

Protein Model Structures for Plant Biology

Organizers:

Krzysztof Fidelis – main contact person

Protein Structure Prediction Center
University of California, Davis
Genome and Biomedical Sciences Facility
451 Health Sciences Drive
Davis, CA 95616-8816
Phone: 530-754-8931, 925-373-0789, 925-980-0870
Fax: 530-754-9658
kfidelis@ucdavis.edu

Krzysztof Fidelis directs Protein Structure Prediction Center at the University of California, Davis. He is also one of the founders (1994) and a member of the Organizing Committee of CASP (Critical Assessment of Protein Structure Prediction). His research is in protein structure analysis and modeling.

Diana Murray

Center for Computational Biology and Bioinformatics
Columbia University
622 West 168th St
New York, NY 10032
Phone: 212-342-5753
Fax: 212-305-8780
dm527@columbia.edu

Diana Murray is an Associate Professor of Pharmacology in the Center for Computational Biology and Bioinformatics at Columbia University, New York. Her research involves modeling of membrane binding proteins.

Ram Samudrala

Computational Biology Research Group
Department of Microbiology, University of Washington, Seattle
UW Micro Box 357242
Seattle, WA 98195-7242
Phone: 206-732-6105
Fax: 206-732-6055
pc@compbio.washington.edu

Ram Samudrala is an Associate Professor at the Department of Microbiology, University of Washington, Seattle. His research involves, among others, prediction of structures, functions, and interactions of proteins in the rice genome.

Likely Participants

The following people have expressed explicit interest in participating in this workshop:

(1) Maqsudul Alam
Professor,
Department of Microbiology
Director
Advance Studies in Genomics, Proteomics
and Bioinformatics (ASGPB),
University of Hawaii at Manoa
2565 McCarthy Mall Keller 301
Honolulu, HI 96822

(2) Prof. Magdalena Bezanilla
University of Massachusetts
Amherst, MA 01003

(3) Prof. Eduardo Blumwald
Professor of Cell Biology and Will W.
Lester Chair
Department of Plant Sciences - Mail Stop 5
University of California
1 Shields Ave, Davis, CA 95616

(4) Janet Braam
Chair and Professor of
Biochemistry and Cell Biology
Rice University
6100 Main Street
Houston TX 77005-1892

(5) Terry M. Bricker, Ph.D.
Moreland Family Professor of Basic
Sciences
Department of Biological Sciences
Louisiana State University
Baton Rouge, LA 70803

(6) Fernando Cabral, Ph.D.
Department of Integrative Biology and
Pharmacology, University of Texas-Houston
Medical School
P.O. Box 20708
Houston, TX 77225

(7) Prof. Alice Cheung
University of Massachusetts Amherst, MA
01003

(8) Prof. Craig Gatto
Illinois State University
North and School Streets
Normal, IL 61790

(9) Wouter Hoff
Associate Professor
Department of Microbiology and Molecular
Genetics
Oklahoma State University
307 Life Sciences East
Stillwater, OK 74078

(10) Toivo Kallas
Professor
Department Biology & Microbiology
University of Wisconsin Oshkosh
Oshkosh, WI 54901

(11) Prof. Christopher Makaroff
Miami University
500 E High Street
Oxford, OH 45056

(12) Prof. Daniel Roberts
University of Tennessee Knoxville
1 Circle Park
Knoxville, TN 37996

(13) Prof. Dimuth Siritunga
University of Puerto Rico Mayaguez
PO Box 9001
Mayaguez, PR 00681

(14) Prof. Daniel B. Szymanski
Genetics and Cell Biology of Plant Growth
Department of Agronomy
Lilly Hall of Life Sciences
915 W. State Street
Purdue University
West Lafayette, Indiana 47907-2054

(15) Hemayet Ullah
Assistant Professor
Biology dept.
Howard University
Washington, DC

(16) Dr. Xiaoqiang Wang
Samuel Roberts Noble Foundation, Inc.
2510 Sam Noble Parkway
Ardmore, OK 73402

Summary

Protein structure models are an essential part of biological data covering the interrelationship between genes, sequence, structure, and function. They also form a rapidly growing share of all available molecular structures (presently there are approximately 6 million model structures available compared with 50,000 experimentally determined structures in the PDB). And yet, in contrast to almost all other biological data, there is no easy, single point access to models being provided at this time. Such access point should comprise information on model accuracy, applicability of models in specific tasks, as well as provide visualization methods and relevant analyses techniques such as structure-based search.

