PROTEINS: Structure, Function, and Genetics 53:917–930 (2003)

Better Prediction of Sub-Cellular Localization by Combining Evolutionary and Structural Information

Rajesh Nair[1,4]* and Burkhard Rost[1,2,3]
[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[2] *Columbia University Center for Computational Biology and Bioinformatics, New York, New York*
[3]*North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[4]*Department of Physics, Columbia University, New York, New York*

ABSTRACT **The native sub-cellular compartment of a protein is one aspect of its function. Thus, predicting localization is an important step toward predicting function. Short zip code-like sequence fragments regulate some of the shuttling between compartments. Cataloguing and predicting such motifs is the most accurate means of determining localization in silico. However, only few motifs are currently known, and not all the trafficking appears regulated in this way. The amino acid composition of a protein correlates with its localization. All general prediction methods employed this observation. Here, we explored the evolutionary information contained in multiple alignments and aspects of protein structure to predict localization in absence of homology and targeting motifs. Our final system combined statistical rules and a variety of neural networks to achieve an overall four-state accuracy above 65%, a significant improvement over systems using only composition. The system was at its best for extra-cellular and nuclear proteins; it was significantly less accurate than TargetP for mitochondrial proteins. Interestingly, all methods that were developed on SWISS-PROT sequences failed grossly when fed with sequences from proteins of known structures taken from PDB. We therefore developed two separate systems: one for proteins of known structure and one for proteins of unknown structure. Finally, we applied the PDB-based system along with homology-based inferences and automatic text analysis to annotate all eukaryotic proteins in the PDB (http://cubic.bioc.columbia.edu/db/LOC3D). We imagine that this pilot method—certainly in combination with similar tools—may be valuable target selection in structural genomics. Proteins 2003;53:917–930.**

Key words: protein sub-cellular localization; protein structure; secondary structure; surface composition; sequence motifs; evolutionary profiles; neural network; bioinformatics; PDB; automatic genome annotation

INTRODUCTION

Sub-cellular localization is important to elucidate protein function. Proteins must be localized in the same sub-cellular compartment to cooperate towards a common function. Therefore, experimentally unravelling the native compartment of a protein constitutes one step on the long way to determining its role. The explosion of sequence information through large-scale sequencing projects has widened the gap between the number of sequences deposited in public databases and the experimental characterization of the corresponding proteins.[1–3] Using high-through-

Abbreviations: 1D structure, one-dimensional (e.g., sequence or string of secondary structure); 3D structure, three-dimensional coordinates of protein structure; ChloroP, prediction of proteins in the chloroplast;[87] DSSP, program and database assigning secondary structure and solvent accessibility for proteins of known 3D structure; [59]PDB, Protein Data Bank of experimentally determined 3D structures of proteins;[88] PHD, profile-based neural networks for predicting secondary structure (PHDsec)[62,66,54], and solvent accessibility (PHDacc); [53,54]PredictNLS, prediction of nuclear proteins through nuclear localization signals;[22,75] HSSP, database of protein structure-sequence alignments;[61] NNPSL, neural networks predicting localization;[31] NLS, nuclear localization signal; SubLoc, support-vector machine-based prediction of localization;[64] SignalP, neural network system predicting signal peptides;[19,20] SWISS-PROT, data base of protein sequences;[16,89] TargetP, combined method predicting chloroplast (ChloroP), extra-cellular (SignalP), and mitochondrial proteins;[21] PSORT, knowledge-based expert system using amino acid composition and sequence motifs;[45,46,40] TrEMBL, translation of the EMBL-nucleotide database coding DNA to protein sequences.[16]

Notations used: "sequence-unique," we refer to sequence-unique sets as those in which no pair of proteins has more sequence similarity than a certain threshold (HSSP-distance < 10, Eqn. 1).

Methods introduced here: LOC3Dnet, combination neural networks trained on PDB sequences using observed (LOC3DnetDSSP) or predicted (LOC3DnetPHD) 1D structure and evolutionary information from multiple alignments; LOC3D, localization prediction based on combination of four different methods; LOCnet, combination of neural networks trained on SWISS-PROT sequences using predicted 1D structure (PHD) and evolutionary information.

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: National Science Foundation; Grant number: DBI-0131168.

*Correspondence to: Rajesh Nair and Burkhard Rost, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032. E-mail: nair@cubic.bioc.columbia.edu, http://cubic.bioc.columbia.edu/

Received 20 February 2003; Accepted 12 May 2003

put methods of epitope-tagging and immunofluorescence analysis Snyder et al.[4] have recently reported localization data for the entire proteome of *Saccharomyces cerevisiae* (bakers yeast). So far, the majority of large-scale experiments suggesting localization have been restricted to yeast. This is primarily due to the fidelity of homologous recombination in yeast and the concomitant ease with which integrated reporter gene fusions can be generated. In contrast, computational tools can provide fast and accurate localization predictions for any organism. [5,6] In fact, attempts to predict sub-cellular localization have become one of the central problems in bioinformatics.[7–9]

Inferring localization by homology is relatively accurate but not always applicable. A variety of approaches have been used to classify proteins with respect to sub-cellular localization. One of the most reliable means is annotation transfer from homologues: [10–13] If a protein of experimentally-known localization is significantly similar in sequence to a query protein U, localization can be inferred for U. However, the level of "significant sequence similarity" varies substantially between localizations, and is much higher than that required for correct inference of folds.[10,14,15] Thus, even when accepting many errors less than 25% of the proteins in SWISS-PROT[16] can be classified by homology into one of ten compartments.[15] Using text analysis of SWISS-PROT keywords to infer localization, we can annotate sub-cellular localization for about 48% of all proteins in SWISS-PROT.[17]

Sequence motifs predict successfully for some compartments. Another way to predict localization is to identify local sequence motifs such as signal peptides[18,19,20,21,8,9] or nuclear localization signals (NLS).[22,8,23,24] Proteins destined for the secretory pathway, the mitochondria and the chloroplast contain N-terminal targeting peptides that are recognized by the translocation machinery.[25,26] Thus, prediction methods use only the N-terminal residues.[19,21] Discriminant analysis has been applied to identify proteins imported into mitochondria.[27] Many proteins destined for the nucleus contain NLS motifs that may occur anywhere in the sequence.[28,29] Recently, we have collected a data set of experimental and potential NLS motifs as an aid to predicting nuclear localization;[22] some of these signals are also used for the export from the nucleus.[22,30] However, the vast majority of proteins have no known motif. Furthermore, a particular problem for methods detecting N-terminal signals is that start-codons are predicted with less than 70% accuracy by genome projects.[31] Overall, known and predicted sequence motifs enable annotating about 30% of the proteins in six entirely sequenced eukaryotic proteomes.[32,33,12]

Ab initio methods predict localization for all proteins at lower accuracy. A third approach to predicting localization has been suggested by the observation that the overall amino acid composition correlates with the native compartment.[34–36] This observation has led to the development of a variety of prediction methods based solely on composition.[37–39,31,40,41] Higher order correlations (residues i and (i+n), n = 2,3,4) have been accounted for by using pseudo-amino acid composition.[42,43] With the availability of many completely sequenced genomes, phylogenetic profiles have been employed to identify sub-cellular localization.[44] So far, this approach has been much less accurate than methods based solely on composition. PSORT II is a knowledge-based expert system that integrates rules based on amino acid composition with known sequence motifs,[45,46,40], and also uses other methods such as NNPSL.[31] Thus, the accuracy of PSORT II somehow depends on the accuracy of the underlying original methods. Drawid & Gerstein have proposed a Bayesian system based on a diverse range of 30 different features.[47] They applied their system to predicting localization of yeast proteins. Using a seven-fold jack-knife procedure on 1342 yeast proteins with known localization, they reported a prediction accuracy of 75% at 67% coverage.[47]

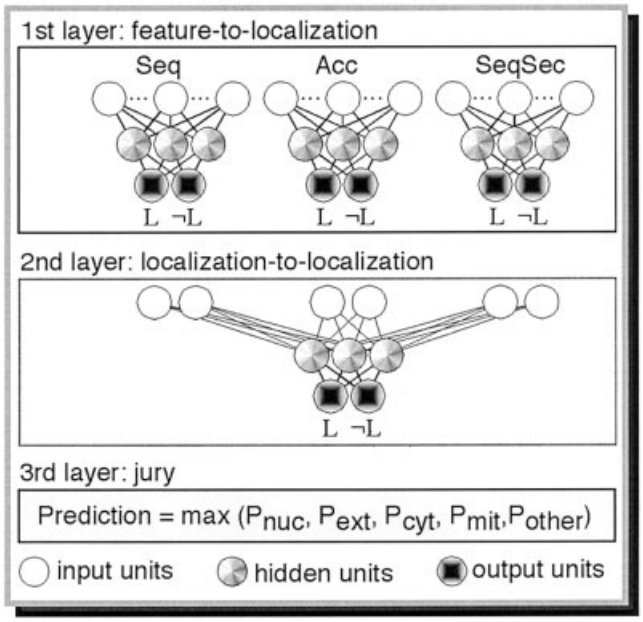

Fig. 1. Final neural network architecture. For the first level of pairwise neural networks we used an architecture of 20–60 input units and 2 output units with a hidden layer consisting of 3–9 units. We used the output from the different first level pairwise neural networks as input to the second level integrating neural network. The second level pairwise networks consisted of 6 input units and 2 output units with a hidden layer consisting of 3 units. The final localization prediction was based on a jury decision of the outputs from the different pairwise integrating networks.

