CHAPTER 9

PROTEIN SUBCELLULAR LOCALIZATION PREDICTION

Paul Horton
National Institute of Industrial Science and Technology
horton-p@aist.go.jp

Yuri Mukai
National Institute of Industrial Science and Technology
yuri-mukai@aist.go.jp

Kenta Nakai
University of Tokyo
knakai@ims.u-tokyo.ac.jp

This chapter discusses various aspects of protein subcellular localization in the context of bioinformatics and reviews the twenty years of progress in predicting protein subcellular localization.

ORGANIZATION.

Section 1. We first provide the motivation for prediction of protein subcellular localization sites, as well as discuss changes being brought about by progress in proteomics.

Section 2. After that, we describe the biology of protein subcellular location. In particular, we explain the principle of protein sorting signals.

Section 3. Then we present several experimental techniques for determining protein subcellular localization sites. The techniques surveyed include traditional methods such as immunofluorescence microscopy, as well as large-scale methods such as green fluorescent protein.

Section 4. Next we mention some of the general issues involved in predicting protein subcellular localization, such as what are the sites? how many sites per protein? how good are the predictions? and so on. We also discuss the distinction between features that reflect causal influences on localization versus features that merely reflect correlation with localization.

Section 5. Lastly, we offer a survey of computational methods and approaches for predicting protein subcellular localization. These include discriminant analysis of amino acid composition, localization process modeling, machine learning, feature discovery, and literature analysis.

1. Motivation

The prediction of the subcellular localization sites of proteins from their amino acid sequences is a fairly long-standing problem in bioinformatics. Nishikawa and Ooi observed that amino acid composition correlates with localization sites in 1982.[617] Around that time early work on the characterization and prediction of secretory signal peptides began.[556, 861, 862] Over a decade later, the publication of a signal peptide prediction program SignalP[611, 614] was awarded a "Hot Paper in Bioinformatics" Award,[613] indicative of the high level of interest in such predictions. The main driving force behind this redoubled interest in predicting protein localization from sequence has been the need to annotate massive amounts of sequence data coming from various genome projects.[227] The importance of the biological phenomenon underlying protein localization was underscored by the 1999 Nobel Prize in physiology or medicine, awarded to Günter Blobel for the discovery that "proteins have intrinsic signals that govern their transport and localization in the cell". More recently, the emergence of proteomic technologies—see Chapter 13—has given birth to terms such as "secretome".[218] Indeed recent experimental studies have determined the localization sites of a large fraction of all the proteins in yeast.[374, 463, 721] Several excellent reviews on protein localization and prediction are available.[232, 244, 599]

Each compartment in a cell has a unique set of functions, thus it is reasonable to assume that the compartment or membrane in which a protein resides is one determinant of its function. This assertion is supported by the fact that localization correlates with both protein-protein interaction data[357, 734, 735, 760] and with gene expression levels.[212, 374] Thus protein localization is a valuable clue to protein function, especially when homologs of known function cannot be found by sequence similarity—a situation that is still common today. A recent study of 134 bacterial genomes and several eukaryotic genomes shows that standard sequence similarity reveals useful functional information in only about half of all proteins, although there is significant variance in that proportional between species.[809] Some applications are interested in a particular localization site. Many industrial applications are concerned with the efficiency of the secretion of non-native proteins in micro-organisms.[854] Proteins on the cell membrane are attractive as potential drug targets because they are accessible from outside the cell.

Recent progress[374, 463] in large-scale experiments to determine protein localization indicates that in the foreseeable future the localization sites of a large percentage of proteins for some model organisms may be experimentally determined. For other organisms, the ability to infer localization by sequence similarity, an approach quantatively analyzed by Nair and Rost[595], will increase significantly. This clearly reduces the practical value of prediction schemes. So you may want to skip to the next chapter...

Still here? Good, because we think there are many excellent reasons for studying localization prediction. For one reason, the rose-colored scenario stated above is not quite upon us yet. There is still less than full coverage and significant error in the proteomic localization data, some of which may be systematic—see Section 3. For example, a recent study[374] covers 75% of the proteome and shows roughly 20% disagreement with data from the *Saccharomyces* Genome Database,[206] which contains data from two other large-scale experiments.[463,721] Also, so far these experiments have been done on yeast, which has many fewer protein encoding genes than, for example, humans—approximately 6,000 for yeast[206,293] versus perhaps 30,000 for human.[467] Many human proteins do not have close homologs in yeast. As mentioned in Chapter 13, the accumulated knowledge from decades of small scale experiments may contain fewer errors than recent large-scale experiments but the coverage is low. We found that, as of October 2003, only around 0.25% of SWISS-PROT entries include an explicit firm localization site assignment. (We searched only the "CC" fields and excluded assignments marked as "potential", "probable", or "by similarity".) However, do see Section 5.6 for automated methods to gain more localization information from SWISS-PROT annotations. Moreover the error rate is still significant. For example, a study of chloroplast signal peptides found roughly 10% of cleavage sites to be incorrect or based on insufficient evidence.[231,614] Thus there is still some utility in predicting localization from amino acid sequence.

This notwithstanding, the coverage of experimental methods is certainly increasing rapidly and the accuracy of information derived from large scale experiments can be increased by comparing the results of multiple experiments—see Chapter 13. Prediction methods can play a role here in identifying outliers that may indicate experimental error. As the number of proteins whose localization has been determined experimentally increases, the role of prediction schemes is certain to change. A black box prediction program for native proteins is going to be of little use in the near future—even if it attains a high accuracy. We believe two qualities are to be demanded of prediction schemes in the future, *viz.*

(1) a high explanatory power, and
(2) the ability to accurately predict the localization of non-native, mutant, and

artificial proteins.

The first is because as scientists we are not satisfied with simply knowing that protein A is localized to site B. We also want to know how it gets there. Prediction schemes encoding our hypothesis about the process can help us gauge how much we really know, and machine learning techniques can help us generate new hypotheses. The second is because when designing new proteins we would like to be able to do experiments *in silico* on proteins that do not yet exist—and therefore clearly do not have experimentally determined localization sites.

