2230–2236 *Nucleic Acids Research, 1998, Vol. 26, No. 9* © 1998 Oxford University Press

Using neural networks for prediction of the subcellular location of proteins

A. Reinhardt* and T. Hubbard

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

Received October 16, 1997; Revised and Accepted March 9, 1998

ABSTRACT

Neural networks have been trained to predict the subcellular location of proteins in prokaryotic or eukaryotic cells from their amino acid composition. For three possible subcellular locations in prokaryotic organisms a prediction accuracy of 81% can be achieved. Assigning a reliability index, 33% of the predictions can be made with an accuracy of 91%. For eukaryotic proteins (excluding plant sequences) an overall prediction accuracy of 66% for four locations was achieved, with 33% of the sequences being predicted with an accuracy of 82% or better. With the subcellular location restricting a protein's possible function, this method should be a useful tool for the systematic analysis of genome data and is available via a server on the world wide web.

INTRODUCTION

Within the last couple of years the complete sequence has been determined for a number of genomes (1,2). This has created the need for fully automated methods to analyse the vast amount of sequence data now available. The assignment of a function for a given protein has proved to be especially difficult where no clear homology to proteins of known function exists (3). Knowing the subcellular location that a protein resides in may give important insights as to its possible function, making an automated method that assigns proteins to a certain subcellular location a useful tool for analysis. For example, a strong location prediction may help to distinguish between a number of alternative functional predictions for a protein. Even when the basic function of a protein is known, knowing its location in the cell may give insights as to which pathway an enzyme is part of. As previous studies have shown (4), intra- and extracellular proteins differ significantly in their amino acid composition and these differences are strong enough to be used as the basis for a prediction method. However, to be useful for genome analysis a larger number of subcellular locations need to be distinguished.

This study examines whether the differences in amino acid composition between other subcellular locations is strong enough to establish a prediction method. As yet only two automatic methods for assignment of the subcellular location are publicly available. One of these (5) does not distinguish intracellular proteins as cytoplasmic or mitochondrial and handles eukaryotic and prokaryotic sequences together, while the other is based on an expert system, strongly relying on the existence of targeting or leader sequences (6,7). In large genome analysis projects genes are usually automatically assigned and these assignments are often unreliable for the 5′-regions. For *Caenorhabditis elegans*, for example, automatic assignment methods alone predict <70% of the start codons correctly (S.J.M.Jones, personal communication). This can lead to leader sequences being missing or only partially included, thereby causing problems for prediction algorithms depending on them. A method based on the amino acid composition should be comparatively stable to this sort of ambiguous assignment.

Initial trials using standard statistical methods for prediction (e.g. Mahalanobis distance; 8) did not yield satisfying results, as cross-validation showed a large variation in prediction accuracy. This method has previously been shown to be sensitive to noise within the data set (9). Neural networks on the other hand have been shown to be reliable tools for protein structural prediction purposes (10), so it was decided to apply them in this study.

MATERIALS AND METHODS

The database

Sequences whose subcellular location was annotated were extracted from release 33.0 of the SWISSPROT database (11). Subcellular location annotation was found for 15 775 out of 52 205 sequences in this release. This set of sequences was filtered to remove: sequences that were annotated as fragments of larger proteins; sequences that contained ambiguities (such as amino acids denoted by X within the sequence); sequences which were annotated as residing in more than one subcellular location; annotations made by similarity or marked as probable or possible concerning the subcellular location; i.e. essentially sequences were only kept if they appeared complete and had what appeared to be reliable location annotations coming direct from experiment. For this study transmembrane proteins were also excluded, as reliable prediction methods for this group of proteins already exist (12). It has also been shown that the extra- and intracellular domains of transmembrane proteins differ in their amino acid composition as do whole proteins (13) and therefore do not need to be considered as a separate compartment. Plant sequences were also excluded, as initial tests showed that their composition appears to be sufficiently different to have a negative influence on prediction accuracy for eukaryotic proteins (plant sequences were

*To whom correspondence should be addressed. Tel: +44 1223 834244; Fax: +44 1223 494919; Email: areinha@sanger.ac.uk

predicted at an accuracy of 20–30% lower than other eukaryotic proteins by a neural network trained on a combined data set). As not enough sequences for plants within the various subcellular locations exist, it was not possible to treat them as an independent group. After these steps 5134 sequences remained (9.8% of the whole release). The sequences were divided into 11 different groups according to their subcellular location and whether they belonged to eukaryotic or prokaryotic species. Within each group the sequence identity was calculated between all pairs and sequences were kept such that none had >90% sequence identity to any other. This was done to avoid a bias towards large sequence families with high similarity. Overall 3420 sequences remained, distributed over the 11 groups as shown in Table 1.

