Published online 27 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W497–W502***
doi:10.1093/nar/gkq477

YLoc—an interpretable web server for predicting subcellular localization

Sebastian Briesemeister[1,*], Jörg Rahnenführer[2] and Oliver Kohlbacher[1]

[1]Division for Simulation of Biological Systems, Universität Tübingen, Tübingen and [2]Department of Statistics, TU Dortmund University, Dortmund, Germany

Received February 1, 2010; Revised May 2, 2010; Accepted May 13, 2010

ABSTRACT

Predicting subcellular localization has become a valuable alternative to time-consuming experimental methods. Major drawbacks of many of these predictors is their lack of interpretability and the fact that they do not provide an estimate of the confidence of an individual prediction. We present YLoc, an interpretable web server for predicting subcellular localization. YLoc uses natural language to explain why a prediction was made and which biological property of the protein was mainly responsible for it. In addition, YLoc estimates the reliability of its own predictions. YLoc can, thus, assist in understanding protein localization and in location engineering of proteins. The YLoc web server is available online at www.multiloc.org/YLoc.

INTRODUCTION

Protein sorting is a complex and still poorly understood process. It is crucial for a protein's function as a protein's location is often correlated with its molecular function. Thus, knowledge of protein localization can help biologists to infer the function of a protein. However, experimental methods for determining a protein's location are expensive and time consuming. In contrast, computational predictions rely only on the protein sequence, are fast, and fairly accurate. Over recent years, various prediction methods have been introduced. Most methods use sequence information, such as known sorting signals and amino acid composition (1–9). More advanced methods incorporate annotation information such as functional domains and motifs (10,11), homologous proteins (12,13), Gene Ontology (GO) terms (14) and textual information (15,16). Predictions based on annotated knowledge are often more accurate, but are less robust in cases where little is known about the protein. Hybrid prediction approaches combine the advantages of both information sources (17–21).

Although the prediction performance of subcellular localization predictors has increased significantly over recent years, their predictions are often not considered to be trustworthy. Very complex machine learning models of state-of-the-art prediction systems make it difficult to understand why a prediction was made. Consequently, the web interfaces of most methods are non-transparent and offer no explanation for a particular prediction. In addition, most methods do not offer confidence estimates for an individual prediction.

We present YLoc, an interpretable web server for predicting subcellular localization. Users are provided with the prediction itself, and also with an explanation why this prediction was made. The features contributing to the prediction are translated into natural language aiming at the most likely explanation of the localization. In addition, a confidence score helps the users to verify whether the prediction is reliable or not. YLoc is available in a low-resolution version, YLoc-LowRes, and a high-resolution version, YLoc-HighRes, covering 5 or 11 eukaryotic subcellular locations, respectively. YLoc+, the most general version, integrates multiple locations sites. All three predictors are available for animal, fungal and plant proteins.

METHODS AND MATERIALS

YLoc-LowRes was trained on the BaCelLo data set (6), which contains only globular proteins. The animal and fungal versions predict four locations: the nucleus (nu), cytoplasm (cy), mitochondrion (mi) and the secretory pathway (SP). The plant version additionally predicts the chloroplast (ch). YLoc-HighRes was trained on the Höglund data set (7). It covers 11 locations: nu, cy, mi, ch, endoplasmatic reticulum (er), Golgi apparatus (go), peroxisome (pe), plasma membrane (pm), extracellular space (ex), lysosome (ly) and vacuole (va). In the training of YLoc+, we used the Höglund data set and additional proteins with multiple locations from the DBMLoc database (22). The extracted 3054 proteins share <80% sequence similarity. Only dual locations

*To whom correspondence should be addressed. Tel: +49 7071 29 70462; Fax: +49 7071 29 5152; Email: briese@informatik.uni-tuebingen.de

with more than 100 representative proteins were included: cy and nu (cy_nu), ex and pm (ex_pm), cy and pm (cy_pm), cy and mi (cy_mi), nu and mit (nu_mi), er and ex (er_ex) and ex and nu (ex_nu). To our knowledge, this is currently the largest data set of proteins from multiple locations.