We propose building a plant biology domain specific data dissemination platform for protein structure models, offering the above capabilities. Since providing general access to models only makes sense if accompanied by a reliable estimation of model quality, a model evaluation system will constitute the key part of the platform. The platform will aim at the plant biology scholars with different backgrounds, interests, experiences, and proficiencies in working with structures. Access will be provided to the global data of publicly available models, including both "human expert" and automatically generated models. The platform designed with plant biology domain in mind will serve as a prototype for a general resource of this kind, and/or specialized resources in other areas of biology.

At this stage we see the platform (1) maintaining a model data management and dissemination resource; (2) performing accuracy assessments and providing them for all models distributed by the platform; (3) providing comprehensive and user-friendly access to data, as well as to model analyses and visualization tools; (4) offering mechanisms initiating collaboration between biologists and expert modelers.

Although the platform falls into the category of foundational tools rather than grand challenges, we submit this GCW proposal with a goal of refining the concept of a protein model platform through (1) consultations among potential platform developers, and most importantly (2) consultations with model users. We estimate the target size of the proposed workshop at 20-40 participants, with a format of two-day four consecutive session meeting.

Biographical Sketches of the Organizers

Curriculum Vitae
KRZYSZTOF A. FIDELIS

Protein Structure Prediction Center
Genome Center
Genome and Biomedical Sciences Facility
University of California
Davis, CA 95616

Telephone: (530) 754 8931
FAX: (530) 754 9658
E-mail: kfidelis@ucdavis.edu
<http://predictioncenter.org>

Education:

University of Warsaw, Poland M.S. 1983 Physics (Biophysics)
University of Oklahoma, Norman Ph.D. 1989 Physical Chemistry

Professional Experience

1983-1983 Graduate Research Assistant, Department of Biophysics, University of Warsaw, Warsaw, Poland. M.S. Thesis Advisor: Bogdan Lesyng. Analysis of NMR spectra.
1983-1989 Graduate Research Assistant, Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK. Ph.D. Dissertation Advisor: Dick van der Helm. Crystallography of siderophores and peptides, electron density studies, simulation techniques.
1990-1993 Research Associate, Center for Advanced Research in Biotechnology, University of Maryland, Rockville, MD. Advisor: John Moulton. Homology modeling, development of modeling methods.
1993-1996 Postdoctoral Fellow, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA. Protein modeling and structure prediction.
1996-2005 Senior Biomedical Scientist, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA. Group leader of protein modeling and protein structure prediction research group.
1996-2005 Director, Protein Structure Prediction Center, Lawrence Livermore National Laboratory, Livermore, CA. Organization of large-scale community-wide prediction experiments.
2005-pres. Research Faculty, University of California, Davis. Group leader of protein modeling and protein structure prediction research group.
2005-pres. Director, Protein Structure Prediction Center, University of California, Davis, Organization of large-scale community-wide prediction experiments.

Honors

American Crystallographic Association Linus Pauling Award, 1987.
LLNL Biology & Biotechnology Research Program Achievement Award, 1999.

Five Most Closely Related Publications:

Critical assessment of methods of protein structure prediction-Round VII. Moulton J, Fidelis K, Krysztafowicz A, Rost B, Hubbard T, Tramontano A. *Proteins, Structure, Function, and Bioinformatics* 2007, 69(S8):3-9.

New tools and expanded data analysis capabilities at Protein Structure Prediction Center. Krysztafowicz A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T, and Fidelis K. *Proteins, Structure, Function, and Bioinformatics* 2007, 69(S8):19-26.

CASP6 data processing and automatic evaluation at the Protein Structure Prediction Center.

A.Kryshtafovych, M.Milostan, L.Szajkowski, P.Daniluk, K.Fidelis. *Proteins* 2005, 61, Suppl 7, 19-23.

A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. Hvidsten, T. R., Kryshtafovych, A., Komorowski, J., and Fidelis, K. *Bioninformatics* 2003, 19, Suppl 2, II81-II91.

