Predicting Protein Subcellular Localization: Past, Present, and Future

Pierre Dönnes and Annette Höglund*

Department for Simulation of Biological Systems, Wilhelm Schickard Institute, University of Tübingen, D-72076 Tübingen, Germany.

Functional characterization of every single protein is a major challenge of the post-genomic era. The large-scale analysis of a cell's proteins, proteomics, seeks to provide these proteins with reliable annotations regarding their interaction partners and functions in the cellular machinery. An important step on this way is to determine the subcellular localization of each protein. Eukaryotic cells are divided into subcellular compartments, or organelles. Transport across the membrane into the organelles is a highly regulated and complex cellular process. Predicting the subcellular localization by computational means has been an area of vivid activity during recent years. The publicly available prediction methods differ mainly in four aspects: the underlying biological motivation, the computational method used, localization coverage, and reliability, which are of importance to the user. This review provides a short description of the main events in the protein sorting process and an overview of the most commonly used methods in this field.

Key words: subcellular localization, prediction methods

Introduction

Large-scale genomic and proteomic efforts worldwide have contributed to a massive amount of sequence data. Annotating these sequences has been a major driving force in molecular and computational biology (*1*, *2*). Functional annotation projects seek to elucidate the potential roles that the proteins play in a cellular context, such as metabolic pathways and interaction networks.

Eukaryotic cells can synthesize up to 10,000 different kinds of proteins, which all are destined for one or more pre-determined target organelles. Proteins have evolved to function optimally in a specific subcellular localization; hence, the correct transport of a protein to its final destination is crucial to its function. The process of directing a newly synthesized protein to its target organelle is often referred to as protein targeting or protein sorting. Failure in transporting a protein has proven to be a key event behind several human diseases, such as cancer and Alzheimer's disease (*3–5*).

Computational methods aiming to assign subcellular localization in an automated and high-throughput fashion provide an appealing complement to experimental techniques. The development of methods for predicting subcellular location has been an area of great activity during recent years (*6*) and has long been seen as the detective work of a bioinformatician (*7*). The enormous complexity of the protein sorting process, alternative means of transportation pathways, and lack of complete data for every organelle, present great challenges to the eager prediction method developers.

In this review, we describe the main events of the protein sorting process, provide an overview of the computational contributions made to this field, and finally give a few guiding words to potential users.

Biological Background

There are at least ten main subcellular localizations in eukaryotes, several of which can be further subdivided into intraorganellar compartments. Bacteria, on the other hand, consist of a single intracellular space and a plasma membrane. The organelles have distinct, well-defined, and complementing functions in the cel-

*** Corresponding author.**
E-mail: hoeglund@informatik.uni-tuebingen.de

lular machinery, and are thought to have evolved from ancestral bacterial endosymbionts of prokaryotic cells (*8*).

Most proteins in the cell are encoded in the nuclear DNA, only a small subset is encoded in the chloroplastic and mitochondrial DNA. An elaborate and highly selective system of sorting and transportation mechanisms provides for the guidance of each protein to its final destination (*9*, *10*).

The cytoplasm surrounds the nucleus and is the place where the mRNA is translated into protein. Proteins in the cytoplasm can enter the secretory pathway (SP), be directed to other non-secretory pathway (nSP) organelles, or remain in the cytoplasm. The intracellular routing of non-cytoplasmic proteins was traced in pioneering experimental studies by George Palade (*11*), who received the Nobel Prize for his work in 1974.

Proteins of the secretory pathway carry a targeting sequence in their precursor protein sequences and are transported co-translationally across the Endoplasmatic Reticulum (ER) membrane. Proteins in the ER are further transported into the Golgi apparatus, plasma membrane, lysosome, vacuole, or the extracellular space, unless they carry an ER-retention sequence. Vesicular carriers are often employed for transporting the proteins and have been shown to shuttle between ER and the Golgi apparatus (*12*, *13*).