Regrettably, most readers of this book will not become specialists in protein localization. Therefore perhaps a more compelling reason to read this chapter is the fact that localization prediction has been intensively studied with a variety of computational techniques and is an excellent vehicle for discussing several issues that apply to many areas of bioinformatics. More specifically there has been significant cross-fertilization of ideas between localization prediction and the related topics of predicting of membrane protein structure and post-translational modifications, *e.g.* lipidification, of proteins. Many groups working on protein localization prediction have also consistently published in these areas.[226, 394, 559, 600, 724]

2. Biology of Localization

Living organisms are classified into two categories: prokaryotes and eukaryotes. Unlike prokaryotes, eukaryotic cells are equipped with many kinds of membrane-bound compartments called organelles—*e.g.*, the nucleus, mitochondrion, endoplasmic reticulum (ER), and vacuole. We also consider some other subcellular sites such as the cytoplasm, plasma membrane, and cell wall; see Figure 1. Each organelle plays some specific cellular roles thanks to the presence of specifically localized proteins.

It is well known that proteins are synthesized based on the genetic information encoded in DNA. Although some information is encoded in the DNA within mitochondria—and chloroplasts in plants—most information is encoded in the nuclear DNA. Even most mitochondrial and chloroplast proteins are encoded in the nuclear DNA. The proteins encoded in the nuclear DNA are first synthesized within the cytoplasm and then specifically transported to each final localization site. Note that the transportation across the membrane of the ER generally starts in a pipelined fashion before synthesis of the amino acid chain is finished. The study of molecular mechanisms on how the final localization site of a protein is recognized and transported—often called "protein sorting"—is one of the central themes in modern cell biology. General textbooks on molecular cell biology typically devote many pages to this topic. As a recent example, we recommend the textbook of Alberts *et al.*[18]

The most important principle of protein sorting is that each protein has the information of its final localization site as a part of its amino acid sequence. In many cases, proteins are first synthesized as precursors having an extra stretch of polypeptide that function as a "sorting signal". They are specifically recognized and transported with some molecular machinery. After they are localized at their final destination, these sorting signals are often cleaved off. Therefore, it should be possible to predict the subcellular localization site of a protein if we can specifically recognize its sorting signal, as the cellular machinery does. This attempt is still challenging because our knowledge is incomplete.

Like all principles in biology, the sorting signal hypothesis allows some exceptions. That is, some proteins do not have sorting signals within their amino acid sequences, but instead are localized by binding with another protein that has the information. Fortunately for the developers of prediction methods, this "hitch-hiking" strategy does not seem to be common—probably because it is difficult for protein complexes to go through the organellar membranes. The nucleus is somewhat special in this respect because its membrane has large nuclear pores that allow small—up to about 60 kDa—proteins to diffuse into and out of the nucleus and also makes hitch-hiking relatively easier.

In many cases, the information of sorting signals is encoded within a limited length of the polypeptide. However, there are some examples where sorting information is encoded by sequence patches that are only recognizable in the 3D structure. The sequence features of sorting signals are variable. Some are represented as relatively well-conserved sequence motifs. Others appear more ambiguous—such as hydrophobic stretches—to our eyes, at least. Usually, at least to some extent, sorting signals can be discriminated with appropriately employed pattern recognition algorithms without the knowledge of their 3D structures.

We should keep in mind that the sorting signal for each localization site is not necessarily unique. For example, many mitochondrial proteins have the mitochondrial transit peptide on their N-terminus, but many others do not have this kind of signal and are instead localized by some different pathways. Indeed, recent developments of cell biology have enriched our knowledge of protein sorting greatly for certain proteins. Nevertheless, such knowledge is often applicable to a very limited set of proteins and is not general enough to raise overall prediction accuracy significantly.

Generally speaking localization signals appear to be well conserved across species. For example, Schneider *et al.* have studied mitochondrial targeting peptides in mammals, yeast, *Neurospora crassa*, and plant proteins. Although they observe some differences between plant and non-plant species, the clustering of

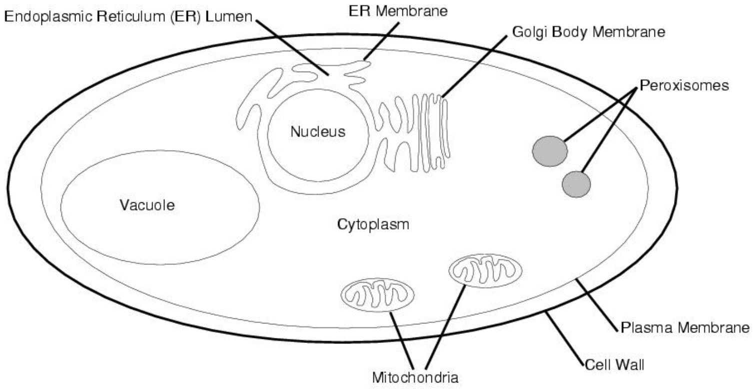

Fig. 1. A cartoon of some of the comparments of a yeast cell is shown.

targeting peptides produced by an unsupervised learning algorithm—Kohonen self-organizing map—do not produce clusters based on species. We expect that prediction methods trained on one species and applied to another species can give reasonable results, even for distant pairs such as yeast and human. Indeed most predictive studies have used training data that include proteins from multiple species. Some differences do exist and have been analyzed at some length.[222, 598]

3. Experimental Techniques for Determining Localization Sites

3.1. *Traditional Methods*

Two traditional methods for determining the cellular localization sites of proteins are immunofluorescent staining and gradient centrifugation. We briefly outline each technique here.

3.1.1. *Immunofluorescence Microscopy*

Immunofluorescence microscopy can be used to determine the localization of a target protein. Cells—treated with detergent to help solubolize their plasma membrane—on a cover glass slide are treated with two antibodies. The first antibody is chosen to selectively bind to the target protein. The second antibody, which is tagged with a fluorescent marker, binds to the immunoglobulin epitope of the first antibody. The advantage of this scheme is that the second antibody does not depend on the target protein.

