W714–W719 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl228

The Path-A metabolic pathway prediction web server

Luca Pireddu, Duane Szafron*, Paul Lu and Russell Greiner

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

Received February 14, 2006; Revised March 6, 2006; Accepted March 27, 2006

ABSTRACT

Pathway Analyst (Path-A) is a publicly available web server (http://path-a.cs.ualberta.ca) that predicts metabolic pathways. It takes a FASTA format file containing a set of query protein sequences from a single organism (a partial or complete proteome) and identifies those sequences that are likely to participate in any of its supported metabolic pathways (currently 10). Path-A uses a number of machine-learning and sequence analysis techniques (e.g. SVM, BLAST and HMM) to predict pathways. Each machine-learned classifier exploits similarity between sequences in the pathways of its model organisms and sequences in the query set. It predicts the pathways that are present in the query organism and annotates each predicted reaction and catalyst, using the appropriate sequences from the query set. Path-A also provides a browsable and searchable database of the pathways for the model organisms that are used to make its predictions. Path-A's predictor sets (using different classifier technologies) have been evaluated using standard cross-validation techniques on a dataset of 10 metabolic pathways across 13 model organisms—a total of 125 organism-specific pathways. The most accurate classifier technology obtained a mean precision of 78.3% and a mean recall of 92.6% in predicting all catalyst proteins, of all reactions, in all pathways present in the dataset. Although Path-A currently only supports metabolic pathways, the underlying prediction techniques are general enough for other types of pathways. Consequently, it is our intent to extend Path-A to predict other types of pathways, including signalling pathways.

INTRODUCTION

Each biochemical pathway describes an identifiable subset of the complex system of reactions that exist in living organisms. Understanding these pathways is essential to understanding the machinery of life. The existence of a large quantity of proteomic sequence data, and the hard work of experimental molecular biologists, has resulted in several on-line pathway databases: KEGG PATHWAY (1), BioCarta (http://www.biocarta.com), aMAZE (2) and BioCyc (3). In addition, tools such as Pathway Tools (4) have appeared, providing some data-mining capabilities that try to correlate protein annotations to pathway templates so that organism-specific pathways can be derived. However, we are unaware of any freely available tools that perform protein analyses at the primary sequence level, without requiring annotations, to predict organism-specific pathway variants, or *pathway instances*.

Pathway Analyst (Path-A) is a web server that takes a set of protein sequences from a single organism as input. First, Path-A predicts which of these proteins catalyse each reaction in a set of selected metabolic pathways. Second, it assembles the predicted catalysts and reactions into an organism-specific variant of each metabolic pathway under consideration. In addition, Path-A makes it easy for users to browse and search known metabolic pathways in model organisms. Path-A is a freely available service at http://path-a.cs.ualberta.ca.

PREDICTION TECHNIQUE

Path-A uses a set of machine-learning techniques to predict pathway instances (5). The basic prediction algorithm is based on the realization that instances of the same pathway in different organisms will share some degree of similarity, in both the reactions they encompass and the primary sequence structure of the proteins that catalyse them. Path-A exploits this fact to predict the reactions and catalysts of pathways unknown to the system based on patterns learned from pathway instances that are already known to the system. The basic algorithm has two inputs: a set of protein sequences from the query organism and a set of model pathways, one for each target pathway. For each model pathway, the algorithm iterates through each reaction. For each reaction, the algorithm tries to determine which proteins from the query set are functionally compatible to any catalyst protein of the model reaction. Functional compatibility is predicted using one of several sequence-based measures described later. If at least one protein from the query set is deemed functionally compatible to a catalyst protein for the given model reaction, the reaction is predicted to exist in the pathway for the query organism. The catalysts of the reaction

*To whom correspondence should be addressed. Tel: +1 780 492 5468; Fax: +1 780 492 1071; Email: duane@cs.ualberta.ca

in the query organism are predicted to be the proteins from the query set that were predicted to be compatible to the catalyst sequences from the model reaction. If no functionally compatible protein is found for any catalyst of that model reaction, the reaction is predicted not to exist in the pathway of the query organism.

