Drug Discovery Today • Volume 12, Numbers 5/6 • March 2007

Computational classification of classically secreted proteins

Eric W. Klee[1] and Carlos P. Sosa[2,3]

[1] Stabile 3-15, Department of Laboratory Medicine, Mayo Clinic, Rochester, MN 55905, USA
[2] IBM, 650 3rd Ave. South, Minneapolis, MN 53402 USA
[3] University of Minnesota Supercomputing Institute, 599 Walter Library, 117 Pleasant St, Minneapolis, MN 55455m USA

The ability to identify classically secreted proteins is an important component of targeted therapeutic studies and the discovery of circulating biomarkers. Here, we review some of the most recent programs available for the *in silico* prediction of secretory proteins, the performance of which is benchmarked with an independent set of annotated human proteins. The description of these programs and the results of this benchmarking provide insights into the most recently developed prediction programs, which will enable investigators to make more informed decisions about which program best addresses their research needs.

Predicting secreted proteins

Determining the subcellular localization of a protein can provide insights into how it functions and the pathways that are involved, as well as highlighting whether the protein could either provide a therapeutic target or act as a biomarker. Proteins that are processed through the secretory pathway and included in the secretome [1] are a niche of the proteome that has been the focus of receptor antagonist-binding-site studies and therapeutic development [2]. Secreted proteins are also targeted in many biomarker studies because they are more likely to be present in body fluids and, ultimately, measured by non-invasive assays [3]. With recent trends towards personalized medicine, the development of diagnostic and theragnostic biomarkers that predict drug efficacy and guide individual treatment decisions are important objectives of the drug-development community [4,5]. Unfortunately, experimental determination of localization is only available for ~30% of the human proteome [6], and often researchers rely on localization-prediction programs to annotate gene products of interest. Deciding which prediction programs to use is a daunting task because, each year, more programs are created and published and are readily available. Programs use different computation methods, predict localization to different locations and, often, are not either independently benchmarked or easily compared. To aid investigators in understanding the recent developments in secreted protein prediction, here, we describe several new prediction programs and measure their performance based on an independent benchmark.

The localization of most proteins occurs through a cascade of sorting events that are directed by either short peptides or motifs that enable site-specific uptake, retention and transport [7,8]. Often, these sorting events vary depending on the kingdom, thus, we concentrate here on protein localization in humans. In eukaryotes, the initial sorting event occurs during translation when proteins that enter the secretory pathway are cotranslationally translocated into the endoplasmic reticulum (ER) and, thereby, differentiated from other intracellular proteins. This process is mediated by an N-terminal signal peptide on the nascent protein sequence. Although the peptide sequence is not conserved strictly between proteins, it does contain three regions with conserved properties; an N-terminal region of basic residues, a central region of hydrophobic residues, and a C-terminal region of polar residues. Following translocation to the ER, the signal peptide is cleaved from the mature protein. Proteins that are processed in this manner are referred to commonly as classically secreted proteins. These proteins use secondary signals, such as the C-terminal 'KDEL' and 'HNEL' motifs for ER retention, to direct final localization to the ER, Golgi, lysosome, cellular membrane and extracellular space [9,10]. Whereas most extracellular proteins are

Corresponding author: Sosa, C.P. (cpsosa@us.ibm.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.01.008

processed through this pathway, a few are exported from the cell by independent mechanisms. These proteins, called non-classically secreted proteins [11], are often not well characterized by localization-prediction programs.

The semi-conserved nature of the N-terminal signal peptide makes it an attractive target for computational algorithms. The earliest methods to predict secreted proteins used decision rules and weighted matrices to identify N-terminal signal peptides and predict classically secreted proteins [7,12]. The predictive accuracy of these pioneering programs was limited, in part, by the number of well annotated proteins that were available to train the algorithms. Subsequent revision and adaptation of von Heijne's weighted matrix method has improved predictive accuracy [13,14], but not matched the accuracy obtained when advanced machine learning algorithms were incorporated into prediction programs. SignalP was one of the first programs to use artificially trained neural networks to identify N-terminal signal peptides and predict secreted proteins [15]. The use of neural networks greatly improved prediction accuracy. By retraining its neural networks with more recent annotated sequences, revising the program architecture and incorporating a Hidden Markov Model (HMM) component in its prediction strategy, SignalP remains a mainstay of secreted protein prediction [16,17]. PSORT, another widely used predictor of secreted proteins, uses a different strategy. Instead of targeting secreted proteins specifically, PSORT uses a set of 'if-then' rules to predict protein localization simultaneously to one of 14 cellular locations [18,19]. Subsequent revisions of this program (PSORT II) replaced the 'if-then' rule set with a kth nearest-neighbor algorithm, and implemented new feature-selection methods (WoLF PSORT) [20–22]. PSORT and SignalP were developed during the infancy of protein-localization prediction. They employ two different strategies for predicting secreted proteins; targeted prediction and global prediction, and are still used routinely in research today. To better understand if the new prediction programs that are reviewed in this article offer improvements over existing methods, WoLF PSORT and SignalP 3.0 are included in the performance testing.

