Nucleic Acids Research, 2005, Vol. 33, Database issue ***D147–D153***
doi:10.1093/nar/gki120

PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization

Paul Lu*, Duane Szafron, Russell Greiner, David S. Wishart, Alona Fyshe, Brandon Pearcy, Brett Poulin, Roman Eisner, Danny Ngo and Nicholas Lamb

Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8

Received August 15, 2004; Revised and Accepted October 20, 2004

ABSTRACT

PA-GOSUB (Proteome Analyst: Gene Ontology Molecular Function and Subcellular Localization) is a publicly available, web-based, searchable and downloadable database that contains the sequences, predicted GO molecular functions and predicted subcellular localizations of more than 107 000 proteins from 10 model organisms (and growing), covering the major kingdoms and phyla for which annotated proteomes exist (http://www.cs.ualberta.ca/~bioinfo/PA/GOSUB). The PA-GOSUB database effectively expands the coverage of subcellular localization and GO function annotations by a significant factor (already over five for subcellular localization, compared with Swiss-Prot v42.7), and more model organisms are being added to PA-GOSUB as their sequenced proteomes become available. PA-GOSUB can be used in three main ways. First, a researcher can browse the pre-computed PA-GOSUB annotations on a per-organism and per-protein basis using annotation-based and text-based filters. Second, a user can perform BLAST searches against the PA-GOSUB database and use the annotations from the homologs as simple predictors for the new sequences. Third, the whole of PA-GOSUB can be downloaded in either FASTA or comma-separated values (CSV) formats.

INTRODUCTION

Biologists need tools and annotated databases to deal with the volume of genomic and proteomic data. There are more than 1200 complete or partially sequenced genomes in public databases (http://www.ebi.ac.uk/genomes/) and this number is growing rapidly. Given the size and complexity of these datasets, many researchers are compelled to use automated annotation systems to filter, identify or classify individual genes/proteins in their genomic data. A number of systems have been developed over the past few years that permit automated genome-wide or proteome-wide annotation. These include GeneQuiz (1), GeneAtlas (2), EnsEMBL (3), PEDANT (4), Genotator (5), MAGPIE (6) and GAIA (7).

As previously reported, the Proteome Analyst (PA) system (8–10) (http://www.cs.ualberta.ca/~bioinfo/PA/) uses machine learning (ML) techniques to predict various characteristics of a protein, including molecular function and subcellular localization. In particular, PA has high accuracy and coverage for both Gene Ontology molecular function (GO MF) (e.g. accuracy of 96.9% on a training set with 102 225 proteins) and subcellular localization (9) across a wide range of organisms and annotation classes (e.g. cell organelles). In fact, PA's subcellular localization predictions are more accurate and have broader coverage than many other well-known systems, including PSORT-B, LOCKey, SubLoc and TargetP (9). Such annotations are important in understanding the role of proteins in cellular processes. Moreover, identifying the destination or localization of a protein is key both to understanding its function and to facilitating its purification.

After PA was made publicly available, our group received several requests to process the entire proteome of a number of model organisms, such as the human proteome. Since a single organism can require tens of CPU hours of processing, we have now pre-computed the GO MF and subcellular localization (GOSUB) annotations of 10 model organisms (so far) and made the results available. The benefits of PA-GOSUB include:

(i) PA-GOSUB significantly extends the coverage of GOSUB annotations compared with existing databases. For the 10 model organisms currently in PA-GOSUB, there are GO MF annotations for 108 784 proteins and subcellular localization annotations for 107 684 proteins

*To whom correspondence should be addressed. Tel: +1 780 492 7760; Fax: +1 780 492 1071; Email: paullu@cs.ualberta.ca

in a database with over 50 GB of information. In contrast, for the same model organisms, the Gene Ontology Annotation Project (GOA, as on March 15, 2004; http://www.ebi.ac.uk/GOA) has 27 285 GO MF annotations and Swiss-Prot v42.7 contains 21 050 subcellular locations. Therefore, PA-GOSUB extends GO MF coverage by a factor of 4.0 and subcellular localization coverage by a factor of 5.1. Of course, the improvement in coverage varies widely from organism to organism. Still, PA-GOSUB provides high accuracy and broad coverage for both GO MF and subcellular localization.

