EDITOR'S CORNER

The Protein Target List of the Northeast Structural Genomics Consortium

The U.S. NIH Protein Structure Initiative (PSI) is a joint government, university, and industry effort, organized and supported by the National Institute of General Medical Sciences, and aimed at reducing the costs and increasing the speed of protein structure determination. Its long-range goal is to make the 3D atomic-level structures of most proteins in nature easily obtainable from knowledge of their corresponding DNA sequences (http://www.nigms.gov/psi). It is the primary U.S. component of a broad international effort in structural genomics, involving at least 20 projects throughout the world.[1–6]

In order to minimize overlap of their efforts, most of these structural genomics pilot projects make their protein target lists and progress reports publicly available. These protein target lists provide dynamic summaries of progress on the production and structure determination of each target protein. These Web-accessible data represent a tremendously valuable new resource to the biological science community, which is only beginning to be widely recognized. As illustrated in the article by Liu et al. in this issue of *Proteins*, much thought and effort, often involving advanced bioinformatics analysis, has gone into developing these protein target lists. The article by O'Toole et al. in this issue of *Proteins* describes some of the features of these protein targets lists, the overlap between these worldwide efforts, and a first pass at the data mining that becomes possible by analyzing success and failure at various points along the structure production pipeline across thousands of protein targets. Such retrospective analysis of structural genomics data has the potential to greatly improve methods for protein expression, sample preparation, functional characterization, and structure determination. In addition, the targets lists themselves provide inventories of protein expression vectors, protein samples, and many other biochemical reagents that are generally freely available to the broader biological community.

The Northeast Structural Genomics Consortium (NESG) is one of the several pilot projects of the PSI. Its primary goals are to develop and refine new technologies for high-throughput protein production and structure determination by both NMR and X-ray crystallography, and to apply these technologies in determining representative structures of the domain sequence families that constitute eukaryotic proteomes. The project (http://www.nesg.org) is developing technology aimed at optimizing each stage of the structure determination pipeline, including intelligent protein target selection, high-throughput, and cost-effective protein sample production, robotics-aided protein crystallization screening, rapid NMR data collection, automated NMR and X-ray diffraction data analysis, and integrated databases for laboratory information management and structure–function annotations. The key short-term goal of the project is to construct a technology platform capable of experimentally determining 100–200 sequence-unique NMR or X-ray crystal structures of proteins per year.

Most structural genomics projects involve collaborative interactions between multiple research groups, coordinated through LIMS. The development and integration of these LIMS are significant challenges that are being addressed both individually and collectively by the structural genomics research community. SPiNE (http://spine.nesg.org)[7,8] is a data warehouse and integrated data tracking tool that holds detailed records about the cloning, expression, purification, biophysical characterization, crystallization, and structure determination by NMR and/or X-ray crystallography of each target under study by the NESG Consortium. The NESG also aims at correlating the structural data produced by the project with the extensive biological data emerging from large-scale functional genomics efforts (e.g., see Goh et al.[8] and Carter et al.[9]).

NESG PROTEIN TARGETS

Most of the current NESG target proteins are full-length polypeptide chains shorter than 340 amino acids, selected from domain sequence clusters generated by bioinformatic analysis (see the Liu et al. article in this issue of *Proteins*). Each cluster represented in the NESG target list consists

Abbreviations: 3D, three-dimensional; CGI, Common Gateway Interface; CLUP, Clusters of Proteins, a database of protein families; LIMS, Laboratory Information Management System; NESG, Northeast Structural Genomics Consortium; NIH, National Institutes of Public Health; ORF, open reading frame; PDB, Protein Data Bank; PEP, Predictions for Entire Proteomes, an SRS-based database of proteomes with predicted structural and functional features; PI, isoelectric pH; PSI, Protein Structure Initiative; SeMet, selenomethionine; SRS, Sequence Retrieval System; SPiNE, Structural Proteomics in the Northeast, a LIMS for the NESG; XML, Extensible Markup Language.

Grant sponsor: Protein Structure Initiative of the National Institutes of Health, Institute of General Medical Sciences; Grant number: P50-GM62413.