We derived about 30 000 features from our protein sequences using amino acid composition and pseudo composition (3) as well as properties such as hydrophobicity, charge and volume of amino acids. In addition, we included PROSITE motifs and GO terms from close homologs. For more details, we refer to Briesemeister *et al.* (26). To guarantee interpretable predictions, we first reduced the number of features using a backward best first search together with correlation-based feature selection (23) implemented in the Weka machine learning library (24). For YLoc-LowRes, we obtained 20 features; for YLoc-HighRes and YLoc+, we obtained 30 features. However, a small number of features is only the first step toward interpretable predictions. To provide meaningful explanations, we manually annotated all selected features in biological terms. Unfortunately, not every feature can be easily mapped to a biological property. In such cases, we carefully inspected the initial feature set and transferred the biological meaning of a highly correlated feature. A list of all selected and annotated features can be found in the Supplementary Data.

YLoc uses naïve Bayes alongside entropy-based discretization (25) to make predictions. Given a set of features $F = \{F_1, \ldots, F_k\}$ and a set of location classes $C = \{C_1, \ldots, C_n\}$, the conditional distribution of class C_j can be expressed by:

$$P(C_j|F) \propto P(C_j) \prod_{i=1}^{k} P(F_i|C_j). \quad (1)$$

The final posterior probabilities $P(C_j|F)$ are calculated by normalizing the right term of Equation (1) such that all posteriors sum up to one. Based on the feature likelihoods $P(F_i|C_j)$, we calculate a discrimination score which provides a simple and transparent understanding of the influence of a feature on the prediction, for details see Briesemeister *et al.* (26). A positive score indicates that this feature is typical for the predicted location, whereas a negative score indicates that this feature alone would suggest a different location. Secondly, the discrimination score shows how strongly a feature influenced the prediction.

For multiple localization prediction, we assume that a protein present in multiple locations is equally distributed among those. Proteins labeled with two locations are assigned to a dual-location class, for example, C_{nu_cy}. YLoc+ then evenly distributes the posterior probabilities of the dual-location classes onto the probabilities of the two individual locations. For example, $P(C_{nu\_cy}|F)/2$ is added to $P(C_{nu}|F)$ and to $P(C_{cy}|F)$. All locations with a probability above a threshold of $1/|C|$ are predicted, where $|C|$ is the number of locations. If a location is less than half as probable as the next most probable one, this location and all less probable locations are not predicted.

The probability of the predicted location shows only how likely a protein is to be found in this location compared with the other locations. A confidence estimate, however, tells how likely it is that this prediction is to be correct. For this purpose, we analyze whether the protein is typical for the predicted class or whether YLoc already extrapolates. If a feature vector is more likely for proteins from the predicted location than for proteins from all locations, i.e. $P(F|C_{\text{pred}}) > P(F|\bigcup C_j)$, we rate a prediction as being reliable. Since predicted locations with only a few training examples are often less reliable, we include the prior class probability in our confidence score:

$$\frac{P(C_{\text{pred}})P(F|C_{\text{pred}})}{P(C_{\text{pred}})P(F|C_{\text{pred}}) + P(F|\bigcup C_j)}. \quad (2)$$

Confidence scores ranges from zero for unreliable predictions to one for very confident predictions. For more details on the YLoc methodology, refer to Briesemeister *et al.* (26).

EVALUATION

We have tested the performance of YLoc on two independent data sets (IDSs). The BaCelLo IDS (27) consists of animal, fungal and plant proteins from the nu, cy, mi and SP which have at most 30% sequence identity to proteins in the BaCelLo data set. The Höglund IDS (20) contains animals proteins from remaining locations, the er, go, pe, pm, ex and ly, and was constructed with the same restrictions as the BaCelLo data set. In addition, proteins from the same location which align with an E-value $>10^{-3}$ are clustered and treated as one instance in the evaluation. We compared the YLoc predictors with five other state-of-the-art subcellular localization predictors: MultiLoc2 (20), BaCelLo (6), LOCTree (4), WoLF PSORT (9) and Euk-mPloc (19). All methods are available as web servers. The individual prediction performance was evaluated using the overall accuracy (ACC), which is the percentage of correctly predicted instances, and the average F_1-score (F_1), which is the average over the harmonic means of precision and recall of each location. Note that YLoc+, WoLF PSORT and Euk-mPloc are evaluated using the generalized ACC and F_1 from multilabel classification (28).

