A data management system for structural genomics

Raymond, Stephane; O'Toole, Nicholas; Cygler, Miroslaw

Publisher's version / Version de l'éditeur:
https://doi.org/10.1186/1477-5956-2-4
Biomed, 10, pp. 1477-5956, 2004

Proteome Science

BioMed Central

Research

Open Access

A data management system for structural genomics

Stéphane Raymond[1,3], Nicholas O'Toole[2,3] and Miroslaw Cygler*[1,2,3]

Address: [1]Biotechnology Research Institute, National Research Council, 6100 Royalmount Avenue, Montréal, Québec H4P 2R2, Canada, [2]Department of Biochemistry, McGill University, Montréal, Québec H3G 1Y6, Canada and [3]Montréal Joint Centre for Structural Biology, Montréal, Québec, Canada

Email: Stéphane Raymond - stephane@BRI.NRC.CA; Nicholas O'Toole - nicholas@BRI.NRC.CA; Miroslaw Cygler* - mirek@BRI.NRC.CA

* Corresponding author

Published: 21 June 2004

Proteome Science 2004, **2**:4 doi:10.1186/1477-5956-2-4

Received: 09 March 2004

Accepted: 21 June 2004

This article is available from: http://www.proteomesci.com/content/2/1/4

Abstract

Background: Structural genomics (SG) projects aim to determine thousands of protein structures by the development of high-throughput techniques for all steps of the experimental structure determination pipeline. Crucial to the success of such endeavours is the careful tracking and archiving of experimental and external data on protein targets.

Results: We have developed a sophisticated data management system for structural genomics. Central to the system is an Oracle-based, SQL-interfaced database. The database schema deals with all facets of the structure determination process, from target selection to data deposition. Users access the database via any web browser. Experimental data is input by users with pre-defined web forms. Data can be displayed according to numerous criteria. A list of all current target proteins can be viewed, with links for each target to associated entries in external databases. To avoid unnecessary work on targets, our data management system matches protein sequences weekly using BLAST to entries in the Protein Data Bank and to targets of other SG centers worldwide.

Conclusion: Our system is a working, effective and user-friendly data management tool for structural genomics projects. In this report we present a detailed summary of the various capabilities of the system, using real target data as examples, and indicate our plans for future enhancements.

Background

Structural genomics (SG) initiatives aim to determine thousands of protein structures at an unprecedented rate [1-4]. These projects have been initiated partly in response to the massive amount of information that continues to be generated by the genome sequencing projects. Since the sequence-structure-function paradigm underpins modern biology, it would be invaluable for the discovery of protein structural information to proceed at a pace comparable to that of the sequence data. Most of the SG initiatives select as protein "targets" hundreds or thousands of the open reading frames (ORFs) of particular genomes. These groups have been developing high-throughput techniques for all steps in the experimental structure determination pipeline (e.g. cloning, expression, crystallization), with the goal that a large proportion of their targets will yield protein structures in a short period of time. It is anticipated that the new high-throughput experimental techniques being developed will also affect "conventional" structural biology laboratories, focusing on particular proteins or biochemical systems of interest, by enabling them to rapidly process many more cloned

constructs or crystallization conditions, for example, than was possible in the past.

In common with the genome sequencing projects, the rapid accumulation of large amounts of data by an SG project renders the conventional method of archiving and tracking experimental data via laboratory notebooks highly inefficient. The problem is particularly acute in structural biology because each step of the experimental pipeline involves different techniques and results. There is therefore a need for computerized experimental data information management systems in structural biology, and for structural genomics projects in particular. Such systems, often called Laboratory Information Management Systems (LIMS), have been developed in the past decade for genomics laboratories involved in sequencing and microarray analysis [5,6]. Structural genomics presents unique requirements for data tracking systems and these are outlined in detail in the following section.

We have developed an experimental data management system for structural genomics, the SPEX Db (Structural Proteomics EXperimental Database). The system serves both as a LIMS and also as a tool for SG target selection and management. It follows a standard three tier client/server architecture, using Oracle 8i for the database tier and Netscape iPlanet Web Sever (Enterprise Edition 4.1) for the server tier. The client tier is expected to be a standard web browser such as Microsoft Internet Explorer or Netscape Navigator. The SPEX Db can accommodate any type of protein targets and is currently used by over 10 structural biology projects throughout Canada. The primary user of the system is the Montreal-Kingston Bacterial Structural Genomics Initiative (M-KBSGI; http://sgen.bri.nrc.ca/brimsg/bsgi.html), which selects as targets the ORFs of bacterial genomes such as *E. coli* K12 and the pathogenic *E. coli* strains O157 and CFT073. In this paper we describe the SPEX Db in detail, with reference to its actual use by the M-KBSGI. Central to the system is the Oracle-based relational database. We illustrate the database schema and describe the user interface to the data. In addition, we describe the interactions of the system with external sources of data. These facets of the system aid SG target selection and enable monitoring of the status of an SG project, both at the level of individual targets and as a whole. The target selection and external interaction facets of the SPEX Db differentiate it from other data management systems recently developed for structural genomics, such as SPINE [7], SESAME [8] or HalX [9], which are primarily devoted to experimental data tracking, as for a traditional LIMS system.

Results and Discussion

Requirements of an SG data management system

An ORF selected as a structural genomics target proceeds along the general experimental pipeline depicted in Figure 1. This flowchart, from target selection to structure determination, is relatively standard for structural genomics groups, up to the "Quality Control" stage. A particular SG lab may not possess the facilities for both the NMR and X-ray methods of structure determination, so that one of these branches after quality control would not be followed. The stage of "Proteolysis" in the pipeline figure indicates the commonly used technique of applying limited proteolysis to identify protein domains suitable for structural analysis [10]. New constructs for the target ORF suggested by the proteolysis will rejoin the experimental pipeline at the "Cloning" stage (hence the dashed arrow). At each of the steps in Figure 1 different experimental procedures are used, resulting in a variety of experimental data unique to the step. This heavily influences the design of the database schema related to the tracking of experimental data, as described in the following section.