What determines sub-cellular localization? Experimental studies of protein targeting have shown that the sub-cellular localization of a protein is determined by its three-dimensional (3D) structure and/or the presence of local sequence motifs. Mutation studies have shown that disrupting the structure of a protein causes aberrant localization.[48–51] One way of incorporating global information about protein structure into predictions of localization is by using secondary structure content. A number of local sequence motifs have been shown to mediate protein targeting.[25,18,29] Miguel Andrade (EMBL, Heidelberg), Sean O'Donoghue (LION, Heidelberg) et al. have previously concluded that the signal for sub-cellular localization is almost entirely due to the surface residues.[10] Here, we have utilized these aspects of protein structure information to aid predicting localization. We describe *LOC3Dnet*

TABLE I. Number of Proteins in Data Set†

Sub-cellular localization	SWISS-PROT all	SWISS-PROT unique	SWISS-PROT new-unique	PDB all	PDB unique
Nucleus	2673	556	178	769	124
Cytoplasm	2137	348	146	1958	94
Extra-cellular space	1936	361	128	1970	99
Mitochondria	914	200	60	504	23
Lysosome	117	28	7	111	6
Endoplasmic reticulum	112	15	13	55	3
Golgi apparatus	11	4	7	7	2
SUM	8976	1512	539	5906	359

†Data sets: *SWISS-PROT*: number of all eukaryotic proteins with annotated experimentally determined sub-cellular localization taken from SWISS-PROT release 40 (Methods); *SWISS-PROT* unique: number of proteins in sequence-unique subset of all *SWISS-PROT* (Methods); *SWISS-PROT* new-unique: number of proteins in sequence-unique subset of all proteins found in SWISS-PROT release 41 and not in release 40 (chosen by same procedure as SWISS-PROT unique); PDB: number of PDB chains that could be assigned the given localization (Methods); *PDB unique*: sequence-unique subset of previous.

and *LOCnet* (Fig. 1), two systems of neural networks that sort proteins into one of four localization classes: extra-cellular, cytoplasmic, nuclear and mitochondrial. One (*LOC3Dnet*) is specialized on sequences from the PDB, the other (*LOCnet*) on sequences from SWISS-PROT. We excluded helical membrane proteins and used proteins from other minor compartments only as "false positives." In particular, proteins from the secretory pathway that are retained in the Golgi apparatus or the Endoplasmic reticulum were treated as "non extra-cellular." The method used 3 + 1 layers to make the final decision. The first layer consisted of four dedicated neural networks that used particular features from protein sequences, alignments, and structure to pre-sort proteins into L/not-L (L = cytoplasmic, nuclear, extra-cellular, mitochondrial). The second layer neural networks combined different input features. The third layer used a simple jury decision[52] to assign one of four localization-states to each protein. We applied the final system to predict the native sub-cellular compartment for all eukaryotic proteins in PDB. Toward this end, we added a fourth layer combining the information from PredictNLS,[22] sequence homology,[15] automatic text analysis[36] and the prediction system introduced here. The neural networks were trained and tested on PDB and SWISS-PROT sequences of experimentally annotated localizations. We distinguished the following input features: overall amino acid composition, the amino acid composition of surface residues, composition of three-state secondary structure (helix, strand, other), and the amino acid composition separated into three secondary structure states (helix, strand, other). We also predicted secondary structure and surface composition using PHDsec and PHDacc respectively.[53,54] Since biased data sets tend to yield over-estimates in prediction accuracy,[55] we took great care to select sequence-unique subsets of the data to estimate performance. All results of the methods described here were based either on four-fold cross-validation experiments or on SWISS-PROT sequences that had no significant sequence similarity to any protein used for development. The LOC3D localization prediction server for protein structures and the results of our annotations for all eukaryotic chains in the PDB can be accessed through the web at http://c2b2.columbia.edu/db/LOC3D/.

Methods

Data sets used for development and evaluation. We selected all eukaryotic proteins with explicit annotations about sub-cellular localization in SWISS-PROT release 40.[16] We excluded proteins annotated as MEMBRANE, POSSIBLE, PROBABLE, SPECIFIC PERIODS, or BY SIMILARITY. We also excluded proteins annotated with multiple localizations. This left 8980 eukaryotic proteins in our SWISS-PROT data set of experimentally annotated localization ("trusted SWISS-PROT set," Table I). Next, we assigned localization to PDB chains[56] by searching for homologues in the "trusted SWISS-PROT set." We transferred the annotated localization for all PDB chains, which were aligned at HSSP-distances (Eqn. 1) above 10 (number of PDB chains found given in Table I). Above this homology threshold sub-cellular localization annotation can be reliably transferred at over 90% accuracy.[17,15] Training, test and validation sets were constructed such that no pair of proteins from any two sets had levels of sequence similarity above HSSP-distances of 5 (Eqn. 1). We picked this value, since below this threshold assigning sub-cellular localization based solely on homology leads to significant errors.[15] Furthermore, the test set was redundancy-reduced at HSSP-distances < 10 using a simple greedy search.[57] This ensured that no two proteins in the test set had greater than 40% sequence identity over more than 100 residues (number of sequence unique chains given in Table I). The reason for this reduction was to find a balance between biased data known to yield over-estimates[58,55] and between data sets that were too small. Note that we did not have to define thresholds for significant sequence similarity between motifs, such as signal peptides,[58] since we used the entire protein information. All non-eukaryotic proteins were also excluded for testing. We identified eukaryotic proteins by using three methods: (1) PDB to SWISS-PROT links in the SWISS-PROT database, (2) using the source and organism entry in PDB and (3) first hit is eukaryotic protein when the chain is aligned to the SWISS-PROT database. Note: all data sets are available at: http://cubic.bioc.columbia.edu/results/2003/localization/.

SWISS-PROT-new set used only for testing. After we completed the development of all our methods, we used an

additional data set to re-examine performance, namely, we collected all proteins that had been added to SWISS-PROT between release 40 and 41 (labelled "SWISS-PROT-new"). We filtered out all of these new proteins that had HSSP-distances > 5 to any previously used protein and found the sequence-unique subset of the new proteins (Table I). We never used any of these proteins for development, and it is rather unlikely that the other methods tested used any of these (see Table IV).

HSSP-distance to measure pairwise sequence similarity. The HSSP-distance is defined as the distance from the HSSP threshold;[14] it is given by:

$$\text{HSSP-DISTANCE} = \text{PIDE} - \text{HSSP\_PIDE}(\theta)$$

$$\text{HSSP\_PIDE}(\theta) = \theta + \begin{cases} 100, & \text{for } L \le 11 \\ 480 \cdot L^{-0.32.\{1+e^{-L/1000}\}}, & \text{for } L \le 450 \\ 19.5, & \text{for } L > 450 \end{cases} \quad (1)$$

where L is the length of the alignment between two proteins, PIDE the percentage of pairwise identical residues, and HSSP_PIDE(θ) the revised HSSP-threshold for the level θ.

Observed and predicted information about protein structure. The observed secondary structure was extracted from the DSSP assignments.[59] Exposed residue composition was calculated from the solvent-accessible surface area[60] in the DSSP database.[59] Three-state secondary structure was predicted using PHDsec. We predicted all residues as exposed that were predicted to have relative solvent accessibility $> 10\%$ by PHDacc.[54] We chose this threshold since it gave good prediction accuracy on a limited subset of the training sets.

Building profiles. Profile-based composition was calculated by aligning the sequences against the SWISS-PROT + TrEMBL database using MaxHom dynamic programming algorithm.[61] The aligned sequences were filtered for redundancy at 95% pairwise sequence identity, i.e., pairs exceeding this limit were removed. Finally, we included only those proteins into the alignment that were above an HSSP-distance of 5 and had a pairwise sequence identity above 50% with respect to the guide protein of known localization. These thresholds were found to be optimal on a limited subset of the training data. Finally the profile composition was calculated by replacing each amino acid residue in the protein by the residue frequencies in the profile.