Archiving structural models of biological macromolecules. Outcome of a Workshop on Biological Macromolecular Structure Models. Berman, H.M., Burley, S.K., Chiu, W., Sali, A., Adzhubei, A., Bourne, P.E., Bryant, S.H., Dunbrack, R.L., Fidelis, K., Frank, J., Godzik, A., Henrick, K., Joachimiak, A., Heymann, B., Jones, D., Markley, J.L., Moul, J., Montelione, G.T., Orengo, C., Rossmann, M.G., Rost, B., Saibil, H., Schwede, T., Standley, D.M., Westbrook, J.D. *Structure* 2006, 14:1211-7.

Five Other Relevant Publications:

A Sliding Clamp Model for the Rec1 Family of Cell Cycle Checkpoint Proteins. Thelen M., Venclovas C., and Fidelis K. *Cell* 1999, 96:769-770.

Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins. Venclovas C., Ginalski K., and Fidelis K. *PROTEINS: Structure, Function, and Genetics* 1999, Suppl. 3:73-80.

Structure-based sequence alignment for the β -trefoil sub-domain of clostridial neurotoxin family provides residue level information about putative ganglioside binding site. Ginalski, K., Venclovas, C., Lesyng, B., and Fidelis, K. *FEBS Letters* 2000, 482:119-124.

Progress over the First Decade of CASP Experiments. A.Kryshtafovych, C.Venclovas, K.Fidelis, J.Moul. *Proteins* 2005, 61, Suppl 7, 225-236.

Progress from CASP6 to CASP7. Kryshtafovych, A., Fidelis, K., and Moul, J. *Proteins, Structure, Function, and Bioinformatics* 2007, 69(S8):194-207.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Diana Murray, Ph.D.		POSITION TITLE Associate Professor	
eRA COMMONS USER NAME dim2007			
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
State University of New York, Stony Brook, NY	B.S.	1989	Physics
State University of New York, Stony Brook, NY	Ph.D.	1994	Physics
State University of New York, Stony Brook, NY		1996-1999	Biophysics
Columbia University, New York, NY		1999-2001	Biophysics

A. Positions and Honors

- 1994-1996** Research Analyst, Institute for Pattern Recognition, Robert Nathans Ltd., Setauket, NY.
1996-1999 Postdoctoral Fellow, Physiology and Biophysics, SUNY Stony Brook.
1996 NIH Institutional Training Grant postdoctoral fellowship, SUNY Stony Brook.
1997-2000 Helen Hay Whitney Foundation Postdoctoral Fellowship.
1999-2001 Postdoctoral Fellow, Biochemistry and Molecular Biophysics, Columbia University, NY.
2000-2001 Alfred P. Sloan Foundation/Department of Energy Fellowship in Computational Biology.
2001-2003 Director, Computational Genomics Core Facility, Weill Medical College of Cornell University
2001-2005 Assistant Professor, Department of Microbiology and Immunology and The Institute for Computational Biomedicine (ICB), Weill Medical College of Cornell University.
2003-2005 Research fellowship from the Alfred P. Sloan Foundation.
2005-2007 Associate Professor, Dept of Microbiology and Immunology and ICB, Weill Medical College
2007-present Associate Professor, Department of Pharmacology and the Center for Computational Biology and Bioinformatics, Columbia University

B. Selected peer-reviewed publications (Total 60)

- Murray, D., Hermida-Matsumoto, L., Buser, C.A., Tsang, J., Sigal, C., Ben-Tal, N., Honig, B., Resh, M.D., and McLaughlin, S. (1998). Electrostatics and the membrane association of Src: Theory and experiment. *Biochemistry* 37:2145-2159.
- Burden, L.M., Rao, V.D., Murray, D., Ghirlando, R., Doughman, S.D., Anderson, R.A. and Hurley, J.H. (1999). The flattened face of type IIb phosphatidylinositol phosphate kinase binds acidic phospholipid membranes. *Biochemistry* 38:15141-15149.
- Murray, D., Arbuzova, A., Mihaly, G., Gambir, A., Ben-Tal, N., Honig, B., McLaughlin, S. (1999). Electrostatic properties of membranes containing acidic lipids and adsorbed basic peptides: Theory and experiment. *Biophysical Journal* 77:3176-3188.
- Arbuzova, A., Wang, L., Wang, J., Hangyas-Mihalyne, G., Murray, D., Honig, B., McLaughlin, S. (2000). Membrane binding of peptides containing both basic and aromatic residues: Experimental studies with peptides corresponding to the scaffolding region of caveolin and the effector region of MARCKS. *Biochemistry* 39:10330-39.
- Provitera, P. Bouamr, F., Murray, D., Carter, C., and Scarlata, S. (2000). Binding of Equine Infectious Anemia Virus matrix proteins to membrane bilayers involved multiple interactions. *J. Mol Bio* 296: 887-898.
- Murray, D., McLaughlin, S., Honig, B. (2001). The role of electrostatic interactions in the membrane association of G protein bg heterodimers. *J. Biol. Chem.* 276:45153-45159.
- Murray, D., and Honig, B. (2002). Electrostatic control of the membrane targeting of C2 domains. *Molecular Cell* 9:145-154.