Table 1. Number of sequences within each subcellular location group

Location	Number of sequences
Cytoplasmic (eukaryotic)	684
Cytoplasmic (prokaryotic)	687
Extracellular (eukaryotic)	325
Extracellular (prokaryotic)	105
Glycosomal	9
Glyoxysomal	21
Lysosomal	15
Mitochondrial	321
Nuclear	1097
Periplasmic	201
Peroxisomal	66

Number of sequences in the 11 different subcellular locations that were distinguished for analysis. The glycosomal, glyoxysomal, lysosomal and peroxisomal groups were considered to contain too few data to be statistically analysed.

As can be seen from Table 1, for four of these groups the amount of data available is too small for a statistical analysis to be performed. As this leads to the exclusion of only 3.2% of all sequences in this database, a distinction between the remaining groups should still prove useful for analysis. Once the number of sequences available for the excluded groups becomes large enough for statistical analysis they can be included in the prediction method. To provide a further independent data set the above procedure was performed on sequences which first appeared in SWISSPROT releases 34 and 35. This yielded another 749 eukaryotic and 243 prokaryotic sequences. A list of the sequences within each group is available on request.

The neural network

The Stuttgart Neural Network Simulator (14) was used to build and train all the neural networks used.

Two different types of neural networks were used in prediction. A simple fully connected architecture with 20 input units, one for the fraction of each amino acid, and two output but no hidden units was used for predictions that distinguish between two possible locations. Each input unit was connected to each output unit. An output scheme of {1, 0} or {0, 1}, indicating one or the other location was selected, which made it possible to use the difference between the values of both output units as a reliability measure.

Two more general neural networks, predicting a sequence as belonging to one of three locations for prokaryotic or one of four locations for eukaryotic sequences, were built with a somewhat more complex architecture. Each consisted of 20 input units and three and four units in a hidden layer for prokaryotic and eukaryotic sequences respectively. Extensive tests showed that this number of units in the hidden layer yields optimal results (see Results). The number of output units matched the number of possible locations. Each input unit was connected to each hidden unit as well as each output unit. Also, each hidden unit was connected to each output unit. Again, a coding scheme for the output was chosen which assigns 1 to the correct location and 0 to all other locations. A standard back propagation algorithm was used during training, with $\eta = 0.2$.

Cross-validation testing

Using neural networks three data sets are needed to perform a jack-knife test. While the neural network learns from a training set, a test set is used to determine when the training process has to be stopped. As during this procedure the information within the test set is implicitly used, a third completely independent data set is needed to evaluate the prediction accuracy of the trained neural network. Accordingly, all data sets were split into three equally sized subsets. To provide cross-validation the sets were used for training, testing and evaluation in every possible combination, yielding six different neural networks. The overall prediction accuracy was determined as the average of the prediction accuracies of all six neural networks.

Weighted training

To prevent a bias of the neural network a weighted training has to be performed. The same number of sequences for each location has to be presented to the network during training. This causes a problem, as some of the groups are considerably larger than others. To include all information given in large groups some of the sequences in small groups have to be used repeatedly. This is done by first splitting into three subsets for training, testing and evaluation and then repeatedly using sequences within the subset.

Applying a reliability index

As the output nodes of the neural networks have values between 0 and 1, the difference between the highest and next highest node (Δ_0) can be used as a reliability index for a prediction. Reliabilities were binned in five groups (with ascending reliability index) for analysis: $0 < \Delta_0 < 0.2$; $0.2 < \Delta_0 < 0.4$; $0.4 < \Delta_0 < 0.6$; $0.6 < \Delta_0 < 0.8$; $0.8 < \Delta_0$.

Calculation of prediction accuracies

The prediction accuracies quoted throughout are the average of the accuracies determined for each subcellular location independently. This procedure is necessary because a weighted training of the neural networks for the cross-validation tests was performed. As a result, their prediction accuracy weights each subcellular location equally, regardless of the number of sequences within the group. Accuracies should therefore be compared with random values of 50% for 2 states, 33.3% for 3 states and 25% for 4 states.