To more accurately determine the localization site of the target protein, colocalization with a reference marker—which is known to localize and fluoresce at a specific localization site—can be measured. The reference marker is chosen so that it emits a different wavelength of light when fluorescing than the second antibody used to detect the target protein. If the spatial pattern of the two wavelengths of light coincide then one may conclude that the target protein has the same localization site as the reference marker. The reference marker does not need to be a protein. Some well known reference markers are DAPI, MitoTracker ®, and LysoTracker®; which are specific for the nucleus, mitochondrion, and lysosome respectively. Unfortunately, it is difficult to accurately measure the quantitative distribution of the target protein with immunofluorescence alone.

3.1.2. *Gradient Centrifugation*

Cell homogenation is the process of disrupting the plasma membrane of cells by mechanical means; for example with a rotating rod placed in a test tube. If carefully done, cells can be homogenated with the nucleus and most organelles intact. Fortunately these different compartments have different densities, allowing them to be separated by density gradient centrifugation.

The contents of each fraction, corresponding to a particular localization site, obtained by centrifugation can be analyzed with a "Western blot"—in which SDS-PAGE is used to separate the fraction on a electrophoretic gel and immunofluorescence is used to detect which band corresponds to the target protein. Once the band is identified it is possible to accurately measure the amount of protein it contains based on its size and darkness. Thus although this approach still requires the preparation of an antibody specific to the target protein, it has the advantage of allowing measurement of the quantity of the target protein at each localization site.

3.2. ***Large-Scale Experiments***

3.2.1. *Immunofluorescent Microscopy*

Traditional immunofluorescent microscopy methods for determining localization required an antibody specific to each target protein. Unfortunately the development of such antibodies is a difficult and costly process. Kumar[463] and colleagues employed a scheme which uses immunofluorescent microscopy but does not require target specific antibodies. For each target protein they constructed vectors that expressed a fusion protein consisting of the target protein fused at its C-terminal to the V5 epitope. The localization of the fused protein could thus be determined with fluorescently labeled V5 antibodies. They also performed similar experiments using a haemagglutinin (HA) tag and HA antibodies instead of V5. They have determined the localization of 2022 and 1083 proteins using V5 and HA respectively. The union of the two sets contains 2744 proteins.

3.2.2. *Green Fluorescent Protein*

Green Fluorescent Protein (GFP) is a valuable tool[286] for studying localization. GFP is a 238 amino acid protein, with known 3D structure,[916] that naturally occurs in the bioluminescent jellyfish *Aequorea victoria*. GFP emits green light upon excitation with blue light. Genetically-engineered variants of GFP with different emission wavelengths such as Yellow Fluorescent Protein, Red Fluorescent Protein (RFP), and Cyan Fluorescent Protein are also available. An important property of GFP and its variants is that, unlike typical bioluminescent molecules, it does not require any co-factors. GFP is used in studying protein localization by creating fusion proteins consisting of the target protein and GFP fused together, usually with GFP connected to the target protein C-terminus. The location of the fused protein can be traced with fluorescence microscopy. The fluorescence of GFP is relatively stable over time and thus lends itself to quantitative measurements. One way to introduce these fusion protein into cells is by transfection with an expression vector containing DNA coding for the fusion protein co-expressed with a gene that can be selected for, such as resistance to a particular drug.

Recently GFP fusion proteins have been combined with homologous recombination in the largest experiment on localization performed so far. Huh *et al.*[374] transfected cells with vectors specific to each yeast protein coding gene (more precisely ORF). The vector for each target gene contains specific sequences that allows the vector to be inserted into the chromosome in the native position with the native promoter of the target gene. This technique solves the problem of over or under expression mentioned in Section 3.2.3. Of a total of 6,234 ORFs, fluorescence from 4,156 fusion proteins was detected and divided by localization site into: cell periphery, bud, bud neck, cytoskeleton & microtubule, cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, vacuole, vacuolar membrane, punctate, and ambiguous. Many proteins showed multiple localization. The localization of some proteins was further investigated using proteins with known localization tagged with RFP. The comparison of the resulting pattern of green and red fluorescence gives detailed information about the localization of the target protein.

3.2.3. *Comments on Large-Scale Experiments*

The large-scale experiments described here are extremely impressive and have radically increased what we know about localization. However GFP is a large tag and may interfere with localization in some cases. Most sorting signals are located near the N-terminal of proteins, but in some cases the C-terminal region can also affect localization. For example, the peroxisomal translocation signal PTS1[230, 574] and ER retention signal "KDEL"[659] are found in C-terminal region. Other signals which are not specific to the C-terminal region may also sometimes occur there.[709, 828] It is also at least conceivable that a GFP tag can interfere with the diffusion of small proteins into and out of the nucleus. The smaller tags used in immunofluoresence microscopy studies have not been expressed with native promoters and may exhibit abnormal expression levels. Over-expression potentially causes false positives in the cytoplasm by overloading localization mechanisms. Conversely, under-expression may lead to false negatives in various compartments due to missing small amounts of the protein. Thus there are still some potential sources of systematic error that may explain part of the 20% disagreement between localization data gained with the large-scale GFP experiment[374] and compilations of previous experimental data.

4. Issues and Complications

Before we delve into specific prediction schemes we mention some of the general issues involved in localization prediction.

4.1. *How Many Sites are There and What are They?*

Predictive studies have divided the cell into anywhere from two to around 12 sites. In a sense the classification of proteins as integral membrane proteins versus soluble or peripheral membrane proteins is a kind of localization prediction. Many tools have been developed for the prediction of membrane spanning regions in proteins.[182, 353, 457, 575] A binary classification problem that is at the heart of understanding protein localization is the prediction of signal peptides. The development of the SignalP program[611] is an example of an influential paper which focused on the binary classification problems of predicting the presence or absence of signal peptides, and distinguishing between cleaved signal peptides versus their uncleaved counterparts (N-terminal signal anchor sequences). PSORT[602] classifies eukaryotic proteins into 14 (17 for plants) localization classes. PSORTII[368] classifies eukaryotic proteins into 10 classes.