8. Ananthanarayanan, B., Das, S., Rhee, S.G., Murray, D., Cho, W. (2002). Membrane targeting of C2 domains of phospholipase C-delta isoforms. *J. Biol. Chem.* 277:3568-3575.
9. Kulkarni, S., Das, S., Funk, C.D., Murray, D. and Cho, W. (2002) A molecular basis of specific subcellular localization of the C2-like domain of 5-lipoxygenase. *J. Biol. Chem.* 277: 13167-13174.
10. Stahelin, R.V., Long, F., Diraviyam, K., Bruzik, K.S., Murray, D., and Cho, W. (2002). Phosphatidylinositol-3-phosphate induces the membrane penetration of the FYVE domains of Vps27p and Hrs. *J. Biol. Chem.* 277:26379.
11. Wang, J., Gambhir, A., Hangyas-Mihalyne, G., Murray, D., Golebiewska, U., McLaughlin, S. (2002). Lateral sequestration of phosphatidylinositol 4,5-bisphosphate by the basic effector domain of myristoylated alanine-rich C kinase substrate is due to nonspecific electrostatic interactions. *J. Biol. Chem.* 277:34401-34412.
12. Stahelin, R.V., Burian, A., Bruzik, K.S., Williams, R.L., Murray, D., Cho, W. (2003). Membrane binding mechanisms of the PX domains of NADPH oxidase. *J. Biol. Chem.* 278:14469-14479.
13. Diraviyam, K. Stahelin, R.V., Cho, W., Murray, D. (2003). Computer modeling of the membrane interaction of FYVE domains. *J. Molecular Biology* 328:721-736.
14. Singh, S. and Murray, D. (2003). Molecular modeling of the membrane targeting of Phospholipase Pleckstrin homology domains. *Protein Science* 12:1934-1953.
15. Evans, J.H., Gerber, S.H., Murray, D., and Leslie, C.C. (2004). The calcium binding loops of the cPLA2 C2 domain specify association with Golgi and ER membranes. *Molecular Biology of the Cell* 15:371-383.
16. Wang, J., Gambhir, A., McLaughlin, S., Murray, D. (2004). A computational model for the electrostatic sequestration of PI(4,5)P2 by membrane-adsorbed basic peptides. *Biophys. J.* 86:1969-1986.
17. Gambhir, A., Hangyas-Mihalyne, G., Zaitseva, I., Cafiso, D.S., Wang, J., Murray, D., Pentylala, S.N., Smith, S.O., McLaughlin, S. (2004). Electrostatic sequestration of PIP2 on phospholipid membranes by basic/aromatic regions of proteins. *Biophys. J.* 86:2188-2207.
18. Yu, J.W., Mendrola, J.M., Audhya, A., Singh, S., Keleti, D., DeWald, D.B., Murray, D., Emr, S.D., Lemmon, M.A. (2004). Genome-wide analysis of membrane targeting by *S. cerevisiae* Pleckstrin Homology (PH) domains. *Molecular Cell.* 13:677-688.
19. Blatner, N.R., Stahelin, R.V., Diraviyam, K., Hawkins, P.T., Hong, W., Murray, D., Cho, W. (2004). The molecular basis of the differential subcellular localization of FYVE domains. *J. Biol. Chem.* 279:53818-53827
20. Bollinger, J.G., Diraviyam, K., Ghomashchi, F., Murray, D., Gelb, M.H. (2004). Interfacial binding of bvPLA2 to membranes occurs predominantly by a nonelectrostatic mechanism. *Biochemistry.* 43:13293-304.
21. Stahelin RV, Ananthanarayanan B, Blatner NR, Singh S, Bruzik KS, Murray D, Cho W. (2004). Mechanism of membrane binding of the phospholipase D1 PX domain. *J. Biol. Chem.* 279, 54819.
22. Malkova, S., Long, F., Stahelin, R.V., Pingali, S.V., Murray, D., Cho, W., Schlossman, M.L. 2005. X-ray reflectivity studies of cPLA2-a domains adsorbed onto Langmuir monolayers of SOPC. *Biophys. J.* 89:1861-1873.
23. Stahelin, R.V., Wang, J., Blatner, N.R., Rafter, J.D., Murray, D., Cho, W. (2005). The origin of C1A-C2 interdomain interactions in protein kinase C-alpha. *J. Biol. Chem.* In Press.
24. Dalton, A., Murray, P., Murray, D., Vogt, V. (2005). Biochemical characterization of Rous sarcoma virus MA protein interaction with membranes. *J. Virology.* 79:6227-6238.
25. Liu, G., Li, Z., Chiang, Y., Acton, T., Montelione, G.T., Murray, D., Szyperski, T. (2005). High quality homology models derived from NMR and X-ray structure of *E.coli* proteins YdgK and SufE suggest that all members of the YdgK/SufE protein family are enhancers of cysteine desulfurases. *Protein Sci.* 14:1597-1608.
26. McLaughlin, S., Murray, D. (2005). Plasma membrane phosphoinositide organization by protein electrostatics. *Nature.* 438:605-611.
27. Murray, P.S., Li, Z., Wang, J., Tang, C., Honig, B., Murray, D. (2005). Retroviral matrix domains share electrostatic homology: A model for membrane targeting function. *Structure.* 13:1521-1531.
28. Diraviyam, K., Murray, D. (2006). Computational analysis of the membrane association of Group IIA secreted phospholipases A2: A significant role for electrostatics. *Biochemistry* 45:2584-2598.
29. Mulgrew-Nesbitt, A., Murray, D. (2006). Electrostatics of protein/membrane interactions: A computational perspective. *BBA special issue on Lipid Binding Domains.* 1761:812-826.
30. Mirkovic, N., Li, Z., Parnassa, A., Murray, D. (2006). Strategies for high-throughput comparative modeling: Applications to leverage analysis in structural genomics and protein family organization. *Proteins.* 66:766-777.