Table 2. Prediction accuracies achieved for the prediction of all subcellular locations against eachother

	Cyto (P)	Extra (P)	Peri	Cyto (E)	Extra (E)	Mito
Extra (P)	93.12 [σ = 2.01]					
Peri	87.77 [σ = 1.89]	82.59 [σ = 4.43]				
Cyto (E)	81.11 [σ = 1.90]	92.91 [σ = 2.23]	85.56 [σ = 1.53]			
Extra (E)	88.28 [σ = 1.82]	86.42 [σ = 2.77]	84.80 [σ = 1.15]	76.72 [σ = 2.44]		
Mito	84.72 [σ = 1.19]	92.60 [σ = 3.09]	85.67 [σ = 2.00]	73.79 [σ = 2.62]	82.64 [σ = 2.57]	
Nuclear	94.20 [σ = 0.55]	94.36 [σ = 2.41]	92.35 [σ = 2.77]	84.86 [σ = 1.09]	83.84 [σ = 0.99]	86.12 [σ = 1.65]

The accuracy achieved by neural networks in predicting the subcellular location using only the amino acid composition as input. For each prediction accuracy the standard deviation in percent is given as yielded from cross validation tests.

Table 3. Summary of the prediction performances of the neural networks for eukaryotic and prolaryotic sequences

	Eukaryotic Proteins	Prokaryotic Proteins
Overall Prediction Accuracy	66.1 [σ = 1.59]	80.9 [σ = 1.99]
Prediction Accuracy Reliability Group 1	51.1 [σ = 6.05]	59.1 [σ = 9.34]
Prediction Accuracy Reliability Group 2	57.9 [σ = 3.04]	71.2 [σ = 11.11]
Prediction Accuracy Reliability Group 3	68.7 [σ = 4.56]	78.1 [σ = 6.55]
Prediction Accuracy Reliability Group 4	82.5 [σ = 2.47]	91.0 [σ = 2.85]
Prediction Accuracy Reliability Group 5	81.9 [σ = 4.33]	84.9 [σ = 2.18]

Summary of the prediction accuracy achieved by the neural networks for eukaryotic and prokaryotic sequences. Shown is the overall accuracy and the accuracy for the various reliability groups together with the standard deviation σ as yielded by cross validation tests.

RESULTS

Pairwise neural network prediction accuracy

The average fraction for each amino acid and its standard error was calculated for all subcellular locations which featured enough data for analysis (data not shown, but can be found at <URL http://predict.sanger.ac.uk/nnpsl/aminoacidcomposition.html >). Only phenylalanine (F), histidine (H), methionine (M) and tryptophan (W) show minor fluctuations, while the other amino acids show strong differences between different subcellular locations. Although the fractions of some amino acids are similar between locations [e.g. A, D, E, F, G, H, L, M, N, T, V, W and Y between extracellular (Eu) and mitochondrial], other amino acids differ substantially (e.g. C, I, K, P, Q, R and S). No uniform behaviour distinguishing eukaryotic from prokaryotic sequences is apparent except for alanine, for which prokaryotic proteins show a clearly higher average fraction than eukaryotic proteins. Overall the differences in the amino acid composition between all groups appear strong for prediction purposes. To determine how effective a measure the amino acid composition is, neural networks were trained to distinguish subcellular locations in a pairwise manner. Their prediction accuracy, as determined by cross-validation, varies from 74 to 94%, as shown in Table 2.

Neural networks distinguishing between a subcellular location of eukaryotic origin on the one hand and of prokaryotic origin on the other tend to achieve a very high prediction accuracy, showing that eukaryotic and prokaryotic organisms exhibit substantial differences with respect to their amino acid composition. This makes it necessary to handle them separately, although other studies imply that this is not the case (5).

Comparing subcellular locations of prokaryotic organisms against each other also shows substantial differences between all compartments, with the lowest prediction accuracy at 82.6%. It is worth noticing that the standard deviation (SD), as determined by cross-validation, is especially high in this case. This is most likely due to the fact that for both groups distinguished (extracellular and periplasmic) only a comparatively small number of sequences could be used (see Table 1). Neural networks are known to improve in performance as the amount of data for training increases, so the considerably smaller number of sequences used for training of this specific neural network may have resulted in the network being less stable.