(ii) PA-GOSUB is searchable by homology. A user can BLAST query sequences against a database containing the sequences of all of the model organisms.

(iii) PA-GOSUB is browsable by annotation. A user can search the model organism database for all proteins that have any particular combination of GO annotation, subcellular localization and words in the FASTA tag line.

(iv) All of the annotations in PA-GOSUB are downloadable in both FASTA and comma-separated values (CSV) format.

(v) The Explain facility, previously described in the context of PA, is also available for the PA-GOSUB results. Consequently, the bioinformatics and machine learning evidence for each of PA-GOSUB's annotations are graphically, intuitively and interactively explained.

In contrast with PA, which is a separate tool, PA-GOSUB is a large database of GO and subcellular localization annotations with web-based search and browsing capabilities that add query functionality to the pre-computed model organisms.

USING PA-GOSUB

The model organisms

The 10 model organisms currently in PA-GOSUB are summarized in Table 1, and other model organisms will be added over time. The number of annotations is shown on a per-organism, per-annotation and per-database basis. For example, of the 4353 protein of the well-annotated *Escherichia coli* proteome, GOA has GO MF annotations for 3524 proteins while PA-GOSUB has annotations for 3772 proteins, for an increased coverage factor of 1.07, which is the lowest increase for a model organism. In contrast, Swiss-Prot v42.7 has only 85 subcellular localization annotations for *Plasmodium falciparum*, while PA-GOSUB has 4275 annotations, for a factor of 50.3 times more coverage. More typically, on a per-organism basis, PA-GOSUB increases GO MF and subcellular localization coverage by a factor of 2–10. As discussed above, over all 10 model organisms, PA-GOSUB increases GO MF and subcellular localization coverage by factors of 4 and 5.1, respectively.

Table 1. Model organisms and annotation coverage in PA-GOSUB

Model organisms	Number of Proteins	GO MF		Subcellular localization	
		GOA	PA-G	SP 42.7	PA-G
Methanobacterium thermoautotrophicum	1868	497	1250	157	1100
Bacillus Subtilis	4105	1534	3187	862	2999
Escherichia coli	4353	3524	3772	2167	3627
Plasmodium falciparum	5257	78	4309	85	4275
Saccharomyces cerevisiae	6195	3017	5049	2024	4978
Drosophila melanogaster	16 371	1535	12 924	1246	12 869
Caenorhabditis elegans	21 821	1459	14 379	933	14 297
Arabidopsis thaliana	26 173	1891	18 338	1528	18 130
Mus musculus	26 556	5520	22 512	4912	22 431
Home sapiens	27 954	8230	23 064	7136	22 978
Total	140 653	27 285	108 784	21 050	107 684

All of the proteomes are from the European Bioinformatics Institute (EBI), except for *Mus musculus* and *Homo sapiens*, which are from the National Center for Biotechnology Information (NCBI). Although there are 140 653 proteins in the model organisms, not all proteins have PA-GOSUB annotations. The main reason for a missing annotation is the lack of relevant homologs in existing bioinformatics databases (e.g. Swiss-Prot), which is important to PA-GOSUB's prediction technique (discussed below). A more detailed breakdown, including proteome-wide statistics, of PA-GOSUB's coverage on a per-organism, per-GO class and per-subcellular localization basis is available at the PA-GOSUB website.

The GOSUB annotations

PA-GOSUB provides annotations for 12 high-level classes (out of the over 7000 possible GO classes) for molecular function: hydrolase activity (0016787), signal transducer activity (0004871), metal-ion binding (0046872), lyase activity (0016829), binding (0005488), structural molecule activity, transporter activity (0005215), transferase activity (0016740), catalytic activity (0003824), nucleic acid binding (0003676), oxidoreductase activity (0016491) and nucleotide binding (0000166). The GO classes have been selected to cover the major branches of the GO hierarchy and to provide sufficient training examples for the ML algorithms used to create PA-GOSUB.

