Nucleic Acids Research, 2003, Vol. 31, No. 11 ***2833–2838***
DOI: 10.1093/nar/gkg397

SPINE 2: a system for collaborative structural proteomics within a federated database framework

Chern-Sing Goh[1], Ning Lan[1], Nathaniel Echols[1], Shawn M. Douglas[1,3], Duncan Milburn[1], Paul Bertone[2], Rong Xiao[4,5], Li-Chung Ma[4,5], Deyou Zheng[4,5], Zeba Wunderlich[4,5], Tom Acton[4,5], Gaetano T. Montelione[4,5,6] and Mark Gerstein[1,3,*]

[1]Molecular Biophysics and Biochemistry, [2]Molecular, Cellular and Developmental Biology and [3]Computer Science, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA, [4]Center for Advanced Biotechnology and Medicine, [5]Department of Molecular Biology and Biochemistry, Rutgers University and [6]Department of Biochemistry, Robert Wood Johnson Medical School, UMDNJ, Piscataway, NJ 08854, USA

Received February 15, 2003; Revised and Accepted April 9, 2003

ABSTRACT

We present version 2 of the SPINE system for structural proteomics. SPINE is available over the web at http://nesg.org. It serves as the central hub for the Northeast Structural Genomics Consortium, allowing collaborative structural proteomics to be carried out in a distributed fashion. The core of SPINE is a laboratory information management system (LIMS) for key bits of information related to the progress of the consortium in cloning, expressing and purifying proteins and then solving their structures by NMR or X-ray crystallography. Originally, SPINE focused on tracking constructs, but, in its current form, it is able to track target sample tubes and store detailed sample histories. The core database comprises a set of standard relational tables and a data dictionary that form an initial ontology for proteomic properties and provide a framework for large-scale data mining. Moreover, SPINE sits at the center of a federation of interoperable information resources. These can be divided into (i) local resources closely coupled with SPINE that enable it to handle less standardized information (e.g. integrated mailing and publication lists), (ii) other information resources in the NESG consortium that are interlinked with SPINE (e.g. crystallization LIMS local to particular laboratories) and (iii) international archival resources that SPINE links to and passes on information to (e.g. TargetDB at the PDB).

INTRODUCTION

The structural genomics effort is generating a vast amount of data, underscoring the need for database systems and servers that can organize this information (1–4). Structural genomics consortia have been formed to consolidate these efforts. These consortia, including the Northeast Structural Genomics Consortium (NESG), are composed of numerous researchers in disparate locations working cooperatively at each step of the structural determination process, from selection of targets through to analysis of the results. The SPINE (Structural Proteomics In the NorthEast) database (5) was designed to coordinate the efforts of these researchers in the NESG as an information management and data analysis resource.

SPINE was created in 1999 as a data repository with associated data mining tools. Through many revisions, its tracking functionality has expanded to accommodate detailed histories for individual samples, thereby presenting a more complete framework for transmitting information through all stages in high-throughput protein production and structure determination. In the original publication, Bertone *et al.* (5) described the system architecture and how it was interlinked with specific tools to enable data mining. There have been previous discussions regarding a role of proteomics ontologies in structural genomics (6) and, at this stage, SPINE can be described as the beginning of an ontology of standardized protein properties. Here we present version 2 of SPINE and describe its overall development over the last 3 years.

Our system encompasses several aspects. First, the core of the SPINE system (core SPINE) is a centralized information management system, which tracks protein targets through the structure determination process, from the cloning of expression constructs to final biophysical and structural characterization and submission of PDB coordinate sets, providing histories for particular samples. Secondly, SPINE sits at the center of a federation of computational resources; there are a number of SPINE-integrated web tools, both local and remote, which allow members of the consortium to post further information related to protein targets. Finally, the database

*To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: mark.gerstein@yale.edu
Correspondence may also be addressed to Gaetano T. Montelione. Tel: +1 732 235 5321; Fax: +1 732 235 5633; Email: guy@cabm.rutgers.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