Model pathways

Path-A's prediction algorithm requires a pathway model for each metabolic pathway that it supports. Typically, a single pathway instance is not a good pathway model for two reasons. First, the reactions that compose each pathway—the pathway's *structure*—vary between organisms. Using only one training pathway increases the chance that the training pathway has a different structure than the target pathway that Path-A is trying to predict. If the target pathway has a different structure, predicting its true structure is impossible since the algorithm does not attempt a prediction on any reaction not found in the model pathway. Second, predictors trained using only a few positive training instances typically have poor accuracy. Using more than one pathway instance to construct a model pathway increases the number of catalysts for reactions in the model pathway. This increase in the number of positive training instances (catalyst sequences) produces more accurate predictors.

Therefore, we create a single abstract model pathway as the union of all available organism-specific instances of that pathway. The *union*, U, of two pathways, A and B, is a new pathway whose structure includes all the reactions occurring in either A or B. For each reaction in pathway U, if that reaction existed in both A and B, then the reaction's protein catalyst set in U is the union of the catalyst sets from the same reaction in A and B.

By using model pathways, the pathway prediction algorithm can even predict instances of a pathway with variations in structure that were never observed in the training pathway set—and perhaps never found in any physical laboratory. Such emergent structures can be computationally predicted before being observed.

A Path-A user can examine each model pathway in the system. The organism-specific pathway instances that were merged to create each model can be viewed, along with each reaction in the model pathway and the pathway instances. The user can drill down to examine the catalyst proteins of each reaction, either as they appear in the abstract model pathway, or in each individual organism-specific reaction that comprises the model reaction. Path-A currently provides abstract models for 10 metabolic pathways, spanning 125 organism-specific pathway instances. Each model reaction is annotated by information about which classifier is used to predict functional compatibility between its catalyst sequences and catalyst sequences in query organisms. As it is typically advantageous to use more training data, the model pathways in Path-A are built using all available instances of the pathway, though this is not required for the correct operation of the algorithm. In addition, since Path-A requires a model of a pathway to make a prediction, it is currently only able to make predictions on these 10 available metabolic pathways. We plan to support more metabolic pathways and modify existing model pathways by augmenting them with reactions and catalysts from additional organism-specific instances of these pathways.

Classifiers

A machine-learned classifier is a component of Path-A's prediction system that predicts whether a protein from a query protein set is a catalyst of a particular reaction. A single classifier is created for each reaction in a model pathway, and trained to recognize functionally compatible proteins associated with this reaction. Training the classifier requires both positive and negative training sequences. We use sequences that are known to catalyse the reaction in any of the organism-specific pathways comprising the model as positive training sequences, and the rest of the sequences in those organisms as negative training sequences. To make a prediction, the trained classifier decides whether a query protein sequence is functionally closer to the positive or the negative class of proteins.

We have evaluated several classifiers as predictors of which proteins are functionally compatible (5). A key research result, established using Path-A, is that no single classifier is best for all pathway reactions. Some are based on sequence alignment techniques such as BLAST (6). Others use machine-learning techniques such as profile hidden Markov models (HMMs) (7) and support vector machines (SVMs) (8). Some classifiers use combinations of these simple classifiers. Complete descriptions of these classifiers are not germane to this paper and can be found elsewhere (5). Table 1 shows a sampling of accuracy measures of six different classifiers we tested. Each entry shows the mean and standard deviation of each accuracy measure for each kind of classifier computed in n-fold cross-validation tests.

