Developing a Database for Proteomic Analysis of Extracytosolic Plant Proteins

Yang Wang[1], Osmar R. Zaïane[1], Randy Goebel[1], Jennafer L. Southron[2], Urmila Basu[2]
Randy M. Whittal[3], Julie L. Stephens[2], Gregory J. Taylor[2]
[1]*Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G 2E8, Canada*
[2]*Department of Biological Sciences, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada*
[3]*Department of Chemistry, University of Alberta, Edmonton, Alberta, T6G 2G2, Canada*
[1]{*wyang, zaiane, goebel*}*@cs.ualberta.ca*
[2,3]{*jsouthro, ubasu, randy.whittal, julieste, gregory.taylor*}*@ualberta.ca*

Abstract

Extracytosolic plant proteins are involved in numerous processes including nutrient acquisition, communication with other soil organisms, protection from pathogens, and resistant to disease and toxic metals. We have developed an on-line database to provide the plant biology community with relevant information about extracytosolic plant proteins. Data for Brassica napus (canola) proteins identified using proteomics tools can be accessed using this database. Several textual and graphical query capabilities allow biologists to populate and query this database. Results are displayed with active links to other databases. The system has an open API allowing other applications to access this database as a Web service. In addition, the database is augmented with a repository of tools that can be used in data analysis and mining tasks.

1. Introduction

A proteome represents the proteins that are expressed in a specific biological unit at a particular time and under a particular set of conditions. Proteomics utilizes a diverse set of tools to display, identify, and investigate the proteins in a proteome. The results of proteomic studies are commonly displayed in on-line databases. Links to many 2D-PAGE database servers and 2-D PAGE related servers and services can be found at WORLD-2DPAGE[1] and efforts have been made to establish a set of federated databases[4] that are maintained independently, but are linked together through the World Wide Web (WWW).

We have developed an on-line database to provide the plant biology community with relevant information about extracytosolic plant proteins. Extracytosolic plant proteins are involved in numerous processes including nutrient acquisition, communication with other soil organisms, protection from pathogens, and resistance to disease and toxic metals. Insofar as these proteins are strategically positioned to play a role in resistance to environmental stress, we are using proteomic tools (two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), liquid chromatography-tandem mass spectrometry (LC-MS/MS), *de novo* sequencing, and bioinformatics) to analyze extracellular proteins. These proteins are collected from *Brassica napus* (canola) plants grown hydroponically in a sterile environment modified from previous work[2, 5].

The database contains the 2D maps showing the protein locations, and descriptions of the identified proteins. Cross references are provided to SWISS-PROT/TrEMBL[6], which is the largest annotated protein database in the world.

What distinguishes our database from other 2D-PAGE databases is that it not only provides a Web interface for querying using a client browser, but also provides Web services that allow other applications to make function calls over HTTP and use XML as a message transfer format to be consumed by the clients.

Our database also aims to provide tools that facilitate more sophisticated data analysis and data mining tasks. To achieve this goal, our database is built as a framework that allows users to submit not only queries, but also data and tools for experimenting with various data analysis and data mining tasks. For example, one tool that we are currently developing is for predicting extracellular proteins from amino acid sequences.

The sections that follow elaborate on the design and implementation of our database. Section 2 introduces the overall architecture of our database. Section 3 discusses the implementation of search functions. Section 4 introduces the Web services implemented as part of our database. Section 5 describes the tools added to our database for experiment-

1 http://ca.expasy.org/ch2d/2d-index.html

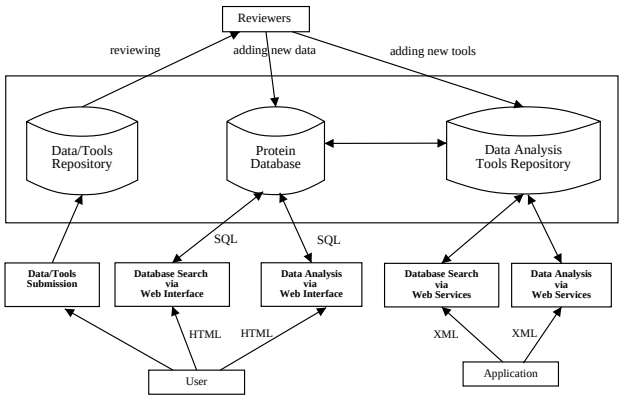

Figure 1. Architecture of our database

ing with various data analysis and data mining tasks. Section 6 concludes the paper and points to future work.