The nSP proteins are synthesized on free cytoplasmic ribosomes and are transported from the cytoplasm post-translationally, if they carry specific N-terminal targeting sequences for the chloroplast (cTP), mitochondria (mTP), or the lysosome (*14*, *15*). These targeting sequences are usually cleaved from the mature protein sequence by specific signal sequence peptidases (*16–18*), once the protein has reached its final destination. Additional intrinsic sequences, such as the hydrophobic stop-transfer sequence, present within the mature protein, can initialize membrane insertion of transmembrane proteins. Secondary targeting sequences occur in chloroplasts and enable further intraorganellar transport (*19*). All nuclear proteins have to be imported from the cytoplasm. This import is facilitated as the nuclear pore complex recognizes a nuclear localization signal (NLS), which is present only in nuclear proteins (*20*, *21*). The NLS is a short stretch of four to eight usually positively charged amino acids, and can be encoded as one fragment (monopartite) or as split into two fragments (bipartite). The NLS has been precisely defined in several nuclear proteins, but there are also nuclear proteins that appear to have no NLS at all. Peroxisomal proteins are also imported from the cytoplasm and carry a short C-terminal signal sequence that facilitates transport across the peroxisomal membrane. Furthermore, post-translational modifications, such as glycosylation (*22*), also play an important role for further protein transport (*23–25*).

Common to all signal sequences is that they show a high specificity and an evolutionary conservation (*26*). The conservation is not necessarily evident within the primary amino acid sequence, rather indirectly at the level of the biochemical properties of the amino acids. Some of the targeting sequences show a tendency to form some degree of secondary structure like an amphiphatic alpha helix or beta sheet (*27*, *28*). The correct cleavage of the TPs is highly dependent on the primary sequence and a few direct sequence motifs have been identified.

The organelles present unique biological conditions to the proteins. During the course of the evolution, the only mutations that have been accepted are the ones from which the cell benefits. It has been observed that proteins from different organelles differ in their overall amino acid composition (*29*); hence the underlying hypothesis that each protein has evolved over time to function optimally in a certain subcellular localization can be formed.

Computational Approaches

Computational methods for predicting protein subcellular localization can generally be divided into four categories: prediction methods based on (i) the overall protein amino acid composition, (ii) known targeting sequences, (iii) sequence homology and/or motifs, and (iv) a combination of several sources of information from the first three categories (hybrid methods).

(i) The pioneering work in using the overall amino acid composition for prediction was done by Nakashima and Nishikawa, who presented a method for discriminating between intracellular and extracellular proteins (*30*). Using the distance between the overall amino acid composition vectors, Cedano *et al* presented ProtLock for predicting five classes of subcellular localizations (extracellular, intracellular, integral membrane, anchored membrane, and nuclear; ref. *31*).

Reinhardt and Hubbard presented NNPSL, an approach using artificial neural networks (ANNs) for predicting four eukaryotic (cytoplasmic, extracellular, mitochondrial, and nuclear) and three prokaryotic (cytoplasmic, extracellular, and periplasmic) subcellular localizations (*32*). Several alternative algorithms have been applied to the data set presented by Reinhardt and Hubbard, including Kohonen's self-organizing maps (*33*), Support Vector Machines (SVMs; ref. *34*), and Markov chain models (*35*). Further developments in the area of using the overall amino acid composition have been made by reflecting sequence order effects. Chou *et al* presented an SVM-based method for predicting twelve different subcellular localizations, taking sequence order effects into account (*36*, *37*). A similar approach was recently presented by Park and Kanehisa as they described the PLOC method (*38*). Fuzzy k-NNs were applied by Huang *et al* to describe the dipeptide composition of the whole protein sequence for eleven different localizations (*39*). The CELLO method enables prediction of five subcellular localizations in Gram-negative bacteria (cytoplasm, inner membrane, periplasm, outer membrane, and extracellular space), based on the composition of peptides of varying lengths (n-peptide composition; ref. *40*). Andrade *et al* were the first to incorporate structural information into the amino acid composition vectors. The surface composition of eukaryotic proteins with known structure was used to distinguish between nuclear, extracellular, and cytoplasmic proteins (*29*). The rationale behind this approach is that the interiors of proteins have stayed fairly constant during evolution, whereas surface residues have adapted to certain biochemical environments.