The training set for PA-GOSUB's classifiers for GO MF is based on a combination of data from Swiss-Prot and GOA. To get the largest possible training set, PA-GOSUB takes each of the 141 681 protein sequences in Swiss-Prot v42.7 and maps it to a set of GO classes using the GOA table, if such a mapping exists. The end result is that PA-GOSUB's training set is based on a total of 102 225 proteins mapped from Swiss-Prot to GOA, instead of only the 9621 sequences that have GO annotations in Swiss-Prot v42.7 itself.

For subcellular localization, there are five different ontologies, depending on the type of organism and possible organelles (9). Specifically, there are different classifiers for predicting the subcellular localization of proteins from animals, green plants, fungi, Gram-negative bacteria (GN bact) and Gram-positive bacteria (GP bact). For example, the possible localizations for animal cells are golgi, nucleus, extracellular, mitochondrion, cytoplasm, plasma membrane, lysosome, peroxisome and endoplasmic reticulum.

Browsing PA-GOSUB

PA-GOSUB has a browsable, pre-computed 'PACard' (pre-computed by PA) for each protein, of each model organism—a summary of the predicted properties of each protein specified in the input. A typical PACard is shown in Figure 1. The PACard concept is based on the *E.coli* cards from the

PA Card for protein "gi|4759240|ref|NP_004608.1| transmem..." Card 1649 of 27954

Definition Line	gi|4759240|ref|NP_004608.1| transmembrane 4 superfamily member 4; intestinal and liver (il) tetraspan membrane protein [Homo sapiens]
Sequence	Protein Sequence MCTGGCARCLGGTLIPLAFFGFLANILLFFPGGKVIDDNDHLSQEIW...
Animal SubcellularPrediction	plasma membrane (78.348%) - (Explain)
General Function Prediction	Not hydrolase activity (0016787) (88.667%) - (Explain) signal transducer activity (0004871) (98.4%) - (Explain) Not metal ion binding (0046872) (81.54%) - (Explain) Not lyase activity (0016829) (99.998%) - (Explain) Not binding (0005488) (85.47%) - (Explain) Not structural molecule activity (0005198) (99.999%) - (Explain) transporter activity (0005215) (82.305%) - (Explain) Not transferase activity (0016740) (99.162%) - (Explain) Not catalytic activity (0003824) (97.023%) - (Explain) Not nucleic acid binding (0003676) (99.992%) - (Explain) Not oxidoreductase activity (0016491) (98.752%) - (Explain) Not nucleotide binding (0000166) (99.073%) - (Explain)
Blast	Results P48230 T4S4_HUMAN Transmembrane 4 superfami... 1.0E-123 P30408 T4S1_HUMAN Transmembrane 4 superfami... 3.0E-58 Q64302 T4S1_MOUSE Transmembrane 4 superfami... 8.0E-52

Figure 1. Sample PACard: protein T4S4_HUMAN from *H.sapiens*.

CyberCell Database (CCDB) (http://redpoll.pharmacy.ualberta.ca/CCDB).

In Figure 1, the 'Definition Line' field of the PACard is taken from the tag (comment) line of the FASTA file that was processed through PA. Note that the name of the protein is not encoded in the 'Definition Line', but this protein is T4S4_HUMAN. As per BLAST convention, we have encoded the NCBI 'gi' accession number, so that a hyperlink to the NCBI entry can be generated for the BLAST-able database (discussed below). The 'Sequence' field provides the first few dozen amino acid residues of the protein, with a hyperlink to the full FASTA information.

The 'Animal Subcellular Prediction' field is the first value-added annotation of PA-GOSUB. Since T4S4_HUMAN is from the *H.sapiens* reference sequence dataset, PA's animal subcellular classifier was used to predict that this protein is localized to the plasma membrane, with a Naïve-Bayes (NB) probability of 78.348%. Note that, for T4S4_HUMAN, the probability is not 100% (which is common with ML techniques and many of PA's annotations); the Swiss-Prot entry for T4S4_HUMAN does not actually specify plasma membrane, but a manual inspection of the bioinformatics evidence strongly suggests that it is indeed localized to the plasma membrane. Furthermore, a hyperlink to 'Explain' provides the evidence for that prediction (data not shown).