The evaluation results are summarized in Table 1. In our benchmark study, we observed that YLoc shows comparable performance to current state-of-the-art methods. For the BaCelLo IDS, YLoc-LowRes and MultiLoc2-LowRes perform best since they are specialized in distinguishing globular proteins. The high-resolution predictors YLoc-HighRes, YLoc+, MultiLoc2-HighRes, WoLF PSORT and Euk-mPloc perform slightly worse on this data set, since they are more general predictors. On the Höglund IDS, MultiLoc2-HighRes, YLoc-HighRes and YLoc+ show comparable performance, whereas WoLF PSORT and Euk-mPloc perform worse. MultiLoc2 shows very good accuracy throughout the study. However, its architecture is very complex and the

Table 1. Performance of the YLoc and other state-of-the-art predictors on the BaCelLo IDS (27) (B) and Höglund IDS (20) (H) concerning F_1 and ACC (in brackets)

Data set	YLoc-LowRes	YLoc-HighRes	YLoc+	MultiLoc2-LowRes	MultiLoc2-HighRes	BaCelLo	LOCTree	WoLF PSORT	Euk-mPloc
B Animals	0.75 (0.79)	0.69 (0.74)	0.67 (0.58)	0.76 (0.73)	0.71 (0.68)	0.66 (0.64)	0.58 (0.62)	0.67 (0.70)	0.54 (0.61)
B Fungi	0.61 (0.56)	0.51 (0.56)	0.51 (0.48)	0.61 (0.60)	0.58 (0.53)	0.60 (0.57)	0.43 (0.47)	0.51 (0.50)	0.56 (0.60)
B Plants	0.58 (0.71)	0.54 (0.58)	0.49 (0.58)	0.64 (0.76)	0.54 (0.62)	0.56 (0.69)	0.58 (0.70)	0.46 (0.57)	0.37 (0.46)
H Animals	– (–)	0.34 (0.56)	0.37 (0.53)	– (–)	0.41 (0.57)	– (–)	– (–)	0.18 (0.36)	0.24 (0.27)

Table 2. Performance of YLoc on the BaCelLo animal IDS (27) for different minimum confidence scores

Predictor	Measure	0.0	0.2	0.4	0.6	0.8	0.9
YLoc-LowRes	F_1	0.75	0.76	0.78	0.80	0.84	0.95
	ACC	0.79	0.79	0.81	0.86	0.91	0.93
	No. of instances	576	467	395	299	189	118
YLoc-HighRes	F_1	0.69	0.74	0.76	0.76	0.77	0.77
	ACC	0.74	0.78	0.80	0.82	0.83	0.84
	No. of instances	576	507	470	428	391	354
YLoc+	F_1	0.67	0.69	0.72	0.77	0.76	0.81
	ACC	0.58	0.60	0.62	0.65	0.65	0.69
	No. of instances	576	494	423	324	219	142

output is not interpretable. In contrast, YLoc uses a very simple model and its predictions are hence interpretable. The detailed location-wise performance of YLoc is shown in the Supplementary Data. When YLoc is applied without the use of GO-term-based features, the performance is only slightly reduced compared with the original predictors. In most cases, the performance drops only by 0.01 to 0.04. However, YLoc-LowRes plants shows a considerable performance loss on the BaCelLo plant IDS. In contrast, on the Höglund IDS, we observe a slight performance gain. For details see supplementary material of (26).

To show that users can benefit from the integrated confidence score, we analyzed the performance enrichment for high confidence scores. We reevaluated the performance of the YLoc predictors on the BaCelLo animals IDS by considering only proteins that could be predicted with a minimum confidence score. For statistical reasons, we excluded classes with less than five instances. The performance of YLoc for different minimum confidence scores is shown in Table 2. For the subset of proteins that can be predicted with high confidence, YLoc shows increased prediction performance. Consequently, predictions made with high confidence scores can be rated as more reliable.

We tested YLoc+'s ability to predict multiple localization sites in a nested 5-fold cross-validation scheme on the DBMLoc data set (22). We found that YLoc yields an ACC of 0.64 and an F_1 of 0.68 using multilabel measures. YLoc+ correctly identifies half of the proteins as multiple targeted and predicts both locations correctly in about one-third of the cases.

WEB SERVER

The YLoc web server requires protein sequences in FASTA format as input. It allows users to predict the location of at most 20 proteins. For large-scale predictions, users can access YLoc via SOAP or HTTP using the Python-based client scripts provided on the YLoc web site. Users can choose between three YLoc predictors, YLoc-LowRes, YLoc-HighRes and YLoc+, and three protein origins, animals, fungi and plants. In addition, they can switch off the use of GO term-based features. In this case, YLoc uses models in which the GO terms from close homologs are replaced by sequence-based features. Consequently, these YLoc models rely less on the presence of close homologous proteins. Every prediction will be assigned with a prediction ID that can be used to retrieve results later on. Alternatively, users can simply bookmark the waiting page or result page to obtain results later. Currently, predictions are saved for 2 weeks. The location prediction of a single protein takes 10–20 s, depending on the protein length.