Development of new localization prediction programs has been prolific in recent years. Several advanced machine-learning algorithms, including, neural networks [15,16,23–30], HMMs [16,17,31–37] and support vector machines (SVMs) [6,38–51], have been used to identify N-terminal signal peptides and predict secreted proteins, In addition., other programs have been developed to predict secreted proteins using methods based on overall and regional amino acid composition, homology to experimentally annotated proteins (BLAST comparisons), co-occurrence of protein domains, and keyword mining of auxiliary protein annotations [52–56]. It is difficult to evaluate which program provides the best prediction because the prediction accuracies reported by authors are often overoptimistic and not comparable. Although several performance-benchmarking studies [57–60] provide independent assessments of the prediction accuracies of older programs, a review of the newly developed prediction programs is needed. Here we review and benchmark eight secreted protein-prediction programs (Table 1) that were published in 2005 and 2006. For each program, we provide a concise description of the implemented architecture, training data, computational methods and predicted outputs. Following the program summaries, we report the prediction accuracies for these eight programs, plus SignalP and PSORT, based on an independent set of annotated protein sequences.

Balanced subcellular localization predictor (BaCelLo)

BaCelLo uses SVMs to predict protein localization to the secretory pathway, cytoplasm, nucleus, mitochondrion and chloroplast [45]. Predictions are based on the amino acid composition of the full length protein, an N-terminal subsequence, a C-terminal subsequence and an amalgamated homolog profile.

TABLE 1

Summary of localization predictors

Program[a]	Locations[b]	Detection features	Method	Training data
Protein Prowler	mit, sigpep	N-terminal peptides	Neural network, SVM	SwissProt 37 and 38
BaCelLo	secpro, cyt, nuc, mit	AA composition	SVM	SwissProt 48
PolyPhobius	tm, sigpep	N-terminal peptides, transmembrane domains, sequence homology	HMM	SwissProt 41
MultiLoc	cyt, mit, secpro, er, gol, lys, nuc, per, pm,	N-terminal peptides, AA composition, signal anchors, protein domains	SVM	SwissProt 42
PredSL	mit, secpro	N-terminal peptides	Neural network, HMM, Markov chains, PrediSi	UniProt 3.5
SignalP 3.0	mit, sigpep	N-terminal peptides	Neural network, HMM	SwissProt 40
LOCtree	ext,org, cyt, mit, nuc	AA composition	SVM, SignalP	SwissProt 40
pTARGET	cyt, er, ext, gol, lys, mit, nuc, pm, per	Protein domains, AA composition	Weighted matrices	SwissProt 45
WoLF PSORT	mit, er, cyt, ext, gol, lys, nuc, per, pm	AA Composition, sequence length, PSORT features, iPSORT features	k-Nearest neighbors, PSORT, iPSORT,	SwissProt 45
HSLpred	cyt, mit, nuc, pm	AA composition, sequence homology	SVM, PSI-BLAST	SwissProt 44.1

[a] The first six prediction programs explicitly identify the general class of classically secreted proteins. The last seven proteins identify the more specific subclass of classically secreted proteins (gol, ext, lys, pm, er and org).

[b] Abbreviations: cyk, cytoskeleton; cyt, cytosolic; er, endoplasmic reticulum; ext, extracellular; gol, Golgi apparatus; lys, lysosomal; mit, mitochondrial; nuc, nucleus; org, organelles; per, peroxisome; pm, plasma membrane; ppl, periplasm; sigpep, signal peptide; secpro, secreted protein; tm, transmembrane proteins.