The tests required the prediction of 125 pathway instances with a total of 1759 reactions. There is no reason why the same classifier must be used for each reaction in a pathway. For example, the test results in the top two rows of Table 1 use classifiers whose parameters were tuned differently for each reaction. The current production version (v1.1) of Path-A allows the user to select a BLAST-based, HMM-based, or combination BLAST–HMM-based classifier for each pathway. This classifier choice and its parameter values are then fixed across all reactions of that pathway. However, we are in the process of replacing these generic classifiers with reaction-specific classifiers to provide best prediction performance. We migrate our best classifiers from the experimental versions to the production version of Path-A as we discover them, so the user need not be concerned with the intricacies of classifier and parameter selection.

Table 1. Different classifiers in Path-A: mean catalyst prediction scores for each classifier type (standard deviation given in parentheses)

Classifier	*F*-measure	Precision	Recall
Opt BLAST	0.837 (0.130)	0.783 (0.170)	0.926 (0.114)
Opt HMM	0.795 (0.141)	0.777 (0.184)	0.848 (0.138)
BLAST–HMM	0.673 (0.152)	0.630 (0.197)	0.784 (0.176)
BLAST	0.667 (0.155)	0.609 (0.205)	0.802 (0.170)
Motif SVM	0.659 (0.155)	0.666 (0.190)	0.692 (0.187)
HMM	0.654 (0.164)	0.704 (0.190)	0.671 (0.221)

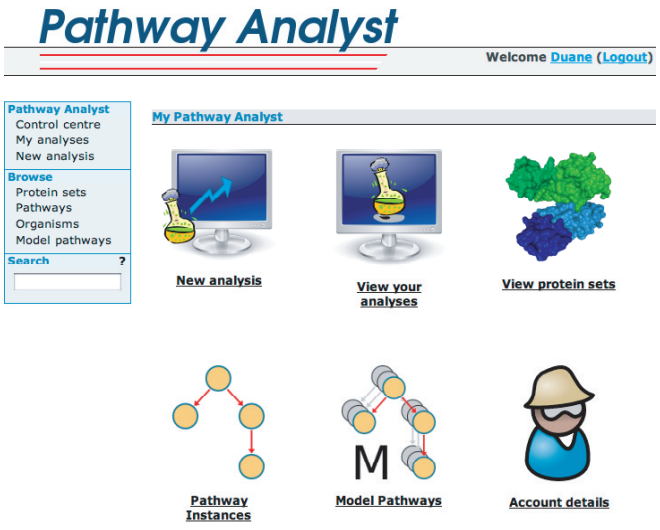

Figure 1. Path-A services in the Control centre.

Accessing pathway analyst

Path-A can be accessed at http://path-a.cs.ualberta.ca. A user can register to obtain a personal account, or login using a Guest account with no registration required. Registering on the Path-A system is free and registered users are provided with a personal space where they can store protein sets and predicted pathways. In addition, Path-A will notify registered users by email when a prediction task is complete (if desired). The Guest account has access to all services available on a regular personal account, except that data in the Guest account are only kept for five days and no personalized features, such as email notification, are available.

After logging into the Path-A system, the user can select from the six services shown in the Path-A 'Control centre' (Figure 1). The 'Account details' service allows a registered user to change user name, password and email contact information. The 'New analysis' and 'View your analyses' services are used to create and view one or more analyses, and the 'View protein sets', 'Pathway instances' and 'Model pathways' services allow the user to browse and search the database. Clicking the appropriate link starts the corresponding service.

NEW ANALYSIS

The user creates a new pathway analysis of a partial or complete proteome from a specific organism in three steps: (1) start by entering an analysis name and optional description, (2) select a protein set and (3) select pathways of interest.

Step 1: Start The user starts by providing a name for the new analysis and an optional description as shown in Figure 2. The road-map at the bottom of the page indicates the current step in the context of the analysis process. Path-A uses *protein sets*, where all the proteins of the set must come from the same organism. The user presses the '*Click* here to select or upload a protein set' button shown in Figure 2 to begin the process of selecting a protein set.