Cross-validation. We separated our data into three sets: training, validation, and testing set. Finally, we rotated through the sets such that each protein was used for testing exactly once. We never used any information from the test set to optimize parameters. In particular, we determined the number of hidden units based on of the validation sets and did not change it when we rotated. We stopped training when the best classification was obtained on the validation set.

Neural network training and architectures for PDB chains. We used three levels of networks (Fig. 1). First, a feed-forward neural network architecture[62] with one hidden layer and two output units trained on class/non-class for each localization. Training was done with standard back-propagation including momentum term (details in Rost and Sander[62]). The two output units represent different strengths of yes/no predictions for each localization class. Only localization classes with sufficient training examples in the PDB were considered. The neural networks were trained on PDB chains using overall amino acid composition, surface residue composition (both twenty input units), three-state secondary structure composition (three input units), and amino acid composition in the three secondary structure states (sixty input units) as input. We applied "balanced training," i.e., examples belonging to the "yes" and "no" classes were alternately presented to the network during training. For networks with three and 20 input units a configuration with three hidden nodes was chosen, while for the network using amino acid composition in a secondary structure state as input, a configuration with 9 hidden nodes was chosen. For the second level, the different first-level networks were combined using a jury decision (sum over all first-level outputs) and combining neural networks (input first-level output). The training, test and validation sets remained the same for the second-level networks. The second-level networks had 6 input units and three hidden units. In the third level the combination networks for the different localizations were combined in a jury to give the final four-state localization prediction.

Neural network training and architectures for SWISS-PROT proteins. We used basically the same architectures as for PDB chains, with two major differences. First, we only trained and tested on predicted secondary structure and solvent accessibility (since structure is not known for most of these proteins). Second, we used additional input units for the final summary networks, namely each network "saw" the composition of the 50 N-terminal (20 units) residues (for proteins shorter than this, we simply used the entire protein for both ends).

Final decision through simple winner-take-it-all on 2nd layer of networks. The second layer networks (see Fig. A6) all have two output units with the values:

out_L and $\text{out}_{\neg L}$ for L = {cytoplasmic, extracellular, nuclear, mitochondrial}

We converted these values into probabilities:

$$p_L = \frac{\text{out}_L}{\text{out}_L + \text{out}_{\neg L}} \text{ and } p_{\neg L} = \frac{\text{out}_{\neg L}}{\text{out}_L + \text{out}_{\neg L}}$$

Then we predicted the protein in the localization L′ with:

$$L' = \max \{p_{\text{nuclear}}, p_{\text{extra-cellular}}, p_{\text{cytoplasmic}}, p_{\text{mitochondrial}}\} \quad (2)$$

The strength of this prediction was measured using the reliability index RI:

$$\text{RI} = \frac{p_{L'}}{\sum_L p_L} \quad (3)$$

Evaluating performance. Four-fold cross-validation was applied to test the neural networks. As a simple measure

TABLE II. Neural Network Performance on Test Set of Sequence-Unique PDB Chains†

	Extra-cellular			Cytoplasmic			Nuclear			Mitochondrial		
Method	oL	σ(oL)	MC	oL	σ(oL)	MC	oL	σ(oL)	MC	oL	σ(oL)	MC
Composition only	82	2.9	0.51	62	1.0	0.29	70	6.7	0.33	85	1.7	0.37
Composition of surface DSSP	79	5.3	0.47	57	7.6	0.22	70	7.9	0.35	73	5.2	0.23
Composition of surface PHDace	76	3.6	0.42	62	2.5	0.23	69	3.5	0.28	65	1.4	0.20
Composition by sec. str. DSSP	84	3.0	0.57	61	3.2	0.22	72	3.0	0.40	78	3.4	0.31
Composition by sec. str. PHDsec	84	2.2	0.58	66	2.5	0.23	70	5.2	0.37	77	6.6	0.32
Sum over DSSP networks	83	3.3	0.57	59	2.2	0.24	73	6.5	0.41	77	5.7	0.31
Sum over PHD networks	83	2.9	0.56	65	2.7	0.28	71	5.4	0.38	77	7.0	0.31
Network on. DSSP networks	81	5.0	0.54	62	0.9	0.26	72	7.6	0.39	77	5.1	0.30
Network on PHD networks	83	4.5	0.55	67	2.0	0.27	69	5.8	0.35	75	9.0	0.29

†Abbreviations used:
Method: name of method/server used to predict localization; note: the first five are first level pairwise neural networks that predict localization and have two output units (is X/is not X), ***Sum*** and ***Net*** mark second level jury decisions by simple majority vote (***Sum***) or by a second level neural network (***Network***). ***DSSP*** marks methods that use data from known structures through DSSP as input, while ***PHD*** marks those based on PHD predictions.
oL: percentage-two-state accuracy on test set (Eqn. 7);
σ(oL): Standard deviation of *oL* for four-fold cross-validation.
MC: Mathew's correlation coefficient[63] (Eqn. 10).

for performance we used the percentage accuracy (Q, number of correctly predicted test proteins as percentage of total number of test proteins). The accuracy/specificity and coverage/sensitivity of the two-state networks were measured using four ratios derived from *TP* (number of proteins predicted to be in localization *i* and observed to be in localization *i*, the true positives), *TN* (number of proteins predicted not to be in localization *i* and observed to be so, the true negatives), *FP* (number of proteins predicted to be in localization *i* and observed not to be in *i*, the false positives) and *FN* (number of proteins predicted not to be in localization *i* and observed to be in *i*, the false negatives). We used:

$$pL = 100 \times \frac{TP}{TP + FP} \quad pX = 100 \times \frac{TN}{TN + FN} \tag{4}$$

$$oL = 100 \times \frac{TP}{TP + FN} \quad oX = 100 \times \frac{TN}{TN + FP} \tag{5}$$

In other words, pL are all correctly predicted in localization L as percentage of all **p**redicted in L, and oL all correctly predicted as percentage of those **o**bserved in L. We combined these two numbers (pL and oL) through the geometric average:

$$\mathrm{gAv} = 1/100 \cdot \sqrt{\mathrm{pL} \cdot \mathrm{oL}} \tag{6}$$

The overall four-state accuracy was measured by the accuracy Q_4:

$$Q_4 = 100 \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$= 100 \times \frac{\text{number correctly predicted}}{\text{number of proteins in data set}}$$

To determine the best two-state networks, we used the Matthews correlation coefficient[63]:

$$\mathrm{MC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} + \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{8}$$

Prediction methods. The prediction accuracy of four publicly available methods was evaluated using the sequence-unique test set of eukaryotic PDB chains (see Table III). The four methods were: (1) NNPSL: neural network based tool predicting localization from amino acid composition;[31] (2) SubLoc: support vector machine prediction of localization from amino acid composition;[64] (3) PSORT II: integrated method based on detecting sorting signals and predictions from other methods like NNPSL and SignalP;[40] and (4) TargetP: neural network based tool predicting localization based on N-terminal sequence information.[21] All methods were run with default parameter settings.

RESULTS[1]

Structural information improves accuracy. The overall amino acid composition, the surface composition, the three-state secondary structure composition, and the combined sequence-structure composition all showed some correlation with sub-cellular localization in two dimensions (see Appendix Fig. A5). The strongest signal was the residue composition separated by three secondary structure states (HEL, Appendix). We trained four different neural networks that were specialized to discriminate between one of four localizations (cytoplasmic, extra-cellular, nuclear, and mitochondrial, Table II) and all others, i.e., each network had two output units. Then, we combined the outputs from each specialist through a statistical jury decision[52] to give the final four-state prediction of localization (Eqn. 2). Networks based on the secondary structure state-specific residue composition reached the highest accuracy (from

[1]Note: all estimates for performance reported were obtained for the test sets or for data sets that had never been used for development. In particular, we never showed results for the training sets to eschew confusion.

