

# Surface profiles predict sub-cellular localisation

Rajesh Nair & Burkhard Rost <sup>□</sup>

CUBIC, Columbia University <sup>§</sup>

## Abstract

The gap between the number of known protein sequence and the knowledge about protein function is rapidly increasing. One important physical aspect of function is the sub-cellular localisation of a protein. Here, we trained two-layered feed-forward neural networks to predict the sub-cellular localisation for proteins of known structure. We introduced two novel key aspects: (1) using evolutionary information, and (2) using surface composition. We also trained networks only on the N-terms. Finally, we combined all our networks. We evaluated sustained levels of performance by four-fold cross-validation. The major single source of improvement was the use of evolutionary information. However, the combination of our various networks yielded the final, significant improvement over previous methods. The final system reached an accuracy above 80% (two-state). This level may suffice to make the method valuable for target selection in structural genomics.

**Keywords:** protein sub-cellular localization, surface composition, sequence motifs, evolutionary profiles, neural network

## Introduction

*Gap between protein sequences and function.* The number of completely sequenced genomes has been rapidly increasing. Currently, we know the full genomes for more than 40 organisms (Fleischmann, et al. 1995, Goffeau, et al. 1996, Blattner, et al. 1997, The C. elegans Sequencing Consortium 1998, Adams, et al. 2000, Domingues, et al. 2000). This sequence explosion has widened the gap between the number of sequences deposited in public databases and the experimental characterisation of the corresponding proteins (Koonin 2000). Computational tools for predicting protein function can help bridge this gap (Eisenberg, et al. 2000, Lewis, et al. 2000). One crucial step toward understanding the function of a protein is elucidating its sub-cellular localisation (Eisenhaber & Bork 1998).

*Prediction by homology relatively accurate but not often applicable.* A variety of approaches have been used to classify proteins with respect to sub-cellular localisation. One of the most reliable ones is annotation transfer from a homologous protein (Bork,

et al. 1998). If a protein with known localisation is sufficiently similar in sequence to a query protein U, the sub-cellular localisation can be inferred for U. However, the level of 'significant sequence identity' varies substantially between localisations, and is much higher than that required for conservation of structure (Andrade, et al. 1998). Less than 22% of the proteins in SWISS-PROT (Bairoch & Apweiler 2000b) can be classified by homology into one of the three classes: intra-cellular, extracellular and membrane associated (Eisenhaber & Bork 1998).

*Prediction through sequence motifs works for some localisations.* Another method to predict localisation is the identification of sequence motifs, namely; signal peptides, nuclear localisation signals (NLS) and other sorting signals. Proteins destined for the secretory pathway, mitochondria and the chloroplast contain an N-terminal targeting peptide that is recognised by the translocation machinery (Rusch & Kendall 1995, Schatz & Dobberstein 1996). A number of neural network-based tools based only on the N-terminal amino acid sequence have been developed (Nielsen, et al. 1997, Emanuelsson, et al. 2000). Discriminant analysis has been applied to identify proteins imported into mitochondria (Claros & Vincens 1996). Proteins destined for the nucleus usually contain NLS motifs (Mattaj & Englmeier 1998, Weis 1998). Recently, we have collected a data set of experimentally known NLS motifs (Cokol, et al. 2000). However, most proteins do not have signal peptides. Furthermore, a particular problem for methods detecting N-terminal signal motifs is that start codons are predicted with less than 70% accuracy by genome projects (Reinhardt & Hubbard 1998). With the availability of large numbers of completely sequenced genomes, phylogenetic profiles have been employed to identify sub-cellular localisation (Marcotte, et al. 2000). So far, this approach has been much less accurate than methods based solely on composition. Other methods have tried to integrate rules based on amino acid composition with databases of known signal sequences. PSORT II, a knowledge-based expert system tries to integrate the two kinds of information (Nakai, et al. 1988, Nakai & Kanehisa 1992, Nakai & Horton 1999). Recently a Bayesian system based on a diverse range of 30 features has been proposed (Drawid & Gerstein 2000).