Received 3 July 2003; Accepted 2 January 2004

Published online 20 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20091

of two or more putative structural domains with at least one representative from a set of 5 eukaryotic “target proteomes” (*Homo sapiens, Arabodopsis thaliana, Drosophila melanogaster, Caenorhabditis elegans,* and *Saccharomyces cerevisiae*). These domain clusters also include homologues from a large set of prokaryotic “reagent proteomes.” One selection criterion is that we cannot build accurate 3D structure models for any members of these domain clusters. Currently, proteins included in the NESG target list are also filtered to exclude those likely to be integral membrane proteins, coil–coil structures, or largely disordered proteins. Protein targets containing signal sequences, indicating that they are likely to be secreted, are flagged, as these often require special treatment in expression and refolding. This target selection process identifies proteins or protein domains that can provide structural data spanning “structure space,” improve the understanding of evolutionary relationships between proteins, and/or aid in developing hypotheses about possible biochemical functions of these proteins. The NESG target list also includes about 100 proteins that were selected early in the project as “technology development targets.” These are small proteins whose 3D structures are not known, and which are not members of our eukaryotic domain family clusters. In addition, the NESG protein target list includes several proteins of known 3D structure that are suspected binding partners of our primary targets of unknown structure. These are reagents used in forming protein complexes that may be more tractable for structural studies than the target protein alone. The phylogenetic distribution of the current NESG Protein Target list is shown in Figure 1.

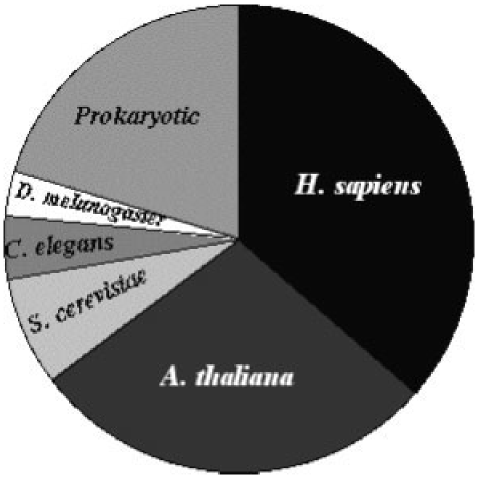


Fig. 1. Phylogenetic distribution of NESG target proteins. The phylogenetic distribution of NESG targets changes over time as new targets are added. Currently, 80% of the NESG target proteins are from the eukaryotic model organisms: *A. thaliana, C. elegans, D. melanogaster, H. sapiens,* and *S. cerevisiae*. As illustrated in the pie chart, the majority of the proteins currently targeted are from the human and *Arabidopsis* proteomes. The remaining 20% of the NESG target proteins are of prokaryotic origin, mostly homologues of eukaryotic domains, including proteins from both archea and eubacteria. These organisms include *Aeropyrum pernix, Aquifex aeolicus, Archaeglobus fulgidis, Bacillus subtilis, Brucella melitensis, Campylobacter jejuni, Caulobacter crescentus, Deinococcus radiodurans, Escherichia coli, Fusobacterium nucleatum, Haemophilus influenzae, Helicobacter pylori, Lactococcus lactis, Methanobacterium thermoautotrophicum, Neisseria meningitides, Pyrococcus furiosus, Pyrococcus horikoshi, Staphylococcus aureus, Streptococcus pyogenes, Streptomyces coelicolor, Thermoplasma acidophilum, Thermotoga maritima, Thermus thermophilus,* and *Vibrio cholera*. In addition, the NESG has also targeted a small number of proteins from two viral genomes, *Human cytomegalovirus* (AD169) and the *Murine* gammaherpesviruses.

ZebaView is a Web-based software that functions as the “Official Target List of the NESG Consortium.” It includes several tools designed to organize efforts across the consortium, to prevent duplication of the work of other structural biology groups and structural genomics consortia, and to provide, in the public domain, a catalog of the biological reagents generated by the project. NESG protein targets are entered, modified, and viewed through SPiNE’s Web-based interface.[7,8] ZebaView (http://www-nmr.cabm.rutgers.edu/bioinformatics/zebaview/) provides a limited view of the SPiNE data archive. It complements SPiNE by organizing and displaying key information essential to the molecular biology, biochemistry, and certain project management components of the project.