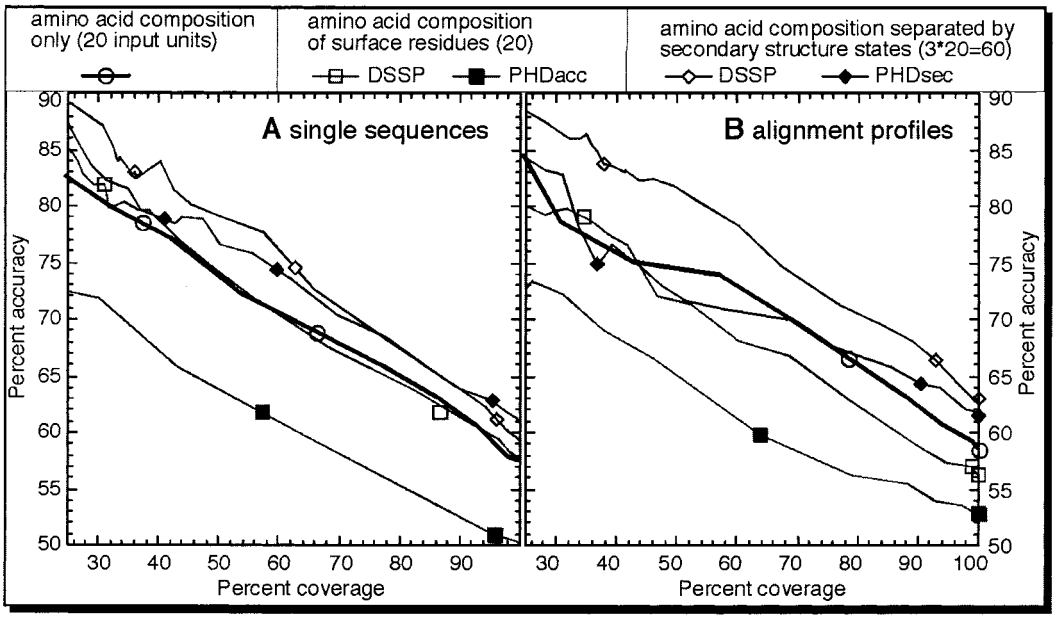

Fig. 2. Structural and evolutionary information improves prediction accuracy. The curves show accuracy (Eqn. 4) vs. coverage (Eqn. 5) for the pairwise prediction of all classes. **(A)** Single sequence information: Of the single pairwise first level neural networks, networks trained on amino acid composition separated into the three secondary structure states (HEL) were the most accurate. Here, *DSSP* represents the observed (from DSSP) and *PHD* the predicted (from PHD) three-state secondary structure. Both training and testing of the neural networks is performed on the observed or predicted features. Using the observed composition of surface residues from DSSP, prediction accuracy improved by up to a few percentage points over using sequence information alone. The three-state secondary structure predicted by PHDsec was accurate enough to preserve most of the gains in accuracy due to introduction of structural information. However, surface composition was not predicted accurately enough: Using exposed surface predicted by PHDacc, prediction accuracy dropped below that obtained by using sequence alone. **(B)** Alignment information: Prediction accuracy increased on average by 2% for the profile based networks. The only exception was the networks based on surface composition for which there was no significant improvement in accuracy. Using sequence profiles as input, networks based on overall composition were more accurate than networks based on observed and predicted surface composition (Table II for estimates of standard deviations).

57% for sequence only to above 59% for secondary structure dependent composition, Fig. 2A). Networks based on predicted surface (Fig. 2A) performed slightly worse than those based on the observed data (Fig. 2A). However, when using only the exposed surface predicted by PHD (Fig. 2A) prediction accuracy dropped below that obtained by sequence alone (Fig. 2A).

Evolutionary information improves accuracy by three percentage points. The advantages of using evolutionary information in the form of sequence profiles have been demonstrated for secondary structure prediction by a number of researchers.[65,62,52,66,54,67–72] Using sequence profiles as input (see Methods), prediction accuracy increased by up to 3% (Fig. 2B) for pairwise networks based on overall sequence and secondary structure. The number of proteins in the alignments was on average similar to that obtained for all PDB proteins solved over the last two years (data not shown,[73,74]), in other words, the set of proteins that we used to evaluate performance did not stand out from what is typical for the PDB.

Nuclear and extra-cellular proteins predicted significantly better. The prediction accuracy for the different localization classes showed similar trends when using different sequence features as input to the neural network. Extra-cellular and nuclear localizations were predicted most accurately while the cytoplasmic and mitochondrial classes were predicted with a much lower accuracy (Fig. 3, Table II).

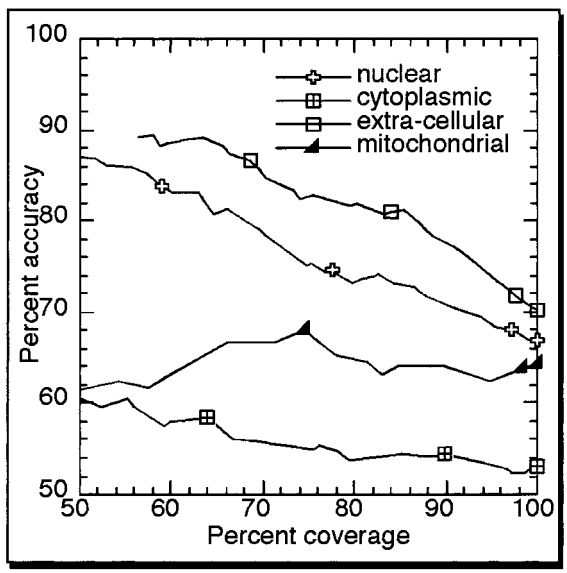

Fig. 3. Pairwise first level neural networks accurate for some localizations. The accuracy versus coverage curve for the four pairwise neural networks trained on amino acid composition separated into the observed secondary structure states shows that extra-cellular and nuclear classes were predicted very accurately (above 80% accuracy at 75% coverage). However, prediction accuracy for the cytoplasmic and mitochondrial classes was much lower (accuracy less than 65% at 75% coverage). The standard deviation in prediction accuracy for each of the localization classes was roughly five percentage points. Networks trained on other composition features showed similar trends in prediction accuracy for the different localization classes (data not shown).

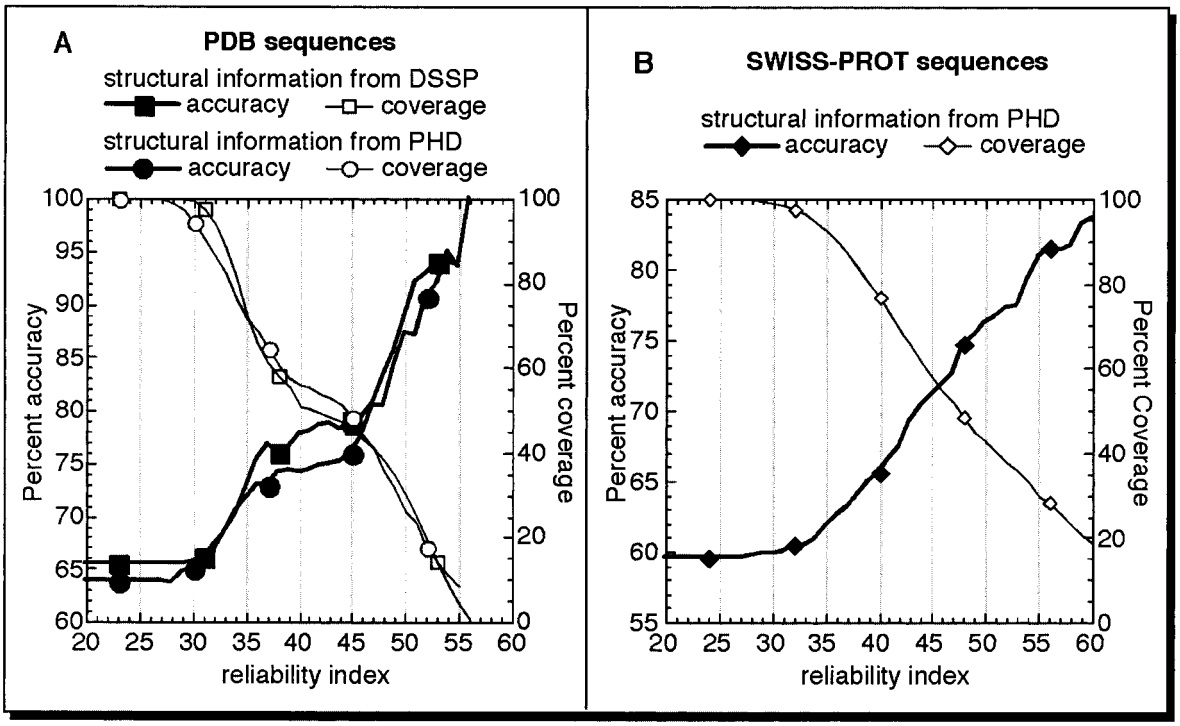

Fig. 4. More than 75% accuracy for the most reliably-predicted half of all proteins. We converted the raw neural network output into a reliability index depending on the strength of the prediction. The left panel **(A)** shows the performance of the systems trained and tested on sequences from PDB, i.e., proteins of known structure, the right panel **(B)** gives the performance of the system trained and tested on sequences from SWISS-PROT, i.e., proteins of unknown structure. For example, at a reliability index of 40, the prediction accuracy was over 74% for the PDB specialists; more than half the proteins in the test set were predicted at this reliability level. While the scale was slightly different for the system trained and tested on SWISS-PROT sequences, the performance was similar: about half the proteins were predicted at a reliability index of 47, about 73% of the proteins predicted at this level were correctly predicted.