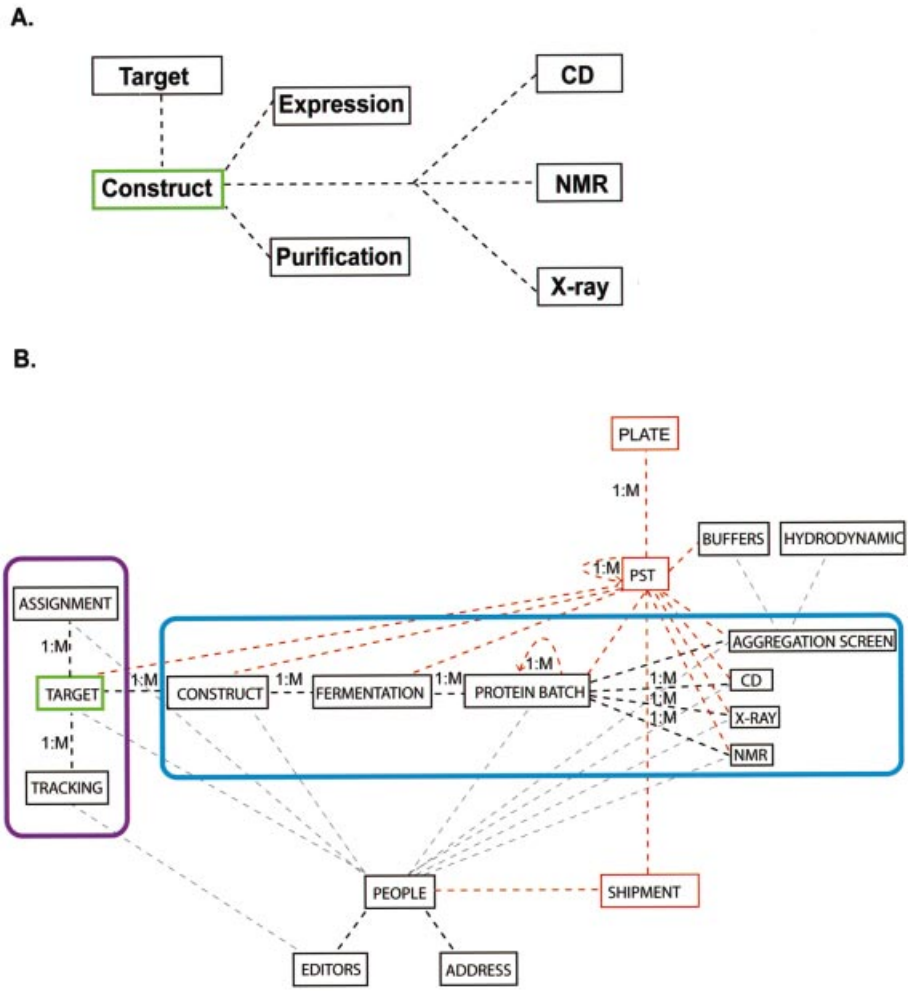

Figure 1. Schema of the SPINE database, showing evolution of tables and data flow. (**A**) The original version of SPINE, in which the construct was the primary object tracked. Table records typically had one-to-one relationships. Each target could be associated with a single expression and purification. (**B**) Current schema. Experimental records (outlined in blue) all have one-to-many relationships branching from target entries. User and tracking information and, most recently, detailed sample tube data can now be tracked in SPINE. The newest additions to the database are highlighted in red.

can be analyzed retrospectively to identify factors that contribute to the ease with which individual proteins may be studied. Although the system is tailored to the needs and goals of the NESG, we expect that many of its features could be readily adapted to similar projects.

