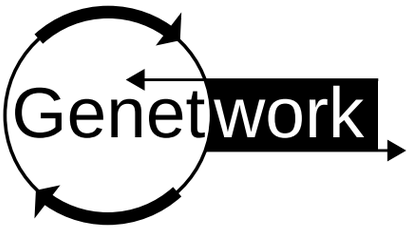

Protein annotation: detective work for function prediction

Computer analysis of genome sequences is currently one of the essential steps for obtaining functional and structural information about the respective gene products. Database searches are used to transfer functional features from annotated proteins to the query sequences. With the increasing amount of data, more and more software robots perform this task[1]. While robots are the only solution to cope with the flood of data, they are also dangerous because they can currently introduce and propagate misannotations[2,3]. On the one hand, functional information is often only partially transferred (underprediction). For example, information is not usually extracted for each functional unit (protein domain) but just taken from the one-line description of the best database match (so multifunctionality is rarely considered). On the other hand, overpredictions are common because the highest-scoring database protein does not necessarily share the same or even similar functions.

Definition and collection of uncharacterized protein families

To avoid unnecessary propagation of poor annotation, we have collected putative, poorly annotated proteins that are usually labeled as 'hypothetical' or just as 'ORF' (open reading frame). We operationally defined uncharacterized protein families (UPFs) to be families of proteins that: (1) contain members in at least three taxonomically distinct (and phylogenetically 'distant') species; and (2) do not contain (to the best of our knowledge) biochemically characterized proteins.

A collection and classification of these proteins should allow: (a) utilization of family information and thus a more detailed characterization; (b) simplification of update procedures for the entire families if functional information becomes available for at least one member; and (c) a careful annotation of functional features that avoids the pitfalls described above.

As the numerous genome sequencing projects progress, more and more of these UPFs emerge in sequence databases. We gave high priority to families that contain members in at least two of the three major kingdoms (archae, eubacteria, eukaryotes). The original 'family' definition was based on significant hits in the statistics provided by FASTA (Ref. 4) or gapped BLAST (Ref. 5).

Annotation of UPFs in SWISS-PROT and PROSITE databases

A serial number has been assigned to each UPF and to each of the corresponding SWISS-PROT (Ref. 6) entries. A SWISS-PROT document file lists all the current UPFs and their members in SWISS-PROT. This document is available on the WWW (Ref. 7). In the majority of cases, PROSITE entries[8] have already been created to document the respective family. Whenever a member of a UPF family is biochemically characterized, that family ceases to be considered as a UPF and is deleted from the list. However, information is provided that allows its history to be traced. For example:

Family: UPF0002 [DELETED]
Taxonomic range: Eubacteria
Comments: Now characterized as a family of pseudouridylate synthases (EC 4.2.1.70).
Prototype: RSUA_ECOLI (Accession No. P33918)
PROSITE entry: PDOC00885

Function prediction for the UPFs

The annotation is handled rather conservatively (see below) because functional overpredictions are most dangerous given the many opportunities for error propagation in sequence database[2,3]. Nevertheless, we intended to retrieve as many functional features as possible for each UPF using comparative analysis. Thus, each UPF was subjected to a variety of sequence analysis methods[9]. In brief, several members of each UPF were compared with a database of non-identical protein sequences, daily updated at the EMBL using PSI-BLAST (Ref. 5) with a conservative expected ratio of false positives (E = 0.001) as a threshold for each iteration. Sequences were preprocessed by filtering for transmembrane[10] and coiled-coil regions[11]. A multiple alignment was constructed for each UPF using ClustalX (Ref. 12). If PSI-BLAST did not identify a relationship to characterized proteins, other iterative methods such as Wisetools (Ref. 13) and Most (Ref. 14) were applied. They also use family information, that is, give more weight to conserved positions and so on, but have the advantage that the underlying multiple alignments can be checked and improved manually (on the cost of speed and the 'easy to use' feature).

Finally, all searches were repeated using a sequence database that only contained sequences from entirely sequenced genomes to reduce noise effects[9,15]. For example, PSI-BLAST E-values depend on the database and a database match might be significant using a small database but becomes insignificant if more background noise (unrelated or redundant sequences) is added.