Second level combination of simple networks improves significantly. We obtained by far the best results for each of the four localizations using combinations of the previously developed networks. We tried two versions: Combine outputs from first level networks through a statistical jury decision, and feed first level network output into a second level network. Amongst the combined networks, using evolutionary information consistently improved over using single sequences by about two percentage points (Appendix Fig. A6 A, B). The combined networks tested on predicted surface and secondary structure were only slightly less accurate than those tested on the DSSP data. The accuracy of our final system was more than six percentage points higher than networks based only on amino acid composition. The profile-based combination networks performed best and were thus used to make the final localization predictions.

More than 73% accuracy for most reliably predicted 50% of all proteins. So far, we reported levels of accuracy valid when we forced a prediction for each protein. However, some of the predictions were "stronger" than others. We translated this prediction strength into a reliability index (Eqn. 3) and investigated the dependency of prediction accuracy on this reliability (Fig. 4). When we made predictions only for the most reliably predicted half of all proteins of known structure, prediction accuracy exceeded 75% for both the networks based on predicted and observed structural information (Fig. 4A). This corresponds approximately to a reliability index of 40 for both the profile-based networks. Similarly, prediction accuracy reached about 73% for the most strongly predicted half of proteins of unknown structure, however, the actual value for this system was slightly different, namely this point was reached for a reliability index of about 47 (Fig. 4B).

Comparison to other methods. Two publicly available methods address a similar general-purpose prediction of localization without sequence motifs or homology, namely the neural network-based program NNPSL[31] and the support vector machine-based program SubLoc.[64] We applied both methods to our test set of 359 sequence-unique eukaryotic PDB chains. Since some of these proteins were used for developing NNPSL and SubLoc, our test of these public methods may slightly over-estimate their performance. On the PDB data, SubLoc reached an overall four-state accuracy around 50%, NNPSL around 44% (Table III). Our networks that used amino acid composition alone reached a similar level (*NetSeq* in Table III). Incorporating predicted surface, secondary structure and evolutionary information the final combination network performed significantly better (*LOC3DnetPHD* in Table III). Our system trained on SWISS-PROT sequences (*LOCnet*) was conceptually identical to the one trained on PDB sequences (*LOC3DnetPHD*). Hence, we were surprised that it performed significantly worse (over ten percentage points reduction in Q_4). This drop suggested that PDB sequences differ so significantly from SWISS-PROT se-

TABLE III. Comparison on Test Set of Sequence-Unique PDB Chains†

		Extra-cellular			Cytoplasmic			Nuclear			Mitochondrial		
Method	Q_4	oL	pL	gAv	oL	pL	gAv	oL	pL	gAv	oL	pL	gAv
NNPSL	43.7	70	52	0.61	47	51	0.50	28	63	0.42	30	8	0.16
SubLoc	50.1	51	56	0.54	67	47	0.56	45	61	0.53	**43**	22	0.31
NetSeq	57.4	65	69	0.67	**74**	45	0.58	53	71	0.62	21	25	0.23
LOC3DnetDSSP	**65.5**	**78**	**78**	**0.78**	55	52	0.54	**76**	**73**	**0.75**	**43**	**34**	**0.39**
LOC3DnetPHD	**63.8**	69	**75**	0.72	**71**	**54**	**0.62**	68	**72**	0.70	34	32	0.33
LOCnet	52.1	**81**	67	0.74	61	**53**	0.57	33	**73**	0.49	39	15	0.24

†Abbreviations used:
Method: name of method/server used to predict localization (methods introduced here in italic); NNPSL: Neural network-based prediction of sub-cellular localization31; *Subloc*: Subcellular localization prediction using support-vector machines 64; *NetSeq*: neural network trained only on amino acid composition of PDB sequences; *LOC3DnetDSSP*: network trained on PDB sequences using observed (DSSP) 1D structure and evolutionary profiles; *LOC3DnetPHD*: network trained on PDB sequences using predicted (PHD) 1D structure and evolutionary profiles; *LOCnet*: network trained on SWISS-PROT sequences using predicted 1D structure and evolutionary profiles.
Missing methods: Note that two general methods are missing in this table, namely PSORT II [18–20, 45, 46, 20] and TargetP [21] since both explicitly use information about signal peptides that are usually not present in PDB sequences (Table IV compares these two based on full SWISS-PROT sequences).
Classes: As described in Methods all test proteins were experimentally annotated in SWISS-PROT as exclusively belonging to one of the four classes shown. (Note that extra-cellular excludes secreted proteins retained in interior compartments.)
Scores: Q_4: percentage four-state accuracy on test set (Eqn. 9); *oL*: two-state accuracy (correctly predicted as percentage of observed, Eqn. 7); *pL*: two-state specificity (correctly predicted as percentage of predicted, Eqn. 6); *gAv*: geometric average between *oL* and *pL* (Eqn. 8).
Significant differences: For our networks the standard deviation in the four-state accuracy was about 5 percentage points. The following estimates for standard deviations were published: *NNPSL*[31], about 2.5 percentage points. No estimates of error have been provided for *SubLoc*.[64] The best methods for each class/score are marked in bold face. The standard deviations given above (2.5 percentage points for *SubLoc*) were used to mark indistinguishable methods.

quences that we need a specialist to predict sub-cellular localization for PDB proteins. When comparing these methods on a data set of sequence-unique SWISS-PROT proteins of known localization that had been added between release 40 and 41 (and were neither used in our development nor in that of the other methods tested), we got different results (Table IV): now SubLoc reached an overall four-state accuracy around 54%, NNPSL around 52%, and our system trained on SWISS-PROT sequences (*LOCnet*) clearly outperformed the system trained on PDB sequences (*LOC3DnetPHD*). Again, we observed that using all the information (predicted 1D structure and alignment profiles) yielded a sustained improvement around eight percentage points (*NetSeq* vs. *LOCnet* in Table III). Our current system was only inferior to NNPSL and SubLoc for mitochondrial proteins. Comparing our system to methods that also utilize sequence motifs (PSORT II) or that specialize on particular general signals (TargetP), we confirmed that our method performed particularly poorly on mitochondrial proteins: TargetP performed clearly best for mitochondrial proteins. In contrast, it appeared that our system implicitly picked up the presence of signal peptides used in TargetP.

Predicting the localizations for all eukaryotic proteins in PDB. Finally, we annotated sub-cellular localization for all eukaryotic protein chains in PDB. The LOC3D system employed toward this end combined four different methods: (1) inferring nuclear localization based on the presence of NLS,[22,75] (2) transferring experimental annotations of from SWISS-PROT through sequence homology,[15] (3) inferring localization through automatic text-analysis of SWISS-PROT keywords,[17] and (4) predictions from the network-based system, described here (*LOC3DnetDSSP*). The final annotation was based on the most accurate prediction from any of the four different methods (winner-take-all). Overall, transfer by homology accounted for 44% of the final annotations (Table V). The other means of explicitly using experimental annotations (automatic text analysis of SWISS-PROT keywords) yielded another 37% of the annotations. Additionally, 130 PDB proteins contained nuclear localization signals.[22,75] Thus, the success of the homology and motif-based methods left only 18% of the PDB proteins un-annotated. For about 40% of these the accuracy of *LOC3DnetDSSP* was above its average of 65%. Secreted proteins were predicted to be the most abundant class in PDB (Table V). We made all predictions available on our web site (http://cubic.bioc.columbia.edu/db/LOC3D/).

DISCUSSION AND CONCLUSION

Significant improvement through combining information. Our major finding was that integrating all sources of information, namely evolutionary information with overall, surface, and secondary structure compositions, yielded by far the best method to predict sub-cellular localization (Table III, Appendix Fig. A6). Hence, all sources of information were crucial in combination. Nevertheless, we could clearly single out the following trends. First, networks using amino acid composition separated by secondary structure state gave the highest prediction accuracy (Fig. 2). Second, the accuracy of secondary structure predictions sufficed to significantly improve predictions of sub-cellular localization. Third, replacing single-sequence composition (Fig. 2A) by profile-composition (Fig. 2B) significantly improved prediction accuracy. The gain in accuracy was maximal (about three percentage points) for the profile-based combination networks that combined all information.

