Nucleic Acids Research, 2006, Vol. 34, Database issue ***D471–D474***
doi:10.1093/nar/gkj036

SGCEdb: a flexible database and web interface integrating experimental results and analysis for structural genomics focusing on *Caenorhabditis elegans*

David H. Johnson[1], Jun Tsao[1,2], Ming Luo[1,2] and Mike Carson[1,3,*]

[1]Center for Biophysical Sciences and Engineering, [2]Department of Microbiology and [3]Department of Biomedical Engineering, University of Alabama at Birmingham, Birmingham, AL, USA

Received August 13, 2005; Revised and Accepted September 26, 2005

ABSTRACT

The SGCEdb (http://sgce.cbse.uab.edu) database/interface serves the primary purpose of reporting progress of the Structural Genomics of *Caenorhabditis elegans* project at the University of Alabama at Birmingham. It stores and analyzes results of experiments ranging from solubility screening arrays to individual protein purification and structure solution. External databases and algorithms are referenced and evaluated for target selection in the human, *C.elegans* and *Pneumocystis carinii* genomes. The flexible and reusable design permits tracking of standard and custom experiment types in a scientist-defined sequence. The database coordinates efforts between collaborators and is adaptable to a wide range of biological applications.

INTRODUCTION

The NIH-NIGMS-sponsored Protein Structural Initiative (1) initially sponsored seven pilot projects in structural genomics, including the Southeast Collaboratory for Structural Genomics (SECSG) (2). The Structural Genomics of *Caenorhabditis elegans* (SGCE) project established a pipeline for high-throughput protein expression, microarray crystallization screening, X-ray data collection and user-friendly bioinformatics. The initial SGCE focus was the nematode worm *C.elegans*, one of the best-studied multicellular model organisms (3) whose complete genomic sequence is known (4). The SGCE project is facilitated by the *C.elegans* ORFeome project (5), which aims at cloning all predicted protein-encoding open reading frames (ORFs) as Gateway Entry clones (6). The 96-well plates of cDNA supplied enabled a high-throughput approach to recombinant protein expression and analysis.