*Prediction from amino acid composition possible.*

---

<sup>□</sup> Corresponding author: rost@columbia.edu, <http://cubic.bioc.columbia.edu/>, Tel: +1-212-305-3773, fax: +1-212-305-7932

<sup>§</sup> Columbia Univ., Dept. of Biochemistry and Molecular Biophysics, 650 West 168<sup>th</sup> Str., New York, NY 10032, USA

A third approach to predicting localisation has been suggested by the observation that the total amino acid composition correlates with the sub-cellular localisation (Nishikawa, et al. 1983a, Nishikawa, et al. 1983b, Nakashima & Nishikawa 1994). This observation has led to the development of a variety of prediction methods based solely on composition (Horton & Nakai 1996, Cedano, et al. 1997, Horton & Nakai 1997, Reinhardt & Hubbard 1998, Nakai & Horton 1999, Yuan 1999). A more detailed study showed that protein surfaces adapt to their physico-chemical environment, so that the signal for sub-cellular localisation is almost entirely due to the surface residues (Andrade, et al. 1998). This observation has not been used to predict localisation, yet.

Here, we describe a system of neural networks using evolutionary profiles and surface composition to predict the major sub-cellular localisations. We train and test the neural networks on PDB proteins with known localisations. In addition to the overall and the surface profiles, we also evaluated the performance of networks trained on N-terminal amino acid compositions. Overall, we found surface residue composition only slightly beneficial. In contrast, evolutionary information improved performance for all localisations. Finally, we discussed the combination of composition-based methods with motif-based methods like TargetP (Emanuelsson, et al. 2000).

## Materials and Methods

### Data sets

We selected all proteins with clearly annotated sub-cellular localisation in SWISS-PROT release 37 (Bairoch & Apweiler 2000a). We excluded sequences annotated as "POSSIBLE", "PROBABLE" or "BY SIMILARITY". We also excluded sequences annotated as being found in more than one localisation (numbers given in Table 1). Then, we assigned localisation to PDB proteins (Bernstein, et al. 1977) by searching for homologues in SWISS-PROT. The criterion for determining homologues was: > 50% pairwise sequence identity over more than 60% of the protein length (numbers of PDB proteins found given in Table 1). Training, test and validation sets were constructed such that no pair of proteins from any two sets had significant sequence similarity. We reduced the redundancy of the test set through a simple greedy-search using our threshold for 'sequence similarity implying structural similarity' (Rost 1999). No two proteins in the test set had more than > 30% sequence identity over more than 100 residues (numbers of unique sets given in Table 1).

Exposed residue composition was calculated from the solvent accessible surface area (Connolly 1983) in the DSSP database (Kabsch & Sander 1983). We

**Table 1: Number of proteins in data set.**

Sub-cellular Localisation	SWISS <sup>a</sup>	PDB <sup>b</sup>	Unique <sup>c</sup>
Extra-cellular	576	508	38
Nuclear	2217	161	65
Cytoplasmic	1987	1768	183
Mitochondrial	659	57	14
Chloroplast	909	154	0
Lysosomal	97	110	15
Peroxisomal	70	33	8
Vacuolar	30	10	3
Periplasmic	2	0	0
Endoplasmic Ret.	112	0	0
Golgi	13	0	0
Total	6672	2801	326

<sup>a</sup> Number of proteins with known localisation from SWISS-PROT; <sup>b</sup> Number of PDB sequences that could be assigned the given location; <sup>c</sup> Number of unique PDB proteins used for testing.

predicted all residues as exposed that were predicted to have a relative solvent accessibility > 16% by PHDacc (Rost 1996). We chose this threshold since it give a good prediction accuracy on a limited subset of the training sets. Profile based composition was calculated by aligning the sequences against the Swissprot database using MaxHom dynamic programming algorithm (Sander & Schneider 1991). Homologues were determined according to the criteria discussed earlier. Henceforth, networks which do not employ profiles will be referred to as single networks (using data from Single protein) to distinguish them from profile networks. Since most PDB proteins have their N-terminal signal peptides cleaved off (Rusch & Kendall 1995), using the functional PDB protein sequence as input to SignalP and to calculate N-terminal amino acid composition can give significantly reduced prediction accuracy. Hence the N-terminal composition was calculated using both the 50 N-terminal residues of the PDB sequence and the 50 N-terminal residues of the closest SWISS-PROT homologue.