IMPLEMENTATION OF CORE SPINE SYSTEM

The core database is implemented in MySQL on a Unix platform, with its user interface written entirely in the Perl programming language and integrated with the Apache web server. (Previous versions of the database used PHP extensively; the current implementation provides a more consistent platform.) This approach offers a considerable speed advantage and allows sharing of libraries with offline programs used in the development of future releases and associated tools. The suitability of Perl for systems programming also allows a wide variety of other modules to be used in the server with minimal set-up and administrative overheads, such as the BLAST package (7), Java and Lisp code for data analysis. The core data elements are stored in a number of tables that record the experimental progress of individual targets (Fig. 1). Auxiliary tables control access to the database and record a history of individual changes.

Initially, the basic unit tracked by SPINE was the expression construct. However, with the evolution by the NESG consortium of a systematic protein target selection process (8), these protein targets, each of which may have multiple associated constructs, have been made the focus of the database. All derivative records from the target onwards comprise one-to-many relationships. At the level of protein purification, we have introduced parent–child relationships for individual protein samples, broadening the data structure instead of compressing multiple purification records into a single instance. Most records include some form of unstructured data, often in the form of analytical images that have been uploaded to the server. This has even been extended to encompass email (see below) related to specific targets.

A key feature of the database is the ability of any registered member of the consortium to add and modify entries via an intuitive web interface. This is regulated by journaling all changes to data (and the user responsible) and by restricting

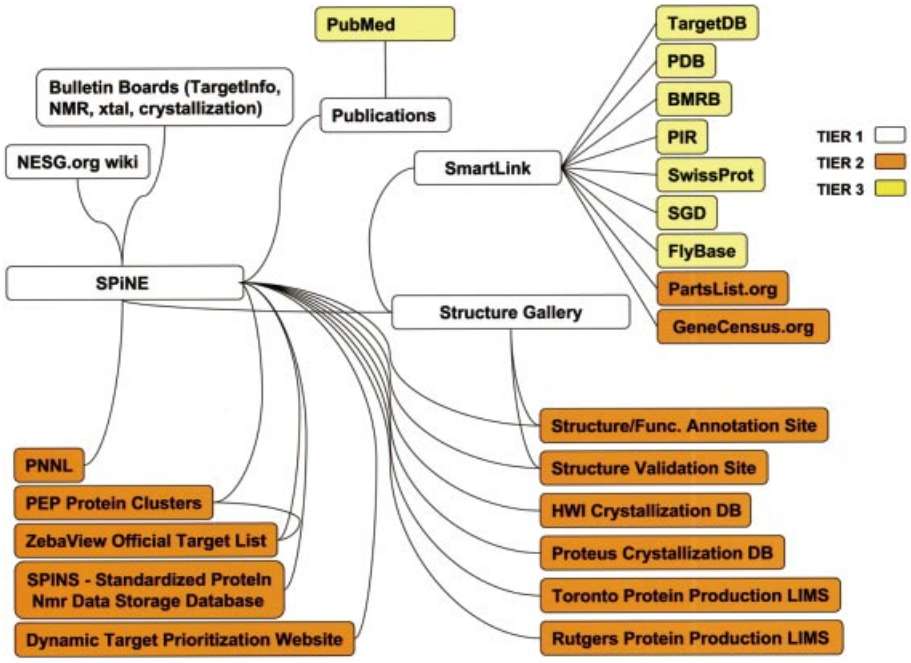

Figure 2. Overview of the SPINE federation. Tier 1 (white) resources are the local resources integrated in SPINE. Tier 2 (orange) resources are other web resources of the NESG project that are linked to SPINE. Tier 3 (yellow) resources are external archival resources that SPINE is connected to.

access to certain entries. More flexible and customized methods of data entry are also possible. The use of direct SQL and the Open Database Connectivity (ODBC) protocols enable a variety of remote interfaces to the server (such as Excel spreadsheets or Java programs), and data interchange uses standard XML or table formats. These features enable bulk uploading of local datasets into SPINE. In the future, development will focus mainly on the schema, the SPINE data dictionary and display functions, leaving data entry as the prerogative of individual users.

