technical notes
Journal of proteome research

SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins

Sebastian Briesemeister,*,† Torsten Blum,† Scott Brady,‡ Yin Lam,‡ Oliver Kohlbacher,† and Hagit Shatkay‡

Division for Simulation of Biological Systems, Center for Bioinformatics Tübingen, Eberhard-Karls-Universität Tübingen, Germany, and School of Computing, Queen's University, Kingston, Ontario, Canada

Received July 24, 2009

Abstract: SherLoc2 is a comprehensive high-accuracy subcellular localization prediction system. It is applicable to animal, fungal, and plant proteins and covers all main eukaryotic subcellular locations. SherLoc2 integrates several sequence-based features as well as text-based features. In addition, we incorporate phylogenetic profiles and Gene Ontology (GO) terms derived from the protein sequence to considerably improve the prediction performance. SherLoc2 achieves an overall classification accuracy of up to 93% in 5-fold cross-validation. A novel feature, DiaLoc, allows users to manually provide their current background knowledge by describing a protein in a short abstract which is then used to improve the prediction. SherLoc2 is available both as a free Web service and as a stand-alone version at http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc2.

Keywords: protein subcellular localization prediction • machine learning • text mining • Gene Ontology

Introduction

The subcellular localization of a protein is highly correlated with its function and is thus an important feature in genome annotation. Computational methods predicting subcellular localization from the amino acid sequence are an attractive alternative to expensive and time-consuming experimental methods. In the past decade, numerous subcellular localization prediction methods have been developed. We distinguish between sequence-based and annotation-based methods.[1,2] Sequence-based predictors make use of known sorting signals,[3−6] amino acid composition information,[2,7−11] or both.[12,13] In contrast, annotation-based predictors use information about functional domains and motifs,[14,15] protein−protein interaction,[16] homologous proteins,[17] annotated Gene Ontology (GO) terms,[18,19] and textual information from Swiss-Prot keywords[20] or PubMed abstracts.[21,22] Hybrid methods[23−28] combine sequence-based information with annotation-based information and, therefore, often achieve excellent prediction performance.

SherLoc2 is a comprehensive hybrid method for subcellular localization prediction which predicts all 11 main eukaryotic locations: nucleus (nu), cytoplasm (cy), mitochondrion (mi), chloroplast (ch), extracellular space (ex), plasma membrane (pm), peroxisome (pe), endoplasmic reticulum (er), Golgi apparatus (go), lysosome (ly), and vacuole (va). It is optimized for animal, fungal, and plant proteins and predicts 9 or 10 locations for each. SherLoc2 combines MultiLoc2,[24] a prediction system based on several sequence-derived features, and EpiLoc,[21] a prediction system based on features derived from PubMed abstracts. It integrates information on amino acid composition, known N-terminal sorting signals, domain motifs, phylogenetic profiles, GO terms derived from the primary sequence, and distinguishing terms that occur in PubMed abstracts to make a final prediction.

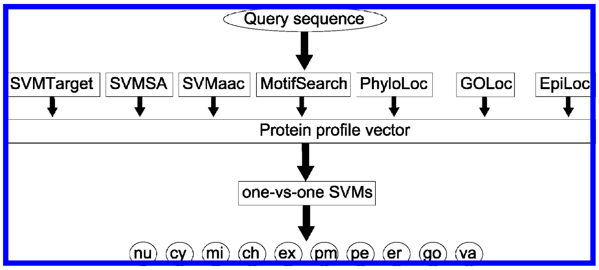

Figure 1. The architecture of SherLoc2 (plant version).

We compared SherLoc2 to current state-of-the-art tools (MultiLoc2,[24] WoLF PSORT,[12] and Euk-mPloc[25]) using independent data sets sharing very low sequence identity with any of the training data sets. SherLoc2 performs considerably better than related predictors for animal and plant proteins and comparably well for fungal proteins. By integrating annotations and textual information, SherLoc2 yields high accuracy predictions that are beneficial for genome annotation.

Methods

SherLoc2 is a support vector machine-based prediction system that integrates features from different sources in a similar way to its predecessor SherLoc.[23] The difference lies in the different feature sources used for SherLoc2. The output of seven subclassifiers is collected in a protein profile vector that in a second step forms the input for a final support vector machine (SVM) classifier (Figure 1).

Each of the seven subclassifiers utilizes a different kind of information: SVMTarget is based on N-terminal targeting signals and uses partial amino acid composition in the N-

[*] To whom correspondence should be addressed. E-mail: briese@informatik.uni-tuebingen.de. Tel: +49 7071 2970462. Fax: +49 7071 295152.
[†] Eberhard-Karls-Universität Tübingen.
[‡] Queen's University.