Prediction results are displayed in three levels of details. The prediction summary presents the predicted location(s), the probability of those and the confidence score for every query protein. The probability of a location is simply how likely the protein is located in this compartment. In contrast, the confidence score is a measure of reliability. A low confidence score implies the possibility that the real probability can differ considerably from the predicted probability. However, a high confidence score signifies that the predicted probability is close to the real probability for being located in the predicted location. Consequently, higher confidence scores imply a higher reliability of the prediction for the individual sequence. In addition, an explanation in natural language clarifies why the prediction has been made. This explanation includes the two most likely reasons for this localization, for example: 'The most important reason for making this prediction is the strong SP sorting signal' or 'Moreover, it is a barely charged protein.' This information can be very

important since it might already give a hint of the underlying mechanism for this protein localization.

The detailed prediction page provides more information on a particular protein prediction. For example, the probability distribution of the locations is provided. It is important to know the runner-up locations, especially for low confidence predictions, since rather ambiguous predictions should be inspected manually. YLoc also provides the most similar protein from Swiss-Prot 42.0 and associated GO terms. More details of how protein attributes influence the prediction are given in a large attribute table (Figure 1). The attributes are expressed in biological terms and ordered according to their absolute discrimination score, which corresponds to its influence on the prediction. A positive discrimination score implies that the attribute value is very typical for the predicted location, but atypical for some other location. In contrast, a negative discrimination score implies that the attribute value is more typical for some other location than the predicted one. A simple +/− encoding shows whether an attribute is typical for a location or not. By simply inspecting only the first lines of the table, it is sometimes already obvious which biological property lead to the prediction outcome and is likely to be responsible for the real localization of the protein. In addition, it gives hints of which parts of the protein should be considered for protein engineering.

How a particular biological attribute is calculated can be found on a detailed attribute page (Figure 2). For example, YLoc-LowRes (animal version) calculates the strength of the SP sorting signal using the 'autocorrelation of every third hydrophobic amino acid within the first 20 amino acids in the N-terminus'. Knowing how the attribute value is calculated is essential to understand which particular amino acids and properties encode for possible sorting signal. Furthermore, the attribute is visualized. Embedded Javascript code displays the distribution of proteins from the different locations regarding this feature. The provided protein distributions are very helpful for understanding how proteins from different locations behave with respect to a biological property or sorting signal.

APPLICATION

The interpretable YLoc web service can be applied to numerous tasks that range from large-scale predictions to the identification of sorting signals. A very interesting application example is supervised protein engineering. YLoc can identify biological properties, e.g. example sorting signals that might be responsible for the localization. For example, human fumerate hydratase (FH, SwissProt AC P07954) is primarily located in the mi. The three YLoc predictors (animal version) detect the correct location and identify a mitochondrial targeting peptide (mTP). After truncating the leading 43 residues, FH lacks an mTP and shows a negatively charged N-terminus which is unfavorable for mitochondrial localization. Consequently, YLoc predicts FH to be cytoplasmic. In fact, the truncated FH protein is a known cytoplasmic isoform of FH encoded by the same gene (29). This example shows that YLoc can by valuable in location engineering of proteins.

YLoc VIA SOAP

For large-scale predictions, YLoc can be accessed via SOAP. The corresponding WSDL can be downloaded from the YLoc web site. In addition, we provide a Python-based script. Alternatively, YLoc can be accessed via an HTTP-based client that is also available for download.

CONCLUSION

As an interpretable web server for predicting subcellular localization of proteins, YLoc explains why a prediction was made and what features are likely to be responsible for the protein localization. This information can be very helpful to understand the localization of a protein and

Attribute		Discrimination Score	Cy	Mi	Nu	SP	Detailed Attribute Information
strong	secretory pathway sorting signal	5.72	--	--	--	++	Attribute Details...
barely	charged protein	2.89	+	--	-	++	Attribute Details...
no	mono NLS sorting signal	2.58	--	+	--	++	Attribute Details...
strong	putative mitochondrial or secretory pathway sorting signal	2.42	--	++	--	++	Attribute Details...
very	hydrophobic protein	2.32	--	-	--	++	Attribute Details...
very	hydrophobic N-terminus	2.06	--	--	--	++	Attribute Details...
absent	GO:0005576 (extracellular region)	-1.77	+	+	+	--	Attribute Details...
no	putative mitochondrial sorting signal	1.68	+	--	+	++	Attribute Details...
barely	negatively charged N-terminus	1.56	--	++	--	++	Attribute Details...
absent	GO term GO:0005739 (mitochondrion)	1.55	+	--	+	+	Attribute Details...