IMPLEMENTATION OF ZEBAVIEW

The ZebaView site combines static and dynamic Web pages produced by Perl-based CGI scripts accessing data stored in XML format. The XML document-type definitions used by ZebaView (Table I) include those that have been developed by the international structural genomics community and described on TargetDB website (http://targetdb.pdb.org/apps/target.txt) of the PDB database,[10] as well as other data items that are shared between ZebaView and the SPiNE database. ZebaView provides a concise summary of the status in the structure production process for each NESG target. It also has a many links to more in-depth information, including tools to evaluate progress on NESG target families by other structural genomics pilot projects, and various maintenance tools. All of the information on target identities, their sequences, and the progress in sample production and structure determination summarized in ZebaView is freely available to the public.

PROTEIN TARGET LIST VALIDATION

Protein targets are provided to ZebaView along with the associated information summarized in Table I. Upon submission of targets, certain features (including the pI and molecular mass) are calculated automatically. ZebaView then performs validations to prevent erroneous data from being included in the database. These validation processes are summarized in Table II. For each protein target, amino acid sequence data are obtained from the SWISS-PROT database[11] and DNA sequence data from the NCBI GenBank[12] database. These amino acid and DNA sequences are stored in ZebaView. Protein targets with conflicts between their reported amino acid sequence and the amino acid sequence coded by their reported nucleotide sequences are flagged in the database. In most cases, the agreement is perfect, but in about 5% of cases, there are one or more discrepancies,

TABLE I. ZebaView Data Dictionary

Data Field	Description
Required and validated data items	
ID	NESG ID, in the format X{x}Y#, where X{x} is an organism code, Y is the institution code, and # is the target number. Each NESG ID is unique and sequentially assigned.
Lab	Lab name. For all NESG targets this is Northeast Structural Genomics Consortium.
Date	Date of most recent update to target entry in YYYY-MM-DD format.
Sequence	Amino acid sequence in IUPAC 1-letter code.
Name	Protein name.
Coding_seq	Nucleotide sequence in IUPAC 1-letter code.
SourceOrganism	Scientific name of the source organism.
Number_of_residues	Number of letters in amino acid sequence.
Predicted_codons	Length of coding_seq divided by 3.
NESG_cluster_id	Assigned cluster ID number.
pI	Calculated pI.
MW, MW_N15, MW_N15_C13, MW_SeMet	Calculated molecular mass of native and labeled proteins.
Extinction	Calculated molar extinction coefficient at 280 nm.
Swiss-prot_id	SWISS-PROT or TrEMBL ID number.
Swiss-prot_acc	SWISS-PROT or TrEMBL accession number.
Optional data items	
Fprimer, Rprimer	Forward and reverse primer sequences in IUPAC 1-letter code.
Rost_frag_id	Assigned fragment ID.
Rost_orf_id	Assigned ORF ID.
Rost_frag_region	Region of the ORF specified by Rost_frag_id.
Leverage	Number of proteins in SWISS-PROT that can theoretically be modeled.
Status	Progress indicators.
Flag	Tags problems with the sequence.