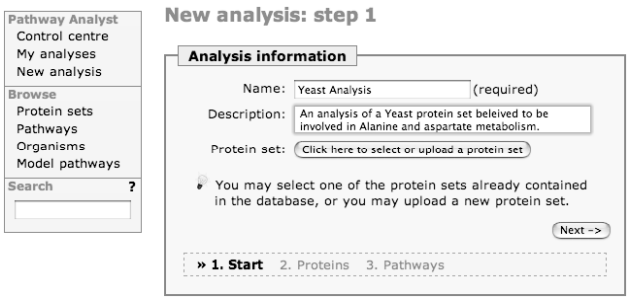

Figure 2. New analysis: Step 1—Start.

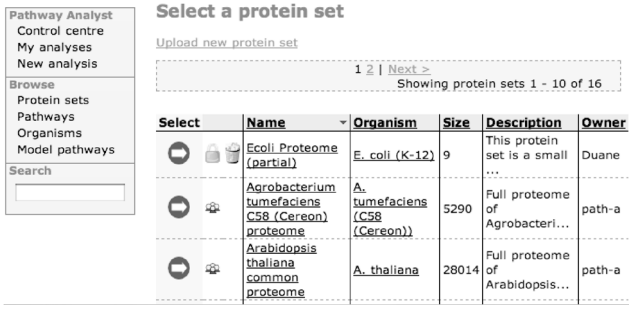

Figure 3. New analysis: Step 2—Proteins.

Step 2: Proteins The user may upload a new protein set, use a protein set that the user previously uploaded or use 1 of the 15 complete proteomes provided in the Path-A database. Figure 3 shows the page used to select a protein set or upload a new one. The protein sets uploaded previously by the user are shown first, followed by the complete proteomes. Some basic information about the protein sets (name, source organism, size, description and owner) are shown for each set. For example, Figure 3 shows a user uploaded protein set named *Ecoli proteome (partial)* on the first line of the table. It is marked with a lock to indicate that it is part of the user's private data and that it is invisible to other Path-A users. To select a protein set, the user clicks on the corresponding selection arrow. To upload a new protein set, the user clicks on the 'Upload new protein set' link.

If the user opts to upload a new protein set, the 'Upload a new protein set' page shown in Figure 4 appears. The user enters a name for the new protein set and (optionally) a brief description that provides useful identification if the protein set is re-used for another analysis in the future. The user then provides the name of the source organism of the protein set for display purposes. This is performed by clicking on the button 'Click here to select an organism'.

A prepared list of almost 81 000 organisms is used. The user can quickly search through the list by typing part of the organism's name into the search field at the top of the page, as shown in Figure 5. Using dynamic AJAX technology, the list's contents are automatically updated to show only the matching organisms. The user clicks on the organism name to select it. The user is then asked to choose the organism's strain from a short list (not shown). If the desired organism or its strain name is not included in the Path-A database, a new organism name

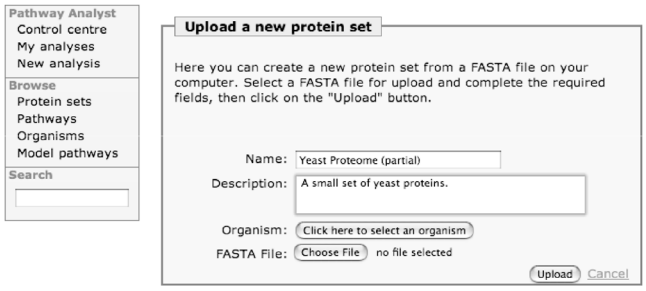

Figure 4. New analysis: Step 2—Proteins: upload a new protein set page.

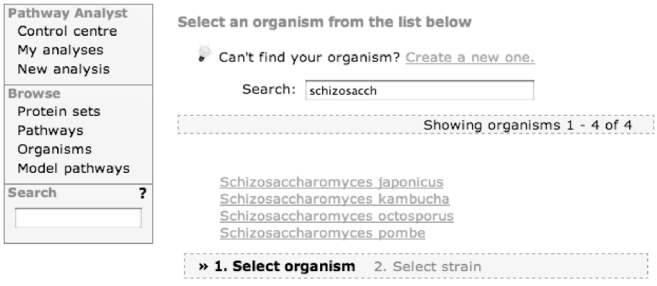

Figure 5. New analysis: Step 2—Proteins: select an organism page.

can be added by clicking on the 'Create a new one' link at the top of the 'Select an organism' page shown in Figure 5.