TABLE IV. Comparison on Sequence-Unique Set of New SWISS-PROT Proteins†

		Extra-cellular			Cytoplasmic			Nuclear			Mitochondrial		
Method	Q	oL	pL	gAv	oL	pL	gAv	oL	pL	gAv	oL	pL	gAv
NNPSL	51.5	62	61	0.61	40	45	0.43	58	66	0.62	68	31	0.46
SubLoc	57.4	52	71	0.61	**57**	46	0.51	71	65	0.68	63	49	0.56
PSORT II	53.2	32	**89**	0.53	51	**52**	0.51	**74**	55	0.64	62	46	0.53
TargetP	—	77	77	0.77	—	—	—	—	—	—	**78**	**57**	**0.67**
NetSeq	56.2	70	74	0.72	55	44	0.49	68	65	0.66	28	29	0.29
LOCnet	**64.2**	**86**	76	**0.81**	**56**	**54**	**0.54**	**73**	**71**	**0.72**	53	45	0.49
LOC3DnetPHD	43.4	59	43	0.50	**58**	39	0.48	33	50	0.40	35	50	0.41

†Abbreviations used as in Table III, with the following exceptions:
Data set: all sequence-unique proteins added between release 41 and 40 of SWISS-PROT (Table I).
Methods: Here the reference method *NetSeq* was a neural network trained only on the amino acid composition of SWISS-PROT sequences. Additional methods: PSORT II: knowledge-based expert system using amino acid composition and sequence motifs,[45, 46, 20]; TargetP: combined method predicting extra-cellular (SignalP), chloroplast (ChloroP) and mitochondrial proteins.[21]
Significant differences: For our networks the standard deviation in the four-state accuracy was about 5 percentage points. The following estimates for standard deviations were published: TargetP[21] about 1 percentage points. PSORT If[40] about 3.5 percentage points (see Table III for the other methods).

TABLE V. Predicted Sub-Cellular Localization for all Eukaryotic PDB Chains†

Confidence	Nprd	Phom	Pkwd	Pnet	Pnls	Pcyt	Pext	Pnuc	Pmit	Pother
100	5015	67	31	0	2	23	49	10	9	9
95–99	182	99	1	0	—	57	2	40	1	2
90–94	674	18	52	31	—	6	60	23	5	6
85–89	296	26	39	35	—	23	45	20	4	7
80–84	589	7	81	12	—	16	69	9	2	3
75–79	566	4	82	15	—	45	28	9	17	1
70–74	359	1	78	21	—	25	26	17	16	16
65–69	118	69	—	31	—	71	2	17	10	0
60–64	195	5	—	95	—	14	23	14	49	0
55–59	236	—	—	100	—	8	13	7	71	0
50–54	94	—	—	100	—	64	14	7	15	0
< 50	469	—	—	100	—	70	9	3	18	0
SUM	8793	44	37	18	1	27	43	12	12	7
				Σ = 100%				Σ = 100%		

†Abbreviations used: *Confidence*: estimated annotation/prediction accuracy at this level of homology/prediction reliability; *Nprd*: number of eukaryotic PDB chains predicted at this accuracy level; *Pnls*: percentage of chains with nuclear localization signal (Nnls/Nprd); *Phom*: percentage of chains annotated through homology (Nhom/Nprd); *Pkwd*: percentage of chains for which localization could be inferred using text-analysis of SWISS-PROT keywords (Nkwd/Nprd); *Pnet*: percentage of sequences predicted using *LOC3DnetDSSP* (Table III), i.e., our final system for proteins of known structure; *Pcyt*: percentage of sequences predicted cytoplasmic; *Pext*: percentage of sequences predicted extra-cellular; *Pnuc*: percentage of sequences predicted nuclear; *Pmit*: percentage of sequences predicted mitochondrial; *Pother*: percentage of sequences predicted in other localizations; the other localizations include: chloroplast, lysosome, peroxysome, endoplasmic reticulum, vacuoles, Golgi apparatus and periplasm. —; no predictions using this method at this confidence level.

Significant improvement over existing methods for three compartments. Overall, our final prediction systems were significantly more accurate than other publicly available general-purpose methods (Table III, Table IV). The only shortcoming of our methods was the relatively poor performance on mitochondrial proteins for which TargetP[21] was significantly, and PSORT II[45,46,40] notably better (Table III). The difference between the performance of PDB-trained and SWISS-PROT-trained systems on mitochondria indicated that one reason for the poor performance was the lack of data. However, our reference system (*NetSeq* in Table IV) was conceptually similar to NNPSL, nevertheless, our system performed significantly worse for mitochondria. We are currently trying to improve this aspect of our method. The problem with mitochondrial proteins arises at the point of combining the pairwise (L/not-L) networks into a four-state prediction: our pairwise networks for mitochondria/other are reasonably accurate (data not shown).

The performance was slightly over-estimated for some methods. On the sequence-unique set of new SWISS-PROT proteins with experimental annotations of localization, we failed to fully verify the published levels of accuracy for some of the methods (Table IV). In particular, SubLoc[64] was estimated to achieve a level of 79% accuracy, while the method reached only about 58% on our data (Table IV). The difference may be explained by the fact that up to 90% pairwise sequence identity was allowed between testing and training set for the original publication of SubLoc. Cai et al.[76] also claim to reach a very high level of accuracy (73%). However, that value is not easy to compare. First, only identical proteins were excluded in training and

testing set, in other words, it is not clear how many of the proteins used for testing are close sequence homologues to the proteins used for training. Second, Cai et al. did not include mitochondrial proteins, instead they included two other classes, namely plasma membrane and chloroplast. More recently this group published even higher estimates using similar data sets with unspecified sequence similarity between testing and training.[77,78] Given the accuracy of our method, we imagine that it will be a good alternative to some methods for all four compartments that we target, and that it constitutes a reasonable complement for others like TargetP and PSORT II.

All methods not specialized on PDB sequences failed on these. Feeding sequences from the PDB directly into public prediction methods is very problematic. Indeed, our specialist for SWISS-PROT proteins (like all other methods tested, Table III) performed significantly worse for sequences taken from the PDB and vice versa (Table IV). On the one hand, this implies that we better use a specialist system when we want to predict sub-cellular localization for proteins of known structure. On the other hand, this result may also suggest that performance for sequences from public genome sequencing efforts may also be reduced, as these may differ significantly from well-characterized, functional sub-units of proteins as deposited in SWISS-PROT. However, we currently have no handle on estimating such a potential error rate.

PDB annotations not representative for entire proteomes. For the majority of eukaryotic proteins in PDB (57%) the sub-cellular localization could be annotated with 100% accuracy through an appropriate parsing of the data and our automated text-analysis program.[17] This left us with 3778 proteins for which we could not annotate localization without errors. For the majority of these (59% = 2219 proteins), sub-cellular localization could be inferred most accurately through (1) known nuclear localization signals (NLS)[22,75] (2) through homology to proteins of experimentally known localization,[15] and (3) through text-analysis of SWISS-PROT keywords taken from homologues.[17] This large proportion of proteins for which localization can be annotated through homology (total of 82%, Table V) is due to the significant amounts of experimental knowledge available for proteins of known structure. Previously, we annotated about 25% of the proteins in six entirely sequenced eukaryotes (human, mouse, fly, worm, weed, and yeast) through either sequence homology, text analysis of keywords, or sequence motifs.[17,13,79] Thus, the number of proteins for which the system introduced here increases the number of annotations will be much higher for entirely sequenced organisms than it was for the proteins of known structure. For PDB, our new method predicted the compartment for 1561 proteins; about 24% of these (382) were predicted at a reliability corresponding to $> 80\%$ prediction accuracy (data not shown). Another aspect of the bias of PDB was the result that over 40% of the eukaryotic proteins of known structure appeared to be extra-cellular. For entire genomes, this number has previously been estimated to be at most half this size.[9,32]

Next step: annotate larger sequence databases and entire proteomes. In future work, we intend to investigate to which extent the GeneOntology database (GO)[80] adds experimental annotations about localization that are not in SWISS-PROT. Such information would be extremely valuable since neural networks are at their best when applied to large data sets. Support vector machines appear to perform better for small data sets.[81] Therefore, we intend to also apply these algorithms to the problem. Finally, preliminary results suggested that it might be possible to increase prediction accuracy, by explicitly incorporating predictions from other methods such as SignalP/TargetP[19,21] or PredictNLS[22] into our neural networks.