In addition to the LIMS-like archiving of experimental data for active SG targets, a data tracking system for structural genomics should serve as a resource for target selection. Methods of target selection for SG projects vary widely between groups [11] but there are certain essential pieces of information on targets, such as the presence of homologous structures already determined, that should be readily available when selecting targets from a pool of potential target ORFs. Most of this information is found in external public databases. The key information stored on targets is also crucial for active targets. For example, the progress of homologous SG targets among other SG laboratories should be monitored regularly to avoid unnecessary work. Experimental difficulties with a particular target may be ameliorated by investigation of its putative domain structure, so that alternative constructs for the selected ORF may be tried. The requirement for up-to-date and extensive information on targets also influences the design of the database underlying our SG data management system.

As important as the database schema is the interaction of its users with the data stored. Workers in the laboratory should be presented with an interface that enables quick inputting of experimental data that is detailed but also as unambiguous as possible. Where there are a discrete number of possible experimental conditions to be entered (such as the name of purification protocol used for a target) drop-down menus should be preferred over manual typing, so that possible ambiguities do not thwart later mining of the stored data. The user interface should also enable easy generation of reports from the database (such as a list of all targets currently at the "Crystallization" stage

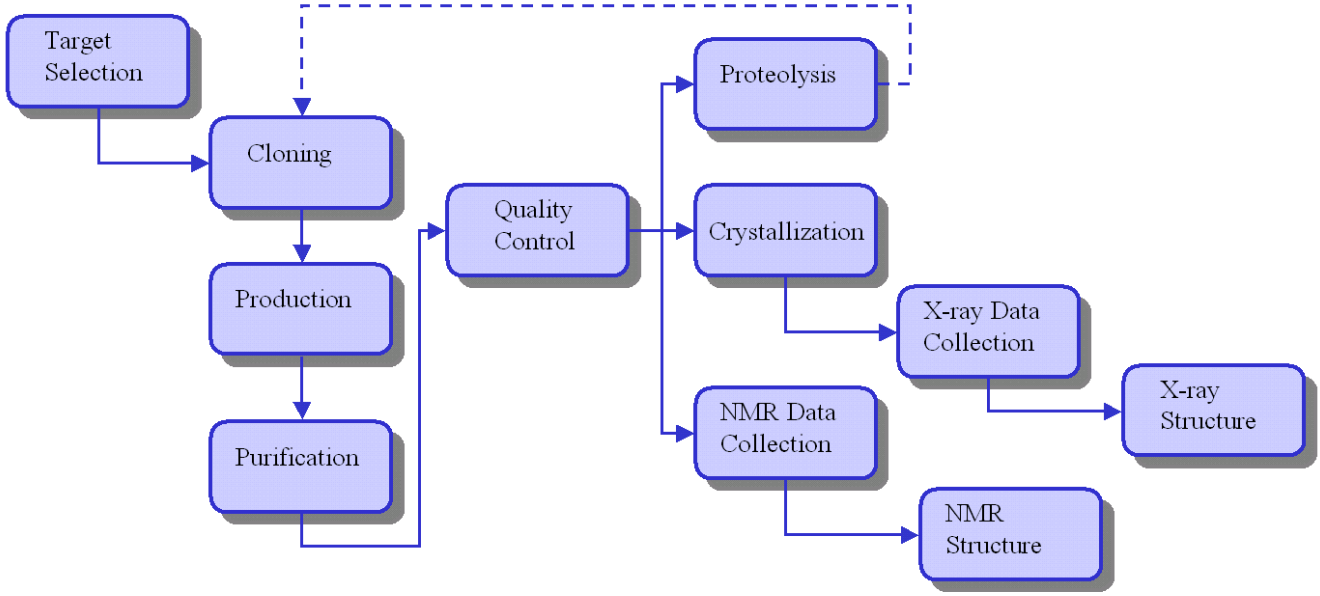

Figure 1
The structural genomics experimental pipeline

from a particular organism) but without the need for manually inputting SQL queries. Our user interface was designed in the light of requirements such as these.

Database design

Figure 2 presents a diagram of the schema of the database underlying the SPEX Db. All tables are linked with foreign keys to ensure data integrity.

For information on targets there are two main tables, PROJECT and TARGET. The PROJECT table contains information on separate projects and is used to group the targets of a project together appropriately. The database is designed to support multiple SG projects and there are over 10 separate projects in our working data management system, although the M-KBSGI project provides a large majority of the targets. The TARGET table is used to hold information on a target, such as source organism, molecular weight, and references to the NCBI and ExPASy databases. These data on targets are useful for target selection and management of active targets, as explained in the previous section. The M-KBSGI has loaded all of the ORFs of several bacterial genomes into the TARGET table. Active targets are selected from these genomes with the help of the up-to-date target information provided by the system. Other projects choose to load only targets that have already been selected. The TARGET table also accommodates user annotation of targets and the status of work (the point along the experimental pipeline) for each target. The table is linked to the PROJECT table with a foreign key and a constraint of one to many on the project_id.

Three tables, PLASMID, OLIGO and GLYCEROL, track the progress of the cloning step of the experimental pipeline. The PLASMID table contains information about the cloning strategy and references to the forward and reverse oligonucleotide primers. Information about each primer is stored in the OLIGO table. The GLYCEROL table is used to keep information about the physical location of frozen glycerol stocks of bacterial cultures carrying the plasmid. The PLASMID table is linked to the TARGET table with a foreign key and a constraint of one to many on the target_id (i.e. for a single target one can create many different plasmids, corresponding to different cloning strategies or alternative constructs.)

