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

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

© 2005, the authors

Nucleic Acids Research, Vol. 33, Database issue © Oxford University Press 2005; all rights reserved
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.

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 (0001666). 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, 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.

Browsing PA-GOSUB

PA-GOSUB has a browsable, pre-computed ‘PACard’ (precomputed 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

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

<table>
<thead>
<tr>
<th>Model organisms</th>
<th>Number of Proteins</th>
<th>GO MF GOA</th>
<th>Subcellular localization GOA</th>
<th>SP 42.7</th>
<th>PA-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanobacterium thermoautotrophicum</td>
<td>1868</td>
<td>497</td>
<td>1250</td>
<td>157</td>
<td>1100</td>
</tr>
<tr>
<td>Bacillus Subtilis</td>
<td>4105</td>
<td>1534</td>
<td>3187</td>
<td>862</td>
<td>2999</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4353</td>
<td>3524</td>
<td>3772</td>
<td>2167</td>
<td>3627</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>5257</td>
<td>78</td>
<td>4309</td>
<td>85</td>
<td>4275</td>
</tr>
<tr>
<td>Saccharomyces cerevisae</td>
<td>6195</td>
<td>3017</td>
<td>5049</td>
<td>2024</td>
<td>4978</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>16371</td>
<td>1535</td>
<td>12924</td>
<td>1246</td>
<td>12869</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>21821</td>
<td>1459</td>
<td>14379</td>
<td>933</td>
<td>14297</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>26173</td>
<td>1891</td>
<td>18338</td>
<td>1528</td>
<td>18130</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>26556</td>
<td>5520</td>
<td>22512</td>
<td>4912</td>
<td>22431</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>27954</td>
<td>8230</td>
<td>23063</td>
<td>7136</td>
<td>22978</td>
</tr>
<tr>
<td>Total</td>
<td>140653</td>
<td>27285</td>
<td>107878</td>
<td>21050</td>
<td>107684</td>
</tr>
</tbody>
</table>
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 Naive-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...
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’).
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 prediction (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
correspond to the presence or absence of a selected, significant token. The absence or presence of a token as 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 ~2^40/76 ~ 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).

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