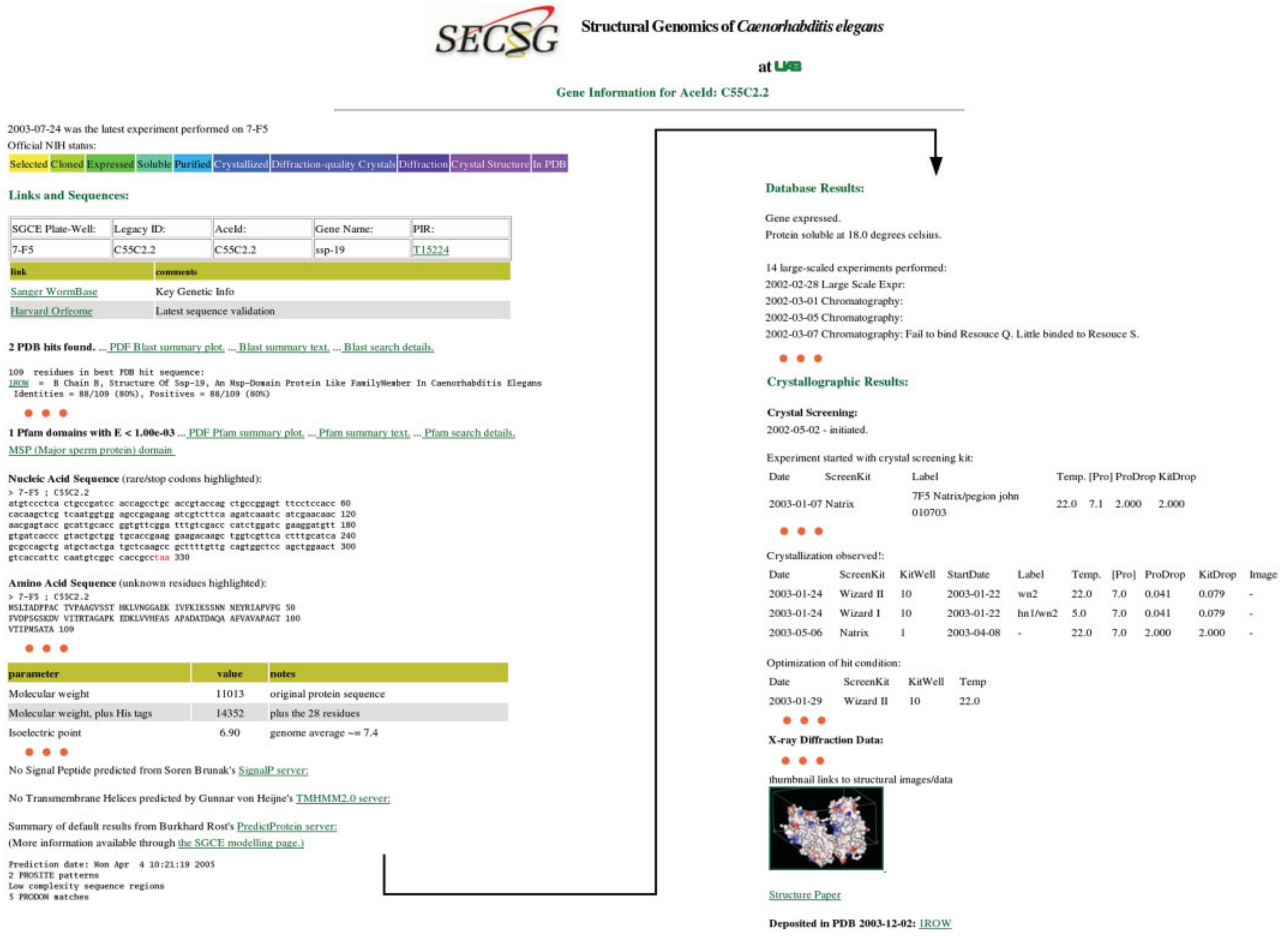
The initial purpose of the SGCEdb was to report progress on the SGCE project. However, a pipeline to express, purify and crystallize over 10 000 *C.elegans* ORFs (7) required more comprehensive tracking for meaningful analysis. Additional requirements for the project included sample tracking and monitoring group progress for a smooth transition between the expression, purification and crystallization groups. Plates of cDNA ORFs for robotic screening needed to be prioritized based on the expectation of results. In order to prioritize, established external databases and analysis algorithms must be applied on a genome-wide basis. Finally, the methods used to accomplish each step were updated on a regular basis; therefore, the database and interface system needed to be flexible enough to rearrange and modify experiments before the data stored became obsolete with respect to the experiments performed. The resulting SGCEdb is a database and interface framework that has been applied to a variety of genomes and heterogeneous experiment methodologies.

THE DATABASE

Browsing the SGCEdb

The website http://sgce.cbse.uab.edu provides the public interface. Selecting the 'Results' button at the top of the homepage allows searching for any protein target. The target ID used for selecting a single protein is generally the WormBase (8) 'AceID', familiar to *C.elegans* researchers. Unlike WormBase, which concentrates on genomic information, the SGCEdb concentrates on information of the putative proteins. This is similar to other structural genomics databases in spirit (9). The SGCE was among the first structural genomics groups to make its protein production data publicly available.

*To whom correspondence should be addressed. Tel: +1 205 934 1983; Fax: +1 205 975 0538; Email: carson@uab.edu
Correspondence may also be addressed to Ming Luo. Tel: +1 205 934 4259; Fax: +1 205 975 9578; Email: mingluo@uab.edu



SECSG Structural Genomics of *Caenorhabditis elegans*

at UAB

Gene Information for AceId: C55C2.2

2003-07-24 was the latest experiment performed on 7-F5

Official NIH status:

Selected Cloned Expressed Soluble Purified Crystallized Diffraction-quality Crystals Diffraction Crystal Structure In PDB

Links and Sequences:

SGCE Plate-Well:	Legacy ID:	AceId:	Gene Name:	PIR:
7-F5	C55C2.2	C55C2.2	ssp-19	T15224

link	comments
Sanger WormBase	Key Genetic Info
Harvard Orfeome	Latest sequence validation