The production step has one table, called PRODUCTION. This table contains information about the cell culture, incubation temperature and various other experimental parameters. The PRODUCTION table is linked to the PLASMID table with a foreign key and a constraint of one to many on the plasmid_id.

The PURIFICATION table contains information about the cell lysis method and subsequent chromatography

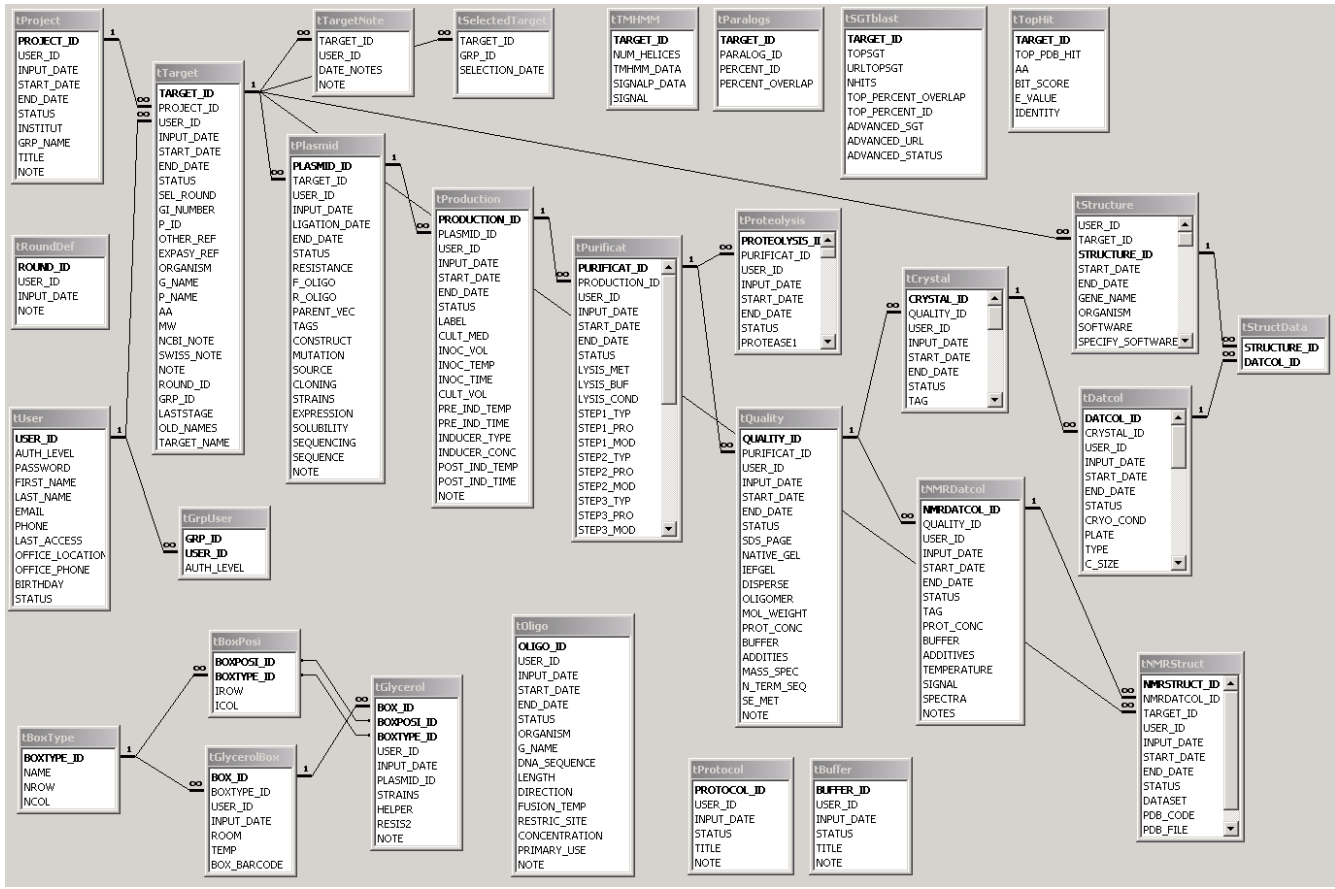

Figure 2
Database schema

experiments used to purify the protein. This table is linked to the PRODUCTION table with a foreign key and a constraint of one to many on the production_id.

After purification, quality control is performed on a protein solution. Experimental results from SDS-PAGE gels, mass spectrometry and dynamic light scattering are stored in a table called QUALITY. The QUALITY table is linked to the PURIFICATION table with a foreign key and a constraint of one to one on the purification_id. Similarly, a PROTEOLYSIS table tracks the limited proteolysis experiments and is linked to the PURIFICATION table with a foreign key and a constraint of one to one on the purification_id.

Along the X-ray crystallography path of structure determination, the CRYSTAL table stores information on crystallization experiments and the DATCOL table contains X-ray data collection parameters, such as the radiation wavelength and cell parameters, etc. Final refinement statistics are stored in the STRUCTURE table. The intermediate STRUCT_DATA table is implemented to enable one structure entry for many X-ray data collections, allowing, for example, complete recording of multiple wavelength anomalous diffraction (MAD) data collection. Experimental information for structure determinations by NMR methods is stored in the analogous tables NMRDATCOL and NMRSTRUCTURE.

Each of these tables containing experimental information for targets along the experimental pipeline has a foreign key to the user_id in a USER table, a listing of laboratory members. The user_id reference is used to check if the user has permission to edit or delete the corresponding entry in an experimental data table (see Access Control, below).

In addition to the experimental tables which track the work in progress on each target, the database contains other tables linked by foreign keys to the TARGET table. For each target in the database, the TOPHIT and

SGTBLAST tables contain weekly-updated information on homologous protein sequences either in the PDB or in progress within other SG laboratories, respectively. Other auxiliary tables exist, such as PROTOCOL and BUFFER, which allow for detailed descriptions of standard experimental protocols and routinely used chemical buffers, respectively.