### 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 sub-cellular localisation group. The total composition vector,  $c_i$ , for a protein  $i$  is defined as the row vector  $c_i = \{c_{ij}\}$ , where  $j=1, \dots, 20$  indicates the amino acid type. The composition of the  $j^{th}$  amino acid,  $c_{ij}$ , is defined as

$$c_{ij} = r_{ij} / \sum_{j=1}^{20} r_{ij} \quad (1)$$

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,  $\mathbf{S}$ , as follows:

$$\mathbf{S} = \{s_{jk}\} = \left\{ \left( \sum_{i=1}^n (c_{ij} - \bar{c}_j)(c_{ik} - \bar{c}_k) \right) \right\} \quad (2)$$

where:

$$\bar{c}_j = \frac{1}{n} \sum_{i=1}^n c_{ij} \quad (3)$$

is the average composition of the  $j$ th 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  $\mathbf{S}$  (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 Fig1.

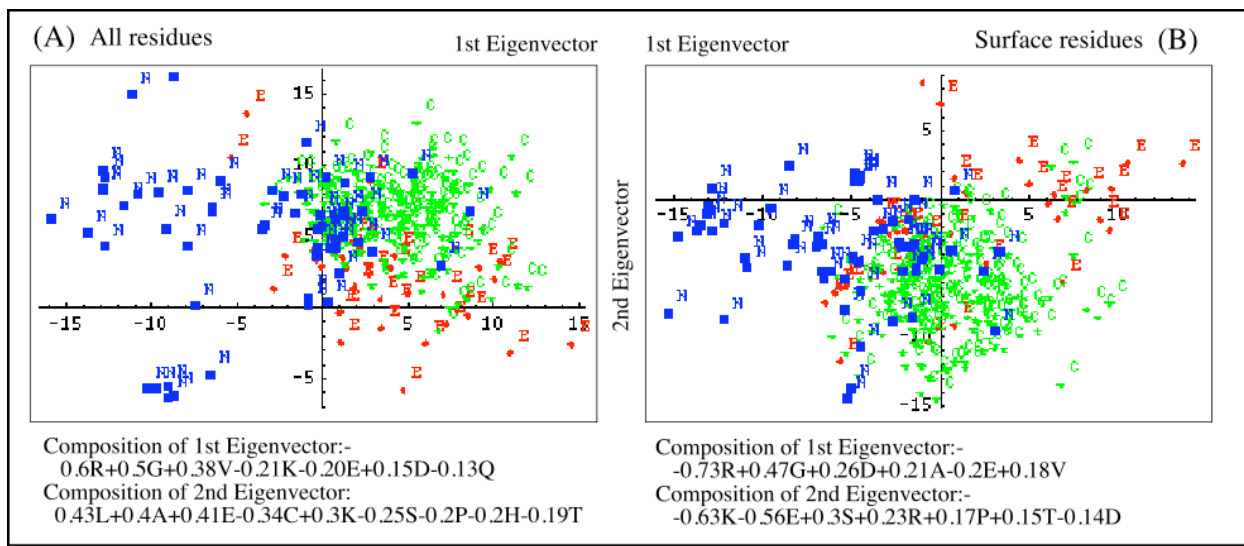
### Neural network architectures

We used two levels of networks. In the first level a feed-forward architecture with 20 input units and one hidden layer and three units was used to train dedicated networks for each localisation. The neural networks were trained on the PDB data sets using full

amino acid composition, exposed amino acid composition and N-terminal amino acid composition as input. A balanced training was performed to prevent bias of the neural networks towards over represented classes. No significant improvement in network performance was gained by increasing the number of hidden nodes. Hence, a network with 3 hidden units was chosen to minimise over-fitting. The second level was an integrating neural network. Outputs from the first level networks trained on full amino acid composition, exposed residue composition and N-terminal amino acid composition were input to the second level. The network architecture in the second level consisted of 12 input units and a hidden layer with 3 nodes.