After selecting the organism, the user is returned to the 'Upload a new protein set' page of Figure 4, except that the 'Organism:' line has its 'Click here to select an organism' button replaced by the text, *Schizosaccharomyces pombe, common (Fission yeast)*, and a different button labelled 'Select a different organism'. The user can then select a local file containing the protein set by clicking on the 'Choose File' button, and browsing to the desired local file (not shown). At this point the 'no file selected' text in Figure 4 is replaced by the file name. The user clicks on the 'Upload' button to upload the protein set. The progress bar shown in Figure 6 informs the user of the status of the file upload.

After the newly uploaded protein set has been stored in the Path-A database, the user can re-use it for future analyses or browsed without the need to upload it again. At this point, Path-A returns to the 'New analysis' page at Step 2 as shown in Figure 7. The user can click on the 'Next' button to proceed to Step 3.

Step 3: Pathways The Path-A algorithm requires a model for every predicted pathway. Path-A currently provides models for the 10 metabolic pathways shown in Figure 8. The user can choose to predict any number of these pathways for the query protein set. Selecting a pathway results in the retrieval of the pathway's model, classifiers and prediction parameters. The retrieved data are then used to perform the prediction. The 'More details' button can optionally be pressed to select a non-default predictor (currently BLAST, HMM, BLAST-HMM or Opt-HMM instead of Opt-Blast) for each pathway. After selecting the desired pathways, the user clicks on the 'Start analysis' button.

After going through the three steps of the 'New analysis' process, Path-A sends the analysis request to its computation

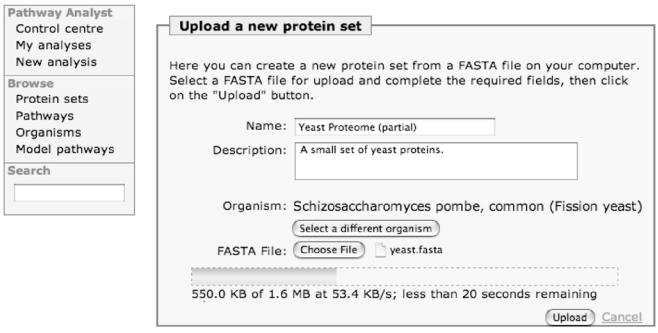

Figure 6. New analysis: Step 2—Upload a new protein set revisited.

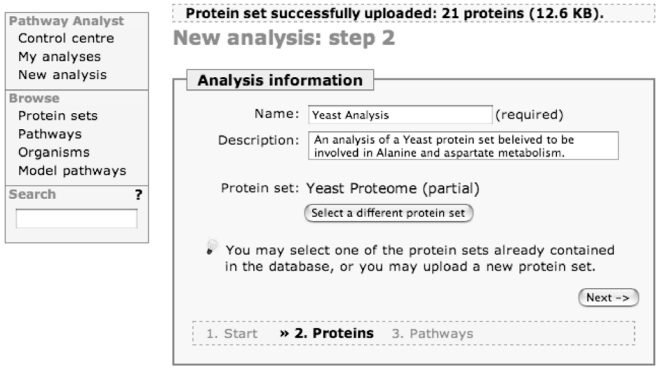

Figure 7. Protein set uploaded.