External interactions of the system

The SPEX Db exchanges information with a variety of external resources. For each target in the database, corresponding entries in the SWISS-PROT/TrEMBL, NCBI, InterPro, and many other external databases are available via HTML links. Information within these databases (and further cross-referenced to others) aids in the selection of targets and enables quick access to potentially useful data on active targets, such as putative domain structures, isoelectric points and cofactors, for example.

It is crucial for structural genomics laboratories to regularly check the Protein Data Bank (PDB) for known structures of homologous proteins. The SPEX Db runs a weekly script for this purpose: for each target sequence in the database, a sequence similarity search using BLAST is made against the latest database of protein sequences in the PDB, downloaded from ftp://ftp.rcsb.org/pub/pdb. The TOPHIT table of the database is updated using the results of the search. Any target for which a new most homologous hit is found in the PDB is listed in an email distributed to all users of the SPEX Db. An E-value homology cut-off of 0.1 is used in the similarity search, corresponding to a minimum sequence identity of approximately 25%. Based on the results of the weekly search and the level of similarity between targets and new structures in the PDB, work may be stopped on a target because the structure of a homologous target has been solved.

The M-KBSGI, the structural genomics project supplying the majority of targets to the SPEX Db, also participates in the global monitoring of structural genomics targets. There is an open exchange of information on the targets of 17 SG laboratories worldwide in a standard XML format. The XML format requires SG groups to report the amino acid sequence and latest experimental status of their targets. Groups can also provide additional information such as the name of the target protein, source organism, and links to corresponding entries in other databases. The data from all groups is available from a central database maintained by the PDB [12] (TargetDB, http://targetdb.pdb.org). Our initiative is one of the 17 SG laboratories reporting target information. Each week, a script harvests information from the database on experimental status for all selected targets and produces an XML file in the required format. The file is available in machine-readable format from the M-KBSGI website and is included in the global XML file assembled by TargetDB.

In order to track the progress of similar SG targets from other groups, the SPEX Db runs another weekly script. A FASTA format database of sequences is assembled from the data in the central XML file, downloaded from TargetDB. Each SPEX Db target sequence is compared against the database of SG sequences with BLAST. For the hits that are found, if any, all information given in the XML format is extracted and presented in an HTML report. Some key information from the hits, such as the status of the most advanced homologous SG target, is stored in the SGTBLAST table of the database and updated after the weekly search. If the most advanced experimental status among homologous SG targets changes for a target sequence in the SPEX Db, the target id and the change in status is reported in a weekly email distributed to all users of the SPEX Db. Like the search for homologous targets in the PDB, an E-value cut-off of 0.1 is used for the search against TargetDB. These weekly global comparisons rapidly identify locally less advanced targets and facilitate stoppage of work that duplicates more advanced efforts around the world.

User interface

Users of the SPEX Db interact through a web-based interface built on top of Netscape iPlanet Web Sever, Enterprise Edition 4.1. We use server side JavaScript code to do all transactions with the database and format the HTML document sent to the client browser.

A detailed description of all facets of the user interface is impractical for this report. In this section we illustrate examples of four commonly used types of web pages in the system.

Target View

Figure 3 is a screenshot for seven targets in the "Target View" mode of the SPEX Db. The smaller screenshot insets in the figure show the pages obtained by clicking on various links and are included for illustrative purposes only. Readers are encouraged to try all aspects of the system available from the Target View (apart from the Tree View, see below) by visiting the M-KBSGI website http://sgen.bri.nrc.ca/brimsg/bsgi.html and clicking on "List all targets" or "Target search". Information on each target occupies a row. The first column shows an index for that target based on the type of search that was performed. Clicking on the globe icon in the "Links" column results in a pop-up window of a wide range of external links related to the target, as referred to in the previous section. The target name is a link to the Target Page for that target, which contains more detailed information on the target than in the rows seen in Figure 3, but on a separate page.