The 'General Function Prediction' field is another value-added annotation that provides the classification of the GO MF from the 12 classes that we currently support. In our example, PA predicts that the query protein is a member of two GO classes: 'signal transducer activity (0004871)' and 'transporter activity (0005215)'. Since it is possible for proteins to have more than one molecular function, as per the GO ontology, PA-GOSUB makes molecular function predictions on a per-class basis. Thus, for each of the 12 GO classes, the prediction is either 'yes' or 'no' and annotated as, e.g. 'Not hydrolase activity (0016787)' when the protein does not belong in that class. For each of the 12 predictions, an 'Explain' hyperlink provides the evidence for the prediction.

Therefore, there are a total of 13 explainable predictions (i.e. exactly 1 for subcellular localization and 12 for GO MF).

The last field, 'Blast', of the PACard shows the top three Swiss-Prot homologs of the query. The top homolog here is the actual query protein itself, T4S4_HUMAN. As discussed below and elsewhere (8), PA-GOSUB relies on homologs of the query protein to provide machine-learning features of the classification computation. A hyperlink in the 'Blast' field provides access to the standard BLAST information.

Searching PA-GOSUB

All the proteins of all the model organisms have been included in a BLAST-able database. PA-GOSUB supports a BLAST search (Figure 2) against this database as a simple way to locate the closest homolog to the user's query protein and as a simple (i.e. nearest neighbor) predictor of the GOSUB properties of the query protein. The user can optionally

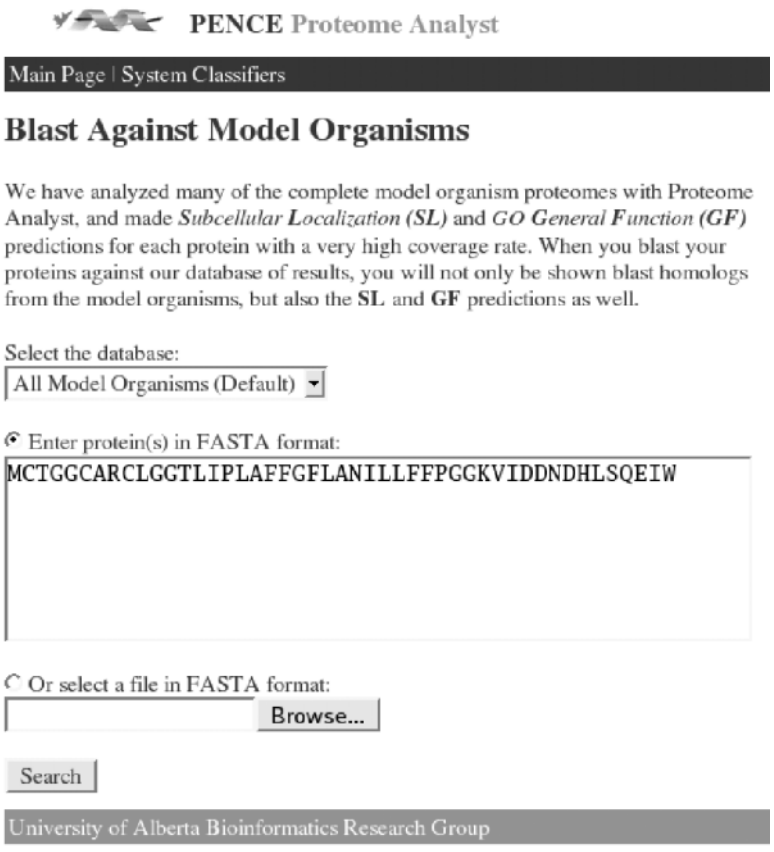


Figure 2. BLAST searches against the model organisms.

BLAST against all model organisms (shown in Figure 2) or specific model organisms (data not shown).