2. System Architecture

The overall architecture of the our database is shown in Figure 1. The protein database contains all the protein entries. The format of protein entries is similar to that in SWISS-PROT/TrEMBL[6] and SWISS-2DPAGE[10]. Each entry is composed of defined lines, used to record various kinds of data. Each line begins with a two character line code, which indicates the type of data contained in the line[2]. An example of a protein entry is shown in Figure 2. Several lines are specific to our database: (i) the DB line lists an accession number specific to our database. Each entry in our database has a unique DB line; (ii) if applicable, the IS line lists the isozyme(s) (identified by their DB accession numbers) of a protein. In addition, the 2-D map associated with a protein entry displays the experimental location of the protein on the chosen map (Figure 3).

The protein entries in the database can be queried in two ways. First, a user can query the database through a Web interface using a client browser ("Database Search via Web Interface" in Figure 1). The Web interface provides the user several textual and graphical query methods. Second, our database can also be queried by another application via the Web services that are provided as part of the system ("Database Search via Web Services" in Figure 1). These two query methods greatly enhance the interoperability of our system.

In addition to database queries, we also augment our database by maintaining a repository of tools for experimenting with various data analysis and data mining tasks, such as characteristic rule mining, sequential pattern analysis and association rules. These tools provide users with the power of intelligently retrieving and analyzing data across a large array of heterogeneous data sets. For example, one tool that we are currently developing is the automatic identification of extracellular proteins from amino acid sequences. Adding a repository of tools to our database is of paramount importance for the long-term usefulness of our system. Similarly, the tools in the repository can be accessed by users through the Web interface or by other applications through the Web services.

Another novel aspect of our database is that it allows

2 http://ca.expasy.org/sprot/userman.html

```
ID Q9M8Y9; PRELIMINARY; 2DG.
AC Q9M8Y9;
DT 30-MAY-2003 (Rel. 01, Created)
DT 30-MAY-2003 (Rel. 01, Last update)
DE Putative trypsin inhibitor.
GN T6K12.5.
OS Arabidopsis thaliana (Mouse-ear cress).
OC Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae;
OC eurosids II; Brassicales; Brassicaceae; Arabidopsis.
OX NCBI_TaxID=3702;
MT <i>Brassica napus</i> Extracellular Proteome.
IM <i>Brassica napus</i> Extracellular Proteome.
RN [1]
RP MAPPING ON GEL.
RA Basu U., et al.;
RT ''Using proteomics to establish an Extracytosolic Plant Proteins Database'';
RL Unpublished observations (MAY-2003).
CC -!- SUBCELLULAR LOCATION: Secretory pathway signal peptide (predicted by
TargetP; RC 1).
CC -!- PTM: SignalP predicts most likely cleavage site to be between pos. 21
and 22 (TSG-VV).
CC -!- MISCELLANEOUS: Predicted pI 5.90.
2D -!- MASTER: BRASSICA_NAPUS_EXTRACELLULAR_PROTEOME;
2D -!- PI/MW: SPOT 00001=5.0/23000;
2D -!- MAPPING: SPOT 00001: LC-MS/MS.
2D -!- PEPTIDE SEQUENCES: SPOT 00001: FANPSKCGESGVWR; VANGEVVLNGVESR;
2D CPHQPVMF; SCKGSLSWETGAAEGN; LLPSSTV.
DR TrEMBL; Q9M8Y9; Q9M8Y9.
DB 00001.
IS 00002;
SQ SEQUENCE 202 AA; 22914 MW; 485A2C8472CD3792 CRC64;
KW
//
```