The subcellular locations in eukaryotic organisms seem to be less distinct from each other. Cytoplasmic and mitochondrial proteins in particular appear to show common features, with the neural network distinguishing them only reaching a prediction accuracy of ~74%, while the standard deviation for prediction is low (2.6), indicating good convergence of the neural networks trained on the data. The same is true for the neural network distinguishing cytoplasmic and extracellular proteins (prediction accuracy 77%, SD 2.4). The fact that the neural network for extracellular and mitochondrial proteins reaches a prediction accuracy of ~83% indicates that cytoplasmic proteins share some features with mitochondrial and some with extracellular proteins, while these features do not overlap. The accuracy for predictions with a high reliability index was considerably higher than the overall accuracy (the prediction accuracy for all sequences) (data not shown).

General prediction of subcellular location

Pairwise networks are interesting to investigate the relative differences between different compartments, however, a practical prediction system requires the ability to distinguish between multiple compartments. Neural networks were therefore built and trained to assign proteins to one of three and four possible subcellular locations for prokaryotic and eukaryotic sequences respectively.

The four different subcellular locations taken into account for eukaryotic proteins were cytoplasmic, extracellular, mitochondrial and nuclear. The overall prediction accuracy reached 66.1%, with individual neural networks scoring between 64.5 and 68.7% (σ = 1.59). The low variation in the accuracy of individual networks indicates that the results are independent of the specific sequences within the training, test and evaluation sets. The accuracy for