The amalgamated homolog profile is constructed from the target protein and all homologs identified by BLAST comparisons (e-value $\leq 1e^{-10}$) to SWISSPROT release 48. By using all four amino acid-composition parameters, BaCelLo incorporates N-terminal and C-terminal motifs into its predictions, and also includes additional sequence information that is relevant to predicting non-classically secreted proteins.

The final protein-localization predictions provided by BaCelLo are based on a cascading series of binary decisions that discriminate: (1) secretory proteins from other intracellular proteins; (2) nuclear and cytoplasmic proteins from mitochondrial and chloroplast proteins; (3) nuclear proteins from cytoplasmic proteins; and (4) mitochondrial proteins from chloroplast proteins (for plants). Protein sequences from SWISSPROT release 48 were used for algorithm training. Proteins that contain qualifying annotation terms, including 'fragment', 'possible', 'probable' and 'by similarity', were excluded. Proteins were filtered to ensure no two training sequences possessed >30% identity. The program is available at (http://www.gpcr.biocomp.unibo.it/bacello/).

HSLpred

HSLpred uses SVMs to predict protein localization to the mitochondria, nucleus, cytoplasm and plasma membrane [46]. Predictions are based on the amino acid and dipeptide composition of the full protein sequence, and PSI-BLAST-determined homology to 3532 proteins with experimentally confirmed localizations. The program uses a two-tiered architecture in its decision making. The first tier generates independent scores for each of the four possible locales. The, second tier combines the independent first tier scores and generates a final prediction of localization. HSLpred was trained using SWISSPROT release 44.1 sequences. Proteins were filtered against qualifying terms and the redundancy reduced so that no two training sequences have >90% identity. HSLpred is available at (http://www.imtech.res.in/raghava/hslpred/).

LOCtree

LOCtree uses SVMs to predict protein localization to one of five locales [6]. This program introduces the concept of mimicking *in vivo* sorting events using a series of algorithmic predictions. Localization is assigned through a cascading series of discriminators that differentiate: (1) secretory proteins from other intracellular proteins; (2a) secretory proteins into extracellular and organellar proteins; (2b) intracellular proteins into nuclear and cytoplasmic proteins; and (3) cytoplasmic proteins into cytosolic and mitochondrial proteins. Predictions are based on the amino acid composition parameters that are derived from the full sequence of the protein, a 50-residue N-terminal subsequence, and three secondary-structure states. These composition parameters incorporate residue-frequency scores that are obtained from a multi-sequence alignment of the target protein to the SWISSPROT + TrEMBL database. In addition, LOCtree incorporates the predictions from SignalP into its decision process [15,16]. The LOCtree SVMs are trained with SWISSPROT release 40 proteins using sequences filtered for qualifying terms, excluded if annotated to multiple locales, and redundancy reduced to ensure no two training sequences share >25% identity. When analyzing protein sequences on the LOCtree server, predictions based on PredNLS, LOCkey, and localization based on protein domains and motifs are also reported. This program is available at (http://www.cubic.bioc.columbia.edu/cgi-bin/var/nair/loctree/query).

MultiLoc

MultiLoc uses SVMs to predict protein localization to one of nine locations, including, the cytoplasm, ER, extracellular space, lysosomes, mitochondria, Golgi apparatus, peroxisomes, nucleus and plasma membrane [40]. The overall localization prediction incorporates output from four submodules that (1) discriminate classically secreted proteins from mitochondrial proteins on the basis of N-terminal peptides, (2) discriminate extracellular proteins from plasma membrane proteins on the basis of uncleaved, N-terminal signal anchor peptides, (3) discriminate global localization on the basis of complete protein sequence amino acid composition, and (4) identify specific PROSITE [61] and NLSdb [62] motifs that differ between locations. MultiLoc was trained with SWISSPROT release 42 proteins, filtered for qualifying terms, and redundancy reduced to ensure no two training sequences share >80% identity. The program is available at (http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc/).

PolyPhobius

PolyPhobius uses a HMM to predict classically secreted proteins and discriminate them from the closely related N-terminal-signal anchor membrane proteins [34]. The unique combination of identifying N-terminal signal peptides and predicting transmembrane domains creates a powerful tool for the identification of N-terminal transmembrane domains (signal anchors) that many prediction algorithms confuse with N-terminal signal peptides. The program bases predictions on N-terminal sequence analysis with sequence homology modeling. PolyPhobius homology profiles use multi-sequence alignments of homologs identified by BLAST searches (e-value $\leq 10^{-5}$) against the SWISSPROT + TrEMBL database. The program was trained using SwissProt 41 proteins that were redundancy reduced to eliminate homologs. PolyPhobius is available at (http://www.phobius.cgb.ki.se/).