BIOGRAPHICAL SKETCH FOR DR. RAM SAMUDRALA, ASSOCIATE PROFESSOR

PROFESSIONAL PREPARATION

Institution	Major/Area	Degree, year(s)
Ohio Wesleyan University	Comp. Sci., Genetics	B.A., 1993
Center for Advanced Research in Biotechnology	Comp. Bio.	Ph.D., 1997
Stanford University	Comp. Bio.	Postdoc, 1997-2000

ACADEMIC/PROFESSIONAL APPOINTMENTS

2006-	Associate Professor in Computational Genomics, Department of Microbiology, University of Washington, Seattle
2001-2006	Assistant Professor in Computational Genomics, Department of Microbiology, University of Washington, Seattle
1997-2001	Postdoctoral Fellow at Stanford University School of Medicine
1993-1997	Graduate Fellow at the Center for Advanced Research in Biotechnology
Summer 1992	Howard Hughes Research Intern at East Carolina University Medical School
Spring 1992	Howard Hughes Research Intern at USDA Laboratories, Delaware, Ohio

FIVE MOST CLOSELY RELATED SCIENTIFIC PUBLICATIONS

- Yu J, Wang J, Lin W, Li S, Li H, Zhou J, ..., McDermott J, **Samudrala R**, Wang J, Wong GK. The genomes of *Oryza sativa*: A history of duplications. *Public Library of Science Biology* 3: e38, 2005.
- Wang K, **Samudrala R**. FSSA: A novel method for identifying functional signatures from structural alignments. *Bioinformatics* 21: 2969-2977, 2005.
- McDermott J, Guerquin M, Frazier Z, Chang AN, **Samudrala R**. BIOVERSE: Enhancements to the framework for structural, functional, and contextual annotations of proteins and proteomes. *Nucleic Acids Research* 33: W324-W325, 2005.
- McDermott J, Bumgarner RE, **Samudrala R**. Functional annotation from predicted protein interaction networks. *Bioinformatics* 21: 3217-3226, 2005.
- McDermott J, Wang J, Yu J, Wong GSK, **Samudrala R**. Prediction and annotation of plant protein interaction networks. *Genomics & Bioinformatics in Plant Biotechnology* 2008. *in press*.