### Evaluating network performance

Four-fold cross-validation was applied to test the neural networks so trained. As a simple measure for performance we used the two-state accuracy (Q2, number of correctly predicted test proteins as percentage of total number of test proteins). The specificity and sensitivity of the networks were measured using four ratios derived from  $TP$  (number of proteins predicted to be in localisation  $i$  and observed to be in localisation  $i$ , the true positives),  $TN$  (number of proteins predicted not to be in localisation  $i$  and observed to be so, the true negatives),  $FP$  (number of proteins predicted to be in localisation  $i$



**Fig. 1: Maximal linear separation of sub-cellular localisation.** Given are the projections onto the first two principal components of: (A) the overall amino acid composition vectors and (B) the surface composition vectors (from DSSP) for the proteins in the test set. The 20 dimensional 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 types that contribute most significantly to the two principal components. Each vector is represented by a coloured letter, n (blue), c (green) or e (red), indicating the sub-cellular localisation of the protein (nuclear, cytoplasmic, or extracellular, respectively). Both figures show clustering by location class.

**Table 2: Prediction accuracy of first level networks.**

Input <sup>(1)</sup>	Localisation <sup>(2)</sup>	NumProt <sup>(3)</sup>	Q <sup>(4)</sup>	Q2 <sup>(5)</sup>	$\square$ <sup>(6)</sup>	MC <sup>(7)</sup>	MI <sup>(8)</sup>	pL <sup>(9)</sup>	pX <sup>(9)</sup>	oL <sup>(9)</sup>	oX <sup>(9)</sup>
compo <sup>a</sup>	cytoplasmic	317	57	78	1.6	0.55	0.23	93	57	74	86
	extra-cellular	311	12	79	3	0.25	0.07	50	83	29	92
	nuclear	319	19	90	2.4	0.63	0.31	58	97	83	90
exposedDSSP <sup>b</sup>	cytoplasmic	317	57	78	1.2	0.55	0.23	86	68	78	78
	extra-cellular	311	12	84	3.3	0.15	0.02	21	92	28	89
	nuclear	319	19	88	1.7	0.6	0.3	60	95	74	90
exposedPHD <sup>c</sup>	cytoplasmic	317	57	73	1.6	0.45	0.15	85	57	72	75
	extra-cellular	311	12	85	3.7	0.17	0.03	21	93	30	89
	nuclear	319	19	89	1.8	0.63	0.32	64	95	75	91
compoN50pdb <sup>d</sup>	cytoplasmic	317	57	70	3	0.37	0.1	85	49	69	70
	extra-cellular	311	12	85	4.3	0.28	0.08	39	91	38	91
	nuclear	319	19	84	2.3	0.46	0.18	55	90	58	89
compoN50swiss <sup>e</sup>	cytoplasmic	317	57	65	1.3	0.29	0.06	68	61	70	61
	extra-cellular	311	12	78	4.3	0.35	0.15	68	80	31	95
	nuclear	319	19	71	3.4	0.16	0.02	40	78	31	84

<sup>(1)</sup> composition vector input to the networks; <sup>(2)</sup> sub-cellular localisations discriminated; <sup>(3)</sup> number of test proteins; <sup>(4)</sup> Percentage of test proteins belonging to the specific sub-cellular localisation; <sup>(5)</sup> two-state prediction accuracy; <sup>(6)</sup> standard deviation of predicted two state accuracy; <sup>(7)</sup> Matthews correlation coefficient (eq. 5); <sup>(8)</sup> Shannon information; <sup>(9)</sup> pL, pX, oL and oX = as defined in eq. 4; <sup>a</sup> full amino acid composition; <sup>b</sup> surface composition taken from DSSP; <sup>c</sup> surface composition predicted by PHD; <sup>d</sup> N-term amino acid composition from PDB sequence; <sup>e</sup> N-terminal (50 residues) composition from SWISS-PROT homologue.