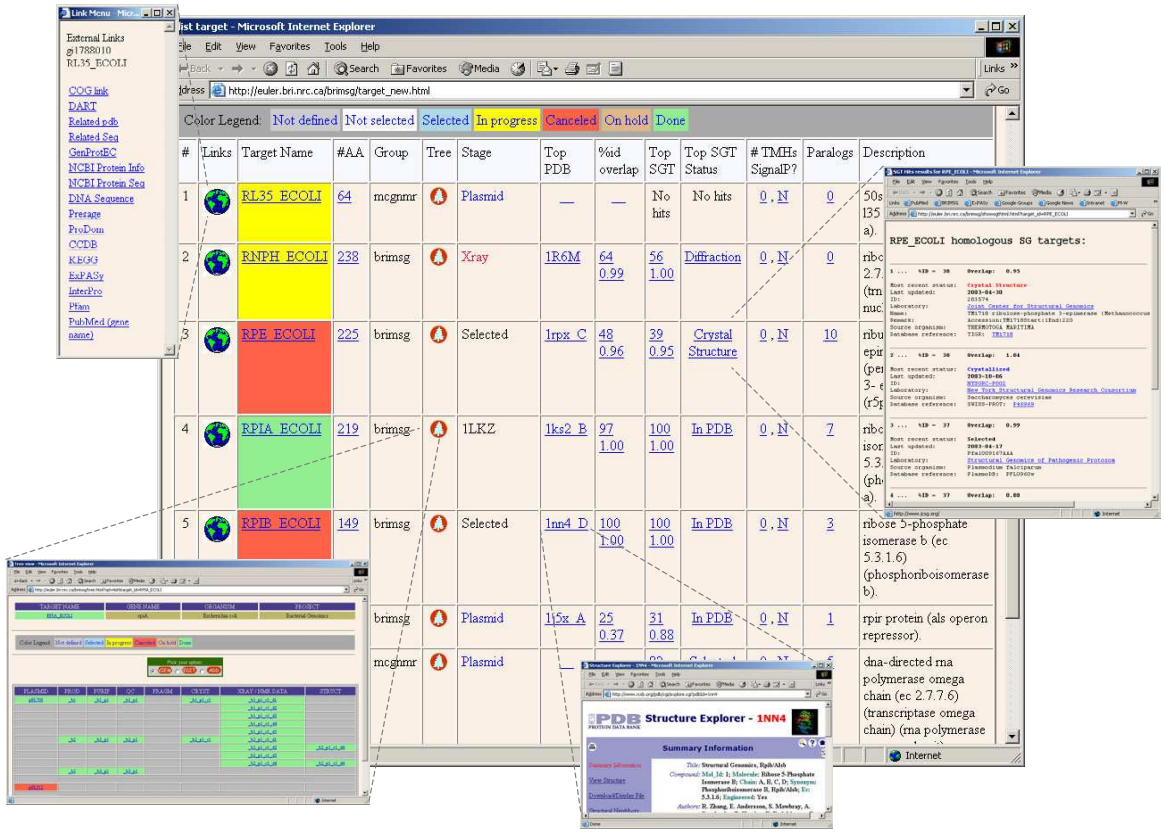

Figure 3
Target view

The background of the target name is colored according to the legend at the top of the page, for the various states of the ORF as a target within the particular SG group using the SPEX Db. The "#AA" column contains the length, in amino acid residues, of the target ORF. Clicking on that number results in the target amino acid sequence from NCBI. The "Group" column displays a code corresponding to the laboratory within the M-KBSGI that is working on that target. Important for targets that are active within the laboratory is the "Tree" column. Clicking on the tree icon yields the Tree View for the target, which is discussed below. The "Stage" column indicates the latest experimental state for the target, in terms of the pipeline in Figure 1. If the structure of the target has already been determined and is deposited in the PDB, the PDB code is displayed.

The following four columns display weekly-updated data from the search for homologous PDB or SG sequences, as described earlier. The PDB code of the most similar PDB sequence found (if at least one is found) is shown in the "Top PDB" column, and it is also a link to the corresponding PDB entry. The next column shows the percentage sequence identity and overlap of this top PDB hit. The "Top SGT" column shows the percentage identity and overlap of the most similar SGT sequence, and the "Top SGT Status" column shows the experimental status (as reported in the worldwide SG XML file) of the most advanced SGT hit.

The "# TMHs/SignalP" column shows data on the predicted presence of transmembrane helices or signal

peptide sequences for the target ORF, calculated using the TMHMM [13] (Krogh *et al.*, 2001) and SignalP [14] (Nielsen *et al.*, 1997) programs. The first number in the column is the number of putative transmembrane helices, and the second symbol, either a "Y" or an "N", corresponds to the presence or absence of a predicted signal sequence, respectively. Clicking on either presents the user with a more detailed page about the predictions. This data is primarily used during target selection to screen out at a glance potential target ORFs that contain membrane-spanning regions, whose structures are notoriously difficult to obtain. The "Paralogs" column indicates the number of homologous sequences within the pool of potential targets for the M-KBSGI. (In this sense the term "paralogs" is used differently from the usual connotation, that of homologous sequences within the same proteome.) Clicking on the number produces a list of the target names for the "paralogous" targets. Finally, the "Description" column contains the description (from the NCBI) of the function of the target ORF.

Tree View

From the Tree View users can access the experimental data that is entered into the experimental data tables described earlier. In Figure 4 we show the Tree View for a target with a solved structure, so that data has been entered in each of the tables from PLASMID to STRUCTURE. (This target structure was solved by X-ray crystallography). From left to right the columns in the tree view follow the stages in the experimental pipeline (Fig. 1). By clicking on the entries in the Tree View, data in the tables corresponding to each of the stages can be viewed from separate web pages, as indicated by the smaller screenshot insets in Figure 4. Tree View entries are coloured according to whether the experiment to which it refers is in progress or completed or cancelled. Because of the one to many constraints in the database schema for sequential experimental data tables, many tree view entries corresponding to one PLASMID entry can accumulate at later stages in the pipeline. Users can view an experimental data entry, edit one (with the correct permission), or add a new entry corresponding to a new experiment, according to the option selected above the Tree View in Figure 4.

Adding a Production entry

The partially completed web form in Figure 5 is an example of the forms used to add experimental data to the Tree View. This form appears if a user wishes to add data corresponding to a Production experiment. (The data entered is actually that of the Production entry leading to the structure in Figure 4.) The layout of the form is identical to that of the pages that simply display data that has been entered. At the top right of the page is the user_id of the laboratory member entering the data. Only this person can further edit the entry. In revisiting past experiments it is useful to know the identity of the laboratory member entering experimental data should any additional information be required. The unit of data required for each numerical entry is explicitly indicated, and note that for three fields (Label, Culture Medium and Inducer Type), drop down menus containing the possible entries are provided. These features serve to make the experimental data entered as unambiguous as possible. Importantly, there is also additional room for the user to insert notes and comments regarding any unusual or noteworthy results of the experiment.

Searching the database

Searches for entries in all experimental data tables (e.g. DATCOL entries, corresponding to X-ray data collections, or proteolysis results) can be performed, resulting in pages listing entries in formats analogous to the Target View described above. Searches can be performed based on the ids of the experiments, their status, or the users who performed them. Searches for targets can be made with combinations of 14 different target criteria via an extensive Target Search page (Figure 6). A smaller number of target search criteria are available for the public via http://sgen.bri.nrc.ca/brimsg/bsgi.html. In addition, there is a "Search All" page in which users can combine criteria from different tables for more extensive searches without the need for manual input of SQL queries. These queries can be stored for future use. The user interface also allows users access to data in the auxiliary tables (such as PROTOCOL) via a search interface.

