for an increasing number of ensemble members.

3.2. *Selection of Base Methods for Ensembles*

Comparing several methods based on amino acid compositions we found an increase of accuracy by adding more sequence-order information. CELLO behaved best no matter for which taxonomy group because it used the most sequence-order information: single amino acid composition, dipeptide composition, n-peptide composition, and even physicochemical properties of amino acids in the updated version that we used. In contrast, PLOC which used only amino acid composition and dipeptide composition had more false predictions

than CELLO, but it was more accurate than SubLoc which used only single amino acid composition. In comparison, the methods based on detecting N-terminal sorting signals performed better than expected, although they have to handle missing N-terminal sorting signals. Of the hybrid methods the two newest, WoLFPSORT (2006) and MultiLoc (2006), had similar prediction ability and their accuracy is higher than that of the others in this category.

Based on the results of our preliminary experimental comparisons and the criteria of usability, reliability, efficiency, coverage, and, for theoretical reasons, as discussed above, diversity in the underlying methods and sources of information, we chose the following methods to build an ensemble: From the methods based on amino acid composition SubLoc was excluded because of its too simple foundation and its lower rank during the preliminary tests. In addition, both PLOC and CELLO use the single amino acid composition too and predict more accurately than SubLoc. iPSORT and Predotar as prominent examples of methods based on sorting signals had similar prediction ability in our preliminary tests but use quite different algorithms, so both of them were chosen for the combination. PA_SUB is a purely homology-based method. The data set used for generating PA_SUB consists of virtually all Swiss-Prot[29] entries that provide a localization annotation. As we evaluate the considered methods and our combination of methods on an up-to-date data set also compiled from Swiss-Prot, we exclude PA_SUB from the experiments, as it is highly overfitted to the data set. Usually, as discussed above, homology-based approaches are combined with other approaches. From the hybrid methods only the method PSORT II was excluded, because we use its extension WoLFPSORT which is more accurate and has a larger taxonomy coverage than PSORT II. HSLPred is used for the human proteins. Although its localization coverage is very narrow, it is still very sensitive for the three localizations within its coverage. Finally we chose 7 methods for the plant, animal and fungi groups and 8 methods for the human group to construct an ensemble method: PLOC, CELLO, iPSORT, Predotar, WoLFPSORT, MultiLoc, ESLPred, and, for human proteins, HSLPred.

3.3. *Ensemble Method Based on a Voting Schema*

Despite a clear theoretical background for ensemble learning in general, the combination of localization prediction methods is not trivial due to the wide range of localization and taxonomic coverage. Imagining a prediction method as a function from some feature space to some class space, the base learners map the proteins into different class spaces. Thus, for unifying the prediction methods, the class spaces must be unified first. The unified class space should contain the classes supported by most of the methods (resulting in the set of ten localization classes as described above). Methods that are unable to predict some of the classes contained in the unified class space must be treated especially. Furthermore, some methods (PLOC, CELLO, WoLFPSORT, and MultiLoc) predict exactly one localization for a query protein while others (iPSORT, Predotar, ESLPred, and HSLPred) predict a range of possible localizations. We define therefore a voting schema as follows: Methods in the first group give their vote to one certain localization at a time if the predicted localization belongs to the 10 localizations in our data set. Otherwise their vote is blanked out. Methods

Table 1. Ranks of different classification methods for the considered taxonomic groups.

Taxonomic group	CELLO	ESLPred	HSLPred	iPSORT	MultiLoc	PA_SUB
Animal	2	10	—	3	6	1
Fungi	4	9	—	1	7	3
Human	4	10	7	2	6	1
Plant	3	2	—	9	8	1

Taxonomic group	PLOC	Predotar	PSORT II	SubLoc	WoLFPSORT
Animal	4	5	8	9	7
Fungi	5	2	8	10	6
Human	5	3	9	11	8
Plant	4	7	—	6	5

in the second group may give their vote to several localizations at a time. If a classifier maps the proteins into a class space containing some of the ten classes and a class 'unknown', a prediction for class 'unknown' can be mapped to the set of the remaining classes. However, if a classifier cannot decide between some classes, this will not mean automatically that the protein belongs to the set of unknown classes. For example, if there is no sorting signal being detected by iPSORT or Predotar, we cannot say that this protein is not localized in chloroplast, mitochondrion, or the secretory pathway, because the N-terminal sequence of this protein may be not complete. In this case, iPSORT and Predotar will give up on voting.