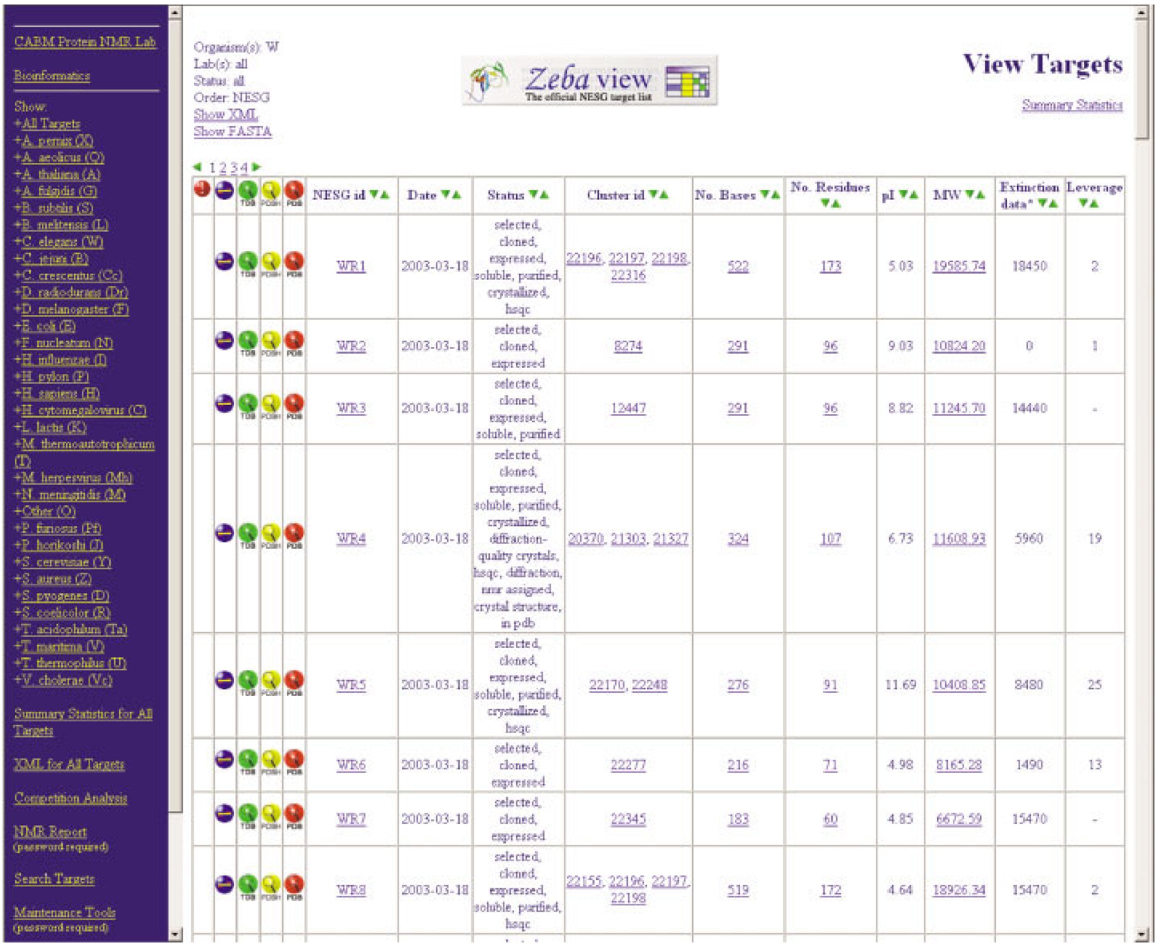
TABLE II. Summary of Protein Target Validations Performed by ZebaView Upon Submission of Targets

Data field	Validation Performed
ID	Must be the next sequential ID that is not currently assigned. If an invalid ID is entered, a valid ID is suggested.
Lab	A lab must be selected.
Date	Must be in YYYY-MM-DD format.
Name	A name must be entered.
Swiss-prot_id	An ID must be entered.
Swiss-prot_acc	An accession number must be entered.
NESG_cluster_id	A cluster ID must be entered.
Sequence	Must consist only of IUPAC 1-letter amino acid abbreviations. Entries will be converted to uppercase.
Number_of_residues	Must equal the number of characters in the sequence entry.
Coding_seq	Must consist only of IUPAC 1-letter nucleotide abbreviations. Entries will be converted to lowercase. Should include stop codon.
Predicted codons	Must equal the number of nucleotides divided by 3. This number must equal the number of residues plus 1 (stop codon included).
Status	At least one status must be selected. More than one status is allowed.
Fprimer, Rprimer	Optional, but if entered must consist only of IUPAC 1-letter nucleotide abbreviations. Case will be preserved.

which are generally resolved by DNA sequencing of the cloned target genes.

VIEWING TARGET INFORMATION

The ZebaView website contains two frames: the side navigation bar and the main information frame (Fig. 2). For each target listed, several pieces of information are available. The first column indicates whether the target is flagged by the target sequence validation process described above, indicating possible errors in the sequence of the ORF. The second column indicates whether cloning primers have been designed for the protein target. If primer information is stored, the icon in the second column is a small disk icon; if primer information is not stored, a round primer icon is displayed. The three "magnifying glass icons" link to searches of each target's protein sequence against different protein databases. Clicking on the green icon activates a search against the PDB TargetDB,[10] a repository of target sequence and progress information for 15 international structural genomics



