doi:10.1016/S0022-2836(02)00379-0 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 319, 1257–1265

JMB

Prediction of Human Protein Function from Post-translational Modifications and Localization Features

L.J. Jensen[1]†, R. Gupta[1]†, N. Blom[1], D. Devos[2], J. Tamames[2]
C. Kesmir[1], H. Nielsen[1], H.H. Stærfeldt[1], K. Rapacki[1], C. Workman[1]
C.A.F. Andersen[1], S. Knudsen[1], A. Krogh[1], A. Valencia[2] and
S. Brunak[1]*

[1]*Center for Biological Sequence Analysis, Biocentrum-DTU Building 208, The Technical University of Denmark DK-2800 Lyngby, Denmark*

[2]*Protein Design Group National Center for Biotechnology, CNB-CSIC Cantoblanco, Madrid E-28049 Spain*

**Corresponding author*

We have developed an entirely sequence-based method that identifies and integrates relevant features that can be used to assign proteins of unknown function to functional classes, and enzyme categories for enzymes. We show that strategies for the elucidation of protein function may benefit from a number of functional attributes that are more directly related to the linear sequence of amino acids, and hence easier to predict, than protein structure. These attributes include features associated with post-translational modifications and protein sorting, but also much simpler aspects such as the length, isoelectric point and composition of the polypeptide chain.

Keywords: functional role; phosphorylation; glycosylation; protein sorting; amino acid composition

Introduction

Out of the 35,000 to 50,000 genes believed to be present in the human genome, no more than 40–60% can be assigned a functional role based on homology to proteins with known function.[1,2] Traditionally, protein function has been related directly to the three-dimensional structure of the polypeptide chain, which currently, for an arbitrary sequence, is quite hard to compute.[3] The method presented here operates in the "feature" space of all sequences, and is therefore complementary to methods that are based on alignment and the inherent, position-by-position quantification of similarity between two sequences. The method does not require knowledge of gene expression,[4] gene fusion and/or phylogenetic profiles.[5–8] Although the latter type of method does not rely on finding direct matches to proteins of known function, it does require sequence similarity to other candidates that can be phylogenetically linked to a protein of known function.

For any function assignment method, the ability to correctly predict the relationship depends strongly on the function classification scheme used. One would, for example, not expect that a method based on co-regulation will work well for a category like "enzyme", since enzymes and the genes coding for their substrates or substrate transporters often may display strong co-regulation. A similar argument holds true for the phylogenetic profile method.

Our approach to function prediction is based on the fact that a protein is not alone when performing its biological task. It will have to operate using the same cellular machinery for modification and sorting as all the other proteins do. Essential types of post-translational modifications (PTMs) include: *N*- and *O*-glycosylation, (S/T/Y) phosphorylation, and cleavage of N-terminal signal peptides controlling the entry to the secretory pathway, but hundreds of other types of modifications exist[9] (a subset of these will be present in any given organism). Many of the PTMs are enabled by local consensus sequence motifs, while others are characterized by more complex patterns of correlation between the amino acids.[10]

† These two authors contributed equally to this work.
Abbreviation used: PTM, post-translational modification.
E-mail address of the corresponding author: brunak@cbs.dtu.dkhttp://www.cbs.dtu.dk

This suggests an alternative approach for function prediction, as one may expect that proteins performing similar functions would share some attributes even though they are not at all related at the global level of primary structure. As several predictive methods for PTMs have been constructed (R.G., S.B., unpublished results),[10–13] a function prediction method based on such attributes can be applied to all proteins where the sequence is known.

Results and Discussion

The ProtFun method described here integrates (using a neural network approach) 14 individual attribute predictions and calculated sequence statistics (out of more than 25 tested for discriminative value). The integrated method predicts functional categories as defined originally by Riley for *Escherichia coli*, that in modified form has been used to describe many entire genomes in recent publications.[1,2,14,15] In addition, it predicts whether a sequence is likely to function as an enzyme, and if so, its category according to the classes defined by the Enzyme Commission.[16,17] The same scheme can be used to predict any other set of functional classes including more narrowly defined ones. We have applied the approach to, for instance, specifically identify hormones, receptors and ion channels in the human genome as defined by the Gene Ontology Consortium.[18]

We have used combinations of attributes in a collection of neural network ensembles for predicting the functional category of a protein. Combinations of attributes were selected by evaluating their discriminative value for a specific functional category, say proteins involved in transcription. Attributes useful in function prediction must not only correlate well with the functional classification scheme, but must also be predictable from sequence with reasonable accuracy.

Interestingly, the combinations of attributes selected for a given category also implicitly characterize a particular functional class in an entirely new way. The method identifies without any *a priori* ranking of their importance, the biological features relevant for a particular type of functionality, see Figure 1. It appears that the use of PTMs is essential for the prediction of several functional classes. In addition to attributes related to subcellular location the most important features for predicting if a protein is, for example, regulatory or not, are PTMs. Similarly PTMs are very important for correct assignment of proteins related to the cell envelope, replication and transcription.