In many cases, the iterations revealed the relationship of the UPFs with other proteins, families or superfamilies. As the main focus here was to assign functional features, the iterations have not been continued when a reasonable prediction could be made. Criteria for the latter were matches to known active site patterns or conserved motifs resembling those in PROSITE as well as the positioning of UPF members within phylogenetic trees. Transmembrane regions were identified in 13 (22%) of the 58 UPFs, although functional predictions for these 13 have not been made. Of the remaining 45 UPFs, 25 could be related to proteins with annotated functional features (Table 1).

Pitfalls in function assignments

The predictions required careful inspection of the functional annotations of the matched database proteins. To illustrate the difficulties, Table 2 shows the result of a Blast search for UPF0002 that includes quite a few proteins with annotations (in addition to the first hits that are labeled as 'hypothetical'). Only one can give a clue about functional features; others are simply wrong, misleading or uninformative.

Another typical assignment error is caused by the sequence similarity of the query to a region that is independent from the one that was the basis for the annotation. For example, the hypothetical protein HI0722 (Accession No. P44842, ID: YIGZ_HAEIN), a member of the UPF0029 family, shows significant similarity to two proteins (GenBank entries gi|2314657 and gi|2688341) in *Helicobacter pylori* and *Borrelia burgdorferi*, respectively, which are wrongly annotated as proline dipeptidases (pepQ). The annotation is based on the N-terminal homology of these two proteins with the C-terminal region of proline dipeptidase (pepQ) (gi|42358) of *E. coli*, which does not harbor the catalytic activity of this enzyme.

There were even examples in which homologs scored best in PSI-BLAST (Ref. 5) that did not have the same catalytic activity because active site residues of the characterized family were not conserved. However, there were significantly lower scoring homologs with perfect matches of their (distinct) catalytic site residues to the query. For example, the UPF0046 family has clear amino acid similarity to proteases that are easily found by PSI-BLAST (Ref. 5) in the fourth iteration; yet, residues involved in metal-binding are only shared with a purple acid phosphatase family that is only picked up in the ninth iteration. The E-value of 1e–5 compared with proteases (E-value of 5e–78) remain considerably higher in subsequence iterations. Such instances have

Copyright © 1998 Elsevier Science Ltd. All rights reserved. 0168-9525/98/$19.00
PII: S0168-9525(98)01486-3

implications for current function prediction programs in which the function of the best hit is transferred. Clearly, another generation of methods is required that include checks for the presence of functionally important residues.

Use of phylogenetic trees

As most of the database proteins with functional annotations were only distantly related to members of the UPFs, transfer of functional information is extremely difficult and arbitrary. The majority of UPFs turned out to be related to enzymes, and based on the conservation of the active site residues one can assume that at least the basic catalytic mechanism remains the same. This, however, is of little predictive value as some families, e.g. those with the α/β hydrolase fold collected in SCOP (Ref. 16) are huge and harbor numerous distinct catalytic activities, such as lipases, esterases, dehalogenases, peptidases, peroxidases and lyases. We have therefore constructed phylogenetic trees of selected members of the UPFs and of related, but distinct families that have been identified during the analysis (Fig. 1). On some occasions, the UPF members clearly clustered with proteins that all performed the same function (Fig. 1a), but in most of the cases the UPFs were of equal distance to distinct enzymatic activities (Fig. 1b), thus not allowing any detailed predictions.

Although the studied protein families were bound to be difficult for function predictions because a considerable number of teams were unable to find functional

TABLE 1. Predicted functional features for 25 UPFs

UPF No.	Family size[a]	Predicted function
02	70	Pseudouridylate synthase
04	60	Methyltransferase
07	15	Cytidyltransferase[b]
08	30	ATPase
09	40	GTPase
10	10	Aldose 1-epimerase
11	10	Methyltransferase[b]
12	25	Nitrilase
17	30	Hydrolase
19	15	Phosphate-binding protein (TIM BARREL)
20	40	N6-adenine-specific methylase
21	50	ATPase
26	30	Two domain protein : iron/sulfur binding and amidotransferase
30	10	Amidotransferase
31	30	Sugar kinase
34	20	Pyrimidin-binding oxidoreductase (TIM BARREL)
35	20	Mutator mutt protein (7,8-dihydro-8-oxoguaninetriphosphatase)
36	70	Hydrolase
37	10	Oxydoreductase
38	35	ATPase[b]
42	10	ATPase
46	15	Phosphatase
49	50	N6-adenine-specific methylase
53	40	CBS domain protein
55	10	Glutaredoxin

[a]The numbers of family members are approximate because of daily changes in databases and loose family definitions.
[b]*E. coli* member also predicted by Koonin *et al.*[17] (UPF0007: nucleotidyltransferase).
Abbreviation: UPFs, uncharacterized protein families.