Based on the votes of all base classifiers, we derive a vector s of scorings for the localizations, where for localization i the score s_i is computed as follows:

$$s_i = \sum_{j=1}^{N} \left(v_j \cdot (N - \mathit{rank}_j + 1) \right),$$

where N is the number of methods used by the ensemble method, rank_j is the rank in accuracy of method j according to our preliminary tests, and $v_j = 1$ if method j votes for localization i (allowing voting for multiple localizations), otherwise $v_j = 0$. This ensemble is therefore built based on prior knowledge concerning the performance of the base classifiers. We also tried a voting without explicitly ranking the votes of the base classifiers, but the results were not acceptable. The ranks we used for the evaluation can be found in Table 1.

3.4. *Ensemble Method Based on Decision Trees*

As requiring prior knowledge to construct a voting schema is not satisfying, we chose to derive the voting schema by decision trees, trained on the predictions of the single base methods and the correct localization classes. Decision trees combine the benefits of generally good accuracy and interpretable models, i.e. the derived voting schema provides further information regarding the performance of the underlying methods on different localization classes. For example, the decision tree for the taxonomic group "plant" learns a rule like *If CELLO predicts class 6 and WoLFPSORT predicts class 4, then class 4 is correct*. We trained decision trees using J48 of WEKA[28] for each taxonomic group.

Table 2. Covered subcellular localizations and corresponding keywords in SWISS-PROT.

ID	Subcellular localization	Keywords in SWISS-PROT	ID	Subcellular localization	Keywords in SWISS-PROT
1	Chloroplast	Chloroplast	8	Peroxisome	Peroxisome, Peroxisomal
2	Cytoplasm	Cytoplasm(ic)			Microsome, Microsomal
3	ER	Endoplasmic reticulum			Glyoxysome, Glyoxysomal
4	Golgi apparatus	Golgi			Glycosome, Glycosomal
5	Lysosome	Lysosome, Lysosomal	9	Extracellular	Extracellular
6	Mitochondrion	Mitochondrion, Mitochondrial			Secreted
7	Nucleus	Nucleus, Nuclear	10	Vacuole	Vacuole, Vacuolar

4. Evaluation

Although more and more prediction methods for subcellular localization have been developed, several limitations exist. First, the coverage of predicted localizations, which ranges from just a few localizations to all possible localizations. While e.g., SubLoc predicts only 4 localizations, PLOC is able to predict 12 localizations. Second, most existing methods were trained by a limited number of sequences from a specific taxonomic category of organisms, so the methods differ in their taxonomic coverage. The third aspect is the so-called sequence coverage, which is the number of sequences the different approaches learn from. Nonetheless, many newly developed methods still use the data set created by Reinhardt and Hubbard in 1998.[24] Thus, we decided to compile an up-to-date data set based on Swiss-Prot.[29] In order to compare methods differing widely in many aspects, we restricted the data set to 10 localization classes which are commonly accepted by most of the methods. These localization classes are listed in Table 2. This selection accommodates most of the available and rather general methods. For methods with a narrower localization coverage we used their reliability indices and assigned query sequences with lower reliability indices to the class "unknown". While their coverage is narrower, these methods often exceed others in their performance for the covered localization classes.

Based on Swiss-Prot (release 53.0), we at first selected all eukaryotic proteins with a unique subcellular localization annotation, where the localization annotation was one of the 10 localization classes listed in Table 2. Then, all proteins with a sequence length smaller than 60 amino acids were removed, as this is the required minimal sequence length for Predotar, the method with the largest required minimal length. Finally we kept only those proteins whose localization annotation was experimentally confirmed and belonged to one of the taxonomic groups "plant", "fungi", "human", or "animal". As the golgi group of plants was too small (7 entries), we complemented this group with 28 proteins whose localization information was not confirmed experimentally. This yielded 4 subsets corresponding to the 4 taxonomic groups. Table 3 lists the final number of proteins for each taxonomic group and each localization class.