The fact that (predicted) PTMs correlate strongly with the functional categories fits well with biological knowledge. For example, predicted *N*-glycosylation sites turn out to be important for prediction of cell envelope proteins. In fact, it has been shown that removal of carbohydrates linked to asparagine residues from a protein normally targeted for the cell envelope retains it in the endoplasmic reticulum.[19]

For proteins with "regulatory function" two of the most important features were S/T phosphorylation and Y-phosphorylation, respectively (Figure 1). It is very satisfying that this correlation was found by the neural networks when considering that reversible phosphorylation is a well known and widely used regulatory mechanism.[20] Glycosylation was also found to be a strong indicator for regulatory proteins. This is true for both *N*-glycosylation and *O*-GalNAc (mucin type) glycosylation of serine and threonine residues. For these proteins, two additional features had significant predictive value: The predicted subcellular location, and PEST regions (rich in proline, glutamic acid, serine, and threonine residues), where the latter targets proteins for degradation. Again, it makes sense that proteins involved in fast regulatory mechanisms should be degraded quickly.[21]

In order to understand further how PTMs correlate with the functional categories, we investigated the effect of alternative representations of the PTM attributes. For example, when phosphorylation of serine and threonine residues was encoded as two separate features the result was a slightly reduced predictive performance. Joining all three types of phosphorylation (S/T/Y) into one single feature led to a much larger drop in performance. Again this makes biological sense, as serine and threonine residues are known often to be phosphorylated by the same group of kinases, while tyrosine residues typically are phosphorylated by different kinases.

The most important single feature for distinguishing between enzymes and non-enzymes turned out to be protein secondary structure as predicted by PSI-Pred.[22] This also makes sense, as enzymes are known to be overrepresented among all-alpha proteins, and more rarely are found to be, e.g. all-beta proteins.

We also trained networks for predicting the enzyme subclasses. Although these networks were trained specifically to discriminate between a given enzyme subclass and all other enzyme subclasses, they implicitly make an enzyme/nonenzyme prediction as well. The enzyme class predictions can thus be used as additional support for the predictions made by the enzyme/nonenzyme networks.

Quantitative description of the ProtFun predictive performance

The selection of category-relevant attributes is based on quantitative assessment of the ability to predict (assign) categories for new sequences nonsimilar to the sequences used to train the method (see below). Figure 2 shows how the ProtFun method fairs for the prediction of functional and enzyme categories in terms of sensitivity and the level of false positives. When the sensitivity is below 40%, the level of false positive predictions