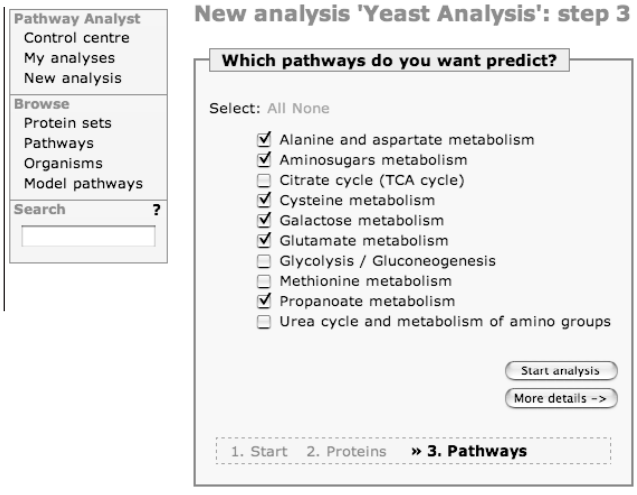

Figure 8. New analysis: Step 3—Which pathways page.

nodes. The time required to perform an analysis depends on the number of pathways to be predicted, the type of classifier being used and the amount of activity on the system. Time to completion can range from minutes to a day. Given the unpredictability in the computation time, users may choose to have Path-A notify them by email when their analyses complete and become ready for viewing. Notification can be set/unset using the 'Account details' service from the 'Control centre' shown in Figure 1.

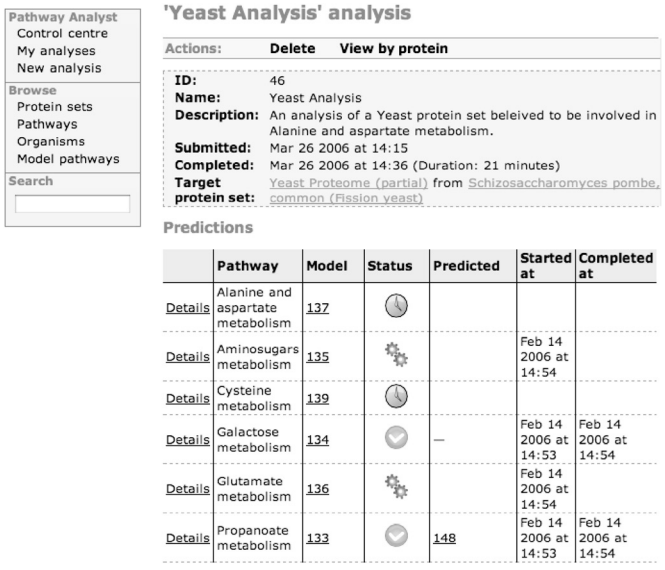

Figure 9. Viewing an analysis.

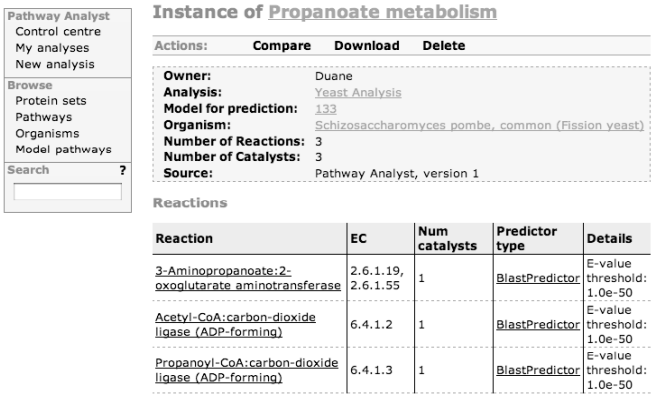

Figure 10. A predicted pathway instance.

ANALYSIS VIEWING

A user can view the results of analyses by clicking on the 'View your analyses' button in the Path-A control centre shown in Figure 1. This can be a new analysis that has just completed or a previously computed analysis. The 'Your Analyses' page (not shown) has two lists. One list contains the analyses that are waiting to be computed or are currently being computed. The second list contains the analyses that are complete. Clicking on the name of a completed analysis shows its details. For example, Figure 9 shows the details of an analysis of a partial Yeast proteome.