Figure 2. An example of a database entry

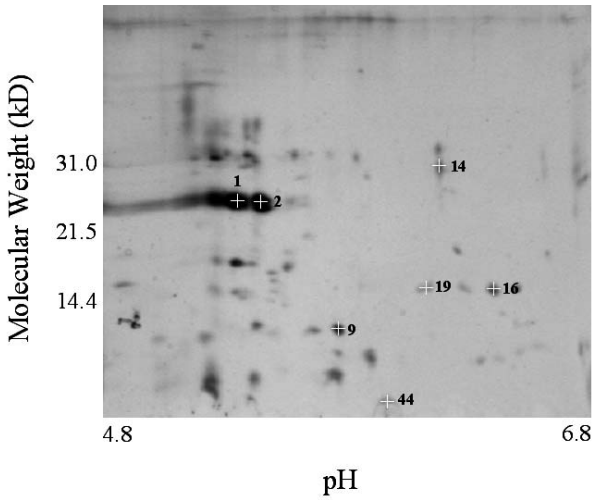

Figure 3. An example of a 2D map showing the locations of proteins identified in a gel

users to submit their new data and new tools. The data or tools submitted by users are maintained in a repository (see Figure 1). These data/tools are reviewed to make sure they do not contain inconsistencies or errors. If the data/tools pass the reviewing process, they are integrated into our database. In this way, our database acts as an "information hub" for biologists all over the world who are working on extracytosolic plant proteins and expedite information exchange and sharing among them.

3. Database Construction

Many existing 2D-PAGE databases use *Make2ddb* package[9] to build a 2D-PAGE database on one's own Web server. The main focus of *Make2ddb* is on ease of use. However, we choose not to use it in our case for the following reasons:

- *Make2ddb* uses text files rather than database management systems to manage the data, which might cause performance problems when the amount of data is huge.
- The queries generated by *Make2ddb* are fixed, i.e., it only allows searching by description (DE or ID line), by accession number (AC line), by clicking on a spot, and by author (RA line). However, in our database, we want to allow users to submit more sophisticated queries.
- *Make2ddb* does not generate an API for the 2D-PAGE database it creates. That means other applications cannot communicate with the database easily. In our database, we provide an API using SOAP technology[3] to allow inter-operations between applications.

We choose to use MySQL[4] as our database management system since it is a free and powerful relational database management system. The query processing is implemented in PHP[5], a scripting language especially suited for Web application development. PHP and MySQL combination is cross-platform and is commonly used for creating data-driven Web sites.

The user can query the database in a variety of ways. The "quick text search" (Figure 4) is currently set up to retrieve entries that contain the specified keyword(s) in the "DE" line. The user can also query the database by clicking on a spot in a gel image (similar to other 2D-PAGE databases). In addition, if a user chooses "advanced text search", the attributes (text lines) to be searched can be explicitly specified.

3 http://www.w3.org/TR/SOAP
4 http://www.mysql.com/
5 http://www.php.net/

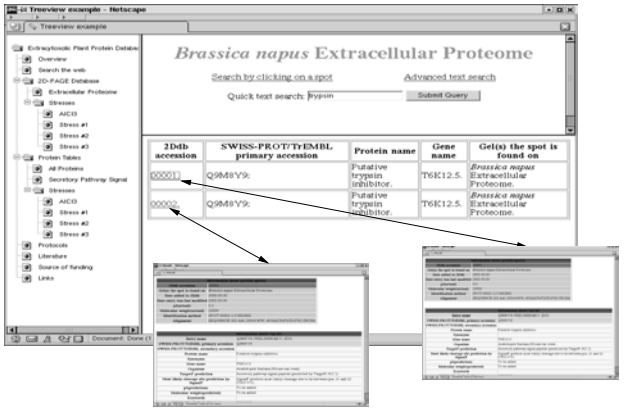

Figure 4. Quick text search

4. Web Services

Most other protein databases focus on building applications that are made globally available through a Web server, defining their user interfaces with HTML, and can be accessed using client browsers. The applications do not take advantage of the Internet to make the services available to a variety of clients. An example is SWISS-PROT. If other protein databases need to link to SWISS-PROT, they must generate pointers to the information related to SWISS-PROT entries. However, SWISS-PROT entries are dynamically generated HTML pages and their layouts could be inconsistent. If proper formats are not used, there is no way for other databases to directly exploit the information provided by SWISS-PROT entries in HTML format.

The idea of Web Services is instead of supplying information via dynamically generated user interfaces (HTML) that are fairly fixed and can only be consumed by client browsers, the server makes available a series of function calls over HTTP and uses XML as a message transfer format to be consumed by the clients. This gives much more freedom for the clients to use the services in whatever way they wish, since XML is a standard format for structured documents and data on the Web that is platform and language independent. In addition, if functions are called using standard HTTP-based protocols over the Internet then the client that calls them can be located anywhere on the Internet. There are no restrictions on what platform it might be running on or in which language it should be written.