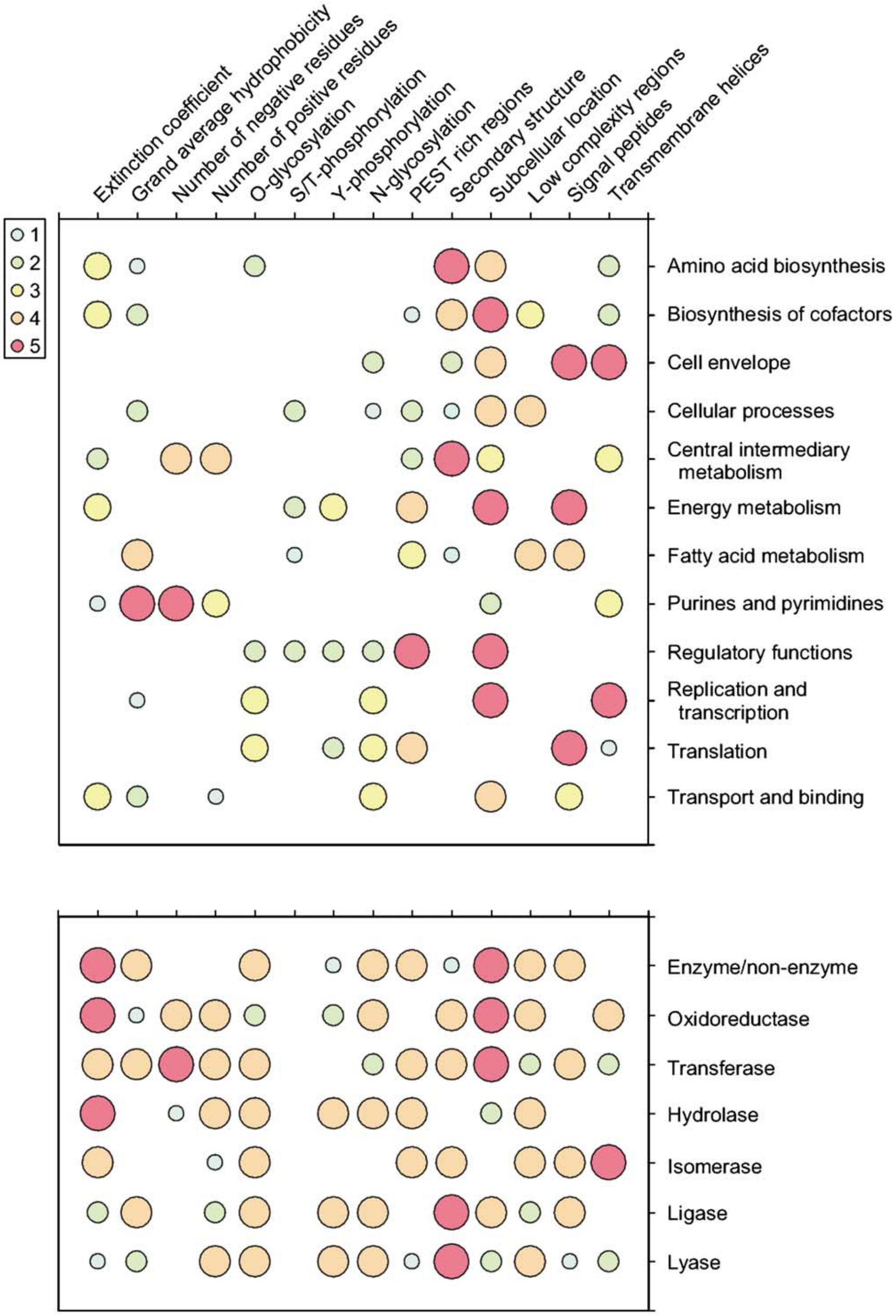

Figure 1. The discriminative impact of features for the different functional categories and enzyme classes. The Figure shows how often a given feature (out of the 14 retained) was included in the five network ensemble performing the classification for a given category and thus its importance. The retained feature predicting methods were: NetNGlyc,[14] NetOGlyc,[11] NetPhos,[10] PEST regions,[21] PSIPRED,[22] SEG filter,[34] SignalP,[12] PSORT,[13] TMHMM.[35] In addition, a number of calculated features were retained: extinction coefficient, grand average hydrophobicity, and the numbers of positively and negatively charged residues. During the feature selection process 11 features were "not" retained due to their low disciminatory value (or their correlation to other features retained): the amino acid composition, the composition of redidues predicted to be buried or exposed, the aliphatic index, instability index, number of atoms, the net charge, the isoelectric point, predicted GlcNAc sites, the sequence length, and predicted coiled coil regions.

