Establishing High-throughput Protein Structure Determination Pipeline for Structural Genomics

Andrzej Joachimiak[1], Rongguang Zhang[1], Youngchang Kim[1], Jerzy Osipiuk[1], Marianne Cuff[1], Changsoo Chang[1], Boguslaw Nocek[1], Andrew Binkowski[1], Marcin Cymborowski[2], Krzysztof Lazarski[1], Maksymilian Chruszcz[2], Roman Laskowski[3], Janet Thornton[3], Norma Duke[1], Frank Rotella[1], Zbyszek Otwinowski[4], Alexei Savchenko[5], Aled Edwards[5,6], Wladek Minor[2], *[1]Midwest Center for Structural Genomics and Structural Biology Center, Biosciences, Argonne National Laboratory, 9700 South Cass Ave. Argonne, IL 60439. [2]University of Virginia, Charlottesville, VA 22908, USA. [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [4]University of Texas, Southwestern Medical Center, Dallas, TX 75390 USA. [5]Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, Ontario M5G, Canada. [6]Clinical Genomics Centre/Proteomics, University Health Network, 101 College St., Toronto, Ontario, M5G 1L7.* E-mail: andrzejj@anl.gov

Genome projects provide comprehensive access to genomic sequence information. The accumulation of sequence data has accelerated significantly, currently 1,386 genome projects are under way, sequences of 256 genomes have been completed, annotated, and available to the public. Many aspects of protein function, including molecular recognition, assembly and catalysis, depend on the 3D atomic structure. Protein structural analysis also contributes to an understanding of the evolutionary and functional relationships among protein families that are not apparent from the genome sequences. However, the structural coverage of proteins coded by new genomes remains low. The structural genomics efforts were initiated to increase structural coverage of proteomes in a rapid and cost-effective manner. Structural genomics programs contribute several tools: (1) comprehensive dictionary of high-resolution protein structures determined experimentally by x-ray crystallography and nuclear magnetic resonance (NMR); (2) comprehensive library of recombinant protein expression clones representing protein structures and functions; (3) methods for automated, HTP protocols of molecular and structural biology; and (4) functional information derived from structure.

Toward these goals the Midwest Center for Structural Genomics (MCSG) has established a protein structure determination pipeline using x-ray crystallography and synchrotron radiation. The current MCSG pipeline integrates all essential experimental and computational processes. Public databases of genomic sequences are being analyzed and targets are selected for structural studies. The MCSG pipeline generates well-characterized protein target expression strains, produces milligram quantities of proteins and heavy-atom labeled crystals. The cryoprotected crystals of x-ray quality are used for data collection at the synchrotron beamlines and structure determination using semi-automated SAD or MAD approach. Structural models are auto-build and structures are refined, verified and analyzed using semi-automated computational tools. Functional analysis is being performed using a newly developed ProFunc server. 3D models of relevant members of the sequence family are generated and their quality is assessed. Majority of the steps in the MCSG pipeline are tracked in near real time by the database. All the structures and their analysis are made available to the public using the MCSG database and web tools. The MCSG structure determination pipeline when combined with data collection facilities at third generation synchrotrons, advanced software and computing resources resulted in significant acceleration of structure determination of novel proteins. Using this pipeline the MCSG has determined 112 novel structures in 2004.

This work was supported by National Institutes of Health Grant GM62414 and by the U.S. Department of Energy, Office of Biological and Environmental Research, under contract W-31-109-Eng-38.

Keywords: high-throughput, structural genomics, stucture determination