FIVE RELATED SCIENTIFIC PUBLICATIONS

- Wang K, **Samudrala R**. Automated functional classification of experimental and predicted protein structures. *BMC Bioinformatics* 7: 278, 2006.
- Chang AN, McDermott J, Guerquin M, Frazier Z, **Samudrala R**. Integrator: Interactive graphical search of large protein interactomes over the Web. *BMC Bioinformatics* 7: 146, 2006.
- Hung L-H, Ngan S-C, Liu T, **Samudrala R**. PROTINFO: New algorithms for enhanced protein structure prediction. *Nucleic Acids Research* 33: W77-W80, 2005.
- Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, **Samudrala R**, Yu J, Wong GK. Neutral evolution of "non-coding" cDNAs from the mouse transcriptome. *Nature* 431, 2004.
- **Samudrala R**, Moulton J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology* 275: 893-914, 1998.

AWARDS AND HONORS

2008	Alberta Heritage Foundation for Medical Research Visiting Scientist Award
2006	NIH Director's Pioneer Award Finalist (25/470 applicants selected as finalists)
2005-2010	NSF CAREER Award
2004	UW New Investigator Science in Medicine Lecture
2003	Named one of the world's top young innovators by <i>MIT Technology Review</i>
2002-2005	Searle Scholar (2002-2005)
1997-2001	NSF PMMB/Burroughs Wellcome Fund Fellow
1993-1997	Ctr. for Advanced Research in Biotechnology Life Technologies Graduate Fellow
1993	Zain-ul-Abedin Memorial Scholarship for Outstanding Graduate Studies
1992	Howard Hughes Internship Award
1990-1993	Dean's List (1990-1993)
1990-1993	Wesleyan Scholar and Honors Student

SYNERGISTIC ACTIVITIES

Bioverse webapplication: The "Bioverse" provides a framework for exploring the relationships among the molecular, genomic, proteomic, systems, and organismal worlds. Our goal is to perform sophisticated analyses and predictions based on genomic sequence data to annotate and understand the interaction of protein sequence, structure, and function, both at the single molecule as well as at the systems levels. <<http://bioverse.compbio.washington.edu>>. **Our work on the rice proteome has resulted in more than 200 media articles in the mont of May, 2008.**

Protinfo protein structure prediction server: enables the general public to predict protein structure and function using methods developed by our group, and resistance/susceptibility of HIV-1 protease mutants based on a novel docking protocol. <<http://protinfo.compbio.washington.edu>>.

Development of software suites to aid in the modelling of protein structure and function: A set of tools to model protein structure and function based on the research performed above, with over 60,000 lines of computer code has been made available without any restrictions at: <<http://software.compbio.washington.edu/ramp/>>.

Sharing scientific knowledge across the world: Examples include an ongoing collaboration with the National Center for Genetic Engineering and Biotechnology in Bangkok, Thailand to share knowledge about bioinformatics, present lectures and tutorials about prediction of protein structure and function, and organise workshops. Besides presenting seminars directly related to research, the investigator also has developed several tutorials on protein structure that have been presented at various conferences, including the Pacific Symposia on Biocomputing and the Intelligent Systems in Biology meetings.

COLLABORATORS AND OTHER AFFILIATIONS

Collaborators/co-authors: Apichart Vanavichit, Kasetsart University, Thailand; Bryan White, University of Massachusetts, Boston; David Baker, Univeristy of Washington (UW); David Teller, UW; Evgeni Sokurenko, UW; Gane Wong, Beijing Genomics Institute (BGI); Gene Nester, UW; James Staley, UW; John Mittler, UW; Joseph Smith, Seattle Biomedical Research Institute; Jun Yu, BGI; Lynn Schnapp, UW; Roger Bumgarner, UW; Ron Stenkamp, UW; and Steve Moseley, UW.