CABM Protein NMR Lab

Bioinformatics

Show:
+All Targets
+A. pernix (X)
+A. aeolicus (Q)
+A. thaliana (A)
+A. fulgidis (G)
+B. subtilis (S)
+B. melitensis (L)
+C. elegans (W)
+C. jejuni (B)
+C. crescentus (Cc)
+D. radiodurans (Dr)
+D. melanogaster (F)
+E. coli (E)
+F. nucleatum (N)
+H. influenzae (I)
+H. pylori (P)
+H. sapiens (H)
+H. cytomegalovirus (C)
+L. lactis (K)
+M. thermoautotrophicum (T)
+M. herpesvirus (Mh)
+N. meningitidis (M)
+Other (O)
+P. furiosus (Pf)
+P. horikoshi (J)
+S. cerevisiae (Y)
+S. aureus (Z)
+S. pyogenes (D)
+S. coelicolor (R)
+T. acidophilum (Ta)
+T. maritima (V)
+T. thermophilus (U)
+V. cholerae (Vc)

Summary Statistics for All Targets

XML for All Targets

Competition Analysis

NMR Report
(password required)

Search Targets

Maintenance Tools
(password required)

Organism(s): W
Lab(s): all
Status: all
Order: NESG
Show XML
Show FASTA

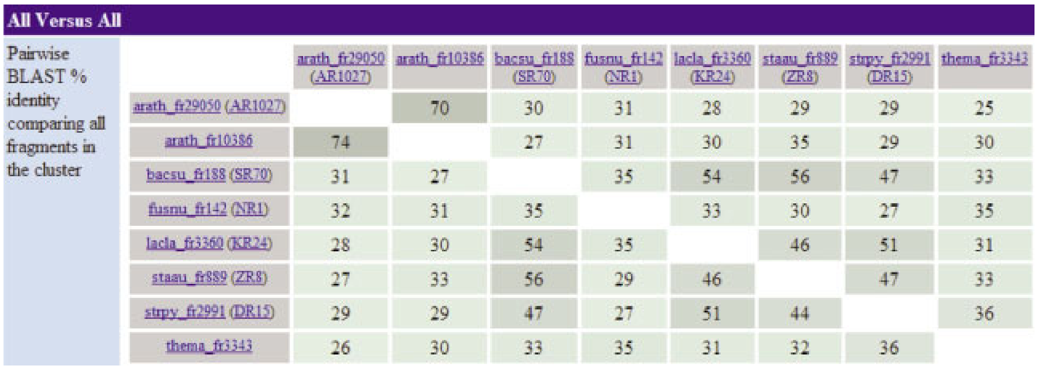
View Targets

Summary Statistics

◄ 1 2 3 4 ►

NESG id ▼▲	Date ▼▲	Status ▼▲	Cluster id ▼▲	No. Bases ▼▲	No. Residues ▼▲	pI ▼▲	MW ▼▲	Extinction data* ▼▲	Leverage ▼▲
WR1	2003-03-18	selected, cloned, expressed, soluble, purified, crystallized, hsqc	22196, 22197, 22198, 22316	522	173	5.03	19585.74	18450	2
WR2	2003-03-18	selected, cloned, expressed	8274	291	96	9.03	10824.20	0	1
WR3	2003-03-18	selected, cloned, expressed, soluble, purified	12447	291	96	8.82	11245.70	14440	-
WR4	2003-03-18	selected, cloned, expressed, soluble, purified, crystallized, diffraction-quality crystals, hsqc, diffraction, nmr assigned, crystal structure, in pdb	20370, 21303, 21327	324	107	6.73	11608.93	5960	19
WR5	2003-03-18	selected, cloned, expressed, soluble, purified, crystallized, hsqc	22170, 22248	276	91	11.69	10408.85	8480	25
WR6	2003-03-18	selected, cloned, expressed	22277	216	71	4.98	8165.28	1490	13
WR7	2003-03-18	selected, cloned, expressed	22345	183	60	4.85	6672.59	15470	-
WR8	2003-03-18	selected, cloned, expressed, soluble, purified, hsqc	22155, 22196, 22197, 22198	519	172	4.64	18926.34	15470	2

Fig. 2. The ZebaView website. On the left, the navigation bar allows users to select target lists by organism, search the targets, modify the target list, or access the help page. The table header summarizes which targets are in the current view, and offers links to XML and FASTA versions of the information displayed. For each target, summary information and links to in-depth details are provided. The details of some of this linked information are described in the text.