Access control

Access to the user interface is controlled by the identification (login name and password) and authority level of a user. Authority levels range from 0 to 9. Users with an authority level of 0 do not have access, and those with level 1 have read only access. Users with authority level 2, the majority of all lab members, can read all pages in the system, but can edit or delete only the entries that they have created. Authority levels higher than 3 are for administrative tasks. For security purposes, if a user logged in to the web server is inactive for an hour, the user is automatically logged out, and must re-enter his or her password to resume the session.

Under the present system users in a particular project with authority levels 2 and above can view and/or add data belonging to another project. We will avoid this unwelcome scenario by implementing a more complex access control scheme under which it will be possible to restrict all access to a subset of the data in the database. The new scheme takes into account in a most general way the differing roles (and differing access requirements) for workers in large collaborative structural genomics projects.

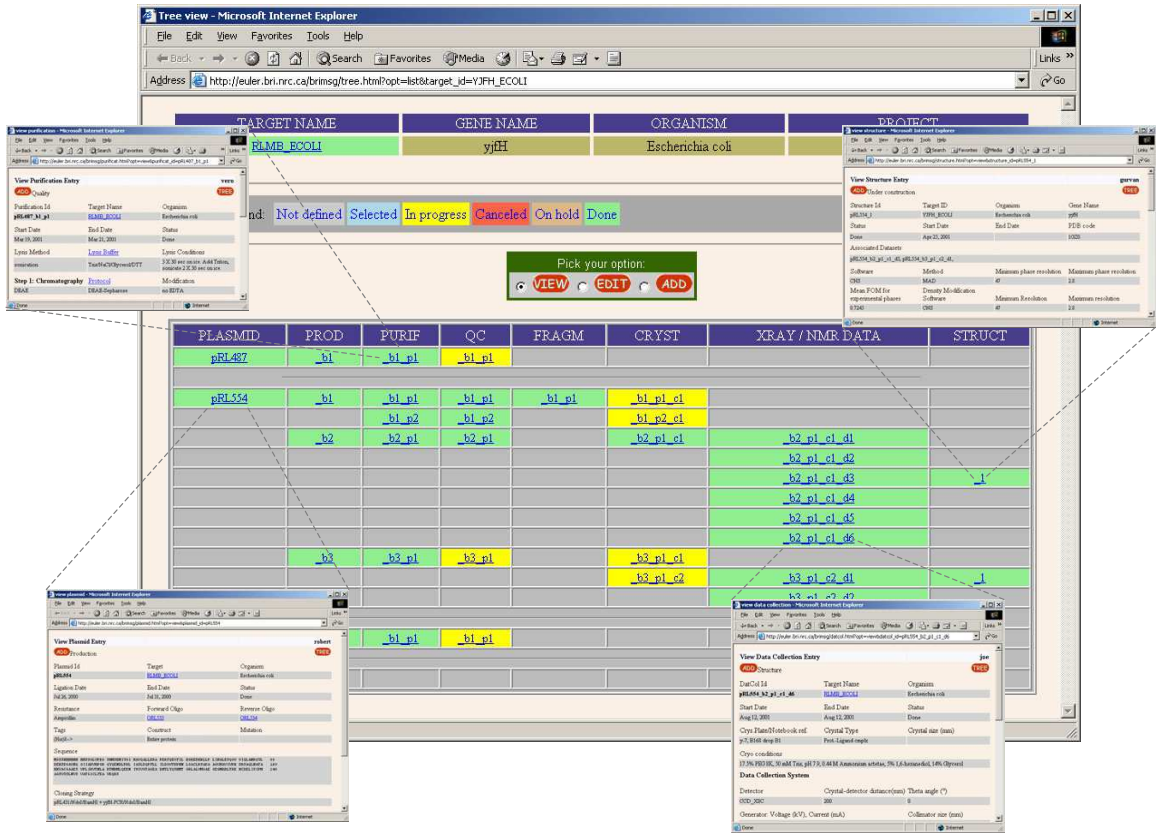

Figure 4
Tree view

Future developments

The design of our database and associated web interface has resulted in a user-friendly system that can be of tremendous benefit to structural genomics initiatives, as it has proven to be for the Montreal-Kingston Bacterial Structural Genomics Initiative. However, as the field of structural genomics evolves, so too must the flexibility of SG data management systems such as ours. Our database schema reflects very well the traditional SG experimental pipeline, but it is apparent that more sophisticated experimental relationships will in the future need to be accommodated by the schema underlying any SG data management system. For example, groups may want to combine different experiments at some point along the pipeline into a new experiment (such as for the co-crystallization of a protein complex). Many-to-one relationships such as these along sequential experimental steps are not easily dealt with in the present system. For another example, a gift of a purified protein from collaborators outside a user of the SPEX Db requires under the present schema the addition of "dummy" entries in the database tables from PLASMID up to PURIFICATION. Other information systems that have been developed to serve SG initiatives, such as SPINE [7], SESAME [8] or HalX [9] are also hampered by similar underlying schema. Database design for SG data management systems should be made as flexible as possible for the future needs of SG groups, so that, for example, many-to-one relationships along the experimental pipeline such as we have mentioned are accommodated. We have participated in the development of a