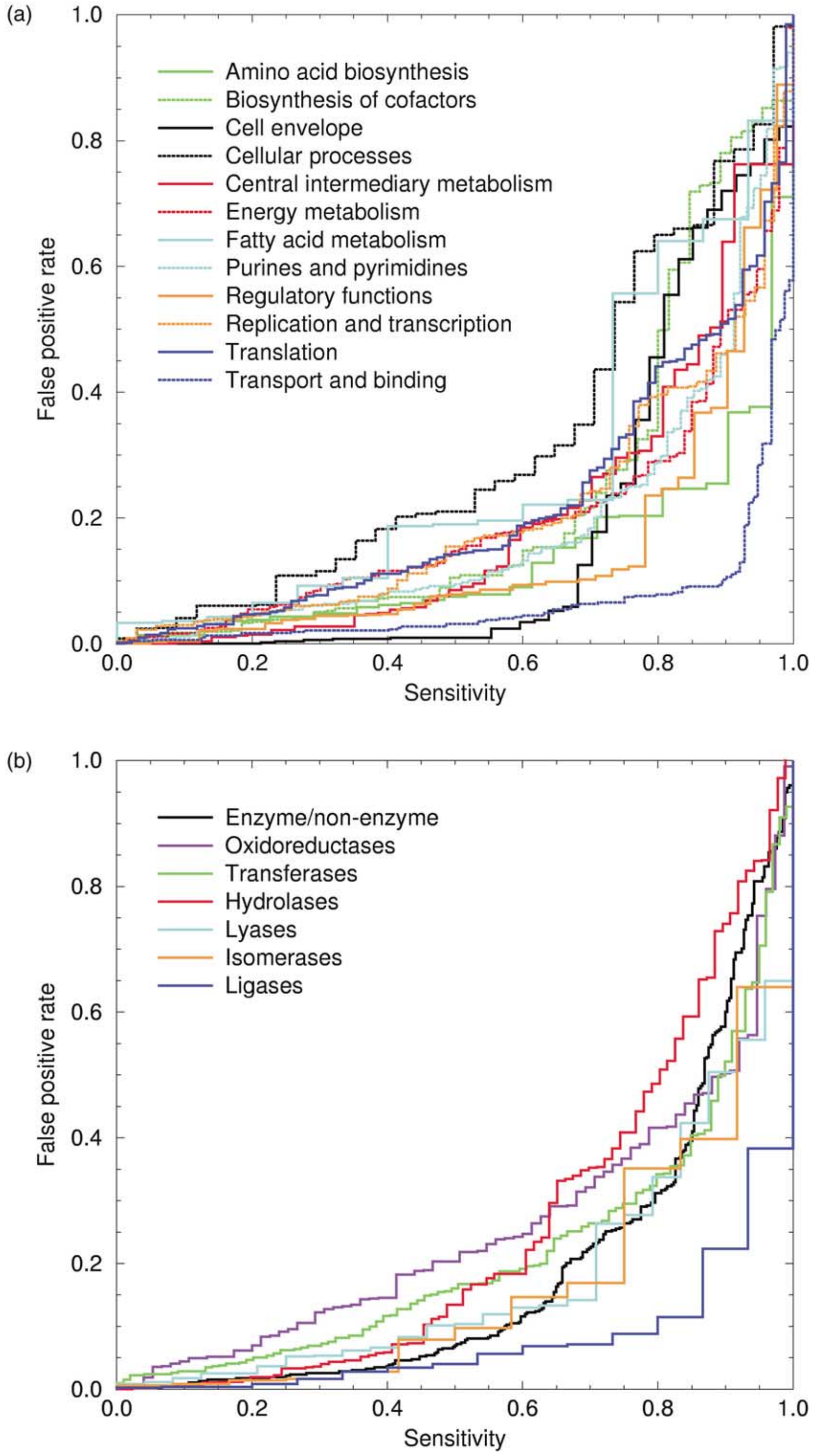

Figure 2. The predictive performance shown as sensitivity *versus* false positive rate for cellular role and enzyme categories. The plot was constructed from results obtained for the independent test set, and corresponds to the expected performance for novel, uncharacterized proteins with no significant match to a protein with known function. For a given category, e.g. transport and binding, a sensitivity of 90% can be achieved with false positive rate of 10% corresponding to 90% correct prediction on both positives and negatives. Random performance would correspond to a line along the diagonal.

is very low. The confidence in the predictions can be used directly to sift out the predictions that almost certainly are correct. The way the probabilities are estimated gives rise to an almost linear relationship between the probability threshold used and the false positive rate. For a given probability threshold, the rate of false positives is essentially 1 minus the threshold. The best performance values are comparable to the upper limits estimated from the consistency in the assignment of the SWISS-PROT keywords.[23]

Relation to the linguistic analysis of SWISS-PROT entries

The classification into functional categories is based on linguistic analysis and clustering of SWISS-PROT keywords by the Euclid method.[24,25] The Euclid method computes scores for placing a protein into each of the 14 categories.[14,26] The classification is not mutually exclusive, i.e. a protein sequence may have scores high enough to be placed into two or more categories.