All Versus All

Pairwise BLAST % identity comparing all fragments in the cluster

	arath_fr29050 (AR1027)	arath_fr10386	bacsu_fr188 (SR70)	fusnu_fr142 (NR1)	lacla_fr3360 (KR24)	staau_fr889 (ZR8)	strpy_fr2991 (DR15)	thema_fr3343
arath_fr29050 (AR1027)		70	30	31	28	29	29	25
arath_fr10386	74		27	31	30	35	29	30
bacsu_fr188 (SR70)	31	27		35	54	56	47	33
fusnu_fr142 (NR1)	32	31	35		33	30	27	35
lacla_fr3360 (KR24)	28	30	54	35		46	51	31
staau_fr889 (ZR8)	27	33	56	29	46		47	33
strpy_fr2991 (DR15)	29	29	47	27	51	44		36
thema_fr3343	26	30	33	35	31	32	36	

Fig. 3. Example of information linked from ZebaView to an NESG protein target cluster. In this case, a particular NESG target in ZebaView is linked to the PEP/CLUP domain cluster viewer.[9] The pairwise sequence identity matrix of PEP/CLUP provides information about relationships among targets within the cluster, with NESG targets indicated by their NESG IDs (e.g., ZR8). While several members of this particular cluster are NESG targets, other cluster members are not on the current target list, usually because these sequence fragments are portions of large proteins, $>$ 340 amino acid residues in length. Following structure determination of one cluster member (e.g., ZR8), this matrix provides guidance in determining the need for determining a second structure from the domain family.

projects (http://targetdb.pdb.org). Clicking on the yellow icon activates a search against sequences of protein structures currently "on hold" in the PDB, and the red icon triggers a search against protein sequences of the entire PDB.[13] The NESG-ID (identifier), date modified, target status, cluster-ID, number of bases, number of residues, pI, molecular mass, extinction coefficient, and "leverage value" are also listed. The "leverage value" is define as the number of homologues in the SWISS-PROT + TrEMBL database,[11] with profile-based PSI-BLAST searches,[14] E value $< 10^{-10}$. These "leverage values" estimate the number of proteins that can be modeled with good accuracy once the 3D structure of the corresponding target protein is determined. The "target status" is a controlled vocabulary summarizing progress across the NESG consortium in cloning, expressing, purifying, and determining the 3D structure of the target. This progress status information is archived in SPiNE and downloaded daily to ZebaView. The "Number of Bases" link shows the nucleotide sequence of the protein and restriction sites present, and the "Number of Residues" link shows the amino acid sequence, as well as the numbers of Met, Cys, Gly, and Pro residues, which are relevant in structural analysis by X-ray crystallography and NMR spectroscopy. The molecular mass link displays the molecular masses of unlabeled, uniformly ^{15}N-, uniformly ^{15}N,^{13}C-, and SeMet-labeled versions of the protein, which are useful for analytical mass spectroscopy and other analyses. The target lists may be sorted by any of the columns by clicking on the arrows next to the column headings.

The complete NESG protein target list, along with all of the associated information stored in ZebaView, can be downloaded using the "XML for all Targets" link in the left menu bar, and specific protein sequences can be searched for in ZebaView with BLAST[14,15] using the "Search Targets" link (Fig. 2). The "Show XML" hyperlink displays the complete document for the targets in the current table, and the "Show FASTA" hyperlink displays a FASTA-formatted list of the targets and their DNA sequences. The "Summary Statistics" link in the upper right corner jumps to the bottom of the ZebaView table, where basic summary statistics are displayed.

LINKS TO OTHER INFORMATION RESOURCES

ZebaView also provides links to some key associated NESG consortium databases. The "NESG ID" links to the corresponding SPiNE target record, providing detailed information about the production and characterization of the sample. This information is maintained under password protection until the corresponding structure is deposited in the PDB, at which point the entire SPiNE record of protein production information becomes publicly accessible. Images of these completed 3D protein structures are also presented in the Structure Gallery of the SPiNE database.[8] The ZebaView "Cluster ID" icon links to information about the corresponding domain sequence cluster (see Liu et al. in this issue of *Proteins*) provided by PEP/CLUP,[9] an SRS-based database of these domain sequence families. In some cases, one protein target may map to multiple domain family clusters (Fig. 2). An example of pairwise sequence similarity information for all members of a target cluster, provided by the PEP/CLUP cluster viewer, is shown in Figure 3.