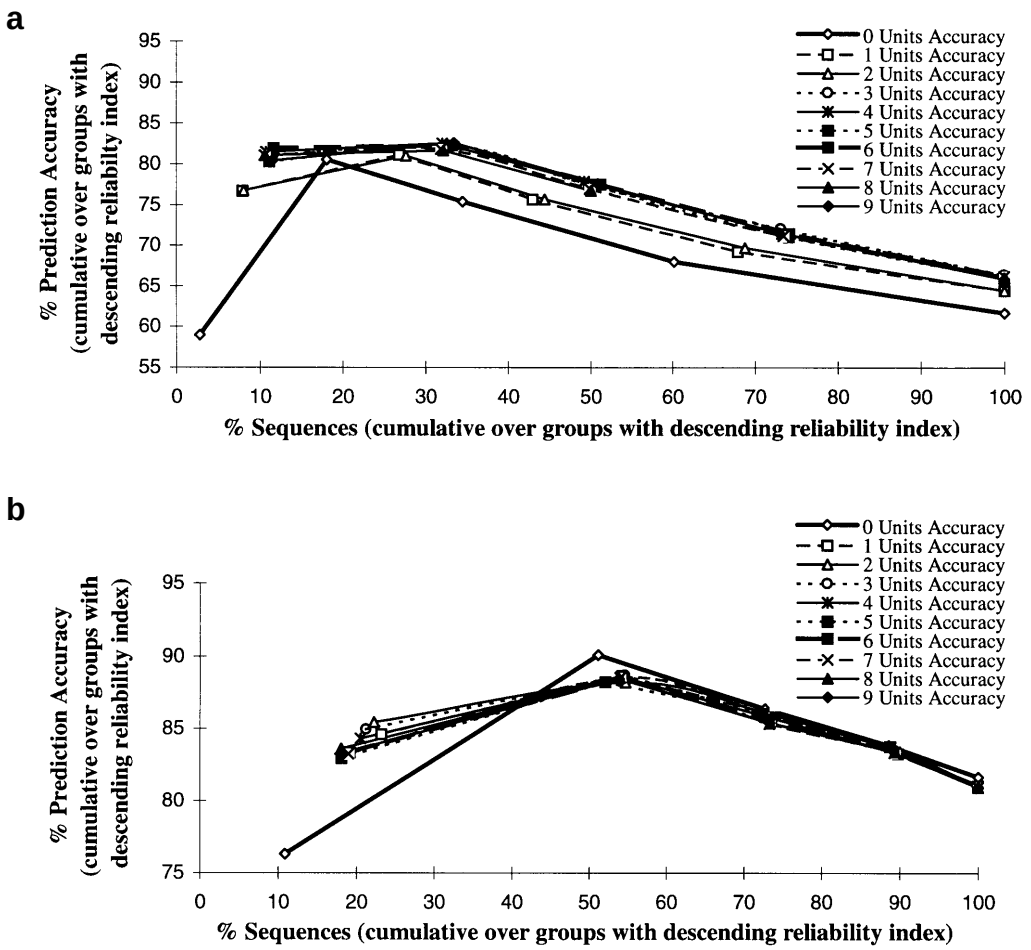

Figure 1. (**a**) The prediction accuracy for eukaryotic proteins was calculated cumulatively with respect to the various reliability indices by starting with sequences with the highest reliability index and then progressively including those with lower indices until the number calculated for the lowest reliability index is equal to the overall prediction accuracy. The percentage of the total number of sequences being considered was similarly calculated cumulatively. Plotting both variables against each other for a number of neural networks with a varying number of units within the hidden layer reveals that the neural network with no hidden units shows a pattern strongly deviating from those of the other neural networks. This indicates that introduction of hidden units leads to further information being picked up by the neural network. (**b**) As for (a) except showing only the results for the best network, with data broken down by location.

predictions with a high reliability index was considerably higher than the overall accuracy (the prediction accuracy for all sequences), as can be seen in Table 3. This compares with an accuracy of 25.0% from random guesses for four locations with a balanced set of data as considered here. For comparison, 66.1% corresponds to a slightly higher real life prediction accuracy of 67.2%, which includes the bias from the different fraction of proteins in each location in the current database.

Testing various neural networks with different numbers of units within the hidden layer reveals that the network with no hidden units not only performs considerably less well (overall prediction accuracy of 61.7%), but also deviates from the behaviour observed for networks with hidden units if the cumulative prediction accuracy for the reliability index groups is plotted against the cumulative number of sequences within each group (specificity against sensitivity), as can be seen in Figure 1a. This indicates that further information is gained through the introduction of hidden units into the neural networks. Networks with hidden units show a uniform behaviour, which converges for networks with three or more hidden units (overall prediction accuracy for three to nine hidden units varies only from 65.8 to 66.3%). However, changes in the distribution of sequences over groups with different reliability indices makes the neural network with four hidden units the best choice, as it performs slightly better for groups with the highest reliability index, achieving a prediction accuracy of 82.5% for 21.5% of all sequences and predicting a further 11.7% of the sequences at an accuracy of 81.9%. As these groups are the most useful for practical prediction purposes, the neural network architecture with four units in the hidden layer was chosen for further use, although it doesn't feature the highest overall prediction accuracy. A further plot of specificity against sensitivity for the best network subdivided by location is shown in Figure 1b. It can be seen that the accuracies are best for extracellular and nuclear proteins, but that this is not an effect of data size, since the extracellular group has the second smallest data size and is more likely to reflect the strength of the signal for different locations.

It was also found that for eukaryotic sequences the correct subcellular location has either the highest or second highest neural network output value in 80–91% of all cases, throughout

the reliability index groups. Attempts were made to better distinguish between these top two locations. Proteins predicted within the lowest reliability group were again predicted with the appropriate pairwise neural network. However, no improvement could be achieved in this way, as all such proteins were predicted within the lowest reliability group of the pairwise network as well.

The three possible subcellular locations for prokaryotic proteins were cytoplasmic, extracellular and periplasmic. This neural network yielded an overall prediction accuracy of 80.9%, with the accuracy of various neural networks lying between 78.2 and 83.4% ($\sigma = 1.99$), which compares with the accuracy for random predictions of 33.3%. Varying the number of units in the hidden layer from zero to nine does not cause any considerable change in the overall prediction accuracy (80.9 to 81.7%). However, plotting the cumulative prediction accuracy for the different reliability groups against the cumulative number of sequences covered by the groups, as shown in Figure 2a, showed that again the neural networks behave in a fairly uniform way once at least one hidden unit is present, while the network with no hidden units clearly differs in its pattern from the others. Although the difference in the overall prediction accuracy between the networks with and without hidden units is not as grave as for eukaryotic sequences, the deviating behaviour of the network without hidden units indicates that additional information is gained by the introduction of hidden units. Overall the architecture with three units within the hidden layer performs slightly better then the rest, achieving 33.4% of all sequences predicted with an accuracy of 91.0% and another 21.3% with an accuracy of 84.9%. As for eukaryotes, a further plot of specificity against sensitivity for the best network subdivided by location is shown in Figure 2b, with similar conclusions.

In the introduction we claimed that a method based on composition would be more robust to errors in 5′ gene annotation than other methods. We tested this by repeating the generation of both eukaryotic and prokaryotic networks with identical training and test data but with the leading 10 amino acids removed, to represent the effect of such uncertainty. The accuracies changed little (63.5 instead of 66.1% for eukaryotic and 80.5 instead of 80.9% for prokaryotic proteins), leading us to conclude that the method is robust in this respect.