PredSL

PredSL uses neural networks, Markov chains, and HMMs to predict protein localization to the mitochondria, secretory pathway and cytoplasm [47]. Predictions are based on N-terminal subsequences of a protein. Multiple methods for discriminating between mitochondrial transit peptides and N-terminal signal peptides of secreted proteins are incorporated into the final predictions. Included in the decision parameters is analysis by the PrediSi prediction program [14]. PredSL was trained using methionine-initiated eukaryotic sequences from UniProt release 3.5. Proteins with multiple annotations were excluded and 40% reduction in redundancy performed. The program is available at (http://hannibal.biol.uoa.gr/PredSL/).

Protein Prowler

Protein Prowler uses neural networks and SVMs to predict proteins that localize to the mitochondria and secretory pathway [50]. It was developed as a modified form of the TargetP program [24] and demonstrates the value of implementing alternative decision-making architectures. Protein Prowler uses a different subclass of neural networks to independently identify N-terminal signaling

peptides of mitochondrial proteins and classically secreted proteins. The program then uses an SVM to combine the independent neural network predictions and derive an overall prediction of localization. Protein Prowler was trained with the training sequence set that was developed for TargetP. This includes methionine-initiated sequences from SwissProt releases 37 and 38 that are filtered for qualifying terms and redundancy reduced. The program is available at (http://www.pprowler.imb.uq.edu.au/).

pTarget

pTarget uses a numeric scoring system to predict protein localization to the cytoplasm, ER, extracellular space, Golgi, lysosomes, mitochondria, nuclei, plasma membrane and peroxisomes [55,56]. The program was designed to allow rapid analysis of a large number of sequences and thereby facilitate -omic-level analysis. Its scoring system integrates numeric values related to either the presence or the absence of location-specific protein domains, and to the amino acid composition of the protein. The scoring system is configured so that a target protein with a location-specific Pfam domain is automatically predicted to that location. In the absence of a location-specific domain, the final prediction depends on a sum of the Pfam domain scores and the amino acid-composition scores. Amino acid composition is calculated for the complete protein sequence and for an N-terminal subsequence of 25 amino acids. These compositions are then compared to *a priori* compute a locale-to-locale value for residues with ≥20% variation in amino acid frequency between locations. Based on these comparisons, two amino acid-composition scores are computed systematically for all nine locations and incorporated into the final prediction. The program was trained using SWISSPROT release 45.0, non-plant, eukaryotic proteins. Sequences were filtered for qualifying terms, excluded if annotated to multiple locales, and redundancy reduced to ensure no two training proteins share >95% identity. pTarget is available at (http://www.bioinformatics.albany.edu/~ptarget/).

Assessing prediction accuracy

Methods

To generate a comparable metric on which to evaluate the accuracy of prediction of these programs, a test-set of human protein sequences was created from the SwissProt database. All test-set proteins have localization annotations in the CommentType= Subcellular Location field, and a 'Sequence was last modified on' date that corresponds to SwissProt Release 49.0 or higher. Test-set membership was restricted to SwissProt entries with recently modified sequences to try and minimize overlap between test-set proteins and program-training proteins; the most recent training sequences for the programs evaluated in this article were from SwissProt release 48. To reduce testing bias from highly homologous sequences, no two testing proteins shared >25% identity across 75% of the query protein. The test-set was also filtered to remove proteins that contain the qualifying terms 'By similarity', 'Probable' and 'Potential' in the localization annotation. Finally, manual review of the Subcellular Location annotations categorized proteins into four broad types: secretory and non-membrane, secretory and membrane, non-secretory and non-membrane, and non-secretory and membrane (Table 2). Secretory proteins consisted of those that were annotated as secreted, extracellular, ER, Golgi and lysosomal. Annotated membrane proteins were included in the 'secretory and membrane' category unless annotated explicitly as localized to a non-secretory compartment. Any protein that could be categorized as either both secretory and non-secretory or containing ambiguous localization annotation was removed from the test-set.