**Table 3: Prediction accuracy for first level networks with profile input.**

Input <sup>(1)</sup>	Localisation <sup>(2)</sup>	NumProt <sup>(3)</sup>	Q <sup>(4)</sup>	Q2 <sup>(5)</sup>	$\square$ <sup>(6)</sup>	MC <sup>(7)</sup>	MI <sup>(8)</sup>	pL <sup>(9)</sup>	pX <sup>(9)</sup>	oL <sup>(9)</sup>	oX <sup>(9)</sup>
compo <sup>a</sup>	cytoplasmic	317	57	79	1.7	0.59	0.28	95	59	75	90
	extra-cellular	311	12	84	4.4	0.18	0.04	26	91	31	90
	nuclear	319	19	88	2.1	0.64	0.34	71	93	70	93
exposedDSSP <sup>b</sup>	cytoplasmic	317	57	80	2.0	0.6	0.28	91	66	78	84
	extra-cellular	311	12	85	4.5	0.21	0.05	31	91	35	90
	nuclear	319	19	87	2.7	0.55	0.24	56	94	68	89
exposedPHD <sup>c</sup>	cytoplasmic	317	57	75	1.9	0.5	0.19	91	54	73	82
	extra-cellular	311	12	87	2.6	0.35	0.12	39	94	48	91
	nuclear	319	19	89	1.3	0.61	0.3	56	97	83	90
compoN50pdb <sup>d</sup>	cytoplasmic	317	57	70	1.9	0.38	0.11	91	41	67	76
	extra-cellular	311	12	83	4.4	0.22	0.07	39	89	32	91
	nuclear	319	19	86	2.7	0.5	0.21	51	94	68	89
compoN50swiss <sup>e</sup>	cytoplasmic	317	57	73	2.6	0.44	0.14	82	61	73	72
	extra-cellular	311	12	83	1.8	0.36	0.14	60	85	36	94
	nuclear	319	19	75	2.7	0.32	0.1	58	79	40	88

Abbreviations as in Table 2.

**Table 4: Prediction accuracy for final integrated networks.**

Input <sup>(1)</sup>	Localisation <sup>(2)</sup>	NumProt <sup>(3)</sup>	Q <sup>(4)</sup>	Q2 <sup>(5)</sup>	$\square$ <sup>(6)</sup>	MC <sup>(7)</sup>	MI <sup>(8)</sup>	pL <sup>(9)</sup>	pX <sup>(9)</sup>	oL <sup>(9)</sup>	oX <sup>(9)</sup>
singleDSSP <sup>a</sup>	cytoplasmic	317	57	77	1.3	0.53	0.22	91	58	74	83
	extra-cellular	311	12	84	2.0	0.4	0.18	63	87	39	94
	nuclear	319	19	85	1.6	0.57	0.29	74	88	59	93
profileDSSP <sup>b</sup>	cytoplasmic	317	57	80	1.4	0.6	0.28	92	65	78	85
	extra-cellular	311	12	79	1.7	0.4	0.19	76	79	33	95
	nuclear	319	19	88	1.0	0.62	0.34	73	91	67	93
profilePHD <sup>c</sup>	cytoplasmic	317	57	80	1.8	0.58	0.26	90	66	78	85
	extra-cellular	311	12	87	1.9	0.44	0.20	60	90	48	94
	nuclear	319	19	88	1.1	0.65	0.37	77	91	67	94

<sup>a</sup> networks trained on overall + surface (from DSSP) + N50 composition; <sup>b</sup> as previous with profile input; <sup>c</sup> as previous, but with predicted (rather than observed) surface.

and observed not to be in  $i$ , the false positives) and  $FN$  (number of proteins predicted not to be in localisation  $i$  and observed to be in  $i$ , the false negatives). The ratios are:

$$\begin{aligned} pL &= 100 \square \frac{TP}{TP + FP} \\ pX &= 100 \square \frac{TN}{TN + FN} \\ oL &= 100 \square \frac{TP}{TP + FN} \\ oX &= 100 \square \frac{TN}{TN + FP} \end{aligned} \quad (4)$$

We also used the Matthews correlation coefficient (Matthews 1975):

$$MC = \frac{TP \square TN \square FP + FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (5)$$

Finally, we measured the mutual information (Rost, et al. 1993, Baldi, et al. 2000) according to:

$$\begin{aligned} I &= \square \frac{TP}{N} \log\left[\frac{TP}{TP + FP}\right] \square \frac{FP}{N} \log\left[\frac{FP}{TP + FP}\right] \\ &\quad \square \frac{TN}{N} \log\left[\frac{TN}{TN + FN}\right] \square \frac{FN}{N} \log\left[\frac{FN}{TN + FN}\right] \\ H &= \square \frac{TP + FN}{N} \log\left[\frac{TP + FN}{N}\right] \square \frac{TN + FP}{N} \log\left[\frac{TN + FP}{N}\right] \\ MI &= 1 \square \frac{I}{H} \end{aligned} \quad (6)$$

## Results

### Linear separation not sufficient

Both, the overall amino acid composition (Fig. 1A), and the surface composition (Fig. 1B) somehow correlated with localisation. However, the linear separation did not suffice to predict localisation from composition. A full principal component analysis (Methods) showed that the Eigen-values of the first eight principal components are of similar magnitude. Hence, projecting the composition vectors onto two dimensions is associated with considerable loss of information.