There is in fact no single accepted "correct" scheme for defining localization classes. Recent experimental techniques have allowed the localization site of proteins in yeast cells to be determined at the resolution of roughly 22 distinct sites.[374] This allows some sites, such as the nuclear periphery and the endoplasmic reticulum, that have been lumped together by most predictive studies to be distinguished. On the other hand, the fluorescence microscopy used in their study does not allow them to distinguish between lumen versus membrane for the mitochondria and endoplasmic reticulum sites. So in this area, the annotation accumulated from small-scale experiments in SWISS-PROT often offers higher resolution.

Indeed the annotations regarding subcellular location in SWISS-PROT[86] strongly reflect the historic lack of a canonical scheme for defining localization sites. Considering only entries containing a subcellular location annotation in the "CC" field and canonicalizing for capitalization and the use of white space, one finds 3214 distinct descriptions for subcellular location. The top 45 such descriptions are shown in Figure 4. Note that the descriptions vary both in terms of localization site and the firmness of the evidence upon which the annotation is based. In fact many annotations—*e.g.*, those containing "by similarity"—do not represent experimental verifications.

The large diversity in annotations has been a practical difficulty to overcome when devising classification schemes for prediction programs. However, less common descriptions often add useful specific information such as conditions—*e.g.*, temperature—under which a localization is observed, or the topology of integral membrane proteins, *etc*. See also Section 5.4. The recent increased use of controlled vocabularies, such as the cellular component vocabulary of the Gene Ontology[35] should make localization site definitions and dataset preparation easier in the future.

4.2. *Is One Site Per Protein an Adequate Model?*

All of the studies that we are aware of use something close to a one-site-per-protein model. Some work has gone a little beyond this; for example, the PSORT-B[272] database includes eight localization classes for gram-negative bacteria, three of which correspond to proteins that localize to two different sites—such as outer membrane and extracellular.

It is known that the function of some transcription binding factors is partially regulated by selective localization to either the nucleus or the cytoplasm—where their action is obviously blocked.[71,568,636] Kumar *et al.*[463] find that approximately 25% of proteins that localize to an organelle also show significant cytoplasmic staining upon immunofluoresence analysis. The distribution of localization sites from another large-scale study[374] gives the dual localization of nucleus and cytoplasm as the number one localization site, as shown in Figure 5. One of the difficulties in accurately predicting the localization of mitochondrial and chloroplast proteins in plants is that their localization signals are very similar, and in fact—in rare cases—the same protein can be localized to both organelles.[137,777]

Although most work to date on eukaryotes has used the one-site-per-protein model, this is a drastic simplification of reality. Some proteins localize to multiple sites simultaneously, some proteins change their localization in a regulated way, and some proteins constantly move—such as proteins found in vesicle membranes which shuttle between the Golgi body and the endoplasmic reticulum membrane. Some simplification of this complicated reality seems necessary to make the problem tractable.

However, given the data in Figure 5, we believe that adding multiple localization to the nucleus and cytoplasm explicitly to prediction schemes is worth considering. On the other hand, for the vast majority of proteins, multiple localization generally appears to be limited to a few pairs of sites. So it is probably unnecessary to require a model to allow proteins to multiply localize to arbitrary combinations of sites.

4.3. *How Good are the Predictions?*

The prediction of integral membrane proteins appears to be the easiest one, with percent accuracies in the high 90s being reported by several studies.[182,353,457,575] A much higher resolution prediction for gram negative bacteria also seems to be relatively easy. Horton and Nakai[368] claim an accuracy of 94% with PSORTII in classifying 336 *E. coli* proteins into 7 localization classes (unfortunately this accuracy estimate is buried in the discussion of their paper). Recently, slightly lower accuracy has been reported by PSORT-B[272] with a much larger data set of 1443 sequences.[a]

Localization in eukaryotic cells has proven harder to predict. For example, PSORTII[368] only achieved a somewhat disappointing accuracy of 60% for ten yeast localization sites in 1997. Interestingly, most of the mistakes are between the cytoplasm and the nucleus. Some of those “mistakes” are likely to have been proteins that, depending on certain conditions, can localize to either site. Some advances have been made since then, and several methods have been published with much higher estimated accuracy for eukaryotic cells.

Some of these methods have been compared on a common dataset by Emanuelsson,[232] who found TargetP to classify plant and non-plant eukaryotic cells proteins into four sites with an accuracy of 85% and 90% respectively. Nair and Rost[595] compared TargetP,[232] SubLoc,[369] and NNPSL[710] on a common dataset. They obtained high coverage (99% and 93%) for extracellular and mitochondrial proteins with TargetP but with low precision (51% and 46%). For

[a] PSORT, PSORTII, and PSORT-B are three distinct prediction programs, although the feature vector and class definitions used by PSORTII are mainly taken from PSORT.

cytoplasmic and nuclear proteins, the study found SubLoc to yield coverages of 67% and 82% with precisions of 60% and 76% respectively.

Many methods not covered in these studies have also claimed high predictions accuracies with various datasets and localization site definitions. We have no intention of doubting any particular estimates, but refer the reader to Section 4.4 for caveats. An independent comparison[562] of signal peptide predictions has compared the accuracy of weight matrix, neural network, and hidden Markov models. Due to different balances of recall versus precision, the results of Table 1 in their study appear insufficient to declare which method is best. However, it seems that an overall accuracy of around 90% is possible.