Simple Object Access Protocol (SOAP) is a universally agreed upon standard protocol for calling up the functions exposed in web services. With SOAP as an XML messaging specification, web services enable developers to target a range of clients and build the services from local and remote resources. A SOAP transaction begins with an application making a call to a remote procedure. The SOAP client encodes the procedure as an XML document and sends it to

a server script. The server parses the request and passes it to a local method, which returns a response. The response is encoded as XML by the server and returned to the client, who parses the response and passes the result to the original function. So as long as other application support XML and SOAP, they can communicate with Web Services in our database and get the answer right away[14].

The search functions involved in our database are encapsulated into Web services. All Web services are implemented and published using NuSOAP[6], a freely available toolkit which provides a simple API for building Web services using Simple Object Access Protocol (SOAP) technology[7]. Other applications can remotely call these functions via XML, without knowing how the search results are presented in HTML format. All the data analysis tools are also encapsulated into Web services. The implementation of Web services provides an appropriate solution for transparently integrating applications from heterogeneous sources.

5. Data Analysis Tools

One unique feature of our database is that it not only provides standard database query capabilities, but also provides domain-specific tools for experimenting with various data analysis and data mining methods. Here we give an example of a data mining tool we are currently developing for our database, which is to predict extracellular proteins from amino acid sequences.

In order to function properly, proteins need to be localized at proper locations within the cell or transported to the extracellular environment. The set of locations depend on the type of a cell. For biologists, a very important question is to determine whether a given protein is an extracellular protein or non-extracellular protein (i.e., intracellular protein). This problem is under the scheme of protein subcellular localization prediction[7]. Most existing localization prediction methods use supervised learning algorithms to learn a classifier from a set of training data containing both intra- and extra-cellular protein sequences. When a new protein sequence comes in, the learned classifier is used to predict the correct label for the sequence.

The problem of extracellular protein prediction poses several challenges for most existing localization prediction algorithms: (i) the effectiveness of most existing algorithms is measured by overall accuracy. However, since extracellular proteins are extremely rare compared with intracellular proteins (less than 1%), predicting every protein as intracellular protein can achieve very high accuracy level of 99%; (ii) biologists are usually interested in those tools that provide some explanations of the prediction, i.e., they want a classifier to let them know why a protein is or is not predicted as extracellular. Many existing algorithms (e.g., Artificial Neural Networks, Support Vector Machines) are "black box" techniques, in the sense that they predict the class label for a given protein sequence without providing any easily interpretable justification. Even if the prediction is correct, biologists may be hesitant in using such non-transparent tools. If the prediction is incorrect, biologists are given no hints to identify the locations in a classifier's reasoning process that might cause the misclassification.

In our system, we use an associative classifier[3] to solve the extracellular protein prediction problem. The basic idea of associative classifiers is to build a classifier model based on association rules. The rules that are built for classification have the form "*set of features⇒class_label*".

There are three main steps in building an associative classifier:

- *Generating the set of association rules*. This step aims at finding a set of rules that satisfy the pre-specified minimum *support* and *confidence*[1], in such a way that the concequent of the rules is always a class label.
- *Selecting rules for classification*. The first step can generated a huge amount of rules. In this step, a subset of rules that are most informative are selected.
- *Building the classifier*. The last step is to build a classifier based on the selected rules. The task is how to use the rules to make an all-round decision.

It has been shown that associative classifiers work well in many applications[3, 12, 13]. One advantage of associative classifiers is that the rules used for classification can be easily understood and modified by human experts, thus providing some explanations about the prediction.

In our application, the classifier is trained on two features of protein sequences, one is amino acid composition[15], the other is frequent subsequences[16]. In order to deal with the problem of rare class, we plan to use the following two strategies:

- Use PNrule induction framework[11]. This framework discovers a set of positive rules (P-rules) that predict the presence of the target class (extracellular proteins), and a set of negative rules (N-rules) that predict absence of the target class. The model is learned in two phases. The first phase discovers a set of P-rules that cover most of the positive cases for the target class, while the second stage discovers a set of N-rules that cover most of the false positive cases covered by the union of P-rules discovered in the first stage. This two-stage rule induction provides a solution for the applications where different classes have different distributions.