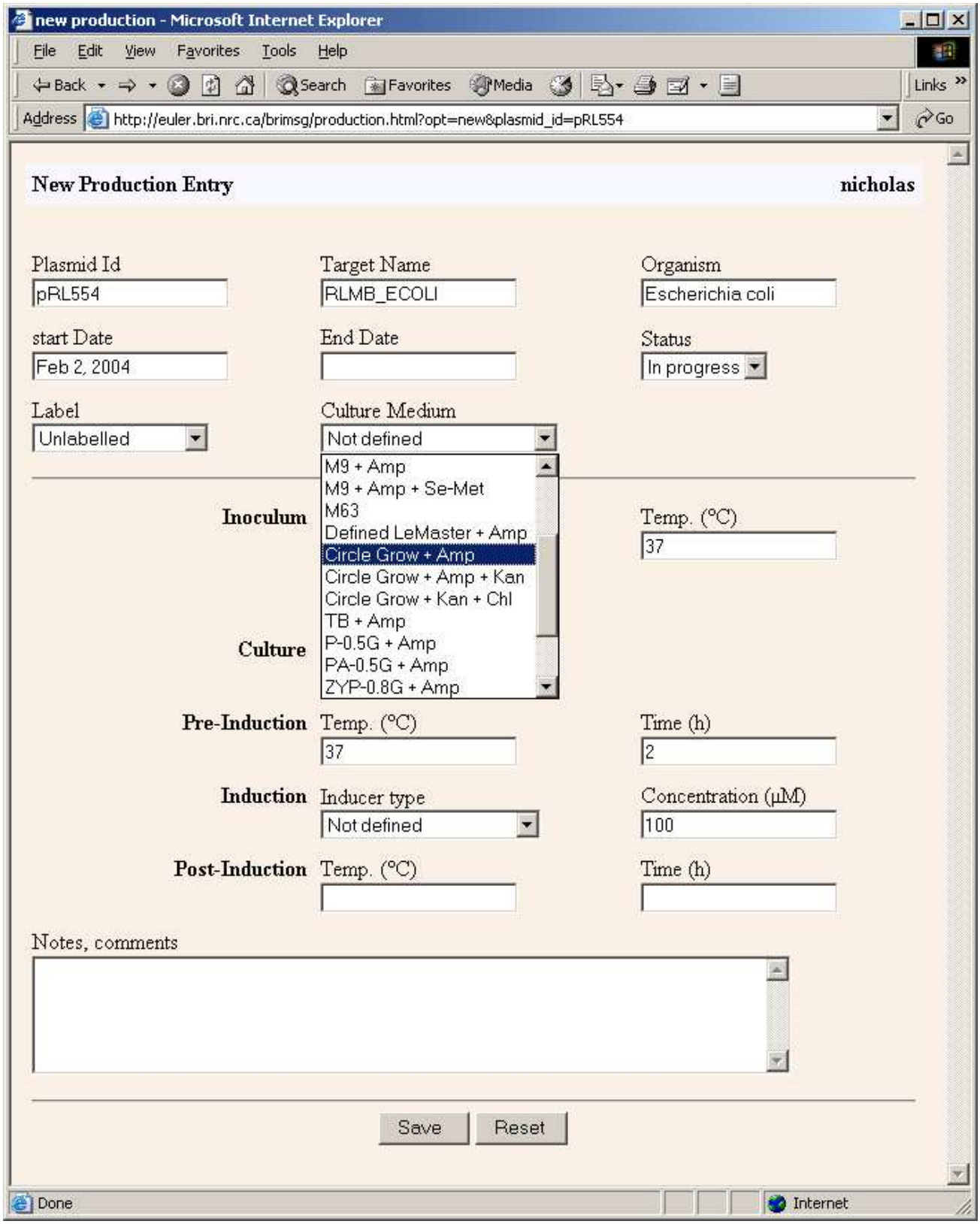

Figure 5
Production entry form

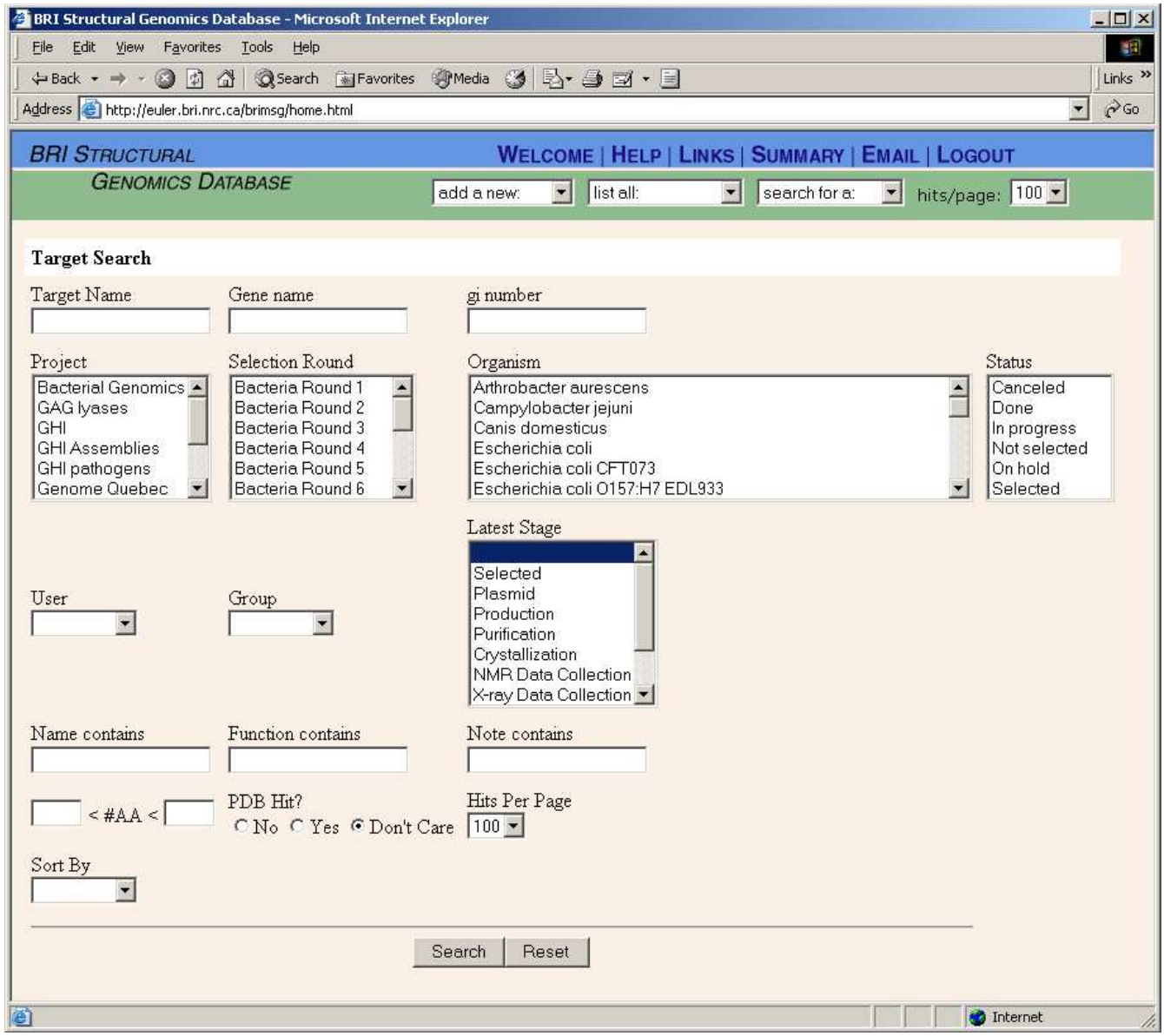

Figure 6
Target search

prototype schema with collaborators at the European Molecular Biology Laboratory and elsewhere that seeks to address such issues (see http://www.ebi.ac.uk/msd-srv/docs/ehtpx/lims/downloads.html for further information).

Other areas of future enhancements of the SPEX Db include automatic interaction with laboratory robots, resulting in real time automatic input of experimental data. The graphical documentation of experiments, such as pictures of gels, will be included in the experimental result pages of the user interface. Users may experience long response times to queries accessing multiple tables from the search interfaces. We are implementing some changes to the code and computer hardware to speed up such searches. Laboratories may wish to use the proteins produced by an SG endeavor for more than just structure determination. For example, purified proteins can be introduced into an appropriate host for polyclonal antibody production. Such experiments correspond to a new

path in the SG pipeline (Fig. 1) after the "Quality Control" stage. Since antibody production is in the planning stage among some users of our system we will introduce the new database tables and web pages required to accommodate the accompanying information. As discussed earlier, we also intend to implement a more sophisticated access-control model.

Availability

Structural biology and structural genomics laboratories are welcome to contact the authors for full access and use of the SPEX Db. In addition, the database underlying the system can be established locally with code for table creation commands and the associated data dictionary, available by request from the authors.

Authors' contributions

SR designed the database and programmed the user interface. NO developed tools for sequence comparisons and other external interactions of the system and drafted the manuscript. MC conceived of the project and participated in its design and coordination. All authors have read and approved the final manuscript.

Acknowledgements

The authors wish to thank Hervé Hogues for assistance in database design and application programming. We also thank Xeuli Li, Nomair Naeem, Gregory Jarrett and Jeffrey Rasmussen for the development of miscellaneous programming tools and graphical images. This work was in part supported by CIHR grant No. 200103GSP-90094-GMX-CFAA-19924 to MC.