TABLE 2

Independent test-set of protein sequences

Category	# of sequences
Secretory and non-membrane	33
Secretory and membrane	111
Non-secretory and non-membrane	237
Non-secretory and membrane	18

Test sequences were analyzed using the default configuration for human proteins on all programs. Predictions were performed using web-server interfaces for all programs except LocTree and PredSL. For these programs the developers assisted in the evaluation by analyzing the protein test-set off-line. The summary of prediction results evaluates only the highest predicted location for a protein. SignalP predictions were interpreted from the D-score coefficient and PolyPhobius predictions were interpreted from the SP score. For each program, the number of true positives, number of true negatives, number of false positives, number of false negatives, sensitivity, specificity, and Matthew's correlation coefficient (MCC) [63] were reported. A summary of prediction results are provided in Table 3, with a more detailed description of the individual sequence predictions in the Online Supplementary Material.

Prediction accuracy assessment

Results

The overall performance of the prediction programs evaluated is strong. Almost universally, the specificity of identification of classical secretory proteins is ≥0.9. Although the sensitivity of predictions is markedly lower for the programs, this might be biased in part by the small number of positives (secretory proteins) included in the test-set. There is considerable variation in the ability of programs to successfully predict classically secreted proteins in the two protein test-sets; membrane-protein included and membrane-protein excluded. This is not surprising because some programs, such as LOCtree, are designed explicitly to not analyze membrane proteins, whereas others, such as HSLpred, are designed to predict membrane proteins but not necessarily to identify classically secreted proteins. Clearly, it is important that users consider the scope of the sequences to be analyzed before selecting which program to use.

The most accurate programs for predicting classically secreted proteins from a non-membrane-sequence set are Protein Prowler (0.90 MCC), BaCelLo (0.88 MCC) and SignalP (0.87 MCC). The accurate predictions by Protein Prowler are rather surprising because this program was trained with one of the oldest sequence sets: it is likely that the program architecture and implementation of alternative decision networks led to this result. The competitive performance of Protein Prowler also raises the question of the how much added benefit is obtained by increasing the number of annotated proteins for algorithm training.

The two legacy programs, SignalP and PSORT, both possess strong predictive accuracies. PSORT generated the most consistent

TABLE 3

Prediction accuracies obtained by analyzing an independent test-set of protein sequences

Parameter	BaCelLo[a]	LocTree	Protein Prowler	SignalP 3.0	pTarget	HSL pred	MultiLoc	WoLF PSORT	PredSL	Phobius
Without membrane proteins										
True positives	29	31	30	30	26	7	29	32	31	31
True negatives	3	2	3	3	7	26	4	1	2	2
False positives	209	224	234	232	219	225	212	224	226	229
False negatives	4	13	3	5	18	12	25	13	11	8
Sensitivity	0.91	0.94	0.91	0.91	0.79	0.21	0.88	0.97	0.94	0.94
Specificity	0.98	0.95	0.99	0.98	0.92	0.95	0.89	0.95	0.95	0.97
MCC	0.88	0.78	0.90	0.87	0.63	0.21	0.63	0.80	0.81	0.84
With membrane proteins										
True positives	81	95	84	86	93	97	113	122	100	80
True negatives	55	49	60	58	51	47	31	22	44	64
False positives	225	241	252	246	235	239	223	239	240	246
False negatives	5	14	3	9	20	16	32	16	15	9
Sensitivity	0.60	0.66	0.58	0.60	0.65	0.67	0.78	0.85	0.69	0.56
Specificity	0.98	0.95	0.99	0.96	0.92	0.94	0.87	0.94	0.94	0.96
MCC	0.65	0.65	0.66	0.63	0.60	0.65	0.66	0.79	0.67	0.60

[a] BaCelLo prediction accuracies were based on a reduced test-set that lacks 57 proteins (40 non-secretory and non-membrane, six non-secretory and membrane, two secretory and non-membrane, and ten secretory and membrane).

predictions regardless of whether membrane proteins are included (0.79 MCC) or excluded (0.80 MCC) from the test-set. This independent benchmarking indicates that older prediction programs should not be disregarded rapidly and replaced by the latest programs. In fact, it is important to note that older programs that have been used routinely for prediction for several years often have more reliable user interfaces and web-servers. While evaluating programs that have been developed during the past two years, we found instances where the public interface of a program was already non-responsive and unsupported.