### Simple first level networks performed poorly

Our first result repeated earlier work (Reinhardt & Hubbard, 1998): simple neural networks achieve a separation of the major three classes of sub-cellular localisation when using the overall amino acid composition as input (Table 2). However, the relatively high levels of two-state accuracy covered problems with the prediction. For example, the networks predicting extra-cellular proteins achieved an amazing level of 79%, but less than 30% of all proteins observed to be extra-cellular were actually predicted correctly (Table 2). Networks trained on

full composition (*compo* in Table 2) and networks trained on the observed surface composition (*exposedDSSP* in Table 2) performed similarly for cytoplasmic and nuclear proteins. In contrast, for extra-cellular proteins, the networks employing full composition performed better. Using the closest SWISS-PROT homologue to calculate the N-terminal composition gave significantly better results than using the PDB sequence for extra-cellular proteins. This was expected, since the N-terminal signal peptide has been cleaved in most PDB proteins.

### Profile-based first level networks did better

Using evolutionary information from sequence profiles improved prediction accuracy significantly (Table 3). We found the gain maximal for cytoplasmic proteins: The Matthews-correlation (MC) increased from about 0.55 to about 0.59. Similarly, networks trained on surface composition (*exposedDSSP* in Table 3) increased the Matthews-correlation from about 0.55 to about 0.60. The schemes exploring the terminal composition also raised accuracy. In contrast, for extra-cellular and nuclear networks profiles did not improve accuracy significantly.

### Profile-based surface composition did better, even

While replacing overall composition by surface composition did not improve the single sequence networks, this transition helped when using profiles as input. All networks trained on surface composition profiles performed consistently better than networks trained on overall composition profiles. When we replaced the observed surface profiles by those predicted with PHD (*exposedPHD* in Table 3), prediction accuracy dropped slightly for cytoplasmic networks, while for nuclear and extra-cellular networks predicted and observed surface profiles yielded similar results (Table 3).

### Combined networks significantly best

We obtained by far the best results for each of the three localisations when combining all previously developed networks (Table 4). Amongst the combined networks, using profiles consistently improved over using single sequences. For example, for cytoplasmic proteins (most populated class) the Matthews-correlation increased from 0.53 (single sequences) to 0.60 (profiles with observed surface), respectively to 0.58 (profiles with predicted surface). Again, the results were surprisingly similar for observed and predicted surface compositions. Interestingly, the combined networks varied much less in prediction accuracy with choice of test set than the single networks (standard deviation in Tables 2-4).

**Table 5: False positives for 6 first-level networks.**

Loc <sup>(1)</sup>	NumProt <sup>(2)</sup>	Compo <sup>(3)</sup>		exposedDSSP <sup>(4)</sup>	
		%pC <sup>(5)</sup>	%pX <sup>(6)</sup>	%pC <sup>(5)</sup>	%pX <sup>(6)</sup>
Cytoplasmic networks					
ext	38	32	68	24	76
nuc	59	36	64	27	73
mit	14	86	14	64	36
lys	14	21	79	21	79
vac	3	0	100	33	67
Extra-cellular networks					
nuc	53	15	85	23	77
cyt	181	3	97	4	96
mit	14	0	100	0	100
lys	14	36	64	7	93
vac	3	67	33	67	33
Nuclear networks					
ext	38	21	79	11	89
cyt	180	4	96	4	96
mit	14	21	79	21	79
lys	14	0	100	0	100
vac	3	0	100	0	100

<sup>(1)</sup> sub-cellular localisation of proteins contributing to false positives; <sup>(2)</sup> number of test proteins belonging to this location; <sup>(3)</sup> input = overall composition; <sup>(4)</sup> input = surface composition (from DSSP); <sup>(5)</sup> percentage of proteins wrongly classified false positives); <sup>(6)</sup> percentage of proteins correctly classified (true negatives).