SELECTING TARGETS TO DISPLAY

ZebaView can be used to display tables in two ways. To view target lists by organism or the entire list, the user can click on the hyperlinks on the navigation bar below the "Show:" heading. To create a more specific view, the targets may be searched by organism, laboratory, and status using the form accessed via the "Search Targets" link on the navigation bar.

COMPETITION ANALYSIS

In addition to the BLAST search tools for each target described above, ZebaView also provides a summary of the current competition via the "Competition Analysis" link. ZebaView creates this report, which includes two sections, daily. The first part of the report is a list of NESG targets not yet "in progress" (those that are "selected" but not yet cloned, or "cloned," "expressed" but not partially or fully soluble) that have BLAST[14] hits ($E < 10^{-3}$) to protein sequences in the PDB or "PDB structures on hold," or to sequences in the TargetDB database, which are reported to be "crystallized," "NMR assigned," or have recently determined 3D structures. This information flags the protein targets in ZebaView that should potentially be terminated. The second part of the report has a list of NESG targets "in progress" (those that are cloned, expressed, and observed to be partially or fully soluble) that have hits to this same set of protein sequences. This information is used to guide consortium members in deciding the value of carrying through structure determinations when 3D structures of homologous proteins have been completed (or nearly completed) by other research groups; for example a protein flagged in the first table as having a homologue that has been "crystallized" by a competing group may be carried through to completion if it also listed in the second table, because the amount of effort already invested in the target is significant. Efforts are in progress to develop standard criteria based on homology modeling that will determine which sequences should no longer be targeted by the project, using 3D protein structure data as it becomes available through the efforts of the NESG and/or other structural biology groups.

MAINTENANCE TOOLS

In the password-protected "Maintenance Tools" section of ZebaView, there are three tools available: two tools to update the ZebaView target list, and one tool for users to request a weekly competition analysis. As described above, the SPiNE update tool is automatically run daily to keep ZebaView current. However, if a user updates a target's status in SPiNE and wishes to see the changes reflected in ZebaView immediately, the user may click the "Update ZebaView Target Status Using SPiNE" link.

A form to request an e-mail copy of the "Competition Analysis" for a subset of specified targets is also available from the Maintenance Tools menu. The user specifies an e-mail address and a list of NESG target IDs, and ZebaView e-mails subscribers reports on the status of these targets and their homologues in PDB, PDB on hold, and the TargetDB database once each week.

ACCESSIBILITY OF NESG BIOLOGICAL REAGENTS

ZebaView provides a summary of all the protein targets cloned and characterized by the NESG Consortium. The corresponding clones and associated reagents are generally available from the consortium for noncommercial uses. Thus, ZebaView provides an important resource to the broader biological community by presenting a listing of available biological reagents.

CONCLUSIONS

The NESG Target List provides the basis for a pilot project in structural genomics and is freely accessible for public analysis. The ZebaView software provides a simple approach for organizing and sharing a key subset of basic protein target information across the NESG. The Web-based tool is open to the public and provides a summary of thousands of biological reagents available to the scientific community. Many of the reagents of the NESG project are freely available for research purposes. ZebaView is a view port of the broader SPiNE data warehouse, which archives a much more complete set of information on each protein target. This data archive is tremendously valuable for retrospective analysis and optimization of the sample production process. ZebaView and SPiNE communicate using a simple XML exchange data dictionary. In addition to its utility for organizing and presenting the Official Protein Target List of the NESG, ZebaView is a valuable tool for organizing such information in traditional structural biology laboratories. The ZebaView software is available from the authors.

ACKNOWLEDGMENTS

This work was carried out in partial fulfillment of Z. Wunderlich's requirements for the Henry Rutgers Scholars Program.