Good enough for annotating proteomes and for structural genomics? Our results for the SWISS-PROT based system might be valid for proteins from genome sequencing projects: about 64% of all proteins were correctly by the profile-based networks using predicted surface and secondary structure. Although, we hope to further improve this level of accuracy, we challenge that the predictions are already good enough to become useful in the context of target selection for structural genomics[33] and to bridge the sequence-annotation gap in entirely-sequenced genomes[82–85,2,86,32,3]

ACKNOWLEDGMENTS

Thanks to Jinfeng Liu (Columbia University) for computer assistance and the collection of genome data sets and to Kazimierz Wrzeszczynski (Columbia University) for proofreading the manuscript. Thanks to Astrid Reinhardt (Baylor College of Medicine, Texas), Tim Hubbard (Sanger Centre, Hinxton), Sujun Hua (Tsinghua University), Zhirong Hun (Tsinghua University), Olof Emanuelsson (Stockholm University), Henrik Nielsen (CBS, Copenhagen), Søren Brunak (CBS, Copenhagen) and Gunnar von Heijne (Stockholm University) for access to their prediction methods. Thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (UCSD), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

REFERENCES

1. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev of Biophys Biomol Struct 1996;25: 113–136.
2. Koonin EV. Bridging the gap between sequence and function. Trends Genet 2000;16:16.
3. Rost B. Did evolution leap to create the protein universe? Curr Opin Struct Biol 2002;12:409–416.
4. Kumar A, et al. Subcellular localization of the yeast proteome. Genes Dev 2002;16:707–19.
5. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823–826.
6. Lewis S, Ashburner M, Reese MG. Annotating eukaryote genomes. Curr Opin Struct Biol 2000;10:349–354.
7. Eisenhaber F, Bork P. Wanted: subcellular localization of proteins based on sequence. Trends Cell Biol 1998;8:169–170.
8. Nakai K. Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 2000;54:277–344.
9. Emanuelsson O, von Heijne G. Prediction of organellar targeting signals. Biochim Biophys Acta 2001;1541:114–119.
10. Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol 1998;276:517–525.
11. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. J Mol Biol 1998;283:707–725.
12. Mott R, Schultz J, Bork P, Ponting CP. Predicting protein cellular

localization using a domain projection method. Genome Res 2002;12:1168–1174.
13. Nair R, Rost, B. Sequence conserved for sub-cellular localization. Protein Sci 2002;11:2836–2847.
14. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
15. Nair R, Rost B. Sequence conserved for sub-cellular localization. Protein Sci 2002;11:2836–2847.
16. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
17. Nair R, Rost B. Inferring sub-cellular localisation through automated lexical analysis. Bioinformatics 2002;18;Suppl 1:S78–S86.
18. von Heijne G. Protein sorting signals: simple peptides with complex functions. Exs 1995;73:67–76.
19. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8:581–599.
20. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
21. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005–1016.
22. Cokol M, Nair R, Rost B. Finding nuclear localisation signals. EMBO Reports 2000;1:411–415.
23. Hodel MR, Corbett AH, Hodel AE. Dissection of a nuclear localization signal. J Biol Chem 2001;276:1317–1325.
24. Nakai K. Prediction of in vivo fates of proteins in the era of genomics and proteomics. J Struct Biol 2001;134:103–116.
25. Rusch SL, Kendall DA. Protein transport via amino-terminal targeting sequences: common themes in diverse systems. Mol Membr Biol 1995;12:295–307.
26. Schatz G, Dobberstein B. Common principles of protein translocation across membranes. Science 1996;271:1519–1526.
27. Claros MG, Vincens P. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 1996;241:779–786.
28. Mattaj IW, Englmeier L. Nucleocytoplasmic transport: the soluble phase. Annu Rev Biochem 1998;67:265–306.
29. Weis K. Importins and exportins: how to get in and out of the nucleus. Trends Biochem Sci 1998;23:185–189.
30. La Cour T, Gupta R, Rapacki K, Skriver K, Poulsen FM, Brunak S. NESbase version 1.0: a database of nuclear export signals. Nucleic Acids Res 2003;31:393–396.
31. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26: 2230–2236.
32. Liu J, Rost, B. Comparing function and structure between entire proteomes. Protein Sci 2001;10:1970–1979.
33. Liu J, Rost B. Target space for structural genomics revisited. Bioinformatics 2002;18:922–933.
34. Nishikawa K, Kubota Y, Ooi T. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J Biochem (Tokyo) 1983;94:981–995.
35. Nishikawa K, Kubota Y, Ooi T. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. J Biochem (Tokyo) 1983;94:997–1007.
36. Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 1994;238:54–61.
37. Horton P, Nakai K. A probabilistic classification system for predicting the cellular localization sites of proteins. In: States D, Agarwal P, Gaasterland T, Hunter L, Smith RF, editors. Fourth International Conference on Intelligent Systems for Molecular Biology. St. Louis: AAAI Press 1996:109–115.
38. Cedano J, Aloy P, Perez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
39. Horton P, Nakai K. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In: Valencia A, editor. Fifth International Conference on Intelligent Systems for Molecular Biology. Halkidiki, Greece: AAAI Press 1997:147–152.
40. Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999;24:34–36.
41. Yuan Z. Prediction of protein subcellular locations using Markov chain models. FEBS Lett 1999;451:23–26.
42. Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 2000;278:477–483.
43. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43:246–255.
44. Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D. Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci U S A 2000;97:12115–12120.
45. Nakai K, Kidera A, Kanehisa M. Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng 1988;2:93–100.
46. Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 1992;14:897–911.
47. Drawid A, Gerstein MA. Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J Mol Biol 2000;301:1059–1075.
48. Lester DS, Orr N, Brumfeld V. Structural distinction between soluble and particulate protein kinase C species. J Protein Chem 1990;9:209–220.
49. Mahalingam S, Collman RG, Patel M, Monken CE, Srinivasan A. Functional analysis of HIV-1 Vpr: identification of determinants essential for subcellular localization. Virology 1995;212:331–239.
50. Kamata M, Aida Y. Two putative alpha-helical domains of human immunodeficiency virus type 1 Vpr mediate nuclear localization by at least two mechanisms. J Virol 2000;74:7179–7186.
51. Vorberg I, Chan K, Priola SA. Deletion of beta-strand and alpha-helix secondary structure in normal prion protein inhibits formation of its protease-resistant isoform. J Virol 2001;75:10024–10032.
52. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proceedings of the National Academy of Sciences 1993;90:7558–7562.
53. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
54. Rost, B. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996;266:525–539.
55. Rost B. Enzyme function less conserved than anticipated. Journal of Molecular Biology 2002;318:595–608.
56. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80:319–24.
57. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
58. Nielsen H, Engelbrecht J, von Heijne G, Brunak S. Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins 1996;24:165–177.
59. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
60. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
61. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
62. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
63. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
64. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17:721–728.
65. Rost B, Sander C. Jury returns on structure prediction. Nature 1992;360:540.
66. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
67. Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. J Mol Biol 1997;268:31–36.
68. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
69. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R.

Predicting protein structure using only sequence information. Proteins 1999;S3:121–125.
70. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
71. Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 2000;9:1162–1176.
72. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 2002;49:154–166.
73. Eyrich V, Martó-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17: 1242–1243.
74. Przybylski D, Rost B. Profile-profile alignments using predicted structure. 2002;46:197–205.
75. Nair R, Carter P, Rost B. NLSdb: database of nuclear localization signals. Nucleic Acids Res 2003;322:397–399.
76. Cai YD, Chou KC. Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol Cell Biol Res Commun 2000;4:172–173.
77. Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 2002;84:343–348.
78. Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002;277:45765–45769.
79. Nair R, Rost B. LOC3D: annotate sub-cellular localization for protein structures. Nucleic Acids Res 2003;31:3337–3340.
80. Ashburner M, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25:25–29.
81. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995; 314 pages.
82. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A. EUCLID: automatic classification of proteins in functional classes by their database annotations. Bioinformatics 1998;14:542–543.
83. Ashburner M. A biologist's view of the Drosophila genome annotation assessment project. Genome Res 2000;10:391–393.
84. Gaasterland T, Sczyrba A, Thomas E, Aytekin-Kurban G, Gordon P, Sensen CW. MAGPIE/EGRET annotation of the 2.9-Mb Drosophila melanogaster Adh region. Genome Res 2000;10:502–510.
85. Gerstein M. Annotation of the human genome. Science 2000;288: 1590.
86. Koonin EV. Computational genomics. Curr Biol 2001;11:R155–R158.
87. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 1999;8:978–984.
88. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
89. Boeckmann B, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.

APPENDIX

Materials and Methods for "Supporting Online Material"

Linear analysis of composition vectors.

A principal component analysis (PCA) was performed on the test proteins to determine whether the data set clusters according to subcellular localization group. The total composition vector, c_i, for a protein i is defined as the row vector $c_i = \{c_{ij}\}$, where $j = 1, \ldots, 20$ indicates the amino acid type. The composition of the j^{th} amino acid, c_{ij}, is defined as

$$c_{ij} = r_{ij} \Big/ \sum_{j=1}^{20} r_{ij}$$

where r_{ij} is the number of residues of amino acid type j in protein i. The surface composition vectors were similarly calculated, with the r_{ij} now representing the number of residues of type j at the surface of the protein. We used these composition vectors to define a sample variance-covariance matrix, **S,** as follows:

$$\mathbf{S} = \{s_{jk}\} = \left\{ \sum_{i=1}^{n} (c_{ij} - c_j)(c_{ik} - c_k) \right\}$$

where,

$$cj = \frac{1}{n} \sum_{i=1}^{n} c_{ij}$$

is the average composition of the jth amino acid type over the n proteins in the data set. The principal components of the set of composition vectors are then the eigenvectors of **S** (e.g., see Anderberg, 1973). The composition vector for each protein was then projected onto the plane defined by the first two principal components using the standard inner product. This provides a two dimensional representation of the clustering of component vectors as shown in Figure 2.

Results for "Supporting Online Material"

Linear separation by principal component analysis not enough.