Both the ensemble methods as well as the single base classifiers were evaluated by 10-fold cross-validations on our data set. The results are illustrated in Figures 1 and 2. Figure 1 shows the total accuracy. The simple weighted voting schema ("Voting") performs slightly better than the base classifiers. The decision tree ensembles ("DT-Ensemble") clearly outperform all other methods (including the voting schema). The most prominent improvement

Table 3. Number of proteins for different taxonomic groups and localization classes.

		Plant	Fungi	Animal	Human	Total
1	Chloroplast	3425	0	0	0	3425
2	Cytoplasm	470	578	1394	511	2953
3	ER	66	170	391	164	791
4	Golgi	35	55	78	55	223
5	Lysosome	0	0	102	56	158
6	Mitochondrion	370	632	1341	347	2690
7	Nucleus	308	899	2221	1094	4522
8	Peroxisome	50	85	181	72	388
9	Extracellular	149	199	596	4723	5667
10	Vacuole	35	68	0	0	103
	Total	4908	2686	10431	2895	20920

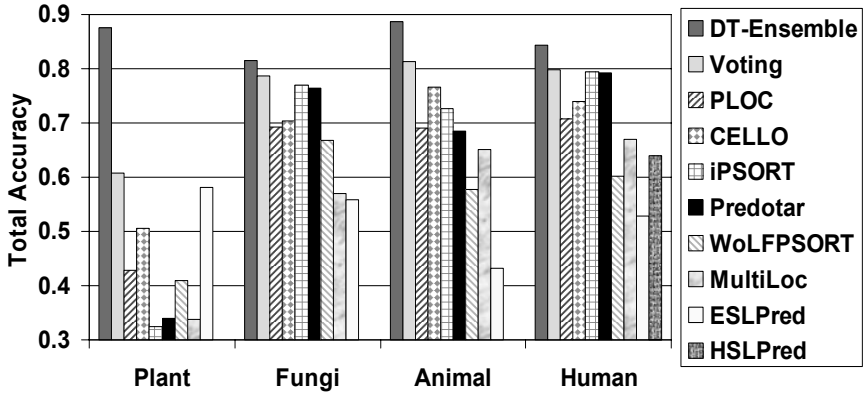

Fig. 1. Comparison between single and ensemble classification methods: Total Accuracy, i.e., the overall percentage of correctly predicted instances.

can be seen in the plant group, were the other methods mostly perform rather weak (at best, ESLPred reaches an accuracy of just below 60%), while the accuracy of the decision tree ensemble is well above 80%.

Most methods perform comparably well in terms of specificity (cf. Figure 2). Again, in the plant group the improvement of both ensemble methods is most prominent. In the remaining taxonomic groups the best base classifiers already reach almost 100%. Thus, no significant improvement can be expected. However, the ensemble methods perform as well as the best base classifiers. The decision tree ensembles even slightly improve over the already very good values.

All our methods are available via a webinterface at `http://www.dbs.ifi.lmu.de/research/locpred/ensemble/`.

5. Conclusions

In this paper, we shortly surveyed some prominent prediction methods for subcellular localization of proteins. The spectrum of underlying information (as amino acid composition,