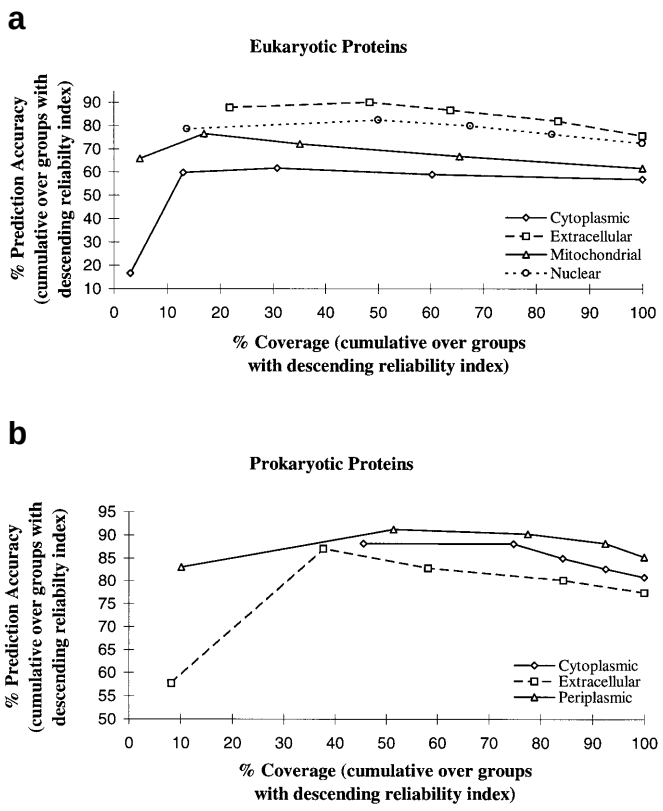

Figure 2. (**a**) The cumulative prediction accuracy and percentage of sequences for prokaryotic proteins in groups with different reliability indices was calculated as for eukaryotic proteins. Plotting both variables against each other for a number of neural networks with a varying number of units within the hidden layer reveals that the neural network with no hidden units shows a pattern strongly deviating from those of the other neural networks. (**b**) As for (a) except showing only the results for the best network, with data broken down by location.

Tests on new data

The specific requirements during training of the neural networks led to only one third of the available data being included in the training set. As it is a well-known fact that neural networks tend to improve with the amount of data presented to them during training, it was interesting to see how a network trained with a much larger number of sequences than the jack-knifing procedure allows will perform on independent data. Final versions of both the eukaryotic and prokaryotic neural networks were created using nine tenths of the available data for training and one tenth as a test set to prevent over-training on the specific data set. These neural networks were trained on ~2.5 times more data than those used for cross-validation. They were used to predict the subcellular location of proteins which appeared as new in SWISSPROT database release 34 or 35. These new sequences can be treated as a randomly chosen and completely independent data set that should have no systematic connection to the training set used.

For eukaryotic proteins the prediction accuracy increases by ~1% (from 66.1 to 67.0%). With at least 50 sequences within each group and 749 sequences altogether this outcome is not very likely to be due to random fluctuations. For prokaryotic proteins the situation is less clear. Although the calculation shows an increased prediction accuracy (from 80.9 to 82.7%), the very small number of sequences within the extracellular (only six sequences) and periplasmic groups (only 25 sequences) makes an influence of random fluctuations on the outcome quite possible.

This method relies on sequence composition, which is fairly orthogonal to sequence homology (data not shown), however, since training and testing sequences share some sequence homology it was felt that an effect from this on prediction accuracy should be checked for. The new sequences from SWISSPROT releases 34 and 35 were therefore grouped according to the highest similarity with a sequence in the training set. The prediction accuracy for each group was calculated and the results are shown in Figure 3a. For prokaryotic proteins sequences with higher similarity to those within the training set do not achieve higher prediction accuracy than less similar ones and for eukaryotic proteins only a slight trend seems to exist. Another possibility is that similar sequences are more likely to be predicted in the same way as the closest training set member as the homology between them increases. Figure 3b shows that for pairs which share the same location there is only a weak correlation between sequence similarity and the chance of the