10.1021/pr900665y CCC: $40.75

Published on Web 09/18/2009

terminus as input. In contrast, the input of SVMaac is the overall amino acid composition. SVMSA scans for signal anchors present in membrane proteins and uses partial amino acid composition as input. MotifSearch searches for other known localization signals as well as for relevant sequence motifs. PhyloLoc uses phylogenetic profiles[29] that encode coinheritance of a protein in different organisms as input. The input of GOLoc is a vector of GO terms inferred from the protein sequence using InterproScan. Thus, we only use GO terms that are associated with InterPro domains and not annotated GO terms from a query protein. EpiLoc, a text-based subclassifier, uses weighted term vectors representing the PubMed abstracts linked to the protein's Swiss-Prot entry. Selected distinguishing terms are used as features in the term vector. All classifiers, except for MotifSearch, are SVM-based. The output of all subclassifiers, a probability distribution over the locations in the case of SVM-based classifiers, forms the input for the final one-vs-one SVM. The probability estimate for locations from the final SVM is used to rank each location and to select the most probable as output. We used one-vs-one SVMs with radial basis kernel function using the LIBSVM software.[30] SVM parameters were optimized by a grid search.

SVMTarget, SVMaac, SVMSA, MotifSearch, and PhyloLoc are purely sequence-based. In contrast, GOLoc and EpiLoc are term-based predictors. In cases where no GO term is found by InterProScan, GOLoc returns a uniform probability vector. In cases where a protein has no Swiss-Prot AC or no PubMed abstract linked to it, HomoLoc,[21] a module of EpiLoc, is applied. If no homologous protein can be found, HomoLoc returns a uniform probability vector. Thus, in this case, SherLoc2 cannot make use of text-derived information. For more details on the remaining subclassifiers, we refer to Blum et al.[24] and Brady and Shatkay.[21]

In cases where no text can be associated with the protein, we offer users a new interactive feature, by incorporating DiaLoc[21] into the system. It allows individual researchers to enter their own short textual description, at least 20 words in length up to the equivalent of an abstract, based on their current knowledge, and, thus, obtain a localization prediction. Like EpiLoc, DiaLoc uses the vector representation of the provided text in order to derive features used for localization prediction.

Results

Cross-Validation Evaluation. For training SherLoc2, we used the original MultiLoc data set,[13] which contains 5959 eukaryotic proteins extracted from Swiss-Prot release 42.0 and covers 11 locations (cy, ch, er, ex, go, ly, mi, nu, pe, pm, va). The prediction performance was measured using overall accuracy (ACC), which is the ratio of correctly predicted proteins, and average sensitivity (AVG), which is the average fraction of called instances from a class. They are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fp} + \mathrm{fn}} \qquad \mathrm{AVG} = \frac{1}{c}\sum_{i=1}^{c} \frac{\mathrm{tp_i}}{\mathrm{tp_i} + \mathrm{fn_i}}$$

where c denotes the number of classes and tp, tn, fp, and fn equal the number of true positives, true negatives, false positive, and false negative instances, respectively. We believe that the AVG is better suited as an evaluation measure since it is not biased toward overrepresented classes. In a 5-fold cross-validation setting, SherLoc2 yields 6−8% higher AVGs as well as ACCs compared to SherLoc and 3−4% compared to MultiLoc2 (Table 1). The gain in performance can be explained by the predictive power of phylogenetic profiles, GO terms, as well as the integration with EpiLoc.

Table 1. 5-Fold Cross-Validation Performance Comparison of SherLoc2, MultiLoc2, and SherLoc with Respect to AVG (ACC)[a]

data set	SherLoc2	MultiLoc2	SherLoc
Animals	**0.94** (**0.93**)	0.89 (0.89)	0.87 (0.86)
Fungi	**0.94** (**0.93**)	0.89 (0.89)	0.85 (0.85)
Plants	**0.94** (**0.93**)	0.89 (0.89)	0.86 (0.85)

[a] The best scoring method regarding each measure is highlighted in bold for all data sets.