Figure 1. The attribute table of the YLoc web service lists all attributes in order of their influence on the prediction outcome. All attributes are expressed in biological terms. The +(+) or −(−) indicates whether that attribute value is (very) typical or (very) untypical for a location.

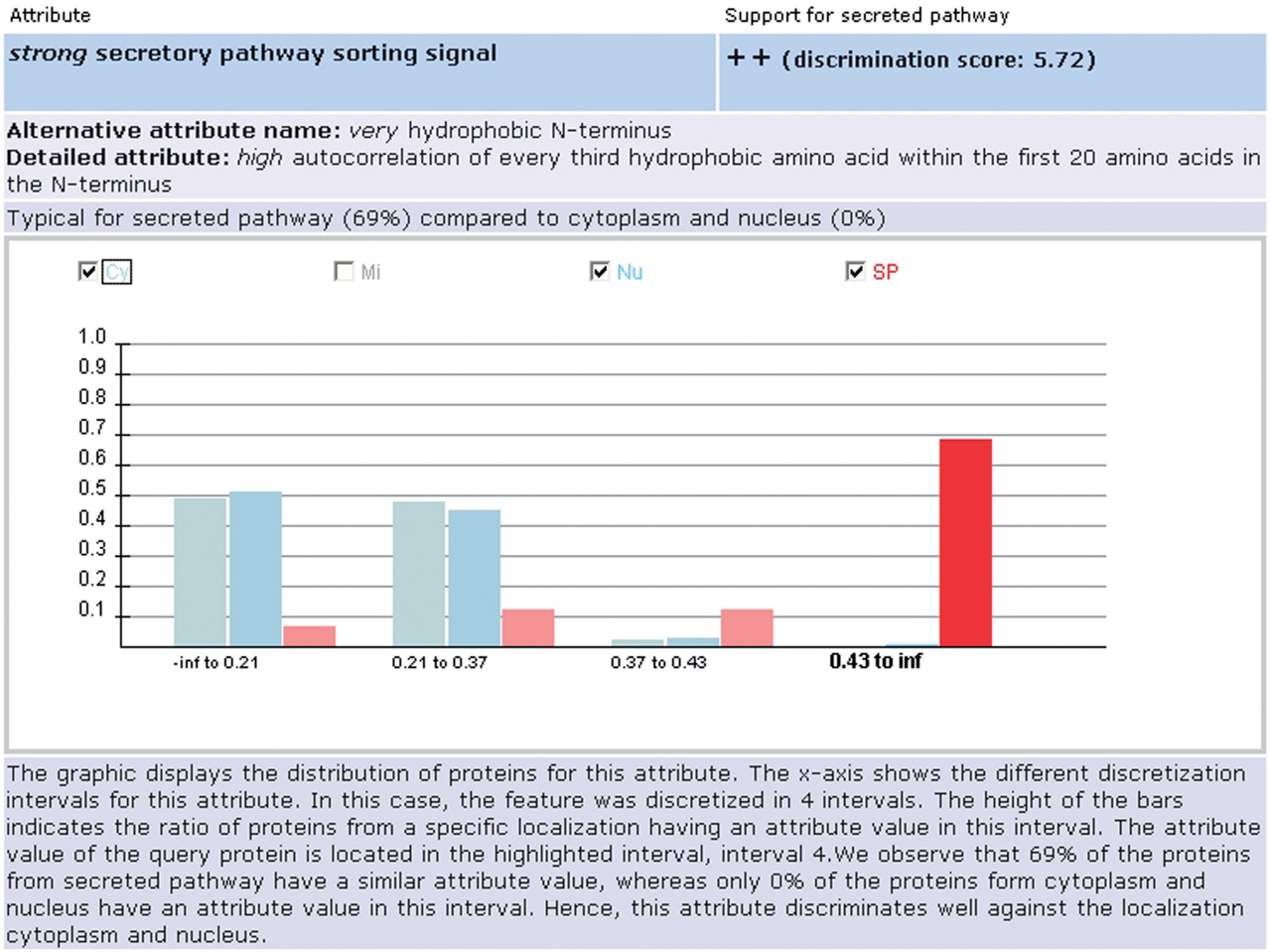

Figure 2. The detailed attribute page of the feature 'secretory pathway sorting signal' in YLoc-LowRes (animal version). The distribution of proteins from the cy, mi, nu and SP over the different attribute intervals is shown.

thus can assist in location engineering of proteins. Furthermore, a confidence score rates the reliability of a prediction. At the same time, it performs comparably with other state-of-the-art predictors. We believe that YLoc is a valuable alternative to experimental methods and current state-of-the-art predictors.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jan Schulze for technical support and Nora Toussaint for comments on the manuscript.