References

1. Terwilliger TC, Waldo G, Peat TS, Newman JM, Chu K, Berendzen J: **Class-directed structure determination: foundation for a protein structure initiative.** *Protein Science* 1998, **7:**1851-1856.
2. Kim S: **Shining a light on structural genomics.** *Nat Struct Biol* 1998, **5(Suppl):**643-645.
3. Montelione GT, Anderson S: **Structural genomics: keystone for a Human Proteome Project.** *Nat Struct Biol* 1999, **6:**11-12.
4. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S: **Structural genomics: beyond the human genome project.** *Nat Genet* 1999, **23:**151-157.
5. Imbert MC, Nguyen VK, Granjeaud S, Nguyen C, Jordan BR: **'LABNOTE', a laboratory notebook system designed for academic genomics groups.** *Nucl Acids Res* 1999, **27:**601-607.
6. Brazma A, Sarkans U, Robinson A, Vilo J, Vingron M, Hoheisel J, Fellenberg K: **Microarray data representation, annotation and storage.** *Adv Biochem Eng Biotechnol* 2002, **77:**113-39.
7. Goh CS, Lan N, Echols N, Douglas SM, Milburn D, Bertone P, Xiao R, Ma LC, Zheng D, Wunderlich Z, Acton T, Montelione GT, Gerstein M: **SPINE 2: a system for collaborative structural proteomics within a federated database framework.** *Nucleic Acids Res* 2003, **31:**2833-8.
8. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips GN Jr, Rayment I, Ulrich EL, Volkman BF, Markley JL: **Project management system for structural and functional proteomics: Sesame.** *J Struct Funct Genomics* 2003, **4:**11-23.
9. **HalX – Free-source Laboratory Information Management System (LIMS) software** [http://halx.genomics.eu.org/]
10. Koth CM, Orlicky SM, Larson SM, Edwards AM: **Use of limited proteolysis to identify protein domains suitable for structural analysis.** *Methods Enzymol* 2003, **368:**77-84.
11. Linial M, Yona G: **Methodologies for target selection in structural genomics.** *Prog Biophys Mol Biol* 2000, **73:**297-320.
12. Westbrook J, Feng Z, Chen L, Yang H, Berman HM: **The Protein Data Bank and structural genomics.** *Nucl Acids Res* 2003, **31:**489-491.
13. Krogh A, Larsson B, von Heijne G, Sonnhammer E: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.
14. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Machine learning approaches for the prediction of signal peptides and other protein sorting signals.** *Protein Engineering* 1997, **10:**1-6.