It is also possible to search the PACards for proteins that match a text string and other criteria. Figure 3 shows part of a PA Card Set (i.e. summary of all PACards that match the search) with the string 'kinase' in the FASTA tag. In the example, only the proteins for *Saccharomyces cerevisiae* are searched. In fact, the search is limited to proteins (Figure 4) that have been annotated as having 'Nucleus' for subcellular localization and 'Nucleic Acid Binding' for GO MF, in addition to having 'kinase' in the FASTA tag. The summaries of the first two PACards (i.e. #364 and #515) are shown, along with their tag information and annotations. Hyperlinks from the PACard Set point to the PACards for the individual proteins, as discussed earlier.

Downloading PA-GOSUB

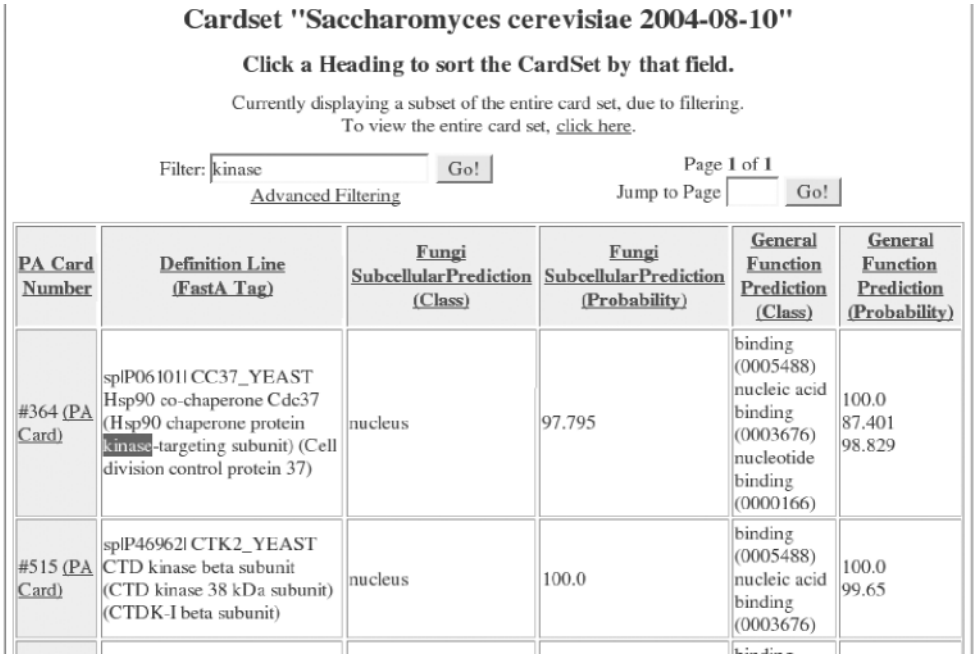
Although PA-GOSUB provides a variety of browsing and search features, researchers may wish to use PA's annotations with other tools. Therefore, PA-GOSUB annotated proteins can be downloaded in FASTA format, where the annotations are encoded in the tag line, or in CSV format so they can be imported into a spreadsheet.

PREDICTION TECHNIQUES IN PA-GOSUB

As discussed in previous publications (8,10), PA-GOSUB and PA make extensive use of machine-learned classifiers to predict annotations. As shown in Figure 5, classification-based prediction is a two-step process: training/learning and prediction. In the training/learning step, a classifier is built using an ML algorithm by analyzing a set of training sequences, each tagged by a known class label. In the prediction step, the generated classifier is used to predict the class label of an unknown query sequence.

In PA, each training item consists of a primary protein sequence and the ontological class it has been assigned by an expert. In general, an ML classifier algorithm requires features to be associated with each training item. Note that PA is given only the primary sequence of the protein; the features are automatically computed by the system. Once built, a classifier takes a protein sequence with unknown class and uses the values of these features (see below) to predict its class. Specifically, PA uses a pre-processing step that maps each sequence to a set of features, as shown in Figure 6.

First, the sequence is compared to the Swiss-Prot database using BLAST. Second, the Swiss-Prot entries of (up to) three top homologs (whose *E*-values are <0.001) are parsed to extract a feature set from the Swiss-Prot KEYWORDS field and any Interpro numbers (11) contained in the DBSOURCE field. The union of the features for the selected homologs forms the feature set. If no homologs match the *E*-value cutoff or if all features are removed by feature selection then the sequence has no features, so no prediction is made. The feature set is then used as input for both the training and classification phases of PA. In essence, PA learns a mapping from feature sets to classes (or 'annotations').