(ii) The most comprehensive method based on N-terminal targeting sequences is TargetP (*41*), which allows for prediction of chloroplast, mitochondrial, secretory pathway, and other proteins. TargetP can be seen as an integration of the SignalP (*42*) and the ChloroP (*43*) methods; all three methods have been presented by the group of Gunnar von Heijne. MitoProt (*44*) and Predotar (http://www.inra.fr/predotar) are two methods both specifically discriminate chloroplast from mitochondrial proteins. Another method in this category is iPSORT (*45*), offering prediction of the same localization categories as TargetP. The iPSORT uses knowledge-based rules for prediction based on protein sequence features derived from the AAindex database (*46*).

(iii) Marcotte *et al* presented a method that assigns the subcellular localization by constructing phylogenetic profiles of the proteins (*47*). Mott *et al* used SMART (*48*) domains for predicting cytoplasmic, secreted, and nuclear proteins (*49*). The method PredictNLS is a method specialized on recognizing nuclear proteins, based on a collection of nuclear localization sequences (NLSs; ref. *50*). A nearest neighbour approach using the composition of functional domains has also been presented and tested on the Reinhardt and Hubbard data set (*51*). Proteome Analyst, presented by Lu *et al*, is based on SWISS-PROT keywords and the annotation of homologous proteins (*52*). This method is similar to the LOCkey (*53*) and LOChom (*54*) methods described by Nair and Rost in 2002. PSLT is a recently presented method that uses Bayesian networks and InterPro motifs for predicting ten subcellular localizations (*55*).

(iv) PSORT, presented in 1992, was one of the first methods developed for predicting the subcellular localization (*56*). PSORT uses the overall amino acid composition, N-terminal targeting sequence information, and motifs, hence considered a hybrid approach. This method uses a set of knowledge-based "if-then" rules and predicts 14 animal and 17 plant subcellular localizations. Extensions of the PSORT method include: PSORT II (a modified decision algorithm; ref. *57*) and PSORT-B (with focus on bacterial proteins; ref. *58*). ESLpred was developed using the Reinhardt and Hubbard data set and is an SVM-based method, which combines the dipeptide composition and PSI-BLAST scores (*59*). Drawid and Gerstein presented a method that incorporates information about sequence motifs, overall sequence properties (*e.g.* isoelectric points and surface composition), and mRNA expression levels (*60*). Their method is based on a Bayesian prediction model and was tested on the yeast genome. MITOPRED is a method specialized for predicting mitochondrial proteins, which is based on Pfam domains (*61*) and amino acid composition (*62*).

Several of the described methods are available as online prediction servers. A compiled list of methods, associated URLs, and references can be seen in Table 1. Since the methods have different localization coverage and different means to assess their accuracy, it is impossible to compare all methods against each other. Accuracy issues of prediction are considered in the Discussion section below.

Discussion

Currently available prediction methods differ in three main aspects critical to the user: the underlying bio-

Table 1 Prediction Methods Available Online

Method	Url	Ref.
Cello	http://cello.life.nctu.edu.tw/	*40*
ChloroP	http://www.cbs.dtu.dk/services/ChloroP/	*43*
ESLpred	http://www.imtech.res.in/raghava/eslpred/	*59*
iPSORT	http://hc.ims.u-tokyo.ac.jp/iPSORT/	*45*
MITOPRED	http://mitopred.sdsc.edu/	*62*
MitoProt	http://ihg.gsf.de/ihg/mitoprot.html	*44*
NNPSL	http://www.doe-mbi.ucla.edu/%7Eastrid/astrid.html	*32*
PLOC	http://www.genome.jp/SIT/ploc.html	*38*
predictNLS	http://cubic.bioc.columbia.edu/predictNLS/	*50*
Predotar	http://genoplante-info.infobiogen.fr/predotar/	
Proteome Analyst	http://www.cs.ualberta.ca/%7Ebioinfo/PA/Sub/index.html	*52*
PSORT	http://psort.ims.u-tokyo.ac.jp/form.html	*56*
PSORT II	http://psort.ims.u-tokyo.ac.jp/form2.html	*57*
PSORT-B	http://www.psort.org/psortb/	*58*
SignalP	http://www.cbs.dtu.dk/services/SignalP/	*42*
SubLoc	http://www.bioinfo.tsinghua.edu.cn/SubLoc/	*34*
TargetP	http://www.cbs.dtu.dk/services/TargetP/	*41*

logical model, the localization *coverage*, and prediction *accuracy*. The biological model can be fairly simple as is the case for predictions based on the overall amino acid composition, or more complex as the model underlying the knowledge-based PSORT prediction system. The localization *coverage* differs immensely and ranges from methods predicting just a few localizations, to all possible localizations.