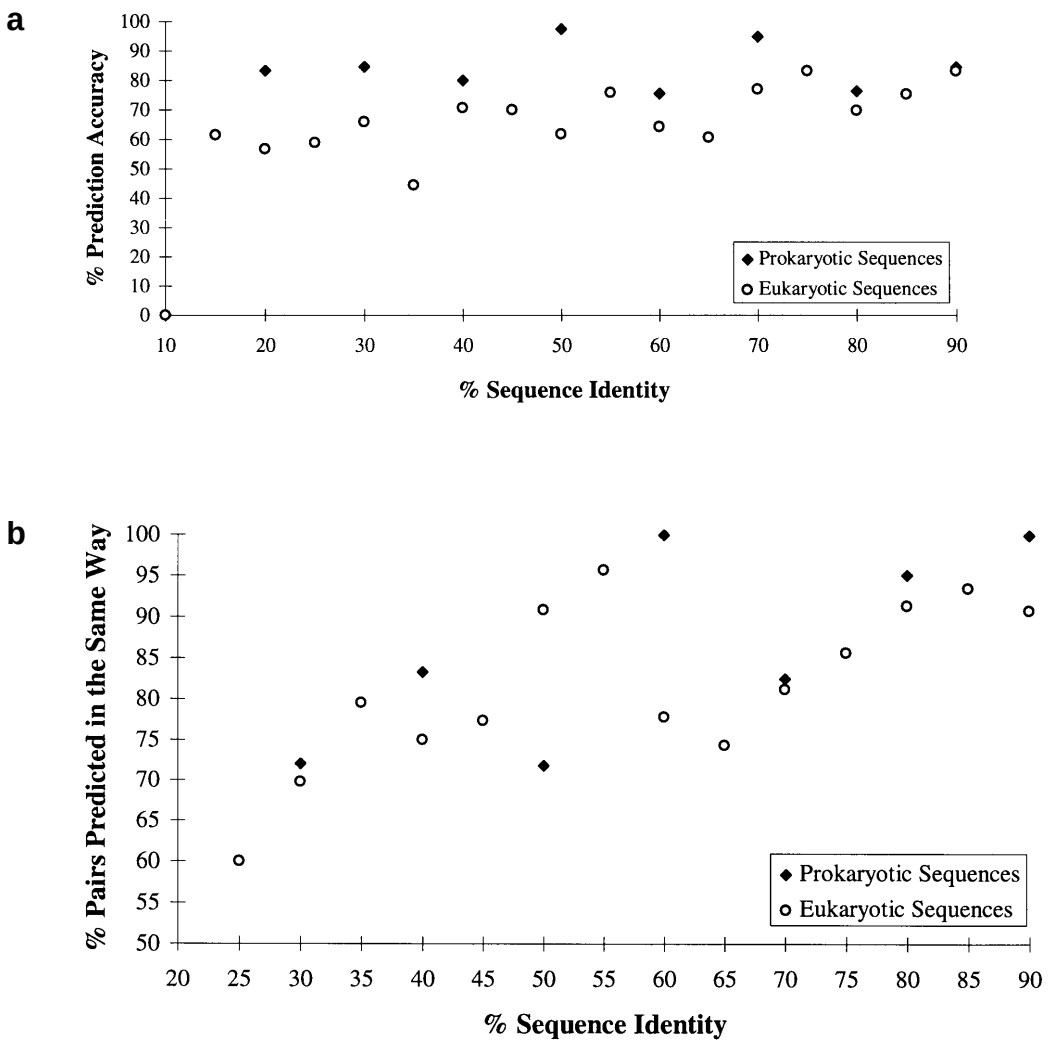

Figure 3. (**a**) For each predicted sequence the sequence with the highest sequence identity within the training set of the neural network was determined. The predicted sequences were then grouped according to this similarity in 5% steps for eukaryotic proteins and 10% steps for prokaryotic proteins (for the latter there were too few sequences to use 5% steps). For each group of sequences obtained in this way the prediction accuracy was determined. (**b**) For each predicted sequence the sequence with the highest sequence identity within the training set of the neural network was determined. Sequence pairs that had the same location were then grouped according to this similarity in 5% steps for eukaryotic proteins and 10% steps for prokaryotic proteins, as for (a). The fraction with identical prediction (both correct or both incorrect) was determined for each group.

same prediction (i.e. both predicted right or both predicted wrong). We conclude that the effect of sequence homology in biasing this prediction is minimal and does not decrease the value of the prediction methods.

DISCUSSION

Amino acid composition alone has been shown to contain sufficient information to distinguish proteins of different subcellular locations at a detailed level.