INTEGRATION OF CORE SPINE WITH A FEDERATION OF OTHER RESOURCES

The NESG consortium is comprised of various database management systems used to store and search critical data. SPINE provides federation technologies to provide a common unified interface for these diverse systems. The core SPINE database system sits at the center of a federation of information resources diagrammed in Figure 2. The core of SPINE is a relational database handling highly standardized information that interoperates with a set of local resources designed to handle more heterogeneous data that is not readily stored in tables. Some of these features include the incorporation of free text fields, the ability to upload data files and the utility of its data mining tools and servers.

SPINE is also associated with external resources that are coupled together in a loose federation associated with the NESG. These resources can be categorized into three tiers.

Local integrated resources

SPINE is integrated with a number of 'local' resources, resident on the same machine and tightly coupled to it. These include the following.

The NESG website. This is built around a wiki (http://wiki.org/) platform that lets users edit or create web pages by using the web browser. This platform allows for easy remote editing of such things as links to related projects and is useful for web-based collaboration.

A structure gallery. This is used for displaying completed 3-dimensional structures of protein targets.

A publication page. This is built on elaborating the NCBI PubMed XML dump to incorporate such things as targets and websites. It allows the direct cross-referencing of targets, URLs and MEDLINE identifiers.

The target info bulletin board. The idea behind this is that there is a lot of information that people would like to track about a particular target that does not fit into standardized tables. This can be easily sent in the form of simple email messages that are copied to this bulletin board. These messages are automatically parsed for specific target identifiers and each instance of a target identifier in the archive is linked to its corresponding record in SPINE, and vice versa.

Other NESG resources

There are a number of other computational resources that are part of the NESG project which are connected to SPINE (Table 1). In particular, the diverse needs of experimentalists have led to the creation of several specialized databases within the consortium, dedicated to aspects of the project such as NMR data collection or crystal screening that are not well served by a single central resource. SPINE is currently being extended to facilitate storage of summary information from these satellite databases and even perform remote queries. Some of the resources that SPINE interoperates with are the

Table 1. Description of other NESG resources

NESG resources	Description
Target resources	
Dynamic Target Prioritization Website	Website that prioritizes targets that can provide useful information for constructing structural models of other proteins. http://maat.med.cornell.edu/nesg.html
PEP: Database for Prediction of Entire Proteomes	Used mainly for target selection, PEP provides clustering sequence identity information of potential targets. It also predicts structural and functional features of the targets to aid in experimental analysis. http://cubic.bioc.columbia.edu/db/PEP/
ZebaView: the official NESG target list	Organizes and reports on the progress of the NESG consortium protein targets. http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/
Annotation resources	
PartsList/Gene Census.org	Database web tools that focus on comparing genomes globally. http://bioinfo.mbb.yale.edu/genome/
Structure/Functional Annotation Website	This provides detailed information on solved structures. http://trantor.bioc.columbia.edu/sharon/Target_list_1.html
Structure Validation Website	A site that summarizes information related to the accuracy and validity of each structure determined. http://www.cabm.rutgers.edu/~aneerban/NESG_sample_reports/
LIMS systems	
HWI Crystallization Database	Database for information repository and tracking of crystallization data for robotic crystallization data generated at the Hauptman Woodward Medical Research Institute, Buffalo, NY.
PNNL	Database for information repository associated with Pacific Northwest National Laboratories, principally related to the progress of NMR structures.
Proteus Crystallization Database	Database for information repository and tracking for crystallization data generated at Columbia University.
Rutgers Protein Production LIMS	Laboratory information management system for Rutgers University.
SPINS: standardizing protein nMR storage	A relational database standardizing protein NMR data storage and submission to public databases. http://www-nmr.cabm.rutgers.edu/bioinformatics/SPINS/SPINSV2.html
Toronto Protein Production LIMS	Laboratory information management system for Toronto Structural Proteomics Initiative at the University of Toronto.