Prediction *accuracy* can be seen either as the overall *accuracy* for a method, or as the individual *accuracy* for each predicted localization. It is often the case that some localizations can be predicted with fairly high accuracy, whereas others not. Furthermore, most methods have been trained using different data sets or training procedures, which makes a fair benchmark comparison a daunting task.

Methods based on targeting sequences, such as TargetP and iPSORT, generally predict only four plant and three non-plant localizations (low *coverage*) but have relatively high prediction *accuracy*. However, it should be pointed out that two of the predicted categories, SP and others, are not subcellular localizations. The SP category contains proteins from at least six different subcellular localizations, and the category of others from at least three. Only if the prediction is chloroplast or mitochondrial, a specific localization can be assigned to the query protein. In these cases it might even be a good choice to use a more specialized method, such as MITOPRED, that predicts mitochondrial proteins with high accuracy. Predictions based on targeting sequences alone are complicated by the fact that it is hard to determine the presence of a targeting sequence (*63*).

Methods based on the overall amino acid composition have great variations in their *coverage*. The underlying biological model is fairly simple, which probably is one of the main reasons for triggering the avalanche of different computational algorithms applied to the data set by Reinhardt and Hubbard, where four localizations are represented (*32–35*, *51*, *59*). This data set has a high level of sequence homology (up to 90%) and essentially all methods perform equally, with only minor deviations. Some of the algorithms used are more prone to overfitting than others, since different types of cross-validation schemes have been used. Other methods in this category with higher localization coverage have comparable overall accuracies and are a better choice if very little is known about the protein at hand.

Methods based on direct sequence homology are in some cases very accurate. These methods rely on finding a highly similar protein with a known subcellular localization annotation. Predictions neglect protein specific features that can be learned from a training data set, hence also the events in the sorting process. The drawback is if there is no homologous protein with annotated localization available, the result is left to chance. Parallels can be drawn to protein structure prediction, which can be fairly reliable if a homologous protein with known structure is known. High sequence homology is often an indication that the proteins are similar in both structure and function, but not necessarily that they share the same localization.

Hybrid methods usually offer prediction of a wider range of subcellular localizations and are the methods of choice, when very little is known about the protein of interest. Methods providing a verbose output of the prediction results can be recommended, since these may give detailed information about potentially detected motifs and targeting sequences.

A sophisticated prediction method should strive towards mimicking the biological process of protein sorting, an important step on the way to simulate a small component of the systems biology of the cell. Machine learning plays an important role in the development and implementation of the complex underlying biological models, which can also be seen from the frequent application within this filed. As the accuracy of the methods continue to increase, it will become more interesting to take a closer look at the relatively small proportion of misclassified proteins. It is likely that these proteins are key players at the interfaces of the organelles and they may provide clues to alternative protein sorting routes. There is a need to constantly update methods and to extract new data sets for training the prediction models. An example is the data set used by Reinhardt and Hubbard extracted from SWISS-PROT release 33.0 (52,205 sequence and 15,775 subcellular localization annotations; ref. *32*), which is still being used to develop new methods. The current release SWISS-PROT 44.1 contains 122,750 protein entries with a total of 80,562 subcellular localization annotations. Hence, the advice to the interested users is to choose a method based on an up-to-date data set. To a general biologist, it is hard to assess the choice of algorithm and whether the accuracy estimation is done in a sound way. The pitfalls here are typically too homologous sequences within the training data and cross-validation schemas to estimate the overall performance of the method.