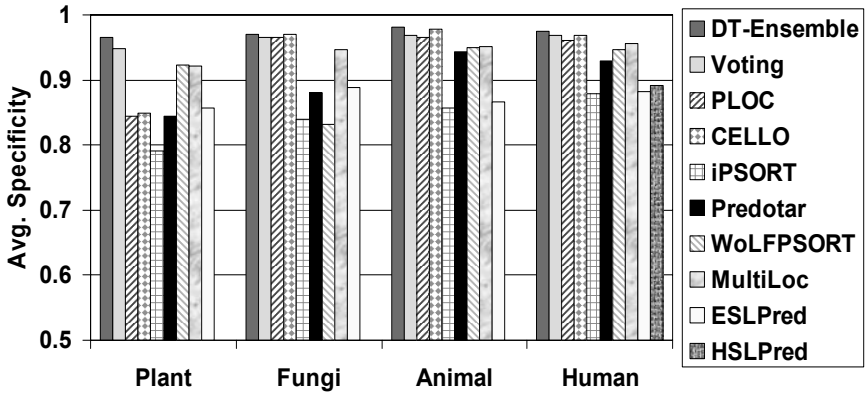

Fig. 2. Comparison between single and ensemble classification methods: Average Specificity, i.e., the percentage averaged over all localization classes to correctly exclude an instance from the corresponding class.

sorting signals, and homology search) makes these methods ideally diverse to expect an ensemble composed of these methods to improve considerably in terms of accuracy. We developed two ensemble methods: First, a simple voting scheme using the votes of the base learners weighted according to their average performance (based on prior knowledge), second, decision trees trained on the prediction values of the base methods (thus learning the weight of the methods on the fly and allowing for a more complex weighting). Both ensembles are shown to improve over the base classifiers in most cases. The decision tree ensemble can even said to outperform the remaining methods.

References

1. H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of n-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305, 2002.
2. J.-D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, 340(4):783–795, 2004.
3. M. Bhasin, A. Garg, and G.-P.-S. Raghava. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21(10):2522–2524, 2005.
4. 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 Res.*, 32(Web Server Issue):W414–W419, 2004.
5. M.-G. Claros and P. Vincens. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, 241(3):779–786, 1996.
6. M. Cokol, R. Nair, and B. Rost. Finding nuclear localization signals. *EMBO Rep.*, 1(5):411–415, 2000.
7. T. G. Dietterich. Ensemble methods in machine learning. In *Proc. MCS*, 2000.
8. T. G. Dietterich. Ensemble learning. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 405–408. MIT Press, second edition, 2003.
9. P. Dönnes and A. Höglund. Predicting protein subcellular localization: Past, present, and future. *Geno. Prot. Bioinfo.*, 2(4):209–215, 2004.
10. J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester, and F. S. L. Brinkman. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21(5):617–623, 2005.

11. J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. L. Brinkman. PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Res.*, 31(13):3613–3617, 2003.
12. A. Garg, M. Bhasin, and G. P. S. Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.*, 280(15):14427–14432, 2005.
13. C. Guda, E. Fahy, and S. Subramaniam. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, 20(11):1785–1794, 2004.
14. P. Horton, K.-J. Park, T. Obayashi, and K. Nakai. Protein subcellular localization prediction with WoLF PSORT. In *Proc. APBC*, 2006.
15. S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.
16. Y. Huang and Y. Li. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, 20(1):21–28, 2004.
17. J.-K. Hwang, C.-J. Lin, and C.-S. Yu. Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Science*, 13:1402–1406, 2004.
18. A. Höglund, P. Dönnes, T. Blum, H.-W. Adolph, and O. Kohlbacher. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, 22(10):1158–1165, 2006.
19. Z. Lu, D. Szafron, R. Greiner, P. Lu, D.S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–556, 2004.
20. K. Nakai and P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, 24(1):34–36, 1999.
21. K. Nakai and M. Kanehisa. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14(4):897–911, 1992.
22. H. Nakashima and K. Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, 238(1):54–61, 1994.
23. K.-J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656–1663, 2003.
24. A. Reinhardt and T. Hubbart. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, 26:2230–2236, 1998.
25. I. Small, N. Peeters, F. Legeai, and C. Lurin. Predotar: A tool for rapidly screening proteomes for n-terminal targeting sequences. *Proteomics*, 4(6):1581–1590, 2004.
26. G. Valentini and F. Masulli. Ensembles of learning machines. In *Proc. Neural Nets WIRN*, 2002.
27. G. von Heijne. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.*, 14(11):4683–4690, 1986.
28. I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.
29. C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, 34:D187–D191, 2006.