Each pathway prediction is an independent task, shown as a separate line on the page. Each line contains the name of the pathway, a link to the model used for the predictions and some information about the computation, including its status (done, computing, pending or error). The user can click on the 'Details link' to see more information about how a pathway prediction was carried out.

The 'Prediction details' page (not shown) explains which predictor type and parameters were used to predict each of the model pathway's reactions, whether the reaction is predicted to exist in the query organism, based on the query protein set. Each predicted reaction also has a link to information about that reaction including a list of proteins from the query protein set that are predicted to catalyse that reaction.

The user can also click on a 'Predicted' link (e.g. 148 for Propanoate metabolism in Figure 9) to learn more about the predicted pathway. A dash (–) in the 'Predicted' column indicates that the pathway does not exist (no reactions in that pathway exist) for the query organism, based on the protein set that was provided. For example, Figure 10 shows a predicted pathway in the query organism (Yeast) that is restricted to proteins from the query set. The top section of the page includes some information about the analysis that generated this pathway: the account name, the name of the analysis, the model pathway used, the organism, the number of reactions predicted and the number of catalysts found in the query set for these reactions. The following section lists each predicted reaction, along with the number of proteins predicted to catalyse it, as well as the predictor that was responsible for the prediction. Clicking on the reaction name (not shown) shows its details, including all the proteins predicted to be its catalysts.

SEARCHING AND BROWSING

Path-A provides a text-based search engine to help locate pathways of interest. Users are able to search through both user-provided and Path-A-provided pathway instances and model pathways. The search mechanism accepts simple text queries, as well as Boolean operators, on names and descriptions of pathways, organisms, reactions, molecules, genes and identifiers from external sites (e.g. UniProt, KEGG, TAIR and FlyBase).

Path-A also provides browsing features to easily organize and navigate the information in different ways. Users can quickly search a list of organisms, and then see which pathways and protein sets associated with it exist in the database. Further, given a protein one can easily see the pathway instances that are associated with it.

CONCLUSION

We have presented Path-A, a web server for the prediction of metabolic pathways. Path-A predicts the pathways supported by arbitrary sets of proteins, using validated prediction techniques based on sequence alignment and machine learning. In our tests, our most accurate predictor achieved a mean precision and recall of 78.3 and 92.6%, respectively. Path-A can also be used as a pathway database, complete with browsing and searching functionality. Path-A is freely available for use at http://path-a.cs.ualberta.ca.

ACKNOWLEDGEMENTS

We thank Jordan Patterson and Stephen Walsh for their contributions to the Pathway Analyst web server implementation, and the entire Proteome Analyst group at the University of Alberta for the continual discussions, advice and ideas. This work is partially funded by research or equipment grants from

the Canadian Protein Engineering Network of Centres of Excellence (PENCE), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Informatics Circle of Research Excellence (iCORE), the Alberta Ingenuity Centre for Machine Learning (AICML), Sun Microsystems, Silicon Graphics, Inc. and the Alberta Science and Research Authority (ASRA). Funding to pay the Open Access publication charges for this article was provided by Alberta Ingenuity Centre for Machine Learning (AICML).

Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
2. Christian,L., Antezana,E., Couche,F., Fays,F., Santolaria,X., Janky,R., Deville,Y., Richelle,J. and Wodak,S.J. (2004) The aMAZE lightbench: a web interface to a relational database of cellular processes. *Nucleic Acids Res*., **32**, D443–D448.
3. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*., **19**, 6083–6089.
4. Karp,P., Paley,S. and Romero,P. (2002) The pathway tools software. *Bioinformatics*, **18**, S225–S232.
5. Pireddu,L., Poulin,B., Szafron,D., Lu,P. and Wishart,D.S. (2005) Pathway analyst-automated metabolic pathway prediction. In *Proceedings of the IEEE 2005 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE Press, La Jolla California, USA, pp. 243–250.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
7. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
8. Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, NY.