TABLE 2. Misleading annotations: PSI-BLAST results for the UPF0002 family (first iteration)

Ranking	Annotation	Probability	Commentary
1	Gnl|PID|e332795 (Z98268) hypothetical protein MTCI125.33 [Mycobacterium tuberculosis]…	(2e–75)	
4	**Sp|P33643|SFHB_ECOLI SFHB PROTEIN**	(1e–67)	SFHB is a gene name (suppressor of the temperature-sensitivity of *ftsh1* mutation) and does not give much functional insight
5	Gnl|PID|e1185138 (Z99112) alternative gene name: *ylmL*; similar to hypothetical proteins [Bacillus subtilis]…	(3e–65)	
37	**Sp|Q12362|RIB2_YEAST DRAP DEAMINASE** >gi|1078332|pir||S50972 RIB2 protein – yeast (Saccharomyces cerevisiae) >gi|642221 (Z21618) DRAP deaminase [Saccharomyces cerevisiae] >gi|1419887|gnl|PID|e252279 (Z74808) ORF YOL066c [Saccharomyces cerevisiae]…	(7e–50)	The homology is not in the catalytic region and does not hold for other deaminases
40	**Sp|P33918|RSUA_ECOLI 16S PSEUDOURIDYLATE 516 SYNTHASE (16S PSEUDOURIDINE 516 SYNTHASE) (URACIL HYDROLYASE)**	(2e–48)	Function prediction based on this protein
41	**Sp|Q47417|YQCB_ERWCA EXOENZYME REGULATION REGULON ORF1** >gi|628643|pir||S45107 hypothetical protein 1 – Erwinia carotovora >gi|496598 (X79474) ORF1 [Erwinia carotovora]…	(7e–48)	Misleading annotation, operon architecture is not conserved between species

Annotations that hamper functional predictions illustrated by the example of the UPF0002 family. Based on the recent experimental characterization of pseudouridylate synthase[18], this family has been deleted from the UPF list (see text). Nevertheless, the various, partly contradictory annotations (bold) are extremely difficult to parse for automatic function prediction programs. For brevity, the PSI-BLAST results have been cut (…).

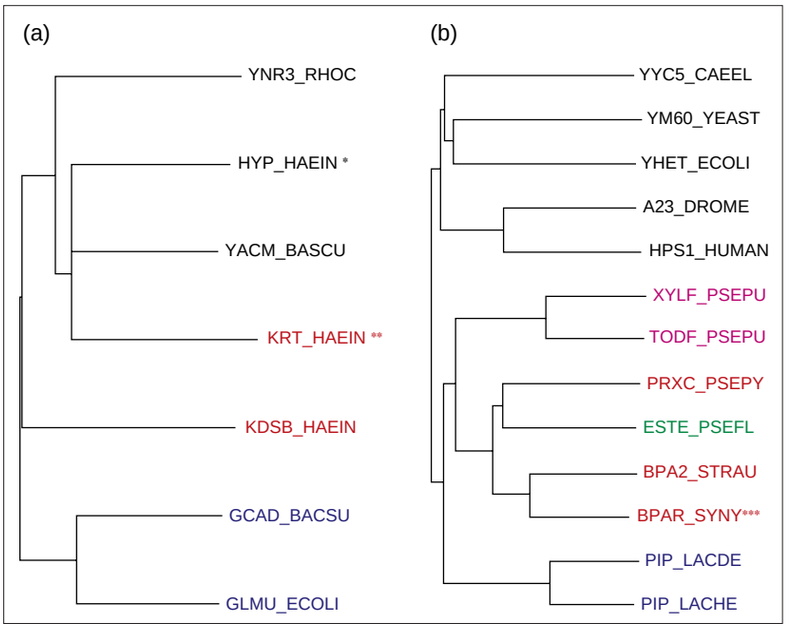

FIGURE 1. (a) Phylogenetic trees of selected members of UPF0007 that indicate a likely function as UPF0007 members with cytidyltransferase activities (red) and related uridilyltransferases (blue) are more divergent (*pir database entry, pir|g64156; **pir database entry, pir|s49238). (b) No clear enzymatic activity can be predicted for UPF0017 members: They clearly have the hydrolase fold but have equal distance to peroxidases (red), esterases (green), peptidases (blue) and other hydrolases (pink) (***GenBank entry gi|1001804). The trees were calculated using CLUSTALX (Ref. 12).

features therein, it is noteworthy that there was not a single case in which we were able to predict the precise mechanism and the substrate specificity. Nevertheless, the information about an enzymatic activity and the likely reaction mechanisms of the 25 UPFs should prove useful for the analysis of upcoming genome sequences.