2 PDB hits found. ... PDF Blast summary plot. ... Blast summary text. ... Blast search details.

```
109  residues in best PDB hit sequence:
1ROW  =  B Chain B, Structure Of Ssp-19, An Msp-Domain Protein Like FamilyMember In Caenorhabditis Elegans
 Identities = 88/109 (80%), Positives = 88/109 (80%)
```

● ● ●

1 Pfam domains with E < 1.00e-03 ... PDF Pfam summary plot. ... Pfam summary text. ... Pfam search details.

MSP (Major sperm protein) domain

Nucleic Acid Sequence (rare/stop codons highlighted):

```
> 7-F5 ; C55C2.2
atgtccctca ctgccgatcc accagcctgc accgtaccag ctgccggagt ttcctccacc 60
cacaagctcg tcaatggtgg agccgagaag atcgtcttca agatcaaatc atcgaacaac 120
aacgagtacc gcattgcacc ggtgttcgga tttgtcgacc catctggatc gaaggatgtt 180
gtgatcaccc gtactgctgg tgcaccgaag gaagacaagc tggtcgttca ctttgcatca 240
gcgccagctg atgctactga tgctcaagcc gcttttgttg cagtggctcc agctggaact 300
gtcaccattc caatgtcggc caccgcctaa 330
```

Amino Acid Sequence (unknown residues highlighted):

```
> 7-F5 ; C55C2.2
MSLTADPPAC TVPAAGVSST HKLVNGGAEK IVFKIKSSNN NEYRIAPVFG 50
FVDPSGSKDV VITRTAGAPK EDKLVVHFAS APADATDAQA AFVAVAPAGT 100
VTIPMSATA 109
```

● ● ●

parameter	value	notes
Molecular weight	11013	original protein sequence
Molecular weight, plus His tags	14352	plus the 28 residues
Isoelectric point	6.90	genome average ~= 7.4

● ● ●

No Signal Peptide predicted from Soren Brunak's SignalP server:

No Transmembrane Helices predicted by Gunnar von Heijne's TMHMM2.0 server:

Summary of default results from Burkhard Rost's PredictProtein server:
(More information available through the SGCE modelling page.)

```
Prediction date: Mon Apr  4 10:21:19 2005
2 PROSITE patterns
Low complexity sequence regions
5 PRODOM matches
```

Database Results:

Gene expressed.
Protein soluble at 18.0 degrees celsius.

14 large-scaled experiments performed:
2002-02-28 Large Scale Expr:
2002-03-01 Chromatography:
2002-03-05 Chromatography:
2002-03-07 Chromatography: Fail to bind Resouce Q. Little binded to Resouce S.

● ● ●

Crystallographic Results:

Crystal Screening:
2002-05-02 - initiated.

Experiment started with crystal screening kit:

Date	ScreenKit	Label	Temp.	[Pro]	ProDrop	KitDrop
2003-01-07	Natrix	7F5 Natrix/pegion john 010703	22.0	7.1	2.000	2.000

● ● ●

Crystallization observed!:

Date	ScreenKit	KitWell	StartDate	Label	Temp.	[Pro]	ProDrop	KitDrop	Image
2003-01-24	Wizard II	10	2003-01-22	wn2	22.0	7.0	0.041	0.079	-
2003-01-24	Wizard I	10	2003-01-22	hn1/wn2	5.0	7.0	0.041	0.079	-
2003-05-06	Natrix	1	2003-04-08	-	22.0	7.0	2.000	2.000	-

Optimization of hit condition:

Date	ScreenKit	KitWell	Temp
2003-01-29	Wizard II	10	22.0

● ● ●

X-ray Diffraction Data:

● ● ●

thumbnail links to structural images/data

Structure Paper

Deposited in PDB 2003-12-02: 1ROW

Figure 1. Web display for the target AceID no. C55C2.2. The left-hand side shows the information available for all targets. The right-hand side shows the experimental data generated. The orange ellipses indicate additional information omitted to create the figure.