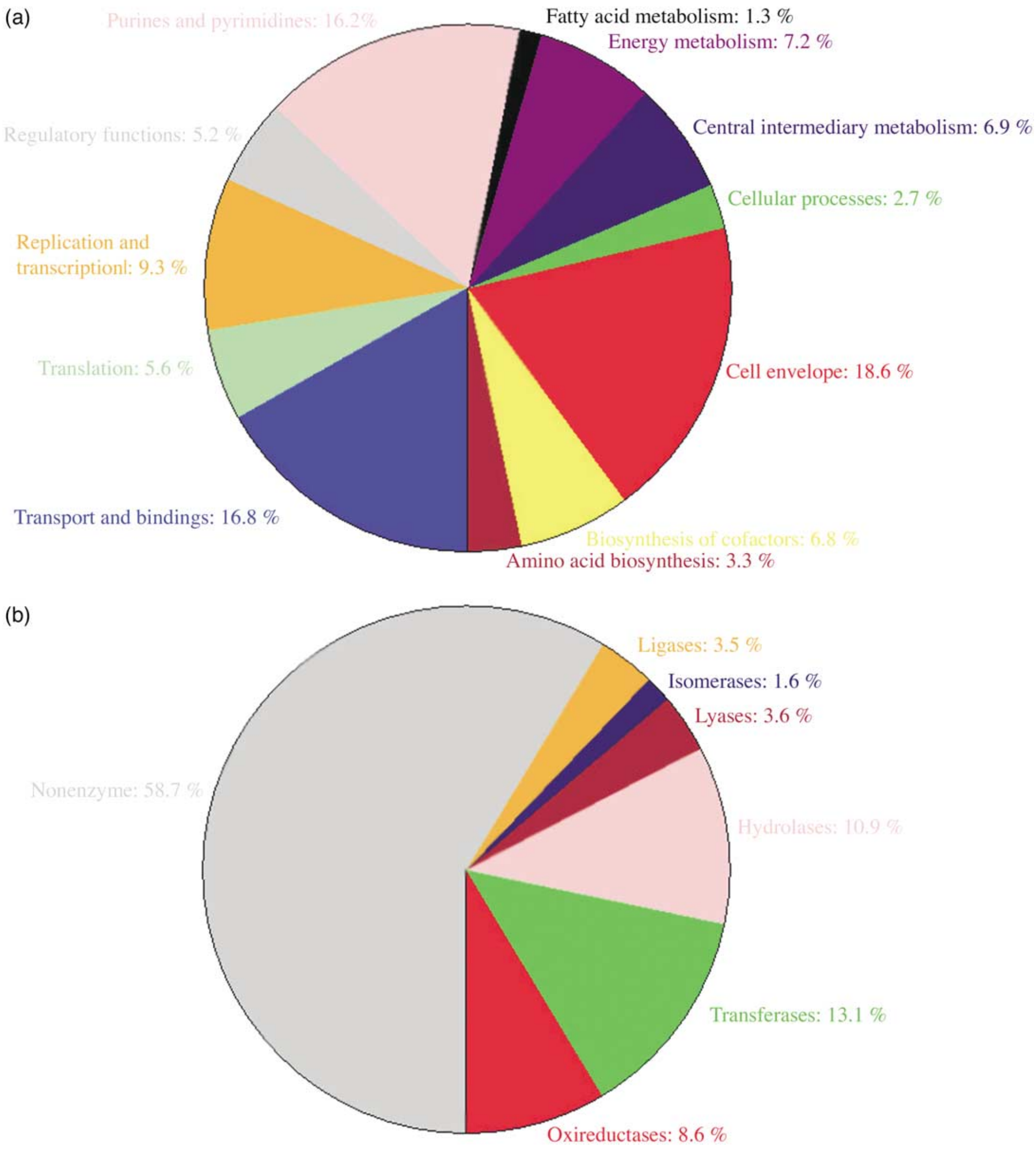

Figure 3. Statistics for the human genome based on the Ensembl gene set.[27] From the probabilistic ProtFun output, the number of proteins belonging to a given category was estimated by summing over all 27,000 sequences. The most striking difference from the Venter *et al.* paper is that a larger fraction of the proteins is predicted to be enzymes, while the distributions over enzyme subcategories agree quite well.

In general we found a strong relation between the prediction quality of Euclid and ProtFun. This should not come as a surprise, considering that the quality of prediction from Euclid determines the quality of the data set on which ProtFun was trained. Two of the worst categories are “cellular processes” and “central intermediary metabolism”, which are both very loosely defined, especially the former, in which many different functions ranging from cell division to chaperones and detoxification are included.

One noticeable exception from this rule is the class of regulatory proteins. Because the class of regulatory proteins tends to overlap other categories a lot, these proteins are hard to categorize correctly by Euclid. However, this has not been a problem for ProtFun, which allows a protein to belong to more than one category. Indeed regulatory proteins are one of the best predicted categories.

Functional characterization of the complete human genome

Using ProtFun it is possible to estimate the breakdown on functional categories of the entire human genome. Ideally a data set with all proteins encoded by the human genome should be used. As no final and highly reliable set is yet available, we have used the database of confirmed sequences made available by the Ensembl initiative.[27] This database consists of ~27,000 protein sequences from the human genome, all of which are supported by EST matches. One should be aware that this database is likely to have a bias towards highly expressed proteins. Using the predicted probability for each category, the number of proteins in each category was subsequently estimated by summing over the probability of the category in question for every protein (Figure 3).

The functional breakdowns in the human genome publications[1,2] are based on function assigned by sequence similarity and are therefore based on approximately 50–70% of the genes (depending on the gene number). Direct comparison to what we predict is also made difficult by the fact that different classification systems are used in the two articles. The most striking difference from the Venter *et al.* paper is that we predict a much larger fraction of the proteins to be enzymes, while the distributions over enzyme subcategories agree quite well. Part of the explanation for the enzyme bias can be that the complete Ensembl data set[27] may have a bias towards highly expressed proteins. We also investigated the spread of functionally related proteins across the different chromosomes (data can be found at the ProtFun Website†). Among several interesting observations (for example that endoplasmic and Golgi proteins are highly abundant at chromosome 6), was the fact that chromosome 11 seems to contain many uncharacterised proteins (belonging to "Other categories"), which falls outside the classification used in this study.

Individual sequence prediction

The ProtFun method is perhaps best suited for obtaining functional hints for individual sequences for later use in assay selection and design. The first example shown here relates to the human prion sequence (PRIO_HUMAN P04156) which is being associated with the Creutzfelt–Jacobs syndrome. The functionality of this protein, which seems to produce no phenotype when knocked out in mice,[28] is still not fully understood. The ProtFun method predicts with high confidence that the human prion sequence belongs to the transport and binding category, and also that it is very unlikely to be an enzyme (Table 1). Indeed prions have been shown to be able to bind and transport copper while no catalytic activity has ever been observed.[29,30] Interestingly, as the prion is a cell surface glyco-protein (expressed by neurones) it has a distinct pattern of post-translational modification, which most likely contain information that can be exploited by the system for functional inference. Incidentally, the cell envelope category is the third-highest scoring category for the prion sequence. The purine/pyrimidine class is second; however, no experimental evidence supports this functionality. Functional information was "not" transferred by sequence similarity from the nearest neighbor, as the maximal similarity between the prion sequence and the data set (training and test) is 14.8% to proline-arginine-rich end leucine-rich repeat protein (PRLP_HUMAN P51888). We believe that predictions like these are very useful when resolving protein function, because they can be used to generate specific hypotheses and direct laboratory experiments.