PEP (9) cluster viewer at Columbia University and ZebaView target list at Rutgers University, where new target entries are automatically downloaded nightly and inserted into SPINE. Other web resources SPINE is linked with include the SPINS database at Rutgers (10), the PartsList and Gene Census databases at Yale University (11), the Proteus crystallization database at Columbia University and a laboratory information management system (LIMS) system at the University of Toronto.

External archival resources

SPINE is also connected to resources outside the NESG through an evolving portal called SmartLink. This system handles much of the difficulty of translating ORF and structure identifiers and dealing with missing or dangling links. Most of the information pertaining to 3-dimensional structure determinations is transferred to the PDB (12). Other resources that SPINE is connected to include SwissProt (13), PIR (14), BMRB (15), TargetDB (12)— the RCSB registry for structural genomics projects—and Wormbase (16).

DETAILED SAMPLE TRACKING AND HISTORY

The requirements of tracking the progress of a target across multiple institutions include the ability to maintain a list of individual samples and their locations. For example, a protein may be purified at one site, shipped to another for crystallization screening and then sent to a third site for structural characterization. Protein production requires greater flexibility, since not only protein samples but also construct stocks and fermentation batches must be stored and tracked. The current system handles all sample types via tube records, whose contents are determined from their 'parent' record (construct, expression or purification). Each 'sample tube' generated in the pipeline of sample production is assigned a unique tube identifier, which eventually will be mapped into a bar coding system. Therefore, a collection of protein samples may be assigned for biophysical analysis without regard for their specific target, since the database automatically determines their history based on tube identifier. This concept has been extended to handle sample plates, which behave as an aggregation of tube records identified by their well number. An example screenshot can be viewed in Figure 3C. With this approach, information pertaining to shipments as well as physical location can be easily associated with sample records. This provides accounting of material transfer between institutions and a more accurate picture of the progress of individual targets.

DISCUSSION

In this paper we have described SPINE version 2, a relational database system that serves as the LIMS for the NESG consortium. This new version serves as the center of a federation of interoperable resources to allow for distributed functionality across the entire consortium. The major goals of the database system include the following.

A distributed LIMS

Constructing a distributed LIMS enables information to be shared among all the members of the consortium. From a vast number of local and external resources, SPINE can incorporate and process data, which can then be made accessible through a simple user browser interface at any location.