The most detailed way of browsing the SGCEdb is by viewing one of the 16 566 *C.elegans* individual proteins. In this view links are provided to WormBase website (10), the ORFeome project (5) and the Protein Information Resource (11) if available. BLAST (12) results against the Protein Data Bank (PDB) structures (13) are updated weekly. BLAST results against the non-redundant database (14) and Pfam (15) results are periodically updated. Theoretical values, such as isoelectric point and hydrophobicity, are calculated from the Expasy site (16). Transmembrane protein (17), signal peptide (18) and prosite (19) results are available. For each of the 11 727 proteins with experiments currently performed, a complete record of the experimental results is also given. All these have at least expression and preliminary solubility data (7). An example web page of a protein view with a completed structure (20) is shown in Figure 1.

Results are also accessible by XML and experimental stage. The XML reports are updated weekly and are intended to provide improved interoperability with external databases. Database queries are used to format all results according to the TargetDB and the Protein Expression, Purification and Crystallization (PEPCdb) XML standards for structural genomics maintained at the PDB (13). These standards include experiment data for each stage from expression through protein structure. Individuals interested in detailed results of a specific step can go directly to that page, e.g. 'Structures'.

The modeling page contains secondary structures and other predictions available from ProteinPredict (21). The modeling page also provides distributions per plate and histograms over the entire database for calculated protein parameters. The parameter distribution per plate is used internally to prioritize efforts. A variety of per plate experimental reports are also available.

Database design

The complete system was developed in close collaboration with the SGCE project scientists and technicians. The infrastructure was constructed entirely with open-source tools including the Apache web server, the Python language and the PostgreSQL database system. A very brief synopsis is given below. A more detailed summary of the design and data entry including figures is provided in Supplementary Data.

The database of SGCEdb is separated into two distinct parts: protein source and experiment tracking. The protein source tables handle external database references, target selection and sequence analysis. Experiment tracking tables store detailed results for expression, solubility, purification and crystallization. Each schema implements advanced database methods to efficiently and comprehensively track proteins, results and analyses throughout the experimental process.

When a protein source is initially received or considered for study, the sequences are organized into plates and wells. The protein source schema has configurable plate geometry allowing it to track protein production, 96-well solubility screens and 384-well crystallization trials. By keeping plates, wells and sequences separate in the protein source schema data attributes can be assigned at an appropriate level. For instance, the Pfam (15) algorithm is initially executed and stored once per sequence. Experiment parameters common to an entire screening plate are stored once per plate whereas individual experiment results are stored at the plate level. This design provides efficient queries of experimental data by eliminating duplicate information and allowing the design to scale to experiments over the entire *C.elegans* genome.

Experiment history is stored in a condensed form, called parenthetical tree notation (22). This notation is exceptionally useful for querying lineage information. Populating the history tree is handled by the database during data entry, with the scientists only having to indicate the source of the experiment being entered. A major benefit of storing an experiment history tree is the ability to efficiently retrieve every experiment that has influenced an experiment of interest.

It is also useful to have a 'generic' experiment table when tracking the sequence of experiments. By tracking experiment history using a generic experiment ID any combination of available experiments can be combined and tracked without a major database re-design. In general the sequence of experiments is fixed by protocol. However, in the event of an unexpected result, quality control experiments such as mass spectrometry and additional chromatography column experiments may be inserted anywhere in the sequence of experiments. All subsequent experiments will include the unplanned mass spectrometry in a query of its history.

SGCEdb uses Zope (23) for data entry forms and PHP for display. A benefit of Zope is the ability to reuse common data entry components to quickly assemble entry forms for the bench scientist.