6 http://dietrich.ganx4.com/nusoap/index.php
7 http://www.w3.org/TR/soap

- Use boosting[8] methods. Boosting is a general method with a theoretically justified ability to improve the performance of any learning algorithm. Boosting algorithms work in iterations. In each iteration, a weak classifier is learned on a weighted distribution of the training examples. After each iteration, the weights of all the training examples are updated to force the weak learner to focus more on the misclassified examples in the next iteration. The algorithm stops after a number of iterations or after some predefined performance measure starts to deteriorate.

6. Conclusion and Future Work

In this paper we have introduced an on-line 2D-PAGE database we have developed for extracytosolic plant proteins. Similar to other 2D-PAGE databases, users are provided with textual and graphical queries to search the database, and the results are displayed with links to other databases (SWISS-PROT/TrEMBL in our case). What is unique about our database is that it also provides Web services that allow other applications to access the database via XML. The implementation of Web services is proved to be an appropriate solution for information exchange and sharing among applications from heterogeneous sources.

We also augment our system by adding tools for experimenting with various data mining methods. With the ever-increasing amount of biological data, data mining will be of great importance for the long-time usefulness of our database.

Our database is still in its early stage of development. Currently we are working on the extracellular protein prediction tool to be added to our database. In the future, we plan to incorporate additional useful tools, e.g., sequence alignment, protein structure prediction, protein function prediction, etc. We also plan to register our database as a member of the Federated 2-D electrophoresis databases[4].

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proceedings of ACM SIGMOD Conference on Management of Data*, pages 207–216, 1993.

[2] V. M. Anoop, U. Basu, M. T. McCammon, L. McAlister-Henn, and G. J. Taylor. Modulation of citrate metabolism alters aluminum tolerance in yeast and transgenic canola overexpressing a mitochondrial citrate synthase. *Plant Physiology*, 132:2205–2217, 1999.

[3] M. L. Antonie. *Categorizing Digital Documents by Associating Content Features*. Master thesis, Department of Computing Science, University of Alberta, 2002.

[4] R. D. Appel, A. Bairoch, J. C. Sanchez, J. R. Vargas, O. Golaz, C. Pasquali, and D. F. Hochstrasser. Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. *Electrophoresis*, 17(3):540–546, 1996.

[5] U. Basu, A. G. Good, T. Aung, J. J. Slaski, A. Basu, K. G. Briggs, and G. Taylor. A 23kd, aluminum-binding, root exudate polypeptide co-segregates with the aluminum-resistant phenotype in *Triticum aestivum*. *Physiologia Plantarum*, 106:53–61, 1999.

[6] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and S. M. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31:365–370, 2003.

[7] K. C. Chou and D. W. Elrod. Protein subcellular location prediction. *Protein Engineering*, 12(2):107–118, 1999.

[8] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[9] C. Hoogland, V. Baujard, J. C. Sanchez, D. F. Hochstrasser, and R. D. Appel. Make2ddb: A simple package to set up a two-dimensional electrophoresis database for the world wide web. *Electrophoresis*, 18:2755–2758, 1997.

[10] C. Hoogland, J. C. Sanchez, L. Tonella, P. A. Binz, A. Bairoch, D. F. Hochstrasser, and R. D. Appel. The 1999 swiss-2dpage database update. *Nucleic Acids Research*, 28:286–288, 2000.

[11] M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining needles in a haystack: Classifying rare classes via two-phase rule induction. *Proceedings of 2001 ACM SIGMOD Conference*, pages 91–102, 2001.

[12] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. *Proceedings of 2001 International Conference on Data Mining*, 2001.

[13] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 80–86, 1998.

[14] B. Marchal. *Applied XML Solutions*. SAMS publishing, 2000.

[15] H. Nakashima and K. Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of Molecular Biology*, 238(1):54–61, 1994.

[16] R. She, F. Chen, K. Wang, M. Ester, J. L. Gardy, and F. S. L. Brinkman. Frequent-subsequence-based prediction of outer membrane proteins. *Proceedings of 2003 International Conference on Data Mining and Knowledge Discovery*, 2003.