Zeba Wunderlich
Center for Advanced Biotechnology and Medicine,
Department of Molecular Biology and Biochemistry, and
Northeast Structural Genomics Consortium (NESG),
Rutgers University,
Piscataway, NJ 08854-5638
Thomas B. Acton
Center for Advanced Biotechnology and Medicine,
Department of Molecular Biology and Biochemistry, and
Northeast Structural Genomics Consortium (NESG),
Rutgers University,
Piscataway, NJ 08854-5638
Jinfeng Liu
Department of Biochemistry and Molecular Biophysics,
Columbia University Center for Computational Biology and Bioinformatics (C2B2), and Northeast Structural Genomics Consortium (NESG),
Columbia University, New York, NY 10032
Gregory Kornhaber
Center for Advanced Biotechnology and Medicine,
Department of Molecular Biology and Biochemistry, and
Northeast Structural Genomics Consortium (NESG),
Rutgers University, Piscataway, NJ 08854-5638 Department of Biochemistry,
Robert Wood Johnson Medical School,
Piscataway, NJ 08854-5638
John Everett
Center for Advanced Biotechnology and Medicine,
Department of Molecular Biology and Biochemistry, and
Northeast Structural Genomics Consortium (NESG),
Rutgers University,
Piscataway, NJ 08854-5638
Department of Biochemistry,
Robert Wood Johnson Medical School,
Piscataway, NJ 08854-5638
Phil Carter
Department of Biochemistry and Molecular Biophysics,
Columbia University Center for Computational Biology and Bioinformatics (C2B2), and Northeast Structural Genomics Consortium (NESG),
Columbia University,
New York, NY 10032
Ning Lan
Department of Molecular, Cellular, and
Developmental Biology,
Department of Computer Science, and Northeast
Structural Genomics Consortium,
266 Whitney Ave.,
Yale University,
New Haven, CT 06520
Nathaniel Echols
Department of Molecular, Cellular, and
Developmental Biology,
Department of Computer Science, and Northeast
Structural Genomics Consortium,
266 Whitney Ave.,
Yale University,
New Haven, CT 06520
Mark Gerstein
Department of Molecular, Cellular, and
Developmental Biology,
Department of Computer Science, and Northeast Structural Genomics Consortium,
266 Whitney Ave.,
Yale University,
New Haven, CT 06520
Burkhard Rost
Department of Biochemistry and Molecular Biophysics,
Columbia University Center for Computational Biology and Bioinformatics (C2B2), and Northeast Structural Genomics Consortium (NESG),
Columbia University,
New York, NY 10032
Gaetano T. Montelione
Center for Advanced Biotechnology and Medicine,
Department of Molecular Biology and Biochemistry, and
Northeast Structural Genomics Consortium (NESG),
Rutgers University,

Piscataway, NJ 08854-5638
Department of Biochemistry,
Robert Wood Johnson Medical School,
Piscataway, NJ 08854-5638

REFERENCES

1. Heinemann U. Structural genomics in Europe: slow start, strong finish? Nat Struct Biol 2000;7:940–942.
2. Terwilliger TC. Structural genomics in North America. Nat Struct Biol 2000;7:935–939.
3. Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, Kyogoku Y, Miki K, Masui R, Kuramitsu S. Structural genomics projects in Japan. Nat Struct Biol 2000;7:943–945.
4. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2:801–809.
5. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. Protein Sci 2002;11:723–738.
6. Gong WM, Liu HY, Niu LW, Shi YY, Tang YJ, Tang MK, Wu JH, Liang DC, Wang DC, Wang JF, Ding JP, Hu HY, Huang QH, Zhang QH, Lu SY, An JL, Liang YH, Zheng XF, Gu XC, Su XD. Structural genomics efforts at the Chinese Academy of Sciences and Peking University. J Struct Funct Genomics 2003;4:137–139.
7. Bertone P, Kluger Y, Lan N, Zheng D, Christendat D, Yee A, Edwards AM, Arrowsmith CH, Montelione GT, Gerstein M. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res 2001;29:2884–2898.
8. Goh C, Lan N, Echols N, Douglas S, Milburn D, Bertone P, Xiao R, Ma L, 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: 2833–2838.
9. Carter P, Liu J, Rost B. PEP: Predictions for entire proteomes. Nucleic Acids Res 2003;31:410–413.
10. Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and structural genomics. Nucleic Acids Res 2003;31: 489–491.
11. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.
12. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23–27.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
14. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.