4.3.1. *Which Method Should I Use?*

Of course we encourage researchers to look at several methods and admit that our coverage in this chapter is only partial. Figures 2 and 3 show some public prediction servers and datasets. For most users we recommend trying the TargetP[232] server because it has a good reported prediction accuracy and the programs it is based on, such as SignalP,[611, 614] appear to have been designed by researchers with a thorough understanding of the current state of knowledge regarding sorting signals and processes. One drawback of SignalP is that it is a commercial piece of software and the program is not freely available to certain non-profit research organizations. However, use of the public server is free. Another minor drawback is that it may not be well suited for proteins that are localized through mechanisms other than signal peptides.[610] These are relatively rare, and we expect such proteins to be hard for any prediction program. Although some of the various PSORT programs are in serious need of updating, such updates are planned and we believe the PSORT web site will continue to be a valuable resource.

In large genome projects, gene finding programs often mispredict the N-terminal region of proteins;[265] see also Chapter 5. Thus methods—many of which are shown in Figure 8—which do not rely on N-terminal signals are especially useful because they can be expected to be relatively robust against start site errors.[369] In any case, since gene finding programs typically can make use of similar sequences to increase accuracy, the accuracy of gene finding is likely to increase along with the rapidly increasing availability of sequences from various organisms.

4.4. *A Caveat Regarding Estimated Prediction Accuracies*

We note here that the prediction accuracy reported in various studies, including our own, may be a bit optimistic. One obvious important difficulty in comparing

Program	URL	Ref.
PSORT	`psort.ims.u-tokyo.ac.jp`	61, 597
PSORTII	`psort.ims.u-tokyo.ac.jp`	368, 597
NNPSL	`www.doe-mbi.ucla.edu/~astrid/astrid.html`	710
TargetP	`www.cbs.dtu.dk/services/TargetP`	232
LOC3d	`cubic.bioc.columbia.edu/db/LOC3d`	596
PLOC	`www.genome.ad.jp/SIT/ploc.html`	648
SubLoc	`www.bioinfo.tsinghua.edu.cn/SubLoc`	369
ProtComp	`www.softberry.com/berry.phtml?topic=proteinloc`	
Predotar	`www.inra.fr/predotar`	

Fig. 2. Some public localization prediction servers are shown with references to the literature where available.

prediction accuracies is the variety of localization site definitions and datasets that have been used in different studies. The majority of the works have used annotations in SWISS-PROT[86] to train and test their methods however, which leads to another more subtle problem in estimating prediction. Strictly speaking, estimating the accuracy based on performance on a test set is only valid if the test set data is used just once.

Consider two arbitrary classifiers—on a particular data set one may classify more accurately than the other simply by chance. Many works have been published using various subsets of the same data. Moreover, since each work generally reflects the results of testing multiple classifiers or one classifier with many parameter settings, the effect is amplified and the same data has been used many times for testing. Thus the results of even rigorous cross-validation studies should be taken with this in mind.

In the machine learning community, the UCI Repository[84] contains datasets—including two for protein localization—that can be used to test classification algorithms. Despite the fact that the repository contains more than 100 datasets, there

Dataset	URL	Ref.
SWISS-PROT	`www.ebi.ac.uk/swissprot`	86
NLSDB	`cubic.bioc.columbia.edu/db/NLSdb`	593
MITOP	`mips.gsf.de/proj/medgen/mitop`	749
YEAST GFP LOCALIZATION DB	`yeastgfp.ucsf.edu`	374
YPL.db	`genome.tugraz.at/ypl.html`	312
TRIPLES	`ygac.med.yale.edu/triples/triples.htm`	463
MIPS CYGD	`mips.gsf.de/genre/proj/yeast/index.jsp`	564

Fig. 3. Some public localization datasets are shown with references to the literature.

is a serious concern that its repeated use has lead to inaccurate conclusions on the general accuracy of classifiers tested on it.[740]

4.5. *Correlation and Causality*

An important issue in evaluating prediction schemes for localization is the distinction between sequence features that reflect causal influences on localization versus those which merely reflect correlation with localization. Figure 6 uses the localization of a transcription factor to the nucleus to illustrate this point. Nuclear localization signals (NLS)[299] in proteins cause them to be selectively imported into the nucleus by importins. Indeed a classic study by Goldfarb *et al.* [295] shows that not only is nuclear localization impaired by mutations in the NLS, but also that non-nuclear proteins are imported into the nucleus when modified to include artificial NLS's. Thus the presence of an NLS naturally correlates with nuclear localization.

The vast majority of the DNA in a eukaryotic cell is found in the nucleus. Thus proteins whose function is to interact with DNA are generally imported to the nucleus, and therefore DNA binding motifs such as the zinc finger binding motif [438] also correlate with nuclear localization. There is however a fundamental difference between these two correlations. The zinc finger binding motif is not believed

Freq.	Description
7307	cytoplasmic (by similarity)
6380	cytoplasmic
5166	secreted
4181	integral membrane protein
3655	nuclear
3251	integral membrane protein (potential)
2177	cytoplasmic (probable)
1862	cytoplasmic (potential)
1542	chloroplast
1241	type i membrane protein
1114	nuclear (potential)
1029	nuclear (probable)
869	nuclear (by similarity)
775	mitochondrial
721	integral membrane protein (probable)
495	integral membrane protein. inner membrane (potential)
484	mitochondrial matrix
435	integral membrane protein. mitochondrial inner membrane
428	secreted (by similarity)
393	extracellular
381	periplasmic
373	integral membrane protein. inner membrane
355	chloroplast thylakoid membrane
352	integral membrane protein (by similarity)
312	secreted (potential)
309	integral membrane protein. inner membrane (by similarity)
296	integral membrane protein. inner membrane (probable)
270	membrane-bound
226	attached to the membrane by a gpi-anchor
216	periplasmic (by similarity)
212	mitochondrial inner membrane
211	membrane-bound. endoplasmic reticulum
204	cytoplasmic and nuclear (by similarity)
202	attached to the membrane by a lipid anchor (potential)
199	membrane-associated (by similarity)
193	inner membrane-associated (by similarity)
180	periplasmic (potential)
175	type i membrane protein (potential)
165	attached to the membrane by a lipid anchor (probable)
162	nuclear; nucleolar
161	mitochondrial (by similarity)
156	type ii membrane protein
156	secreted (probable)
153	integral membrane protein. chloroplast thylakoid membrane
148	lysosomal

Fig. 4. The 45 most frequent descriptions of subcellular localization in SWISS-PROT.

to exert a causal influence on nuclear localization. For example, Mingot *et al.*[568] has created a mutant form of a nuclear protein in which DNA binding is abolished but nuclear localization is retained.