Graduate and postdoctoral advisors: John Moult, Center for Advanced Research in Biotechnology (graduate); Michael Levitt, Stanford University (postdoctoral).

Thesis advisor and postgraduate scholar sponsor (total 24; 13 graduate and 11 postgraduate (overlap of 3)): Aaron Chang; Brady Bernard; Ekachai Jenwitheesuk; Ersin Emre Oren; David Nickle; Duangdao Wichadakul; Duncan Milburn; Gong Cheng; Jason McDermott; Jeremy Horst; Kai Wang; Ling-Hong Hung; Marissa LaMadrid; Rosalia Tugaraza; Sarunya Suebtragoon; Shing-Chung Ngan; Sirphan Manochewea; Stewart Moughon; Tianyun Liu; Weerayuth Kittichotirat; and Yi-Ling Chen, all at the UW.

Statement of the scientific problem

Impact of modeled structures

Model structures are rapidly becoming an important complement of structures obtained experimentally, primarily due to their much broader availability (10-100 fold over those derived through experiment) and an increasing familiarity with what can and cannot be achieved when working with models (for the most recent review see: Comparative modeling in structural genomics. J. Moult, Structure. 2008 Jan;16(1):14-6).

Protein structures predicted automatically by the primary (i.e. not meta-servers) public model providers SwissModel and ModBase (Modeller), are available respectively for up to 30% of the UniProt, and 50% of the Modeller sequence datasets. If fold assignments are included, the estimated fraction of sequences for which models can be built reaches 60%. However not all models generated automatically can be used in reaching biologically relevant conclusions, and some require further refinement. Such refinement can be performed by a human expert equipped with specialized software. Both “expert” and automatic models can accurately represent the predicted structure, depending on their quality and prediction method. They can be resource-intensive, i.e. require significant resources to build, human or computational, and may not be easily reproducible. At present a large proportion of biologically usable models generated for specific projects are not searchable and not easily accessible due to lack of proper data management. Based on 1350 models deposited in the PDB it is reasonable to expect the number of existing "expert" and other resource-intensive models held by groups and individual researchers to be at least an order of magnitude higher. In addition, the number of models produced by the Structural Genomics effort in the near future may reach hundreds of thousands to millions. To utilize this valuable body of data a system capable of searching, analyzing, and delivering the data to biologists is necessary.

Preliminary design principles for the data platform:

Development should be limited to the key problems, namely powerful intuitive data access system and model quality assessment, while making full use of the already developed and available software to be incorporated as platform components. This already available software includes molecule viewers and data visualization tools, sequence analysis and alignment tools, and structure comparison tools.

(1) accessing the models data

A flexible, intuitive, and easy to use data access system is critical. It should be possible to make queries through a single-field search accepting keywords, sequences, structure coordinates, fold type designations, databases entry identifiers, EC numbers, and other terms. To allow researchers accessing the models data from their preferred field of

interest or angle of study, a number of models data representations (data views) should be enabled, with different default configurations of the data search, access and visualization options. The sequence- and structure- centered views would form the basic representations. The platform should support data hosting for models and other predictions, and offer a wide range of deposition options. The platform should provide a single access point to the publicly available global models dataset covering the models deposited with the Platform and the data integrated and linked from public modeling data resources, including server-obtained models. Additionally, experimental structures at the PDB should be integrated. The value of a structure depends on the ability to use it in research, and the advantage the platform will offer is that it will greatly enhance the currently available access to the structures data by the capacity to instantly query the combined dataset of expert and automatic models, curated for accuracy, for the plant biology domain specific model structures. The platform should be open to data sharing and should support integration of models into complex diverse datasets required for systems biology and similar analyses.

(2) user aspect and interface

The goal is to accommodate a wide range of users from those who have no experience of working with structures to experts. Accordingly, the data access system should be based on the following principles. (1) Easy to use and intuitive – in its basic single-field-query configuration almost all but the most sophisticated options should be accessible without the necessity of consulting the help pages. (2) Flexible – enabling data analysis ranging from basic to complex and offering manual user control. (3) User-configurable – offering several preset search and data delivery options, and an advanced fully configurable option defined, and if required saved by the user. (4) "Intelligent" – analyzing the query to determine better search and results delivery parameters, and allowing interactive modification of these parameters. When examining the data, tools should be provided for assembling and manipulating in quick succession large diverse subsets of structures and sequences.