Table 2. Performance of SherLoc2, MultiLoc2, WoLF PSORT, Euk-mPloc, and the BLAST Predictor on the BaCelLo IDSs and the Höglund Animal IDs with Respect to AVG (ACC)[a]

data set	SherLoc2	MultiLoc2	WoLF PSORT	Euk-mPloc	BLAST
BaCelLo Animals	**0.76** (**0.71**)	0.75 (0.68)	0.69 (**0.71**)	0.48 (0.58)	0.35 (0.37)
BaCelLo Fungi	0.61 (**0.59**)	0.59 (0.53)	**0.62** (0.51)	**0.62** (0.57)	0.34 (0.39)
BaCelLo Plants	**0.69** (**0.69**)	0.65 (0.62)	0.46 (0.57)	0.44 (0.41)	0.58 (0.61)
Höglund Animals	**0.39** (0.54)	0.38 (**0.57**)	0.24 (0.56)	0.18 (0.22)	0.06 (0.13)

[a] The best-scoring method regarding each measure is highlighted in bold for all data sets.

Independent Data Set Evaluation. The cross-validation results cannot be used for fair comparison with other prediction methods, since training data of subcellular localization predictors is often not comparable due to different protein sources and different preprocessing. To ensure fair comparison, we applied SherLoc2 to two independent data sets (IDSs). The BaCelLo IDS[31] covers five main eukaryotic locations (nu, cy, mi, secretory pathway, ch for plants). The Höglund IDS consists of animal proteins[24] and covers the remaining main eukaryotic locations (ex, pm, pe, er, go, ly). Both IDSs consist of proteins that were added to Swiss-Prot after release 42.0, whereas SherLoc2 was trained only on proteins from Swiss-Prot release 42.0. Moreover, proteins with a sequence similarity of more than 30% to a protein in the training data set were removed. Note that SherLoc2 does not use information annotated to these proteins since it is restricted to Swiss-Prot release 42.0. Unfortunately, there are insufficient fungal and plant proteins from locations along the secretory pathway to construct independent data sets of reasonable size.

We assess the performance of SherLoc2 by comparing it against three other prediction systems: MultiLoc2,[24] WoLF PSORT,[12] and Euk-mPloc.[25] All three prediction methods are high-resolution predictors that distinguish locations along the secretory pathway. Moreover, they are widely used since they are available as a Web service. Because transferring annotations from homologous, already annotated proteins is a common approach for predicting subcellular localization, we also assign to proteins the location of the top-ranked annotated nonambiguous BLAST hit. To guarantee fair assessment, HomoLoc and BLAST are restricted to Swiss-Prot version 42.0.

In this comparison, the performance of SherLoc2 is superior to other methods for animal and plant proteins and comparable for fungal proteins (Table 2). The integration of EpiLoc leads, in most cases, to performance gains in both AVG and ACC. For plant proteins, the performance gain is 7% in ACC and 4% in AVG. This is because text from homologous proteins was available for all plant proteins (see Supporting Information). SherLoc2 often shows a higher sensitivity than other predictors, particularly for cytoplasmic, nuclear, and chlorplast proteins. Compared to MultiLoc2, SherLoc2 correctly recovers up to 20% more cytoplasmic proteins. The performance of all predictors is relatively low for the Höglund IDS. This is due to the limited number of available training data for the peroxisome and the secretory pathway locations. Nevertheless, for this data set, SherLoc2 performs considerably better than WoLF PSORT and Euk-mPloc. Since the number of protein sequences of the Höglund IDS is comparably low, the performance results should be seen as a trend. As expected,[32] predictions based on homology alone are inferior to those based on other classifiers. For example, the transmembrane adapter protein PAG (Swiss-Prot AC Q9NWQ8) is located in the plasma membrane. However, the most similar protein from Swiss-Prot 42.0 is isoleucyl-tRNA synthetase (Swiss-Prot AC P09436), a cytoplasmic protein. This demonstrates the value of predictions based on both sequence and annotation features. Because of excellent prediction performance, we believe that SherLoc2 is suitable for subcellular localization prediction in the context of automatic genome annotations. More details concerning the performance of all predictors can be found in the Supporting Information.

Conclusion

SherLoc2 is a hybrid subcellular localization predictor that combines sequence-based and text-based information. It outperforms other prediction methods by benefiting from the predictive power of both information sources.

In the future, we plan to integrate text sources additional to PubMed abstracts. Moreover, incorporating proteins localized to multiple cell compartments will be an interesting and challenging task for the future.

Availability. SherLoc2 is available both as a free Web service and as a stand-alone version at http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc2. The Web service offers a user-friendly and straightforward interface for predictions with up to 20 protein sequences. In addition, DiaLoc is available as a stand-alone Web service at http://epiloc.cs.queensu.ca/DiaLoc.html.