The correlation between zinc finger binding motifs and nuclear localization is

Freq.	Description	Freq.	Description
827	cytoplasm, nucleus	13	bud neck, cell periphery
823	cytoplasm	12	cytoplasm, nucleolus, nucleus
496	nucleus	11	punctate composite, early Golgi
485	mitochondrion	11	early Golgi
266	ER	11	ambiguous, bud neck,cell periphery, bud
157	ambiguous	10	microtubule
121	vacuole	10	cell periphery, bud
73	punctate composite	10	bud neck, cytoplasm,cell periphery
73	nucleolus, nucleus	10	bud neck, cytoplasm
70	nucleolus	10	ambiguous, bud neck, cytoplasm, cell periphery, bud
57	cell periphery	9	ER, cytoplasm
54	vacuolar membrane	8	nucleus, spindle pole
53	nuclear periphery	8	mitochondrion, punctate composite
39	spindle pole	8	cytoplasm, nucleolus
34	endosome	8	ambiguous, bud neck, cytoplasm, bud
33	late Golgi	8	ER, vacuole
27	actin	6	ER to Golgi
21	peroxisome	5	mitochondrion, nucleus
21	cell periphery,vacuole	5	early Golgi, late Golgi
19	lipid particle	5	cytoplasm, vacuole
18	cytoplasm, punctate composite		
18	cytoplasm, mitochondrion		
18	Golgi, early Golgi		
15	bud neck		
15	Golgi		

Fig. 5. The 45 most frequent descriptions of subcellular localization in the yeast GFP[374] database.

real and useful for the prediction of native proteins. Especially since NLS's are relatively difficult to detect from primary sequence information—Nair and Rost reported that NLS's were detected in only 1% of eukaryotic proteins, using sequence analysis with a strict precision requirement.[596] However it should be kept in mind that non-causal correlations such as the one between DNA binding motifs and nuclear localization may not be robust when applied to mutant or non-native proteins. The understanding of which correlations reflect causal influences is critical to the ability to design novel protein sequences with some desired behavior. A classic example of this issue from another area of bioinformatics is the use of codon usage in gene finding. Indeed the question of how to treat causal and

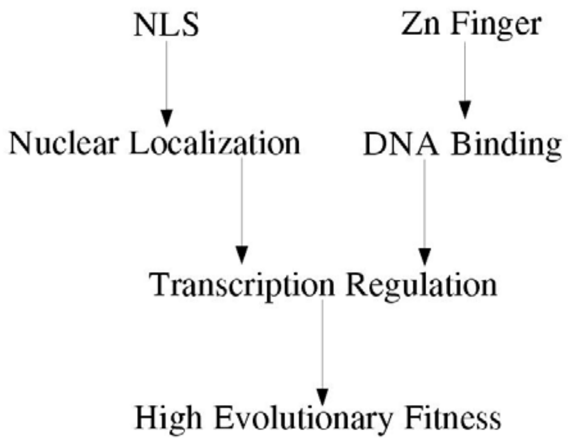

Fig. 6. A schematic diagram illustrating some causal influences is shown. NLS is the nuclear localization signal. Many omitted variables also exert causal influence on function and fitness.

non-causal correlations is not specific to protein localization, but is important in any application of machine learning or statistical analysis. We refer the reader to Pearl[652] for an in depth analysis of causality in the context of statistical inference.

5. Localization and Machine Learning

An impressive number of learning and knowledge representation techniques have been applied to the problem of predicting protein subcellular localization. Indeed the list is fairly representative of the techniques from AI, pattern recognition, and machine learning that have been applied across the entire field of bioinformatics. We have attempted to organize the work in this field by classifier (Figure 7) and by the kind of input used (Figure 8) to classify. We must admit that these two tables are incomplete and imperfect. In particular, many works that are specific to localization to a particular site are omitted—although many are included—and the categorization of classifier and input type is rather arbitrary in some cases. For papers that cover a variety of classifiers or input techniques we are likely to have made some errors of omission as well. Nonetheless we believe that these figures are a useful way to organize the existing body of work on localization prediction.

Many of the methods in Figure 7 are briefly described in Chapter 3. We do not describe them in detail here. Instead we outline some of the common approaches and we give our thoughts on the strengths and weaknesses of these methods when applied to localization prediction.

Prediction Schemes: Classifier Technique Used	Ref.
Rule-Based Expert System	228, 601, 602
Discriminant Analysis	136, 159, 165, 245, 604, 617, 923
Principle Component Analysis	25, 923
Bayesian Network	272
Naïve Bayes	213, 368
Decision Trees	368
Nearest-Neighbor Methods	368, 594, 595
Support Vector Machines	131, 161, 272, 369, 648
Feed-Forward Neural Network	89, 134, 165, 231, 232, 389, 589, 611, 710, 755
Kohonen Self-Organizing Map	755
(Hidden) Markov Models	134, 267, 272, 612, 922
Human-Designed Structured Model	134, 367, 601, 602
(Generalized) Weight Matrix	160, 862, 863
Feature Discovery/Data Mining	61, 272, 366

Fig. 7. A partial list of classifying methods and programs that employ them to predict localization is shown. The Bayesian network framework is general enough to encode any probability distribution, thus including many other categories, but here we reserve the Bayesian network category for authors who presented their methods as Bayesian networks. Weight Matrix includes methods that couple a few fixed columns to allow for interdependence between sites. We include Markov chain models in (hidden) Markov models. Discriminant analysis is used as a catch all for discriminant methods that don't fall into more specific categories.