It is important to note that this benchmarking study has been designed to evaluate the ability to predict classically secreted proteins. Six of the programs, Protein Prowler, BaCelLo, PolyPhobius, MultiLoc, PredSL and SignalP, are designed to report this classification of protein localization. The remaining programs generate predictions to more specific localizations (e.g. extracellular space, Golgi, ER, plasma membrane, lysosomes and organelles). For the purpose of this study, the more specific location predictions were grouped into general secretory and non-secretory categories. This was done, in part, because of limitations in the number of test proteins available for the assessment and, in part, to allow comparison of all prediction algorithms on a single metric. However, an unfortunate side-effect of this grouping is the trivialization of the true predictive capacity of several of these programs to identify more specifically localized proteins, which should be considered when interpreting the benchmarking results. Finally, it should be noted that PolyPhobius might perform better in the analysis of the test-set that includes membrane proteins if its transmembrane-prediction parameter is considered in addition to its signal peptide-prediction parameter. However, to maintain consistency in how the results were interpreted for both sets of test sequences, this was not done.

Conclusions and perpsectives

Here, we have reviewed and evaluated the latest programs that can predict classically secreted proteins. The programs reviewed use a variety of algorithms and program architectures to predict protein localization. For a fair comparison of the predictive capacity of these programs, we performed an independent evaluation using experimentally annotated protein sequences that are not included in any program-training set. Results from this analysis vary widely in their prediction accuracies, based on the inclusion and exclusion of membrane domain-containing proteins. Therefore investigators should understand the nature of the protein sequences that they wish to analyze before selecting the program to perform the analysis. This benchmarking study also illustrates that well designed older prediction programs maintain competitive predictive accuracies compared with newly developed programs.

Identifying secreted proteins is an important component of today's therapeutic and diagnostic development initiatives. The lack of experimentally verified localization annotation for a significant proportion of the human proteome increases the importance of accurate prediction algorithms. Investigators are best served by selecting a prediction program or set of prediction programs that best match the nature of the analysis to be performed. Prediction programs also provide a valuable resource for investigating protein localization, but are equally valuable in that the algorithms are used in conjunction with protein database mining. This hybrid approach ensures that the most comprehensive profile of the localization of a protein is obtained.

Acknowledgements

We thank Lynda Ellis, Steve Ekker and George Vasmatzis for their valuable guidance and contributions to discussions on secretory protein prediction. C.P.S. thanks Carl Obert for his support

throughout this work. We also thank the program developers, Naresh Nair and Evangelia Petsalakis, for aid in our assessment of their programs.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.drudis.2007.01.008.

References

1 Tjalsma, H. *et al.* (2000) Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.* 64, 515–547

2 Damas, J.K. *et al.* (2001) Cytokines as new treatment targets in chronic heart failure. *Curr. Control. Trials Cardiovasc. Med.* 2, 271–277

3 Klee, E.W. *et al.* (2006) Bioinformatics methods for prioritizing serum biomarker candidates. *Clin. Chem.* 52, 2162–2164

4 Batchelder, K. and Miller, P. (2006) A change in the market–investing in diagnostics. *Nat. Biotechnol.* 8, 922–926

5 Ozdemir, V. *et al.* (2006) Shifting emphasis from pharmacogenomics to theragnostics. *Nat. Biotechnol.* 8, 942–946

6 Nair, R. and Rost, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* 348, 85–100

7 McGeoch, D.J. (1985) On the predictive recognition of signal peptide sequences. *Virus Res.* 3, 271–286

8 Doudna, J.A. and Batey, R.T. (2004) Structural insights into the signal recognition particle. *Annu. Rev. Biochem.* 73, 539–557

9 Cabrera, M. *et al.* (2003) The retrieval function of the KDEL receptor requires PKA phosphorylation of its C-terminus. *Mol. Biol. Cell* 14, 4114–4125

10 Bu, G. *et al.* (1997) ERD2 proteins mediate ER retention of the HNEL signal of LRP's receptor-associated protein (RAP). *J. Cell Sci.* 110, 65–73

11 Nickel, W. (2003) The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. *Eur. J. Biochem.* 270, 2109–2119

12 von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* 14, 4683–4690

13 Ladunga, I. (1999) PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. *Bioinformatics* 15, 1028–1038

14 Hiller, K. *et al.* (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* 32, W375–W379

15 Nielsen, H. *et al.* (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6

16 Bendstsen, J.D. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340, 783–795

17 Nielsen, H. and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (Glasgow, J. *et al.* eds), In pp. 122–130, AAAI Press