In conclusion, understanding and predicting protein subcellular localization is a field of research where a lot has happened recently both experimentally and computationally. It is clear that several challenges lie ahead, especially when translating known facts from experimental biology into reasonable computational models and simultaneously to avoid simplifications. A further challenge is the prediction of proteins known to shuttle between compartments, which in principle can be ascribed to two subcellular localizations (*64*, *65*). There is no doubt that numerous new approaches, both in terms of algorithms and biological motivations, will be presented in this field in the near future. Prediction of subcellular localization is very likely to be one building block in systems biology approaches, aiming to understand the broader aspects of molecular biology.

References

1. Eisenberg, D., *et al.* 2000. Protein function in the post-genomic era. *Nature* 405: 823-826.
2. Koonin, E.V. 2000. Bridging the gap between sequence and function. *Trends Genet.* 16: 16.
3. Shurety, W., *et al.* 2000. Localization and post-Golgi trafficking of tumor necrosis factor-alpha in macrophages. *J. Interferon Cytokine Res.* 20: 427-438.
4. Bryant, D.M. and Stow, J.L. 2004. The ins and outs of E-cadherin trafficking. *Trends Cell Biol.* 14: 427-434.
5. Hartmann, T., *et al.* 1996. Alzheimer's disease betaA4 protein release and amyloid precursor protein sorting are regulated by alternative splicing. *J. Biol. Chem.* 271: 13208-13214.
6. Nakai, K. 2000. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* 54: 277-344.
7. Doerks, T., *et al.* 1998. Protein annotation: detective work for function prediction. *Trends Genet.* 14: 248-250.
8. Dyall, S.D., *et al.* 2004. Ancient invasions: from endosymbionts to organelles. *Science* 304: 253-257.
9. Cline, K. and Henry, R. 1996. Import and routing of nucleus-encoded chloroplast proteins. *Annu. Rev. Cell. Dev. Biol.* 12: 1-26.
10. Schatz, G. 1998. Protein transport. The doors to organelles. *Nature* 395: 439-440.
11. Palade, G. 1975. Intracellular aspects of the process of protein synthesis. *Science* 189: 347-358.
12. Lee, M.C., *et al.* 2004. Bi-directional protein transport between the ER and Golgi. *Annu. Rev. Cell Dev. Biol.* 20: 87-123.
13. Neumann, U., *et al.* 2003. Protein transport in plant cells: in and out of the Golgi. *Ann. Bot.* 92: 167-180.
14. Rusch, S.L. and Kendall, D.A. 1995. Protein transport via amino-terminal targeting sequences: common themes in diverse systems. *Mol. Membr. Biol.* 12: 295-307.
15. Schatz, G. and Dobberstein, B. 1996. Common principles of protein translocation across membranes. *Science* 271: 1519-1526.