Cardset "Saccharomyces cerevisiae 2004-08-10"

Click a Heading to sort the CardSet by that field.

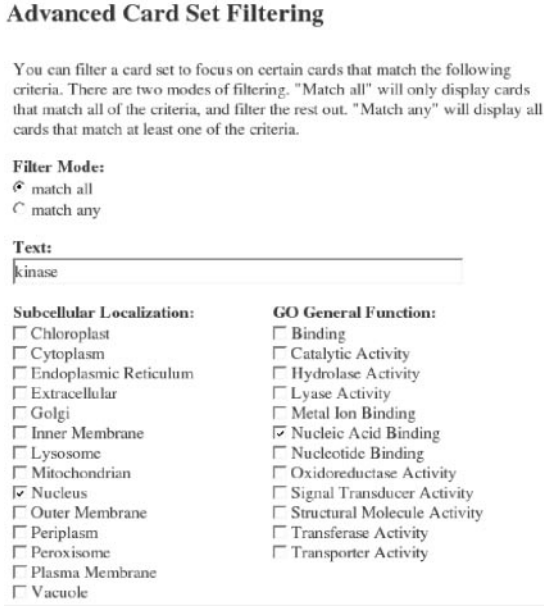
Currently displaying a subset of the entire card set, due to filtering.
To view the entire card set, click here.

Filter: kinase Go! Advanced Filtering

Page 1 of 1 Jump to Page Go!

PA Card Number	Definition Line (FastA Tag)	Fungi SubcellularPrediction (Class)	Fungi SubcellularPrediction (Probability)	General Function Prediction (Class)	General Function Prediction (Probability)
#364 (PA Card)	sp|P06101|CC37_YEAST Hsp90 co-chaperone Cdc37 (Hsp90 chaperone protein kinase-targeting subunit) (Cell division control protein 37)	nucleus	97.795	binding (0005488) nucleic acid binding (0003676) nucleotide binding (0000166)	100.0 87.401 98.829
#515 (PA Card)	sp|P46962|CTK2_YEAST CTD kinase beta subunit (CTD kinase 38 kDa subunit) (CTDK-I beta subunit)	nucleus	100.0	binding (0005488) nucleic acid binding (0003676)	100.0 99.65

Figure 3. Searching and filtering: part of a PACard Set matching the criteria.



Advanced Card Set Filtering

You can filter a card set to focus on certain cards that match the following criteria. There are two modes of filtering. "Match all" will only display cards that match all of the criteria, and filter the rest out. "Match any" will display all cards that match at least one of the criteria.

Filter Mode:
(•) match all
( ) match any

Text:
kinase

Subcellular Localization:
- [ ] Chloroplast
- [ ] Cytoplasm
- [ ] Endoplasmic Reticulum
- [ ] Extracellular
- [ ] Golgi
- [ ] Inner Membrane
- [ ] Lysosome
- [ ] Mitochondrian
- [x] Nucleus
- [ ] Outer Membrane
- [ ] Periplasm
- [ ] Peroxisome
- [ ] Plasma Membrane
- [ ] Vacuole

GO General Function:
- [ ] Binding
- [ ] Catalytic Activity
- [ ] Hydrolase Activity
- [ ] Lyase Activity
- [ ] Metal Ion Binding
- [x] Nucleic Acid Binding
- [ ] Nucleotide Binding
- [ ] Oxidoreductase Activity
- [ ] Signal Transducer Activity
- [ ] Structural Molecule Activity
- [ ] Transferase Activity
- [ ] Transporter Activity

Figure 4. Searching and filtering: selecting criteria.