The overall amino acid composition (Fig. 5A), the surface composition (Fig. 5B), the three-state secondary structure composition (Fig. 5C), and the combined sequence-structure composition (Fig. 5D) all showed some correlation with sub-cellular localization in two dimensions (the first two principal components). However, in contrast to previous studies,[10] we could not fully discriminate between the three major classes by a linear separation on any single feature. The full principal component analysis (Methods) revealed that the Eigen-values of the first eight principal components were of similar magnitude for overall amino acid composition. Hence, projecting the composition vectors onto two dimensions resulted in a considerable loss of information. In order to fully resolve the signal for sub-cellular localization present in the different global composition features, we implemented a neural network based machine-learning algorithm.

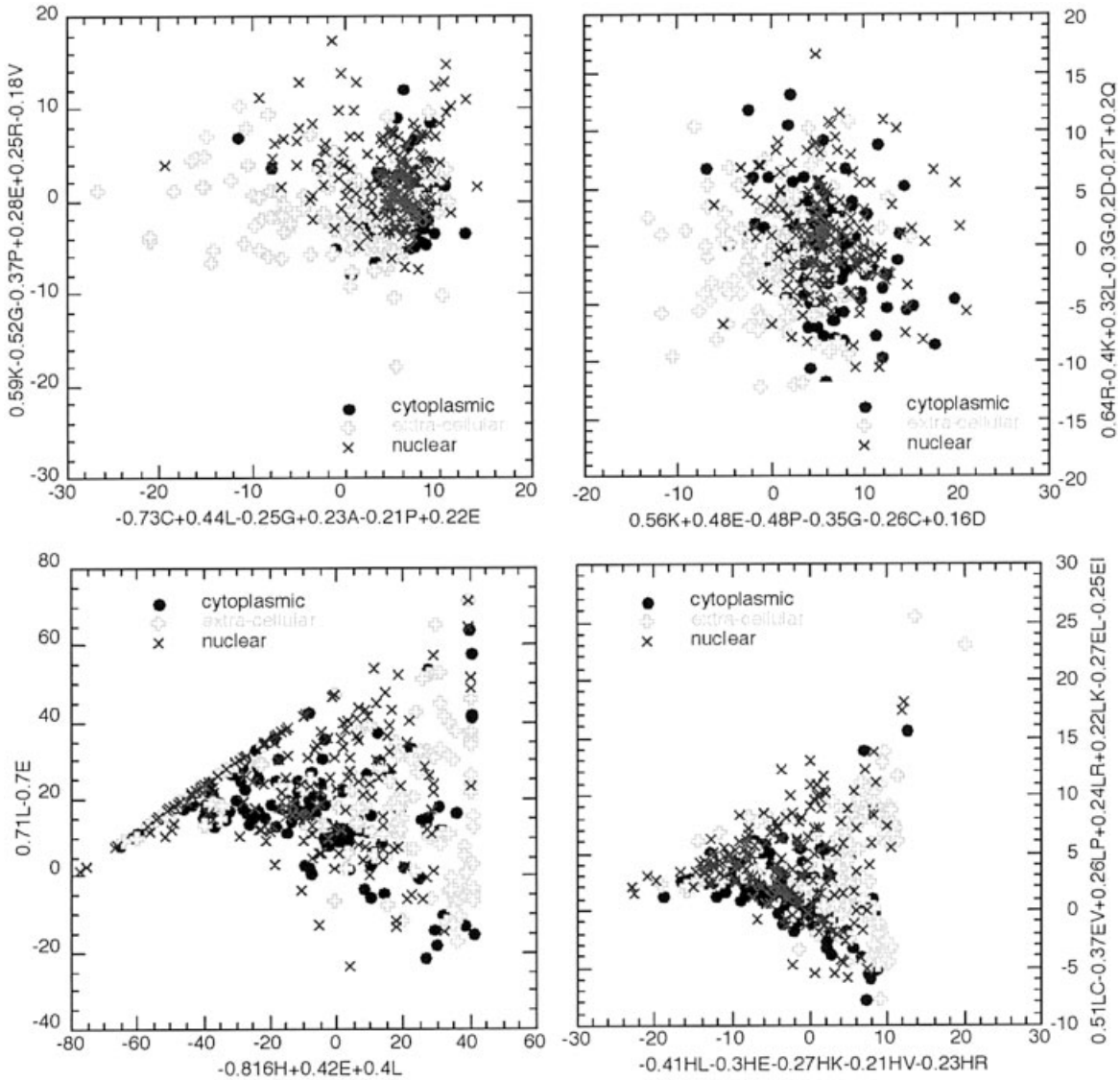

Fig. 5A. Nair & Rost: Fig. 5: Maximal linear separation of sub-cellular localization. Given are the projections onto the first two principal components of: (**A**) the overall amino acid composition vectors, (**B**) the surface composition vectors (from DSSP), (**C**) the three state secondary structure composition vectors (from DSSP) and (**D**) the product sequence-structure composition vectors (from DSSP) for the proteins in the test set. For all four features the composition vectors have been projected onto the plane defined by the first two principal components, respectively represented by the x and y-axis. The axis labels indicate the amino acid/secondary structure types that contribute most significantly to the two principal components. The extra-cellular class (open plusses) is the best resolved and the cytoplasmic class (shaded circles) the worst resolved for all four features.

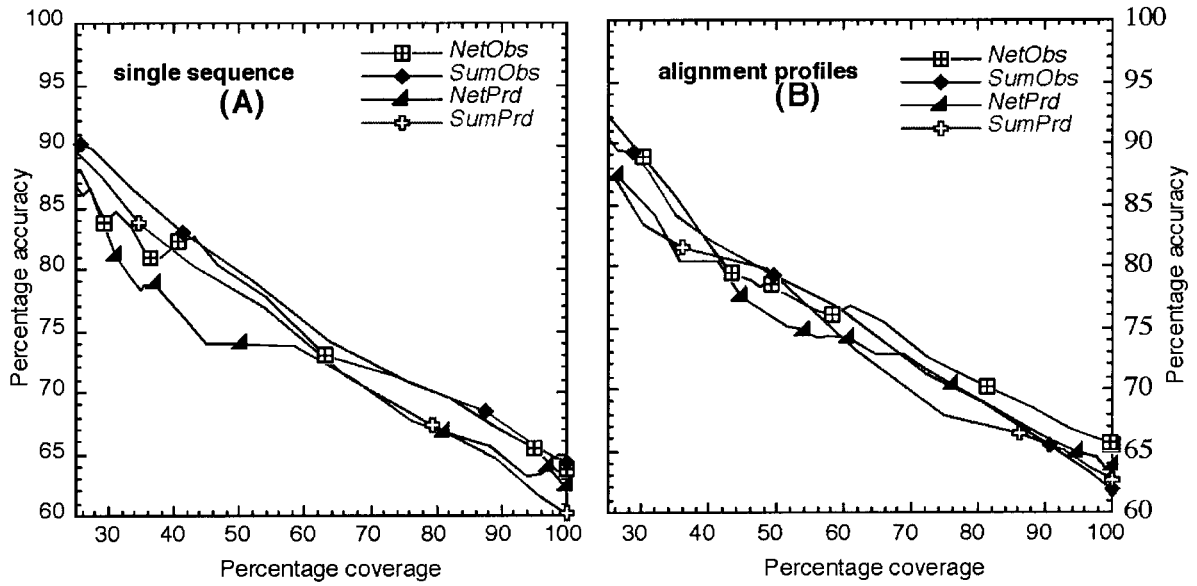

Fig. 6A. Better prediction through combining neural networks. Combining the various sources of information (amino acid composition, surface composition and amino acid composition separated into the three secondary structure states) yielded by far the best results. Prediction accuracy increased by up to three percentage points over simple first-level networks. The different sources of information were combined in two different ways; using a statistical jury decision on predictions from the first-level networks (marked *Sum* in figure), and using predictions from the first level networks as input to a second level neural network (*Net* in figure). (**A**) For the single networks (i.e., using only single sequences and no evolutionary information), combining the networks in a simple jury (*SumObs* and *SumPrd*) performed as well as the neural network combinations (*NetObs* and *NetPrd*). Here, *Obs* and *Prd* represent networks based on the observed and predicted surface and secondary structure of the protein respectively. The standard error in prediction accuracy was approximately $\pm$ 0.25% points. (**B**) For profile-based networks using evolutionary information, the second-level neural network combinations performed best. For the combination networks, using profiles rather than single sequences as input improved prediction accuracy by up to 2%. Profile based *NetObs* (*the final LOC3DnetObs system*) networks gave the best overall localization prediction (accuracy over 65%).

REFERENCES FOR "SUPPORTING ONLINE MATERIAL"

1. Anderberg MR. Cluster analysis for applications. Academic Press: New York, 1973.