Annotation with the right level of precision helps in future projects

In summary, we were able to provide some functional annotation for more than 700 of about 1300 proteins clustered in 25 of the 58 distinct UPFs. Most of them are currently named 'hypothetical protein' so that their annotation adds enormous value to these sequences. For another 13 UPFs currently containing about 250 proteins, the presence of transmembrane regions was recorded. This annotation is now being incorporated into PROSITE and SWISS-PROT so that these features can be assigned to newly sequenced genes as well.

The difficulties we faced in assigning functions by sequence similarity also indicate that many of the automatic predictions by most of the software robots are probably erroneous. Because of the current policies of most of the sequence databases, correction of annotations is very hard to realize. Thus, there should be a combined effort by the database teams, the authors of the current entries, and the community, to work towards a careful functional annotation of all the sequences that become publicly available.

References

1 Boguski, M. and McEntyre, J. (1994) *Trends Biochem. Sci.* 19, 71
2 Bork, P. and Bairoch, A. (1996) *Trends Genet.* 12, 425–427
3 Bhatia, U., Robinson, K. and Gilbert, W. (1997) *Science* 276, 1724–1725
4 Pearson, W.R. and Miller, W. (1992) *Methods Enzymol.* 210, 575–601
5 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
6 Bairoch, A. and Apweiler, R. (1998) *Nucleic Acids Res.* 26, 38–42
7 httpd://www.expasy.ch/cgi-bin/lists?upflist.txt
8 Bairoch, A., Bucher, P. and Hofmann, K. (1998) *Nucleic Acids Res.* 25, 217–221
9 Bork, P. and Gibson, T. (1996) *Methods Enzymol.* 266, 162–184
10 Von Heijne, G. (1992) *J. Mol. Biol.* 225, 487–494
11 Lupas, A., van Dyke, M. and Stock, J. (1991) *Science* 252, 1162–1164
12 Thompson, J.D. *et al.* (1997) *Nucleic Acids Res.* 25, 4876–4882
13 Birney, E., Thompson, J.D. and Gibson, T.J. (1996) *Nucleic Acids Res.* 24, 2730–2739
14 Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 91, 12091–12095
15 Bork, P. and Koonin, E.V. (1998) *Nat. Genet.* 18, 313–318
16 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chotia, C. (1995) *Mol. Biol.* 247, 536–540
17 Koonin, E., Mushegian, A., Galperin, M. and Walker, D. (1997) *Mol. Microbiol.* 25, 619–637
18 Wrzesinski, J. *et al.* (1995) *Biochemistry* 34, 8904–8913

Tobias Doerks
doerks@embl-heidelberg.de

EMBL, Meyerhofstrasse 1, 69012 Heidelberg and Max-Delbruck-Center for Molecular Medicine, Berlin-Buch, Germany.

Amos Bairoch
amos.bairoch@medecine.unige.ch

Swiss Bioinformatics Institute and University of Geneva, Switzerland.

Peer Bork
bork@embl-heidelberg.de

TECHNICAL TIPS ONLINE

FREE ACCESS on the entire site

http://www.elsevier.com/locate/tto ♦ http://www.elsevier.nl/locate/tto

Editor Adrian Bird,
Institute for Cell and Molecular Biology at the University of Edinburgh

Protocols are now featured in *Technical Tips Online*, in addition to peer-reviewed Technical Tips articles (novel applications or significant improvements on existing methods). Protocols incorporate all the features that are currently available in Technical Tips articles: comment facility; links to Medline abstracts; product information and so on.

New Core Protocol articles published recently in *Technical Tips Online* include:

- Mitchell, T.J. and Morely, B.J. (1998) **Isolation of RNA and analysis by northern blotting and primer extension** *Technical Tips Online* (http://www.elsevier.com/locate/tto) P01286