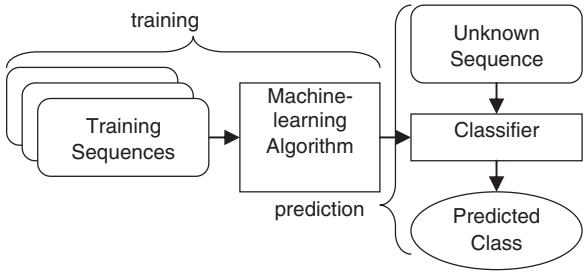


Figure 5. The training and predicting phases of classification.

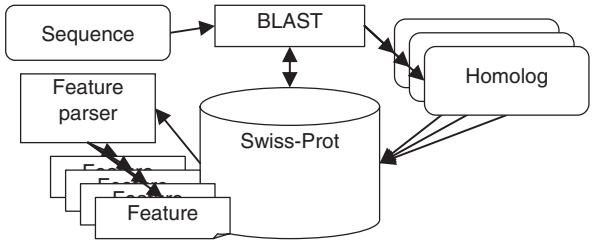


Figure 6. The feature extraction algorithm for a protein sequence in PA.

EXAMPLE OF EXPLAINABILITY

While it is necessary for a protein prediction tool to be accurate, it is also important that it can clearly explain its predictions to the user. A clear and intuitive explanation helps biologists to develop confidence in the annotations in the database.

Explaining a Prediction/Classification

PA-GOSUB provides an explanation mechanism to help users understand why a classifier makes a particular classification (10). In this discussion, we will use the T4S4_HUMAN protein as an example.

If the user clicks the Explain hyperlink for the 'signal transducer activity (0004871)' annotation of the PACard (Figure 1), then an Explain page is displayed. Although there are many elements to an Explain page, an important part is the bar graph (Figure 7).

First, note that the two stacked bars in the graph represent the evidence for both an 'yes' (bottom bar) and a 'no' (top bar) prediction for the class. Each of its five colored sub-bars

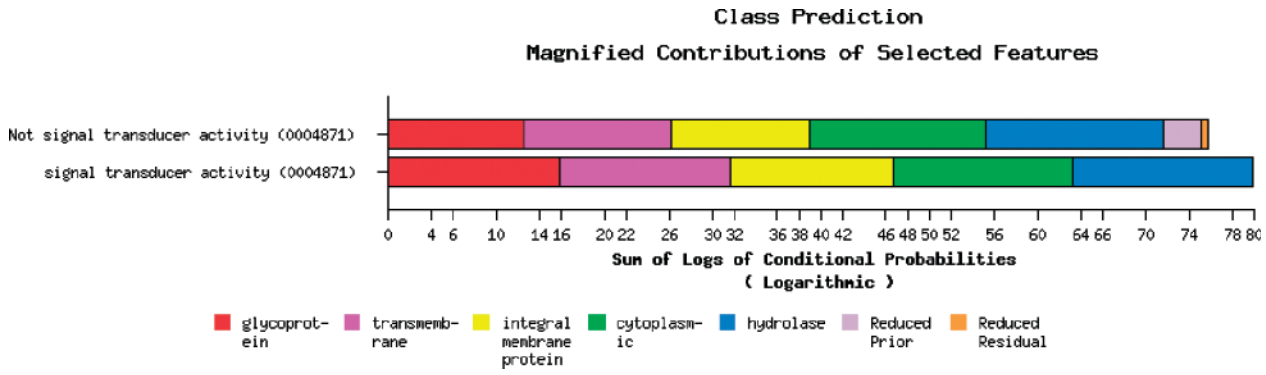


Figure 7. Part of the explain page for T4S4_HUMAN, signal transducer activity annotation.

correspond to the presence or absence of a selected, significant token. The absence or presence of a token is known as a feature. In this example, the tokens 'glycoprotein', 'transmembrane' and 'integral membrane protein' are present for T4S4_HUMAN, but 'cytoplasmic' and 'hydrolase' are absent. Again, note that the absence of a token can also be evidence for or against a particular classification.