(3) estimation of model quality

In many cases models allow reaching valuable biological conclusions. However, their accuracy varies widely, often leading to erroneous expectations when used in research. Consistent model accuracy data, expressed in the form that does not require expert knowledge to interpret is prerequisite for large-scale dissemination of models envisioned in this initiative. The model accuracy assessment system, capable of high-throughput evaluation, should stand the center of the platform. Model quality estimates should be provided using a simple self-explanatory scale indicating the range of possible biological applications. Detailed evaluation results should also be available. Collaboration in the general area of model evaluation has already been started through with the Working Group on Model Quality Assessment (current members include Arne Elofson, Barry Honig, David Jones, Marc Marti Renom, Jarek Meller, William S. Noble, and Silvio Tossato; involvement in this group is open to all scientists working in this area).

Off the shelf solutions will not suffice

Specifically, there are two problems that need to be solved to reach the goals of this initiative. First, to facilitate access to protein structure models, a series of components need to be strung into a single cohesive system. These components include:

- (1) Access to protein structure models deposited in existing databases.
- (2) Access to automated modeling servers.
- (3) A possibility of initiating collaboration with a human expert modeler.
- (4) Evaluation of model quality along with an assessment in terms of possible applications.
- (5) Navigation tools designed to facilitate access to and analysis of plant protein models.
- (6) Visualization tools capable of user controlled display and multiple structure comparisons.

Second, evaluation of model quality above is not a fully solved problem. A sub-system dedicated to this task needs to be engineered basically from scratch, including assembling existing methods and implementing perpetual benchmarking of existing and new techniques.

Goals for the proposed GCW

During the workshop, we plan to focus on model quality assessment and on the contributions to research that models at various levels of accuracy may bring. From this perspective, the main goal of the workshop is to learn much more precisely about the needs of the plant community regarding the use of models. For example if a program increasing the general familiarity with modeling issues is needed? What level of model curation and user support is required? Which protein families necessitate particular attention? What level of emphasis should be placed on membrane proteins? For this prototype initiative to succeed it is critical that issues specific to the plant biology domain are communicated and addressed effectively.

We also would like to use the meeting to develop communication channels between protein structure modelers and those who use structure data in their research. In addition we hope to recruit scientists who would benefit from the use of 3D models of protein structure in their research, but who had not used them before.

Similar initiatives

Torsten Schwede of SwissModel, who also heads model database development for the NIH Protein Structure Initiative is part of this proposal. Synergistic actions are possible. Torsten is willing to share the database management software he has developed for the PSI, thus minimizing some of the software development tasks. In addition, a meeting on Applications of Protein Models in Biomedical Research, organized by Torsten and Andrej Sali is to be held at the University of California, San Francisco, July 11-12, 2008. Outcomes of that meeting should be useful also for this initiative.

Preliminary Meeting Agenda

Four sessions are planned. The first session will be dedicated to the applications of protein structure models in modern biology with the goal of providing an overview of what can be achieved by using models. Krzysztof Fidelis will present followed by 1-2 speakers presenting their own research. Difficulties in modeling as well as successes will be discussed. The second session will focus on methods of model quality assessment. We plan to invite speakers active in this area such as Arne Elofsson, and David Jones. Third session will be dedicated to the design aspects of the platform. Planned speakers include Torsten Schwede, Alexei Adzhubei, and others. Finally the last session will be dedicated to discussion and recommendations.

Suggested Reviewers

Dr. Tim Hubbard
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambs CB10 1SA, UK
Tel: +44 1223 496886
Fax: +44 1223 494919
Email: th@sanger.ac.uk
Structural Genomics

Dr. Dietlind Gerloff
University of California, Santa Cruz
Center for Biomolecular Science & Engineering
Engineering 2, Suite 501, Mail Stop CBSE/ITI
UC Santa Cruz, Santa Cruz, CA 95064
Tel: (831) 459 4833
Email: gerloff@soe.ucsc.edu
Structural /Functional Genomics