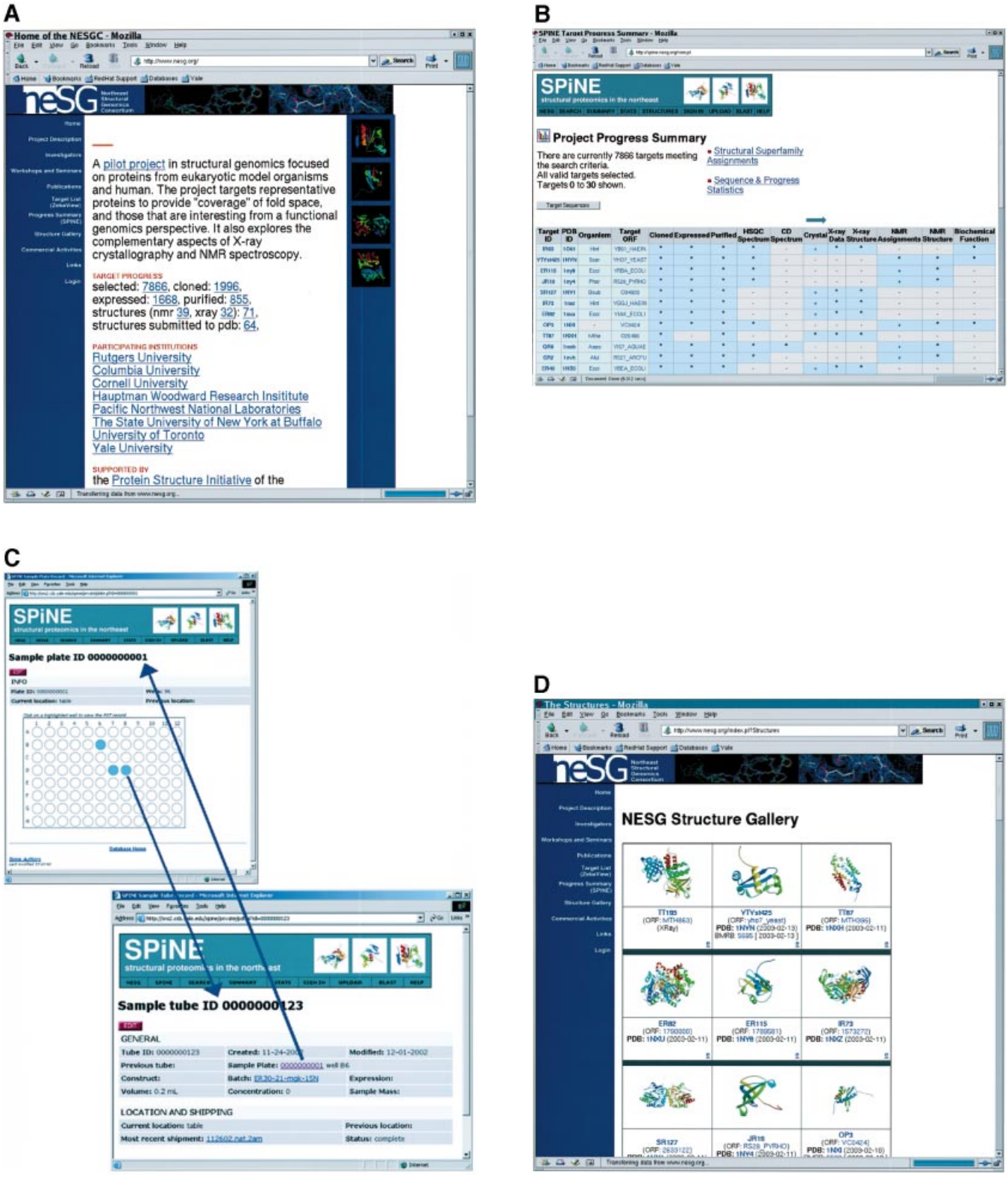

Figure 3. Assorted screenshots from the NESG/SPINE web server. (**A**) The NESG home page, providing links to all the institutions involved and to the gallery of structures produced by them. (**B**) The SPINE target summary page showing the data recorded in SPINE at each stage of the protein production and structure determination process (blue shading indicates data are present) for every target in the project. (**C**) One of the new additions in version 2 of SPINE is the ability to track in detail a given sample tube down to a well, plate and physical location. Any changes can be made by clicking on the desired sample entry (blue circle). (**D**) The NESG structure gallery showing all the structures that have been determined by the NESG consortium. The structure gallery also contains links to the PDB, the BMRB, the Structure Validation Website and the Structure/Functional Annotation Website.

Standardized proteomics ontology

An improved representation of protein properties that is standardized will allow for more efficient retrospective analysis of structural and functional information, which can be mined using an incorporated generalized data mining web tool.

Data mining

The schema for SPINE has been designed to facilitate analysis of the collected data to further optimize target selection criteria. Bertone *et al*. (5) demonstrated the potential data mining capabilities of the SPINE database by developing a decision tree algorithm that was used to infer whether a protein was soluble, and which biochemical properties contributed most to solubility, from a dataset of 562 *Methanobacterium thermoautotrophicum* protein expression constructs. The decision tree analysis indicated that protein characteristics such as protein length, hydrophobicity, and percent composition of charged residues were the strongest determinants in inferring the solubility of a protein. At present, the number of targets in the NESG is 7866, roughly 15 times the size of the original dataset, and presents a wealth of new information for mining (C.-S.Goh *et al*., in preparation).

Future directions in interoperation

We envision enhancing the integration with other databases by establishing a tighter interoperation with external databases, particularly those in the second tier.