Table 1. ProtFun output for the human prion (PRIO_HUMAN P04156) and for an interacting pair of proteins, the amyloid A4 protein (A4_HUMAN P05067; P09000; Q16011) and transthyretin (TTHY_HUMAN P02766), which at the sequence level are entirely unrelated. Both of these proteins are, with high confidence, predicted to be cell envelope-related as well as transport and binding proteins in agreement with the known functionality of these proteins

	Prion	A4	TTHY
Amino acid biosynthesis	0.011	0.011	0.011
Biosynthesis of cofactors	0.041	0.161	0.034
Cell envelope	0.146	0.804	0.698
Cellular processes	0.027	0.027	0.051
Central intermediary metabolism	0.047	0.139	0.059
Energy metabolism	0.029	0.023	0.046
Fatty acid metabolism	0.017	0.017	0.023
Purines and pyrimidines	0.528	0.417	0.153
Regulatory functions	0.013	0.014	0.014
Replication and transcription	0.020	0.029	0.040
Translation	0.035	0.027	0.032
Transport and binding	0.831	0.827	0.812
Enzyme	0.233	0.367	0.227
Non-enzyme	0.767	0.633	0.773
Oxidoreductase (EC 1.–.–.–)	0.070	0.024	0.055
Transferase (EC 2.–.–.–)	0.031	0.208	0.037
Hydrolase (EC 3.–.–.–)	0.101	0.090	0.208
Isomerase (EC 4.–.–.–)	0.020	0.020	0.020
Ligase (EC 5.–.–.–)	0.010	0.010	0.010
Lyase (EC 6.–.–.–)	0.017	0.078	0.017

Using function prediction in conjunction with protein–protein interaction data

The method is also relevant for obtaining additional evidence on protein–protein interactions, where database information may contain many false negatives (as not all possible interactions have been screened). We did, as an example, predict the function of both sequences in all interaction partners found in the database of interacting proteins (DIP).[31] If the functional categories of the interacting proteins are predicted to be the same for otherwise unrelated sequences, that should increase the likelihood of the prediction being correct (as well as the validity of the interaction). Others have successfully used a similar approach based on subcellular localization to lower the rate of false positives for yeast two-hybrid data.[32]

An interesting example of such an interacting protein-pair is the Alzheimer's disease amyloid A4 protein and transthyretin, which at the sequence level are entirely unrelated. Both of these

† The ProtFun method is made available online at the URL www.cbs.dtu.dk/services/ProtFun/

proteins are, with high confidence, predicted to be "cell envelope"-related as well as transport and binding proteins, see Table 1. Amyloid A4, a neural receptor, and transthyretin, a thyroid hormone binding protein believed to transport thyroxine into the brain, have functionalities that are in full agreement with the prediction. The two sequences have a maximal similarity to the data set of 12.4% (to citrate synthase CISY_HUMAN O75390).

When evaluating the functional category profiles for all interacting pairs in DIP (*versus* non-interacting pairs) we found that interacting pairs indeed more often tend to have the same functional categorization (data not shown). However, while interacting non-enzymes in many cases will have the same functional role (belong to the same pathway), it may be more typical for enzyme–substrate pairs to belong to different categories.

Conclusion

The method presented here has the ability to transfer functional information between sequences that are far apart in sequence space. Not even the primary structures of the individual features (which are integrated by the method) need to be alike, or be related by evolution. The ProtFun method performs its non-linear classification in the feature space defined by 14 predicted and calculated attributes, which have been selected by the approach (out of more than 25 different attributes considered initially for discriminative value). The mapping between the space of all sequences and this feature space is also highly non-linear as very different sequences, by the individual feature predictors, may be converted to the same patterns of, for example, post-translational modification or secondary structure. Here it is demonstrated that it is indeed possible to transfer functional information from the knowledgebase accumulated by experimental biology, even to proteins that are completely isolated in sequence space.

Materials and Methods

Data sets and functional class assignment

Classes of cellular function were defined after the 14 class classification originally proposed for the *E. coli* genome[14] and later extended by the TIGR group. The automatic class assignment to sequences was made by an extension of the Euclid system performing linguistic analysis of SWISS-PROT keywords.[25] The system detects sequences similar to a query sequence by a BLAST search in the SWISS-PROT database and extracts common keywords from the entries. As we work with sequences from SWISS-PROT (with known function) we used the keywords directly and include no alignment step. For each functional class the informative weight (*Z*-score) of each keyword was extracted from a dictionary.[25] For each sequence a keyword sum leads to scores for the 14 classes.