### False positives dominated by 'protein highways'

For cytoplasmic and nuclear proteins the major contribution to the false positives came from mitochondrial proteins (Table 5). In contrast, for extra-cellular proteins the largest contribution to the false positives came from vacuolar and lysosomal proteins. The sub-cellular location of eukaryotic proteins is determined by a trafficking system that is reasonably well understood (Pfeffer & Rothman 1987). The system has two main branches that divide at the first stage of protein synthesis on the ribosomes. On one branch, proteins are synthesised in the cytoplasm, and from there can go on to the nucleus, the mitochondria or to the peroxisomes. The second branch leads to the endoplasmic reticulum, then to the Golgi apparatus, and from there to lysosomes, secretory vesicles, or to the cell surface. The false positive contributions showed a similar distribution: proteins of one trafficking branch were the major source of false positives for networks within that branch. Hence, our data showed that proteins falling onto the same branch of the trafficking system have similar compositional properties.

## Discussion

### All results qualitative rather than quantitative

The major problem in analysing our results was the limited size of the data sets available. In particular, our data sets contained only very few extra-cellular proteins (Table 1). Nevertheless, we could clearly single out the following trends. (1) Using only single sequence composition yielded seemingly high levels of two-state accuracy (as reported previously). However, in detail those predictions were rather poor (low correlation indices, in particular for extra-cellular proteins, Table 2). (2) Replacing overall composition by surface composition alone did not improve performance. (3) Replacing single-sequence composition by profile-composition improved all networks. (4) Integrating all sources of information (evolutionary information with overall, surface, and N-term compositions) yielded by far the best method. Hence, all sources were crucial in combination.

### Similar results for proteins of unknown structure?

Here, we focused - in training and testing - on proteins of known localisation and known structure. The database of protein structures is biased. Hence, will the results hold for proteins of unknown structure? Encouragingly, prediction accuracy hardly differed between using observed surface composition (DSSP) and using predicted surface composition (PHD). Hence, will prediction accuracy hold for applying the final system to entire proteomes? This remains to be investigated.

### Future work: extension to SWISS-PROT data sets

In future work, we intend to apply our findings to a much larger data set of localization extracted from SWISS-PROT. Furthermore, preliminary results suggested that it may be possible to increase accuracy, by explicitly incorporating predictions from SignalP/TargetP (Nielsen, et al. 1997, Emanuelsson, et al. 2000) or PredictNLS (Cokol, et al. 2000) into our neural networks. In one implementation, we used SignalP output as the 13<sup>th</sup> input to our second level networks (data not shown). Results from preliminary investigations have been encouraging, but further research needs to be done.

### Good enough for structural genomics / proteomes?

Will we find prediction accuracy decrease or to increase with respect to the values reported here when training on SWISS-PROT proteins rather than on PDB proteins? On the one hand, we anticipate predictions of exposed residues to be less accurate for SWISS-PROT than for PDB proteins. On the other hand, we realised in the detailed behaviour of the neural networks that the training sets were very small. Hence, we expect a significant improvement from training on the SWISS-PROT data. Are predictions

already accurate enough to apply the system to whole proteomes, in particular in the context of structural genomics? 90% of all proteins predicted as cytoplasmic were predicted correctly, and 77% of all proteins predicted as nuclear were correct (Table 4). Nevertheless, we hope to improve prior to applying the system for large-scale annotations of whole proteomes.

**Acknowledgements:** Thanks to Jinfeng Liu (Columbia Univ.) for computer assistance, to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego Univ.), and their crews for maintaining excellent databases. Last, not least, thanks to all those who enabled this analysis by making their experimental data available.