To this end, we are currently investigating automatic data integration tools to span our heterogeneous federation of databases into what appears as a single (but virtual) database. Such an approach seems like a logical progression since not only does it remove implementation dependencies for each site, but it also manages the interconnections between sites. Users at different project locations and people from the outside could then interface, query and interact with the uniform interface at a single site, rather than considering the details of all the different laboratory databases. This will allow us to adopt a hub and spoke topology of database connections rather than a fully interconnected network, simplifying the interface as compared to an all-with-all federation. It also provides a graceful transition from a very loose federation to a single unified database.

A number of technologies that may be useful in this include the SDSC Storage Resource Broker (SRB, http://www.npaci.edu/DICE/SRB, a piece of client–server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets), Bio-Kleisli (17) and the Object-Protocol Model (18).

OBTAINING AND USING THE SOFTWARE

The complete code for SPINE (2.0), including table creation commands and the associated data dictionary, may be downloaded at http://spine.nesg.org/download/. Currently a number of the targets in SPINE are publicly available. These correspond to any protein targets whose structure has been solved. A demo version of the SPINE database including this public data generated by the NESG consortium can be accessed at http://spine-dev.nesg.org/. In the future, it is planned that all of SPINE and its data will be made publicly available.

ACKNOWLEDGEMENTS

We thank Michael Baran, Natalia Denissova and Chi Kent Ho for helpful discussions. All of the authors belong to the Northeast Structural Genomics Consortium.

REFERENCES

1. Burley,S.K. (2000) An overview of structural genomics. *Nature Struct. Biol.*, **7** (Suppl.), 932–934.
2. Berman,H.M., Bhat,T.N., Bourne,P.E., Feng,Z., Gilliland,G., Weissig,H. and Westbrook,J. (2000) The Protein Data Bank and the challenge of structural genomics. *Nature Struct. Biol.*, **7** (Suppl.), 957–959.
3. Christendat,D., Yee,A., Dharamsi,A., Kluger,Y., Savchenko,A., Cort,J.R., Booth,V., Mackereth,C.D., Saridakis,V., Ekiel,I. *et al*. (2000) Structural proteomics of an archaeon. *Nature Struct. Biol.*, **7**, 903–909.
4. Brenner,S.E., Barken,D. and Levitt,M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Res.*, **27**, 251–253.
5. Bertone,P., Kluger,Y., Lan,N., Zheng,D., Christendat,D., Yee,A., Edwards,A.M., Arrowsmith,C.H., Montelione,G.T. and Gerstein,M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.*, **29**, 2884–2898.
6. Lan,N., Montelione,G.T. and Gerstein,M. (2003) Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Curr. Opin. Chem. Biol.*, **7**, 44–54.
7. 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.
8. Liu,J. and Rost,B. (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.
9. Carter,P., Liu,J. and Rost,B. (2003) PEP: predictions for entire proteomes. *Nucleic Acids Res.*, **31**, 410–413.
10. Baran,M.C., Moseley,H.N., Sahota,G. and Montelione,G.T. (2002) SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. *J. Biomol. NMR*, **24**, 113–121.
11. Qian,J., Stenger,B., Wilson,C.A., Lin,.J, Jansen,R., Teichmann,S.A., Parks,J., Krebs,W.G., Yu,H., Alexandrov,V., Echols,N. and Gerstein,M. (2001) PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.*, **29**, 1750–1764.
12. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
13. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
14. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al*. (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
15. Seavey,B.R., Farr,E.A., Westler,W.M. and Markley,J.L. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**, 217–236.
16. Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al*. (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
17. Davidson,S.B., Overton,C., Tannen,V. and Wong,L. (1997) Bio-Kleisli: a digital library for biomedical researchers. *Int. J. Digit. Libr.*, **1**, 36–53.
18. Chen,I.A. and Markowitz,V.M. (1995) An overview of the Object-Protocol Model (OPM) and OPM data management tools. *Inf. Syst.*, **20**, 393–418.