Article ID:1672 - 058X(2012)12 - 0076 - 07

Research Progress in Structural Genomics

JIANG Huai-chun[1], LI Hong[2]

(1. Academic Periodical Office, Chongqing Technology and Business University, Chongqing 400067, China;
2. Department of Life Science and Chemistry, Chongqing Education University, Chongqing 400067, China)

Abstract: North America has made great progress in structural genomics research, Europe is a slow start but strong finish in structural genomics research, Japan is one of the first countries to initiate structural genomics and is making efforts to conduct it, and Australia is also participating in structural genomics research. Nucleic acid-binding proteins are important to clarify the mechanism of nucleic acid-protein interaction for clearly explaining the mechanism of the pathogenesis of some diseases by structural genomics research results so that the diseases, such as cancer, diabetes and so on, can be cured by the medicines synthesized according to the information of the related protein structures.

Key words: structural genomics; nucleic acid-protein interaction; disease mechanism; North America; Europe; Japan; Australia

CLC number: Q71 **Document code**: A

Human Genome Project has been accomplished for over a decade but how to use the information of the Genome to serve human beings is still an issue for biologists to search for and to explore, meanwhile, the genomes of other organisms have been mapped and some of their proteins have been crystallized and characterized. The scientists are poised to embark on a systematic program of high-throughput X-ray crystallography and NMR spectroscopy aimed at developing a comprehensive view of the protein structure universe. Tens of thousands experimental protein structures and millions of calculated comparative protein structure models will be yielded[1]. By comparing phylogenetic (evolutionary) profiles and expression patterns and by analyzing domain fusions, the new methods are used to identify proteins that are functionally linked through a metabolic pathway, a signaling pathway or a structural complex[2]. The medical structural genomics has been applied to discovering new drugs for infectious diseases[3,4]. Since the late 1990s, North America, Europe and Japan have initiated large-scale genomics research.

1 Structural Genomics Research Progress in North America

Structural genomics in North America has moved remarkably quickly from ideas to pilot project, from protein crystallization and NMR spectroscopy to protein modeling, from structural determination to drug discovery.

Received date: 2012 - 09 - 05; **Revised date**: 2012 - 09 - 12.

Biography: JIANG Huai - chun (1961 -), male, native place: Sichuan, Associate Professor of biology, Tel: 023 - 62769715, Email: jianghc123@ yahoo. com. cn.

The pivotal point for the North America efforts was the 1998 Argonne Meeting for a timeline of structural genomics in North America, this Meeting brought together over 80 researchers and representatives of funding agencies who thought that improvement in technology, combined with the successes of the genome sequencing projects, had set the stage for a large-scale structure determination project[5]. However, the discussions at the Argonne Meeting indicated that small scale testing of the ideas of structural genomics and considerable additional technology development were necessary before a full-scale project could be carried out[5]. At first, the pilot projects of structural genomics were sponsored by US Department of Energy, and additional funding came from National Institute of Health (NIH), Ontario Cancer Institute, New Jersey Commission on Science and Technology Initiative and University of California[5]. Most of these pilot projects had two principal goals to demonstrate the overall feasibility of structural genomics and to develop some of the technologies necessary for large-scale structure determination. Target proteins were chosen from thermophilic organisms or from full-sequenced organisms such as yeast or *Haemophilus* influenzae and were hoped to have novel folds or to yield important functional information from protein structures[5]. The strategy for pilot projects is to clone a gene which expresses well in a simple system, to produce protein in soluble form, to crystallize the protein or study the structure of the protein by NMR or X-ray crystallization, to analyze the crystal structure of the protein and then to determine the three-dimensional structure of the proteins, to compare the protein structure with well-known protein structures.

Although the pilot projects have generated a number of structures, most people in this field feel that substantial improvement in technology is required to make structure determination a high-throughput process, since then, many research groups are focusing on technology development, the Rutgers group, for example, is focusing on development of high-throughput methods for NMR structure determination, the Argonne group on developing high-throughput methods for X-ray crystallography, the Scripps/GNF group has been entirely working for technology development, the National Institute of General Medical Sciences (NIGMS) announced and conducted its Protein Structure Initiative to organize a large, cooperative effort in structural genomics by determining and analyzing protein structures and fold families to link sequence, structural and functional information and to enable the prediction of unknown structures through homology modeling[6], and the Seattle Structural Genomics Center for Infectious Diseases (SSGCID) is concentrating on a multi-pronged serial escalation approach for streamlining protein production and structure determination[5,7].

1.1 The Leading Role of National Institute of Health

The NIH held a series of workshops to discuss the possibility of a large-scale publicly-funded effort in structural genomics initially including the scope of a possible structural genomics project, determining a representative set of a few thousand protein structures for understanding the structures and functions of most other proteins, the development of infrastructure and technologies for structural genomics research, two general approaches to selecting targets for structural genomics such as to organize protein sequences into families at a level of ~30% sequence identity and to determine just one representative of each family, and to focus on the proteins with clear biological importance. A particularly influential meeting was held in the fall of 1998 in Avalon, New Jersey, in this meeting, after debates, it became clear that for many participants the real goal of structural genomics was to know the structures of all proteins, and, in this context, the targeting of current small-scale efforts could be understood as the strategies for this long-term and ambitious goal[5].