16. Jarvis, P. and Robinson, C. 2004. Mechanisms of protein import and routing in chloroplasts. *Curr. Biol.* 14: R1064-1077.
17. Hawlitschek, G., *et al.* 1988. Mitochondrial protein import: identification of processing peptidase and of PEP, a processing enhancing protein. *Cell* 53: 795-806.
18. Arretz, M., *et al.* 1991. Processing of mitochondrial precursor proteins. *Biomed. Biochim. Acta* 50: 403-412.
19. Shackleton, J.B. and Robinson, C. 1991. Transport of proteins into chloroplasts. The thylakoidal processing peptidase is a signal-type peptidase with stringent substrate requirements at the -3 and -1 positions. *J. Biol. Chem.* 266: 12152-12156.
20. Nigg, E.A., *et al.* 1991. Nuclear import-export: in search of signals and mechanisms. *Cell* 66: 15-22.
21. Dingwall, C. and Laskey, R.A. 1991. Nuclear targeting sequences—a consensus? *Trends Biochem. Sci.* 16: 478-481.
22. Scheiffele, P. and Fullekrug, J. 2000. Glycosylation and protein transport. *Essays Biochem.* 36: 27-35.
23. Bergeron, J.J., *et al.* 1994. Calnexin: a membrane-bound chaperone of the endoplasmic reticulum. *Trends Biochem. Sci.* 19: 124-128.
24. Paulson, J.C. 1989. Glycoproteins: what are the sugar chains for? *Trends Biochem. Sci.* 14: 272-276.
25. Silhavy, T.J., *et al.* 1983. Mechanisms of protein localization. *Microbiol. Rev.* 47: 313-344.
26. Clausmeyer, S., *et al.* 1993. Protein import into chloroplasts. The hydrophilic lumenal proteins exhibit unexpected import and sorting specificities in spite of structurally conserved transit peptides. *J. Biol. Chem.* 268: 13869-13876.
27. Endo, T., *et al.* 1989. N-terminal half of a mitochondrial presequence peptide takes a helical conformation when bound to dodecylphosphocholine micelles: a proton nuclear magnetic resonance study. *J. Biochem.* 106: 396-400.
28. Hammen, P.K., *et al.* 1994. Structure of the signal sequences for two mitochondrial matrix proteins that are not proteolytically processed upon import. *Biochemistry* 33: 8610-8617.
29. Andrade, M.A., *et al.* 1998. Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* 276: 517-525.
30. Nakashima, H. and Nishikawa, K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* 238: 54-61.
31. Cedano, J., *et al.* 1997. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* 266: 594-600.
32. Reinhardt, A. and Hubbard, T. 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* 26: 2230-2236.
33. Cai, Y.D. and Chou, K.C. 2000. Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. *Mol. Cell. Biol. Res. Commun.* 4: 172-173.
34. Hua, S. and Sun, Z. 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17: 721-728.
35. Yuan, Z. 1999. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* 451: 23-26.
36. Chou, K.C. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43: 246-255.
37. Cai, Y.D., *et al.* 2002. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell Biochem.* 84: 343-348.
38. Park, K.J. and Kanehisa, M. 2003. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* 19: 1656-1663.
39. Huang, Y. and Li, Y. 2004. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* 20: 21-28.
40. 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.
41. Emanuelsson, O., *et al.* 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300: 1005-1016.
42. Nielsen, H., *et al.* 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10: 1-6.
43. Emanuelsson, O., *et al.* 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* 8: 978-984.
44. Claros, M.G. and Vincens, P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* 241: 779-786.
45. Bannai, H., *et al.* 2002. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18: 298-305.
46. Kawashima, S., *et al.* 1999. AAindex: amino acid index database. *Nucleic Acids Res.* 27: 368-369.
47. Marcotte, E.M., *et al.* 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 97: 12115-12120.
48. Schultz, J., *et al.* 2000. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28: 231-234.

49. Mott, R., *et al.* 2002. Predicting protein cellular localization using a domain projection method. *Genome Res.* 12: 1168-1174.
50. Cokol, M., *et al.* 2000. Finding nuclear localization signals. *EMBO Rep.* 1: 411-415.
51. Cai, Y.D. and Chou, K.C. 2003. Nearest neighbour algorithm for predicting protein subcellularlocation by combining functional domain composition and pseduo-amino acid composition. *Biochem. Biophys. Res. Commun.* 305: 407-411.
52. Lu, Z., *et al.* 2004. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 20: 547-556.
53. Nair, R. and Rost, B. 2002. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* 18: S78-86.
54. Nair, R. and Rost, B. 2002. Sequence conserved for subcellular localization. *Protein Sci.* 11: 2836-2847.
55. Scott, M.S., *et al.* 2004. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.* 14: 1957-1966.
56. Nakai, K. and Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14: 897-911.
57. Horton, P. and Nakai, K. 1997. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5: 147-152.
58. Gardy, J.L., *et al.* 2003. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* 31: 3613-3617.
59. Bhasin, M. and Raghava, G.P. 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32: W414-419.
60. Drawid, A. and Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* 301: 1059-1075.
61. Bateman, A., *et al.* 2000. The Pfam protein families database. *Nucleic Acids Res.* 28: 263-266.
62. Guda, C., *et al.* 2004. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* 20: 1785-1794.
63. Frishman, D., *et al.* 1999. Starts of bacterial genes: estimating the reliability of computer predictions. *Gene* 234: 257-265.
64. Corbett, A.H. and Silver, P.A. 1997. Nucleocytoplasmic transport of macromolecules. *Microbiol. Mol. Biol. Rev.* 61: 193-211.
65. Kirchhausen, T. 2000. Three ways to make a vesicle. *Nat. Rev. Mol. Cell. Biol.* 1: 187-198.