18 Nakai, K. and Kanehisha, M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* 11, 95–110

19 Horton, P. and Nakai, K. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36

20 Bannai, H.T. *et al.* (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18, 298–305

21 Gardy, J.L. *et al.* (2003) PSORT-B improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* 31, 3613–3617

22 Horton, P. *et al.* (2006) Protein subcellular localization prediction with WoLF PSORT. In *Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference* APBC06, Taipei, Taiwan (Jiang, T. *et al.*, eds), pp. 39–48, APBioNet

23 Bendtsen, J.D. *et al.* (2004) Feature-based prediction of nonclassical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356

24 Emanuelsson, O. *et al.* (2000) Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016

25 Fariselli, P. *et al.* (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* 19, 2498–2499

26 Reinhardt, A. and Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* 26, 2230–2235

27 Small, I. *et al.* (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 6, 1581–1590

28 Nair, R. and Rost, B. (2004) LOCnet and LOCtarget: sub-cellular localization for structural genomics targets. *Nucleic Acids Res.* 32, W517–W521

29 Reczko, M. *et al.* (2002) Finding signal peptides in human protein sequences using recurrent neural networks. In *Algorithms in Bioinformatics: Second International Workshop* (Guigó, R. and Gusfield, D., eds), pp. 60–67, Springer

30 Jagla, B. and Schuchhardt, J. (2000) Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics* 16, 245–250

31 Eddy, S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.* 6, 361–365

32 Krogh, A. (1998) An introduction to hidden Markov models for biological sequences. In *Computational Methods in Molecular Biology* (Salzberg, S. *et al.* eds), pp. 45–63, Elsevier

33 Kall, L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338, 1027–1036

34 Kall, L. *et al.* (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21, i251–i257

35 Zhang, Z. and Wood, I. (2003) A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* 19, 307–308

36 Chou, K.C. (2001) Prediction of signal peptides using scaled window. *Peptides* 22, 1973–1979

37 Yuan, Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* 451, 23–26

38 Vapnik, V. (1995) *The Nature of Statistical Learning Theory,* Springer

39 Vapnik, V. (1998) *Statistical Learning Theory*. John Wiley & Sons

40 Hoglund, A. *et al.* (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* 22, 1158–1165

41 Cui, Q. *et al.* (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinform.* 5, 66–73

42 Hua, S. and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17, 721–728

43 Xie, D. *et al.* (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* 33, W105–W110

44 Bhasin, M. and Raghava, G.P.S. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32, W414–W419

45 Pierleoni, A. *et al.* (2006) BaCeILo: a balanced subcellular localization predictor. *Bioinformatics* 22, e408–e416

46 Garg, A. *et al.* (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositios, their order, and similarity search. *J. Biol. Chem.* 15, 14427–14432

47 Petsalakis, E.I. *et al.* (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization *Geno. Prot. Bioinform.* 4, 48–55

48 Sarda, D. *et al.* (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinform.* 6, 152–164

49 Yu, C.S. *et al.* (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.* 13, 1402–1406

50 Hawkins, J. and Boden, M. (2006) Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinform. Comput. Biol.* 4, 1–18

51 Nair, R. and Rost, B. (2002) Inferring subcellular localization through automated lexical analysis. *Bioinformatics* 18, S78–S86

52 Mott, R. *et al.* (2002) Predicting protein cellular localization using a domain projection method. *Genome Res.* 12, 1168–1174

53 Marcotte, E.M. *et al.* (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12115–12120

54 Scott, M.S. *et al.* (2006) Predicting subcellular localization via protein motif co-occurrence. *Genome Res.* 14, 1957–1966

55 Guda, C. and Subramaniam, S. (2005) pTARGET: a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics* 21, 3963–3969

56 Guda, C. (2006) pTARGET: a web server for predicting protein subcellular localization. *Nucleic Acids Res.* 34, W210–W213

57 Nielsen, H. *et al.* (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* 1, 3–9

58 Gardy, J.L. and Brinkman, F.S. (2006) Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 10, 741–751
59 Donnes, P. and Hoglund, A. (2004) Predicting protein subcellular localization: past, present, and future. *Genom. Proteom. Bioinform.* 4, 209–215
60 Schneider, G. and Fechner, U. (2004) Advances in the prediction of protein targeting signals. *Proteomics* 6, 1571–1580
61 Hulo, N. *et al.* (2006) The PROSITE database. *Nucleic Acids Res.* 34, D227–D230
62 Nair, R. *et al.* (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res.* 31, 397–399
63 Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451