Table 2. The number of sequences included in the data sets when training networks for the various categories

Category	Positive	Negative
Amino acid biosynthesis	85	3691
Biosynthesis of cofactors	240	2964
Cell envelope	173	2599
Cellular processes	259	3339
Central intermediary metabolism	216	3127
Energy metabolism.id	330	3310
Fatty acid metabolism	53	3846
Purines and pyrimidines	535	1649
Regulatory functions	586	3037
Replication and transcription	746	2591
Translation	174	3677
Transport and binding	1461	2126
Enzyme	1620	4038
Oxidoreductase	319	1213
Transferase	529	1003
Hydrolase	485	1047
Isomerase	72	1460
Ligase	49	1483
Lyase	78	1454

The central point of the Euclid system is the dictionary. The primary version of this dictionary was generated from an initial set of carefully, hand-annotated proteins from different organisms spanning every kingdom of life. From this initial set, a first dictionary was defined that was used to assign all SWISS-PROT proteins and the process of dictionary definition and assignment was reiterated until convergence.

This final dictionary was used to assign functional classes to around 5500 human proteins from SWISS-PROT. In the selection we omitted SWISS-PROT sequences representing fragments and hyphothetical proteins and therefore the "high-quality" subset is smaller that the entire set of human proteins in this database.

The values obtained from the method were then compared to two thresholds: if the score for a category was above 3 it was considered a positive example, while examples with a score below 0 were used as negative examples. Examples scoring between 0 and 3 were considered unclear and were thus not used. By labeling our data set in this way we eliminate the most uncertain functional annotations thereby improving the quality of our data set. The composition of the data set obtained is shown in Table 2.

Enzyme class assignment

SWISS-PROT provides enzyme class information for most enzymes in the DE field. For those without an EC assignment, the suffix ASE and the presence (or absence) of the words INHIBITOR and PRECURSOR were additional considerations when assigning proteins into the categories Enzyme, Non-enzyme or Neither. The Neither category comprised ambiguous cases that were excluded from training. For the six enzyme classes, only proteins with EC assignments were used. The negative set for each class contained enzymes assigned to other enzyme classes.

Similarity screening of test sets *versus* training sets

To generate a training set A, and a test set B, in which the similarity between the two sets was minimal, the following heuristic algorithm was used: a similarity measure $D(a, b)$ between all pairs of sequences (a, b) in the original set was calculated using the Smith–Waterman score for the optimal local alignment between each sequence pair. A similarity measure $H(A, B)$ was defined as the sum of similarities between sequences in set A and sequences in set B. $H(A,B) = \mathrm{sum}(D(a,b)|a \in A,\ b \in B)$

The algorithm for generating the two sets A and B started by having all sequences in A. The algorithm selected a sequence $x \in A$ that maximized the value $H(A,B) - H(A\backslash\{x\}, B \cup \{x\})$. New sequences were selected from A until set B had the desired size.

Feature prediction and encoding

A number of different prediction methods were used as input features for the method (see Figure 1). All the servers were run on all 5500 sequences constituting the training and test sets. The output was parsed and the scores obtained from the different predication servers were normalized and/or converted into probabilities.

Encoding positional feature information (e.g. phosphorylation sites) for proteins of variable length is non-trivial. We tested a number of encoding schemes for the positional information in the input to the neural networks. One approach was to divide each sequence into a number of equally sized bins, where the bin size was dynamically calculated for each sequence. The most important disadvantage of this approach is that each bin does not represent the same number of residues for sequences of different length. An alternative method is to define a fixed number of bins of fixed size. Because the sequences have different lengths this can result in overlapping bins and redundant information in the encoding. For very long sequences the fixed size binning scheme gives rise to gaps between the bins in which case the method will fail to encode all the features fully. Finally a combined approach was tried with one fixed size bin in either of the sequence and dynamic bins to encode the rest of the sequence. Full details of the feature encoding can be found at the ProtFun Website†.

Each of these encodings was tested with a different number of bins on all features having positional information. The performance of each binning scheme was evaluated by training a neural network on each feature separately. For each feature the binning scheme that gave the highest test set correlation coefficients across the different functional categories on most categories was chosen.

Feature combinations and network ensembles

Optimal combination of parameters for each of the different categories were found using a boot-strap strategy. First, for every category a simple network with one fully connected hidden layer was trained on each separate feature. Details can be found elsewhere,[33] while information about specific network architectures can be found at the ProtFun Website†. On the basis of the test set performance of these networks we judged which features were potentially useful for prediction of at least one category. Networks were then trained for every pair of these features, to obtain information on the correlations between features. Many networks using increasing numbers of these features were then trained, and the best five were picked as an ensemble.