5.1. *Standard Classifiers Using (Generalized) Amino Acid Content*

The correlation between amino acid composition and protein localization was observed as early as 1982.[617] As can be seen in Figure 8, many methods have been developed based on amino acid composition or slightly generalized features such as the composition of amino acid pairs separated by 0-3 amino acids, including recent studies[369, 648] using support vector machines.[178, 856]

These methods achieve competitive accuracy. Moreover they have the advantage that they may be accurate even for proteins that have sorting signals which are too subtle to be found with current prediction techniques, or that localize without the direct use of sorting signals—see Section 2.

As mentioned above, for genome projects in which putative amino acid chains

Prediction Schemes: Information Used as Input	Ref.
(Generalized) Amino Acid Composition	$131, 136, 165, 245, 369, 604, 710, 798, 922, 272, 648$
N-terminal Sequence Region	$164, 165, 367, 368, 601, 602, 611, 862, 863, 61, 134, 160, 231, 272, 755$
Entire Amino Acid Sequence	$134, 367, 368, 601, 602$
Sequence Periodicity	$159, 648$
Sequence Similarity	$272, 595$
Sequence Motifs	$61, 213, 231, 366$
Protein Signatures	$130, 161, 272$
Physiochemical Properties	$61, 164, 213, 245, 755$
mRNA Expression Data	213
Knockout lethality	213
Integral α-helix Transmembrane Region Prediction	$134, 272, 367, 368, 601, 602$
Surface Residue Composition	25
Text Descriptions	$228, 594, 798$
Fluorescence Microscope Images	$89, 589$
Meta Localization Features	$272, 367, 368, 601, 602$

Fig. 8. A partial list of types of input information and references for programs that use that input to predict localization is shown. Generalized amino acid composition may include slightly higher-order inputs such as the composition of adjacent pairs of amino acids. Protein Signatures—some of which can also be classified as sequence motifs—are taken from PROSITE[238], SBASE-A[590], and InterPro.[582] Meta localization features are the results of localization prediction programs; for example, PSORT(II) uses a modified version of McGeoch's signal peptide prediction program[863] as an input feature.

may have incorrectly predicted N-terminal regions, the lack of a strong dependence on N-terminal sorting signals is also a practical advantage. However, we speculate that some of the amino acid composition bias utilized by these methods reflects an adaptation to functioning effectively in the different chemical environments found in different compartments, as discussed in Andrade *et al.*,[25] rather than a causal factor in their localization.

We do not dismiss all of the bias as being mere correlation. For example, hydrophobicity can certainly affect the integration into or transition through membranes. Still we feel that this approach is relatively prone to relying on non-causal correlations. The potential drawbacks of such a reliance is discussed in Section 4.4.

5.2. *Localization Process Modeling Approach*

PSORT[601, 602] uses a tree-based reasoning scheme designed to roughly reflect the localization process. Rules are supplied for each decision node in the tree in which a feature is primarily chosen from biological knowledge. For example, a weight matrix designed to detect signal peptides is used at the node representing transport through the endoplasmic reticulum membrane.

Unlike the decision trees mentioned in Section 5.5, the tree architecture is designed from prior knowledge rather than induced from statistical properties of a training set. The advantage of this system is that it is not only a prediction scheme but is also a kind of knowledge base. To the extent that biologists understand the mechanisms of localization it should be possible to design prediction schemes that model the process. Given the dependence on prior knowledge we expect that most of the correlations PSORT uses are causal in nature.

One disadvantage of this approach is the labor intensive process of updating the rules to reflect the ever growing body of knowledge regarding localization processes. Another disadvantage is the lack of a good way—such as cross-validation, but see Section 4.4—to estimate the accuracy on unkown sequences. It is of course not feasible to remove the influence of a set of randomly chosen test sequences on the body of knowledge regarding localization. Finally this method has less ability to fully leverage all correlations between sequence features and localization to maximize prediction accuracy.

5.3. *Sequence Based Machine Learning Approaches with Architectures Designed to Reflect Localization Signals*

A series of works by Nielsen, Emanuelsson, and colleagues have taken an approach in which sophisticated sequence-based classifiers are used—but knowledge of localization is employed to select the classifier architecture and input sequence region.[231, 232, 611, 612, 614] In particular, many of those works use feed-forward neural networks[57, 82, 687, 730] to predict sorting signals. Although provably optimal learning procedures are not known, in principle feed-forward neural networks can learn non-linear interactions between distant amino acids that affect localization.

Applied without careful analysis of the learned weights, neural networks generally have a tendency to produce “black box” classifiers. However the works describe in this section do provide some such analysis and their restriction of the sequence region input should greatly reduce reliance on spurious correlations. In other words, they use prior knowledge to limit the complexity of the input but rely on machine learning techniques to determine the actual function of the input that is used for classification. We feel this general approach should be effective for many bioinformatics applications.

5.4. *Nearest Neighbor Algorithms*

PSORTII[368] uses the PSORT features. But instead of using the PSORT reasoning tree, it uses the k nearest neighbors classifier (k-NN).[214] To classify a sequence, PSORTII simply considers the k sequences in the training data whose feature vector most closely matches, by euclidean distance, the feature vector of the sequence to be classified. This classifier was found to be more effective than decision tree induction, Naïve Bayes, and a structured probabilistic model roughly based on the PSORT reasoning tree[368]. Predicting localization by sequence similarity search for close homologs of known localization is another commonly employed method,[272, 595] which amounts to a kind of nearest neighbor classification.

Nearest neighbor classifiers do not summarize the training data, and thus in a machine learning sense have very little explanatory power. They do however naturally provide the particular examples in the training data which are most similar to the data to be classified. This is useful in problems such as localization where much valuable *ad hoc* annotation information can be given in addition to the predicted class. For example, even if the localization site definition used does not give nucleolar proteins their own distinct class, if the nearest neighbors to a sequence are annotated as “nuclear; nucleolar” in SWISS-PROT, that gives a valuable clue that the sequence may localize to the nucleolus.