## References

- Adams, M. D., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185-2195.
- Andrade, M. A.; O'Donoghue, S. I. and Rost, B. 1998. Adaptation of protein surfaces to subcellular location. *J Mol Biol* 276:517-25.
- Bairoch, A. & Apweiler, R. 2000a. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28:45-8.
- Bairoch, A. & Apweiler, R. 2000b. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28:45-48.
- Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. and Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16:412-24.
- Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T. and Tasumi, M. 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* 80:319-24.
- Blattner, F. R., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453-1474.
- Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M. and Yuan, Y. 1998. Predicting function: from genes to genomes and back. *J Mol Biol* 283:707-25.
- Cedano, J.; Aloy, P.; Perez-Pons, J. A. and Querol, E. 1997. Relation between amino acid composition and cellular location of proteins. *J Mol Biol* 266:594-600.
- Claros, M. G. & Vincens, P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* 241:779-86.
- Cokol, M.; Nair, R. and Rost, B. 2000. Finding nuclear localisation signals. *EMBO Reports* 1:411-415.
- Connolly, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221:709-13.
- Domingues, F. S.; Koppensteiner, W. A. and Sippl, M. J. 2000. The role of protein structure in genomics. *FEBS Lett* 476:98-102.
- Drawid, A. & Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* 301:1059-75.
- Eisenberg, D.; Marcotte, E. M.; Xenarios, I. and Yeates, T. O. 2000. Protein function in the post-genomic era. *Nature* 405:823-6.
- Eisenhaber, F. & Bork, P. 1998. Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol* 8:169-70.
- Emanuelsson, O.; Nielsen, H.; Brunak, S. and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300:1005-16.
- Fleischmann, R. D., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496-512.
- Goffeau, A., et al. 1996. Life with 6000 genes. *Science* 274:546-567.
- Horton, P. & Nakai, K. 1996. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Fourth International Conference on Intelligent Systems for Molecular Biology*, 109-115. St. Louis, M.O., U.S.A.: AAAI Press.
- Horton, P. & Nakai, K. 1997. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In *Fifth International Conference on Intelligent Systems for Molecular Biology*, 147-152. Halkidiki, Greece: AAAI Press.
- Kabsch, W. & Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-637.
- Koonin, E. V. 2000. Bridging the gap between sequence and function. *Trends Genet* 16:16.
- Lewis, S.; Ashburner, M. and Reese, M. G. 2000. Annotating eukaryote genomes. *Curr Opin Struct Biol* 10:349-54.
- Marcotte, E. M.; Xenarios, I.; van Der Blik, A. M. and Eisenberg, D. 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A* 97:12115-20.
- Mattaj, I. W. & Englmeier, L. 1998. Nucleocytoplasmic transport: the soluble phase. *Annu Rev Biochem* 67:265-306.
- Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405:442-51.
- Nakai, K. & Horton, P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* 24:34-6.
- Nakai, K. & Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14:897-911.
- Nakai, K.; Kidera, A. and Kanehisa, M. 1988. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng* 2:93-100.
- Nakashima, H. & Nishikawa, K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol* 238:54-61.
- Nielsen, H.; Engelbrecht, J.; Brunak, S. and von Heijne, G. 1997. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* 8:581-99.
- Nishikawa, K.; Kubota, Y. and Ooi, T. 1983a. Classification of proteins into groups based on amino

- acid composition and other characters. I. Angular distribution. *J Biochem (Tokyo)* 94:981-95.
- Nishikawa, K.; Kubota, Y. and Ooi, T. 1983b. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *J Biochem (Tokyo)* 94:997-1007.
- Pfeffer, S. R. & Rothman, J. E. 1987. Biosynthetic protein transport and sorting by the endoplasmic reticulum and Golgi. *Annu Rev Biochem* 56:829-52.
- Reinhardt, A. & Hubbard, T. 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res* 26:2230-6.
- Rost, B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* 266:525-39.
- Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* 12:85-94.
- Rost, B.; Schneider, R. and Sander, C. 1993. Progress in protein structure prediction? *Trends Biochem Sci* 18:120-3.
- Rusch, S. L. & Kendall, D. A. 1995. Protein transport via amino-terminal targeting sequences: common themes in diverse systems. *Mol Membr Biol* 12:295-307.
- Sander, C. & Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56-68.
- Schatz, G. & Dobberstein, B. 1996. Common principles of protein translocation across membranes. *Science* 271:1519-26.
- The *C. elegans* Sequencing Consortium 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012-2018.
- Weis, K. 1998. Importins and exportins: how to get in and out of the nucleus. *Trends Biochem Sci* 23:185-9.
- Yuan, Z. 1999. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett* 451:23-6.