The outputs of these networks were subsequently transformed into probabilistic scores. Based on the predictions performed by each network on the test set, the network output distributions for positive ($f_{\mathrm{pos}}(x)$) and negative examples ($f_{\mathrm{neg}}(x)$) were estimated using Gaussian kernel density estimators on the output activities with no squashing function applied. From these density estimates and the number of the positive and negative examples (N_{pos} and N_{neg}) the probability that an example is positive ($P(x)$) can be calculated from the network output:

$$P(x) = \frac{N_{\mathrm{pos}} f_{\mathrm{pos}}(x)}{N_{\mathrm{pos}} f_{\mathrm{pos}}(x) + N_{\mathrm{neg}} f_{\mathrm{neg}}(x)}$$

To calculate the combined prediction of an ensemble of networks we simply take the average of their probabilistic predictions. It is these values that are reported by the method.

Chromosomal gene location

Chromosome locations for the human SWISS-PROT sequences were obtained by web-linking through SWISS-PROT references to the OMIM database (Online Mendelian Inheritance in Man) maintained at NCBI‡. From OMIM, by further linking to LocusLink§, one could obtain the chromosome number for the gene being considered. Not all proteins could be tracked down to their chromosome number in this fashion. For the remaining sequences, BLAST-ing them against the human genome database at NCBI revealed the chromosome number in most cases.

References

1. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Lesk, A., Conte, L., Hubbard, T. (2001). Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures and interresidue contacts. *Proteins: Struct. Funct. Genet.*, **45** (suppl. 5), 98–118.
4. Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
5. Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
6. Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
7. Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C. *et al.* (2000). Functional

† The ProtFun method is made available online at the URL www.cbs.dtu.dk/services/ProtFun/

‡ http://www3.ncbi.nlm.nih.gov/Omim/

§ http://www.ncbi.nlm.nih.gov/LocusLink/

discovery *via* a compendium of expression profiles. *Cell*, **102**, 109–126.

8. Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. & Yeates, T. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
9. Garavelli, J. S., Hou, Z., Pattabiraman, N. & Stephens, R. M. (2001). The RESID database of protein structure modifications and the nrl-3d sequence-structure database. *Nucl. Acids Res.* **29**, 199–201.
10. Blom, N., Gammeltoft, S. & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362.
11. Hansen, J., Lund, O., Tolstrup, N., Gooley, A., Williams, K. & Brunak, S. (1998). NetOglyc: prediction of mucin type *O*-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.* **15**, 115–130.
12. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
13. Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34–36.
14. Riley, M. (1993). Functions of the gene products of *Escherichia coli*. *Microb. Rev.* **57**, 862–952.
15. Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
16. Enzyme Nomenclature (1965). *Recommendations (1964) of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes, Together with their Units and the Symbols of Enzyme Kinetics*, Elsevier, Amsterdam.
17. Nomenclature, E. (1992). *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*, Academic Press, New York. Supplements at http://www.chem.qmw.ac.uk/iubmb/enzyme/supplements/
18. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J. *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29.
19. Chen, C. & Colley, K. (2000). Minimal structural and glycosylation requirements for ST6Gal I activity and trafficking. *Glycobiology*, **10**, 531–583.
20. Cohen, P. (2000). The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem. Sci.* **25**, 596–601.
21. Rechsteiner, M. & Rogers, S. (1996). PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* **21**, 267–271.
22. Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
23. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.
24. Blaschke, C., Andrade, M., Ouzounis, C. & Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein–protein interactions. In *Proc. Seventh Int. Conf. Intelligent Syst. Mol. Biol.*, pp. 60–67, AAAI Press, Menlo Park, CA.
25. Tamames, J., Ouzounis, C., Casari, G., Sander, C. & A, V. (1998). EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**, 542–543.
26. Andrade, M., Ouzounis, C., Sander, C., Tamames, J. & Valencia, A. (1999). Functional classes in the three domains of life. *J. Mol. Evol.* **49**, 551–557.
27. Birney, E., Bateman, A., Clamp, M. & Hubbard, T. (2001). Mining the draft human genome. *Nature*, **409**, 827–828.
28. Collinge, J., Palmer, M., Sidle, K., Hill, A., Gowland, I., Meads, J. *et al.* (1995). Unaltered susceptibility to BSE in transgenic mice expressing human prion protein. *Nature*, **378**, 779–783.
29. Brown, D., Qin, K., Herms, J., Madlung, A., Manson, J., Strome, R. *et al.* (1997). The cellular prion protein binds copper in vivo. *Nature*, **390**, 684–687.
30. Brown, D. (2001). Copper and prion disease. *Brain Res. Bull.* **55**, 165–173.
31. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X., Thompson, M., Marcotte, E. & Eisenberg, D. (2001). DIP: the database of interacting proteins: 2001 update. *Nucl. Acids Res.* **29**, 239–241.
32. Schwikowski, B., Uetz, P. & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnol.* **18**, 1257–1261.
33. Brunak, S., Engelbrecht, J. & Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220**, 49–65.
34. Wootton, J. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285.
35. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

Edited by B. Honig

(Received 17 October 2001; received in revised form 12 April 2002; accepted 16 April 2002)