Acknowledgment. S. Briesemeister is supported by the LGFG Promotionsverbund "Pflanzliche Sensorhistidinkinasen". H. Shatkay is supported by an NSERC Discovery Grant 298292-04, CFI New Opportunities Award 10437, and an Ontario Early Researcher Award #ER07-04-085. We thank Prof. Dorothea Blostein for her help and support.

Supporting Information Available: The Supporting Information includes detailed information on the used data sets, the BLAST predictor, and the IDS predictions as well as insights how to use HomoLoc and DiaLoc. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Emanuelsson, O.; Brunak, S.; von Heijne, G.; Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2007**, *2* (4), 953–971.
(2) Nair, R.; Rost, B. Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* **2005**, *348* (1), 85–100.
(3) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **2004**, *340* (4), 783–795.
(4) Emanuelsson, O.; Nielsen, H.; Brunak, S.; von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **2000**, *300* (4), 1005–1016.
(5) Bannai, H.; Tamada, Y.; Maruyama, O.; Nakai, K.; Miyano, S. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **2002**, *18* (2), 298–305.
(6) Boden, M.; Hawkins, J. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics* **2005**, *21* (10), 2279–2286.
(7) Hua, S.; Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **2001**, *17* (8), 721–728.
(8) Park, K. J.; Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **2003**, *19* (13), 1656–1663.
(9) Xie, D.; Li, A.; Wang, M.; Fan, Z.; Feng, H. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* **2005**, *33*, W105–W110.
(10) Pierleoni, A.; Martelli, P. L.; Fariselli, P.; Casadio, R. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* **2006**, *22* (14), e408−e416.
(11) Chou, K. C.; Cai, Y. D. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.* **2003**, *90* (6), 1250–1260.
(12) Horton, P.; Park, K. J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C. J.; Nakai, K. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **2007**, *35*, W585–W5857.
(13) Höglund, A.; Dönnes, P.; Blum, T.; Adolph, H. W.; Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* **2006**, *22* (10), 1158–1165.
(14) Chou, K. C.; Cai, Y. D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277* (48), 45765–45769.
(15) Scott, M. S.; Thomas, D. Y.; Hallett, M. T. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.* **2004**, *14* (10 a), 1957–1966.
(16) Shin, C. J.; Wong, S.; Davis, M. J.; Ragan, M. A. Protein-protein interaction as a predictor of subcellular location. *BMC Syst. Biol.* **2009**, *3*, 28.
(17) Garg, A.; Raghava, G. P. S. ESLpred 2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinf.* **2008**, *9*, 503.
(18) Huang, W. L.; Tung, C. W.; Ho, S. W.; Hwang, S. F.; Ho, S. Y. ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinf.* **2008**, *9*, 80.
(19) Lei, Z.; Dai, Y. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinf.* **2006**, *7*, 491.
(20) Nair, R.; Rost, B. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* **2002**, *18Suppl* (1), S78−S86.
(21) Brady, S.; Shatkay, H. EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.* **2008**, 604–615.
(22) Fyshe, A.; Liu, Y.; Szafron, D.; Greiner, R.; Lu, P. Improving subcellular localization prediction using text classification and the Gene Ontology. *Bioinformatics* **2008**, *24* (21), 2512–2517.
(23) Shatkay, H.; Höglund, A.; Brady, S.; Blum, T.; Dönnes, P.; Kohlbacher, O. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* **2007**, *23* (11), 1410–1417.
(24) Blum, T.; Briesemeister, S.; Kohlbacher, O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular localization prediction. *BMC Bioinf.* **2009**, *10*, 274.
(25) Chou, K. C.; Shen, H. B.; et al. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6* (5), 1728–1734.
(26) Lee, K.; Chuang, H. Y.; Beyer, A.; Sung, M. K.; Huh, W. K.; Lee, B.; Ideker, T. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res.* **2009**, *36* (20), e136.
(27) Scott, M. S.; Calafell, S. J.; Thomas, D. Y.; Hallett, M. T. Refining protein subcellular localization. *PLoS Comput. Biol.* **2005**, *1* (6), e66.

(28) Chou, K. C.; Cai, Y. D. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun.* **2003**, *311* (3), 743–747.

(29) Marcotte, E. M.; Xenarios, I.; Van der Bliek, A. M.; Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (22), 12115–12120.

(30) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

(31) Casadio, R.; Martelli, P. L.; Pierleoni, A. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Briefings Funct. Genomics Proteomics* **2008**, *7* (1), 63–73.

(32) Eisenhaber, F.; Bork, P. Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.* **1998**, *8* (4), 169–170.

PR900665Y