Second, note that the scale on the *x*-axis is logarithmic, where each composed bar on a single line represents the logarithm (base 2) of the combined probability that the protein is either in the class or not. For example, the 'no'/top bar is ~76 units long and the 'yes'/bottom bar is ~80 units long. Thus, the prediction is that T4S4_HUMAN is in the class 'signal transducer activity (0004871)'. The difference of 4 units means that the ratio of the probabilities is $\sim 2^{(80-76)} \approx 16$, which is correct based on other quantitative information on the Explain page (data not shown). The logarithm is used so that the contributions to the probabilities represented by each feature can be added. Additive quantities can be visualized using stacked bar graphs.

The (red) 'glycoprotein' sub-bar occurs in both bars, but it is significantly longer (especially considering the logarithmic scale) for the 'yes'/bottom bar. The same observation is true for the (yellow) 'integral membrane protein' sub-bar. Both 'glycoprotein' and 'integral membrane protein' are examples of features, extracted from the top three BLAST homologs (Figures 1 and 6), that support the prediction of 'signal transducer activity (0004871)'. Further details on the mathematics behind and interpretation of the Explain page can be found elsewhere (8,10).

SUMMARY

Annotating proteins using bioinformatics and computational techniques can be an important aid in filtering the vast amounts of raw genomic and proteomic data. Annotations for the general function or subcellular localization of specific proteins can help in hypothesis generation and in selecting a specific protein isolation protocol. PA-GOSUB extends the coverage of existing databases, by a factor of over 5.1 (and growing) with respect to subcellular localization, by annotating all of the proteins for 10 model organisms. New model organisms are regularly added to PA-GOSUB. A total of over 107 000 proteins have been annotated, for both GO molecular function and subcellular localization, and there are plans to add more proteins and model organisms, as they become available.

In addition, PA-GOSUB is browsable, searchable, and downloadable. A simple, web-based interface provides access to the PACards of all the proteins, including all of the annotations, explanations for the annotations and information about the homologs to each protein. A special BLAST database has been constructed so that new, unknown query proteins can be compared with the proteins of all the model organisms. The tags and annotations can also be searched. Lastly, it is possible to download (and thus use any other analysis tool on) the PA-GOSUB database in either FASTA or CSV formats. PA-GOSUB is publicly available at http://www.cs.ualberta.ca/~bioinfo/PA/GOSUB.

ACKNOWLEDGEMENTS

We are grateful to Fiona Brinkman and Jennifer Gardy for many helpful pointers, ideas and data for subcellular localization. This research was partially funded by research or equipment grants from the Protein Engineering Network of Centres of Excellence (PENCE), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Alberta Ingenuity Centre for Machine Learning (AICML), Sun Microsystems, Silicon Graphics Inc., and the Alberta Science and Research Authority (ASRA).

REFERENCES

1. Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
2. Kitson,D.H., Badretdinov,A., Zhu,Z.Y., Velikanov,M., Edwards,D.J., Olszewski,K., Szalma,S. and Yan,L. (2002) Functional annotation of proteomic sequences based on consensus of sequence and structural analysis. *Brief. Bioinformatics*, **3**, 32–44.
3. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res*., **30**, 38–41.
4. Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.
5. Harris,N.L. (1997) Genotator: a workbench for sequence annotation. *Genome Res*., **7**, 754–762.
6. Gaasterland,T. and Sensen,C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet*., **12**, 76–78.

7. Overton,G.C., Bailey,C., Crabtree,J., Gibson,M., Fischer,S. and Schug,J. (1998) The GAIA software framework for genome annotation. *Pac. Symp. Biocomput.*, 291–302.
8. Szafron,D., Lu,P., Greiner,R., Wishart,D.S., Poulin,B., Eisner,R., Lu,Z., Anvik,J., Macdonell,C., Fyshe,A. and Meeuwis,D. (2004) Proteome analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res.*, **32**, W365–W371.
9. Lu,Z., Szafron,D., Greiner,R., Lu,P., Wishart,D.S., Poulin,B., Anvik,J., Macdonell,C. and Eisner,R. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**, 547–556.
10. Szafron,D.R., Greiner,P., Lu,D., Wishart,C.MacDonell,J., Anvik,B., Poulin,Z., Lu,Z. and Eisner,R. (2003) Explaining naive Bayes classifications. TR03-09, Department of Computer Science, University of Alberta.
11. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.