At the present level of prediction accuracy the method is not reliable enough for eukaryotic proteins to be used for blindly assigning a subcellular location to large numbers of potential proteins. It can, however, be used to give initial clues for further analysis. A further improvement in prediction accuracy may be achieved by passing ambiguous cases to an expert system for a final decision. This approach appears to be especially promising, as sequences of one of the two subcellular locations which are hard to distinguish (cytoplasmic and mitochondrial proteins) feature targetting sequences, although the previously mentioned problem of incorrect assignments in the 5′-region of automatically annotated genes is an issue when doing this.

As was shown through a test on independent data, the prediction accuracy can be improved by including more sequences in training of the neural network, although the increase of only ~1% in prediction accuracy for the eukaryotic neural network indicates that an upper limit may have been reached in this case. Taking into account that the cross-validation test for this network showed a very good convergence (σ = 1.55%), it seems likely that the amount of data used for initial training was already sufficient, which explains why only fairly small improvements can be achieved by including more sequences in the training. The neural

network for prediction of prokaryotic sequences on the other hand converged somewhat less (σ = 1.99%), indicating that additional sequences in training may result in a further improvement in prediction accuracy.

The prediction method is available on the world wide web at location <URL: http://predict.sanger.ac.uk/nnpsl >. An Email-based service for large numbers of sequences will be made available at the same address shortly. The site includes a link to the world wide web-based service for prediction of transmembrane proteins (12), to make it easier for users to test sequences for transmembrane regions before making a subcellular location prediction. Large non-membrane spanning domains of transmembrane proteins can be predicted by the described method, by handling them as independent protein chains. Also, work is in progress to add predicted subcellular location annotations to TREMBL (15,16) entries based on a combination of transmembrane prediction and this method.

ACKNOWLEDGEMENTS

A.R. thanks the Stiftung Stipendien Fonds des Verbandes der Chemischen Industrie e. V. for support by way of a Kékule scholarship and the Wellcome Trust and ZENECA for support.

REFERENCES

1 Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.-C. and Herrmann,R. (1996) *Nucleic Acids Res*., **24**, 4420–4449.
2 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., Fitzgerald,L.M., Clayton,R.A., Cocayne,J.D., *et al*. (1996) *Science*, **273**, 1058–1073.
3 Bork,P., Ouzounis,C. and Sander,C. (1994) *Curr. Opin. Struct. Biol*., **4**, 393–403.
4 Nakashima,H. and Nishikawa,K. (1994) *J. Mol. Biol*., **238**, 54–61.
5 Cedano,J., Aloy,P., Perez-Pons,J.A. and Querol,E. (1997) *J. Mol. Biol*., **266**, 594–600.
6 Nakai,K. and Kanehisa,M. (1992) *Genomics*, **14**, 897–911.
7 Nakai,K. and Kanehisa,M. (1991) *Protein Struct. Function Genet*., **11**, 95–119.
8 Chou,K.-C. (1995) *Protein Struct. Function Genet*., **21**, 319–344.
9 Eisenhaber,F., Frömmel,C. and Argos,P. (1996) *Protein Struct. Function Genet*., **25**, 169–179.
10 Rost,B. and Sander,C. (1994) *Protein Struct. Function Genet*., **19**, 55–72.
11 Bairoch,A. and Boeckmann,B. (1993) *Nucleic Acids Res*., **21**, 3093–3096.
12 Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) *Protein Sci*., **4**, 521–533.
13 Nakashima,H. and Nishikawa,K. (1992) *FEBS Lett*., **303**, 141–146.
14 Zell,A., Mamier,G., Vogt,M., Mache,N., Hübner,R., Döring,S., Herrmann,K.-U., Soyez,T., Schmalzl,M., Sommer,T., Hatzigeorgiou,A., Posselt,D., Schreiner,T., Kett,B., Clemete,G. and Wieland,J. (1995) University of Stuttgart, Institute for Parallel and Distributed High Performance Systems (IPVR), Stuttgart, Germany.
15 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res*., **25**, 31–36.
16 Apweiler,R., Gateau,A., Contrino,S., Martin,M.J., Junker,V., O'Donovan,C., Lang,F., Mitaritonna,N., Kappus,S. and Bairoch,A. (1997) In Gaasterland,T.P.K, Karplus,K., Ouzonis,C., Sander,C. and Valencia,A. (eds), *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Greece, pp. 33–43.