FUTURE DEVELOPMENTS

Currently the SGCEdb framework has the capability to handle individual and plate-based experiments with external database reference and algorithm analysis. A modified version of the system is in use at Beijing University. Plans for the SGCEdb framework include an experiment creation tool and modular visualization methods. An experiment creation tool would generate table definitions, entry forms and standard html views based on data types specified in a graphical interface. This tool would allow end users to rapidly add new experiments into the framework. The modular visualization methods would utilize Java display layers to allow the user to superimpose additional information where it is most useful—for instance, secondary structure under the sequence and information content over a 3D structure (24) or gel markers by a gel image. These future development plans are geared toward making customization easier for independent laboratories.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The SGCEdb was made possible by the work of Chi-Hao Luan, Shihong Qiu, Norbert Schormann, Jindrich Symersky, Yun-Jia Chen, Rita Gray, Wen Ying Huang, Zhuhua Cao, Monica Stinnet, Qiao Shang, Jennifer Zhou, Jim Finley, Robert Bunzel, Danlin Luo, Alireza Arabshahi, Yun-Jia Chen, Annette McKinstry, Alexis Stanton, Guanda Lin, Pamela Pruett, Elizaveta Karpova, Jianli An, Hongli Chen, Songlin Li, Shanyun Lu, John Zhang, Lisa Nagy, Debbie McCombs, Ben Sha, Amber Vickers, Philippe Vaglio, Jérôme Reboul, David Hill, Marc Vidal, Bi-Cheng Wang and Larry DeLucas. SGCEdb development is funded by a Protein Structure Initiative grant (NIGMS 1P50-GM62407) from the United States National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by the NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Norvell,J.C. and Zapp-Machalek,A. (2000) Structural genomics programs at the US National Institute of General Medical Sciences. *Nature Struct. Biol.*, **7**, 931.
2. Adams,M.W.W., Dailey,H.A., DeLucas,L.J., Luo,M., Prestegard,J.H., Rose,J.P. and Wang,B.C. (2003) The Southeast collaboratory for structural genomics: a high throughput gene to structure factory. *Acc. Chem. Res.*, **36**, 191–198.
3. Brenner,S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics*, **77**, 71–94.
4. The *C.elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C.elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
5. Reboul,J., Vaglio,P., Rual,J.-F., Lamesch,P., Martinez,M., Armstrong,C.M., Li,S., Jacotot,L., Bertin,N., Janky,R. *et al.* (2003) *C.elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet.*, **34**, 35–41.
6. Walhout,A.J., Temple,G.F., Brasch,M.A., Hartley,J.L., Lorson,M.A., van den Heuvel,S. and Vidal,M. (2000) GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.*, **328**, 575–592.
7. Luan,C.H., Qiu,S., Finley,J.B., Carson,M., Gray,R.J., Huang,W., Johnson,D., Tsao,J., Reboul,J., Vaglio,P. *et al.* (2004) High-throughput expression of *C.elegans* proteins. *Genome Res.*, **14**, 2102–2110.
8. Stein,L.D., Sternberg,P., Durbin,R., Thierry-Mieg,J. and Spieth,J. (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **29**, 82–86.
9. 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.
10. Chen,N., Harris,T.W., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Canaran,P., Chan,J., Chen,C.-K. *et al.* (2005)

WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.*, **33**, D383–D389.

11. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z-Z., Ledley,R.S., Kourtesis,P., Suzek,B.E. *et al.* (2003) The protein information resource. *Nucleic Acids Res.*, **31**, 345–347.
12. 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.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
15. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
16. Gasteiger,E., Hoogland,C., Gattiker,A., Duvaud,S., Wilkins,M.R., Appel,R.D. and Bairoch,A. (2005) Protein identification and analysis tools on the ExPASy server. In Walker,J.M. (ed.), *The Proteomics Protocols Handbook*. Humana Press, Totowna, NH, USA, pp. 571–607.
17. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
18. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
19. Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.
20. Schormann,N., Symersky,J. and Luo,M. (2004) Structure of sperm-specific protein SSP-19 from *Caenorhabditis elegans. Acta. Crystallogr. D Biol. Crystallogr.*, **60**, 1840–1845.
21. Rost,B., Yachdav,G. and Liu,J. (2003) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.
22. Celko,J. (2004) *Trees and Hierarchies in SQL for Smarties*. Morgan Kaufmann, San Francisco, CA, USA.
23. Grizel,P.-J. (2003) *ZOPE*. John Wiley and Sons, Hoboken, NJ, USA.
24. Li,S., Finley,J., Liu,Z.J., Qiu,S.H., Chen,H., Luan,C.H., Carson,M., Tsao,J., Johnson,D., Lin,G. *et al.* (2002) Structural genomics of *Caenorhabditis elegans*: the structure of the cytoskeleton associated protein (CAP-Gly) domain of F53F4.3. *J. Biol. Chem.*, **277**, 48596–48601.