FUNDING

LGFG Promotionsverbund 'Pflanzliche Sensorhistidinkinasen' of the Universtity of Tübingen (S.B.). Funding for open access charge: LGFG Promotionsverbund 'Pflanzliche Sensorhistidinkinasen' of the Universtity of Tübingen.

Conflict of interest statement. None declared.

REFERENCES

1. Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.
2. Horton,P. and Nakai,K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbor classifier. *Intell. Syst. Mol. Biol.*, **5**, 147–152.
3. Chou,K. and Cai,Y. (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.*, **90**, 1250–1260.
4. Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.
5. Petsalaki,E., Bagos,P., Litou,Z. and Hamodrakas,S. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.
6. Pierleoni,A., Martelli,P., Fariselli,P. and Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
7. Höglund,A., Dönnes,P., Blum,T., Adolph,H. and Kohlbacher,O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.
8. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.

9. Horton,P., Park,K., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
10. Chou,K. and Cai,Y. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **277**, 45765–45769.
11. Scott,M., Thomas,D. and Hallett,M. (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, **14**, 1957–1966.
12. Garg,A. and Raghava,G. (2008) ESLpred 2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics*, **9**, 503.
13. Lin,H.N., Chen,C.T., Sung,T.Y., Ho,S.Y. and Hsu,W.L. (2009) Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics*, **10**, S8.
14. Huang,W., Tung,C., Ho,S., Hwang,S. and Ho,S. (2008) ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, **9**, 80.
15. Brady,S. and Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pacific Symposium on Biocomputing*. World Scientific, pp. 604–615.
16. Fyshe,A., Liu,Y., Szafron,D., Greiner,R. and Lu,P. (2008) Improving subcellular localization prediction using text classification and the Gene Ontology. *Bioinformatics*, **24**, 2512–2517.
17. Chou,K. and Cai,Y. (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating Gene Ontology. *Biochem. Biophys. Res. Commun.*, **311**, 743–747.
18. Scott,M., Calafell,S., Thomas,D. and Hallett,M. (2005) Refining protein subcellular localization. *PLoS Comput. Biol.*, **1**, e66.
19. Chou,K. and Shen,H. (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.*, **6**, 1728–1734.
20. Blum,T., Briesemeister,S. and Kohlbacher,O. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.
21. Briesemeister,S., Blum,T., Brady,S., Lam,Y., Kohlbacher,O. and Shatkay,H. (2009) SherLoc2: a high-accuracy hybrid method for predicting protein subcellular localization. *J. Proteome Res.*, **8**, 5363–5366.
22. Zhang,S., Xia,X., Shen,J., Zhou,Y. and Sun,Z. (2008) DBMLoc: a database of proteins with multiple subcellular localizations. *BMC Bioinformatics*, **9**, 127.
23. Hall,M. (2000) Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufman Publishers, pp. 359–366.
24. Whitten,I. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman Publishers, San Fransisco, CA.
25. Fayyad,U.M. and Irani,K. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, pp. 1022–1027.
26. Briesemeister,S., Rahnenführer,J. and Kohlbacher,O. (2010) Going from where to why – interpretable prediction of protein subcellular localization. *Bioinformatics*, **26**, 1232–1238.
27. Casadio,R., Martelli,P. and Pierleoni,A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic Proteomic*, **7**, 63–67.
28. Tsoumakas,G. and Katakis,I. (2007) Multi-label classification: an overview. *Int. J. Data Warehousing Min.*, **3**, 1–13.
29. Tolley,E. and Craig,I. (1975) Presence of two forms of fumarase (fumarate hydratase E.C. 4.2.1.2) in mammalian cells: immunological characterization and genetic analysis in somatic cell hybrids. Confirmation of the assignment of a gene necessary for the enzyme expression to human chromosome 1. *Biochem. Genet.*, **13**, 867–883.