We are currently designing a successor to PSORTII but plan to retain at least some form of nearest neighbor classification. It has been reported[603] that the accuracy of PSORTII can be improved by using more sophisticated variants of nearest neighbor classifiers such as discriminant adaptive nearest neighbor classifiers.[327]

5.5. *Feature Discovery*

In this section we introduce two studies, Horton[366] and Bannai *et al.*[61], that focused on trying to automatically discover simple sequence features that are relevant to protein localization. Although neither study was able to surpass previous methods in estimated prediction accuracy, we feel their feature discovery approach merits mention. Since the methods used in these studies are less established than many of the classification algorithms mentioned in Figure 7, we briefly describe them here.

Horton[366] first identified substrings in the protein sequences that correlate significantly to localization site and then built decision trees with the standard decision tree algorithm using those substrings as potential features. Decision trees were chosen because decision trees are relatively easy to interpret and the standard decision tree induction algorithm[692] includes feature selection. The leaves of the decision trees are localization sites and the internal nodes are binary nodes that represent tests of the number of occurrences of a particular substring in the input sequence. Such trees were induced on several random subsets of the localization data with the idea that consistently selected substring features would be important. Figure 9 shows one of the induced trees. The prediction accuracy of this tree is far from competitive with the best methods and many discovered features appear to simply reflect amino acid composition bias—but we were pleased that a test for the presence of a carboxy-terminal phenylalanine was consistently selected from an *E.coli* dataset. Although we were not aware of the fact until after the feature discovery experiment was conducted, the presence of a carboxyl-terminal phenylalanine is an experimentally verified factor in localization to the outer membrane in bacteria.[807]

Bannai *et al.*[61] considered a broader class of features, including substrings pattern that only require a partial match and patterns that group amino acids based on published amino acid indexes, many of which reflect various chemical properties of amino acids. Instead of the decision tree induction algorithm they used an extensive search of possible rules to build a decision list[472] which is a special case of a decision tree with a linear structure. Their method was able to "discover" the known fact that mitochondrial targeting signals have an amphiphilic α-helix structure and predict signal peptides with competitive accuracy using only a simple average of hydrophobicity over the appropriate sequence region.

5.6. *Extraction of Localization Information from the Literature and Experimental Data*

This chapter has focused on predicting protein localization from information directly derivable from the amino acid sequence of the protein, which is more or less the same problem nature is faced with. There is of course another source of information available to use—experimental data and the vast literature based upon it. Ultimately all of the methods mentioned in this chapter are based on information gained by human interpretation of experimental data. But recently there has been progress in developing computer programs to automatically extract such information. Eisenhaber and Bork[228] have developed a rule-based method for classify-

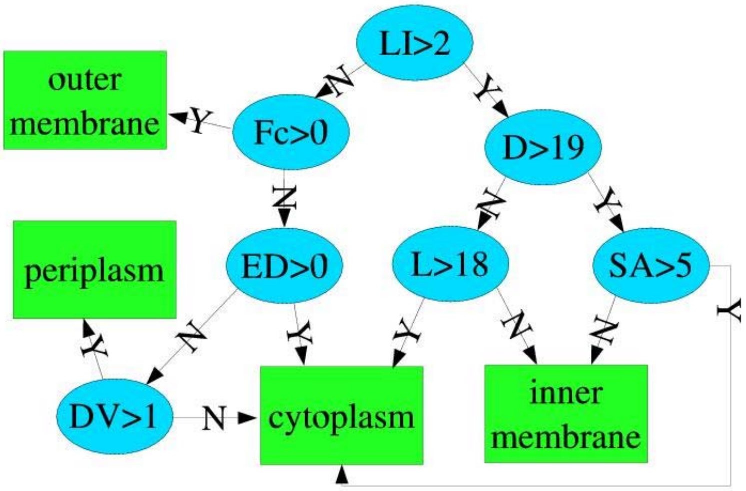

Fig. 9. A decision tree induced for localization in *E.coli* is shown. The leaf nodes (green rectangles) represent localization sites. To simplify presentation, leaves representing the same localization site have been merged. Internal nodes (blue ovals) represent binary tests. The root node test for whether the substring "LI" appears more than twice or not. A lower case "c" represents the C-terminal of the amino acid sequence. Thus the condition "Fc$>$ 0" tests for the presence of a carboxy-terminal phenylalanine. (*Image credit: Adapted from Figure 15.3 of Horton.*[366])

ing SWISS-PROT entries that lack an explicit "subcellular localization" tag but contain sufficient information in their description to determine their localization. For example, they observe that the functional description of "cartilage protein" is sufficient to infer extracellular localization. Support vector machines have also been applied to extracting localization information from the literature.[798] Nair and Rost[594] also used machine learning to predict localization from SWISS-PROT keywords. Murphy and colleagues have developed methods for classifying localization from fluorescence microscopy images,[89, 589] and methods to extract and analyze such images automatically from figures and captions in the literature.[449]

6. Conclusion

In this chapter we have briefly discussed many aspects of protein localization in the context of bioinformatics. We hope that the overview of the biology and some experimental techniques presented here will be valuable background for computer scientists seeking to develop new prediction algorithms. For biologists who are primarily users of such algorithms, we hope that our brief summary of common approaches to prediction will give helpful insights as to what is going on "under the hood" in many commonly used prediction tools, as well as providing a starting point for algorithm developers.

We have stated our views regarding how priorities will shift in the age of proteomics—namely towards methods that reflect the biology in a robust way. As we have mentioned earlier, some aspects of localization cannot be understood statically, since the localization of some proteins is dynamic or conditioned upon the state of the cell. With proteomic-scale experiments vastly increasing our body of knowledge, we believe that more sophisticated models incorporating some form of time or cell state may soon become feasible. The roughly 20 years of research reviewed in this chapter is impressive—but at the current rate of innovation we believe the next decades will prove even more exciting.