1.2 NIH-supported Centers and Technology Development

NIH funds seven structural genomics centers out of eleven applicants in the United States located in Northeast,

New York, Southeast and Midwest, and the seven NIH-funded centers (the Midwest Center for Structural Genomics, the Southeast Collaboratory for Structural Genomics, the Northeast Structural :Genomics Consortium, the New York Structural Genomics Research Consortium, the Joint Center for Structural Genomics, the Berkeley Structural Genomics Center and the TB Structural Genomics Consortium) have substantially varying emphases such as obtaining hundreds of new protein structures representing families of protein structure that previously had no representatives with known structure, and developing new technologies allowing a full-scale structural genomics effort to succeed. NIH-funded TB Structural Genomics Consortium plans to emphasize protein expression, expects to use in vitro evolution-based methods to engineer its protein targets to increase solubility. Almost all seven centers plan to develop automated procedures for X-ray crystallography and NMR data collection and analysis[5].

1.3 The Efforts and Achievements Made by the Seattle Structural Genomics Center for Infectious Diseases (SSGCID)

The Seattle Structural Genomics Center for Infectious Diseases (http://www.ssgcid.org home/index.asp) is devoted to the application of state-of-the-art structural genomics technologies to structurally characterize targeted proteins from NIAID Category A-C pathogens and organisms and its goal is to create a collection of three-dimensional protein structures that are widely available to the broad scientific community and serve as a blueprint for structure-based drug development for infectious diseases. The SSGCID developed escalating tier approach to high-throughput protein determination pipeline to minimize the cost for the procedures such as target selection, cloning, expression and screening, protein production, crystallization, data collection and structure solution[7]. SSGCID also made efforts to answer the challenge at the all points of protein production pipeline by using combined gene engineering and structure-guided constructs design to overcome challenges at the levels of protein expression and protein crystallization through a multi-pronged serially escalating approach to protein production[8], established a robust protein-purification pipeline designed to purify 400 proteins per year at a rate of eight purifications per week by using two AKTAexplorer 100s and four AKTAprimes to perform immobilized metal-affinity chromatography and size-exclusion chromatography[9] as well as developed a high-throughput screening protocol for the measurement of protein recovery from immobilized metal-affinity chromatography[10]. SSGCID has made progress in using protein fragment-based information for drug discovery, although the failure in bacterial enzyme inorganic pyrophosphatase, it is a good start for using protein structural information to synthesize medicines and so it is an achievement[3,4]. Penicillin destroys the cell wall of a bacterium, if only a protein is destroyed, the bacteria DNA will make more the same kind of proteins, so the researchers must choose the targets for drug development which are critical to the survival of the bacteria or the viruses[11].

1.4 Major Recent Advances in Structural Genomics in North America

Since the Protein Structure Initiatives (PSI) was established in 2000 by NIGMS and with the advancing of genomic sequencing bioinformatics, structural genomics takes advantage of completed genome sequences in several ways to have determined a lot of protein structures. The gene sequences of the target proteins have been compared to a known sequence and structural information which is then inferred from the known protein's structure. Structural genomics has been used to predict novel protein folds based on other structural data. So far, the protein-modeling methods include *de novo* methods, *ab initio* modeling, sequence-based modeling and threading[12-15]. One progress is that a mount of 3D structures of proteins (or domains) have been deposited into the Protein Data Bank (PDB) and have been made integrative database analysis[16], novel structures are defined as those which have < 30% sequence identity with any structure in the PDB at the time of deposition. In the second stage of PSI (2005-

2010), the researchers achieved their goal in 2005 by depositing more than 3,000 Distinct Structures (two protein sequences are distinct if they share $<98\%$ sequence identity over the full-length of the shortest sequence of the pair) into the PDB, most of which were also Novel Structures[15]. Over the full of ten years of the PSI program, over 5 000 3D protein structures including protein-ligand complexes and pairs of X-ray and NMR structures have been studied and deposited[15,17], PSI Centers are also involved in collaborative projects aimed at accelerating the field of protein X-ray and NMR structure analysis and computational protein design[18-20]. Some scientists have begun to try to develop medicines by using protein structural genomics information[3,4].

2 Structural Genomics Research in Europe

European researchers are focused on multi-centered projects aiming at the development and interfacing of the methods for X-ray diffraction and NMR structure analysis because the Fifth Framework Program of the European Commission (EC) initiated a thematic program "Quality of Life and Management of Living Resources", expecting to contribute to the international efforts toward high-throughput structure analysis methods. In 2000, British Wellcome Trust supported the conference on structural genomics, the First International Structural Genomics Meeting in Hinxton, UK, the scientists from the UK may also turn to the Biotechnology and Biological Sciences Research Council (BBSRC) for funding and BBSRC is interested in Structural Molecular Analysis of Rational Targets (SMART)[21]. In France, money has been committed by the Ministry of Research to structural genomics by way of integration into the national Genome Project. In 2000, one structural genomics program was funded and operational and more funding was going to be available for the research in Germany. The Swiss National Fund (SNF) is concentrated on the methods for structural analysis of membrane proteins and large protein assemblies[21]. A considerable number of laboratories in central and western Europe are involved in structural genomics research activities, such as Aston University, Imperial College of Science, Technology and Medicine, Douglas Instruments Ltd, Molecular Dimensions Ltd, and Farfield House, Southmere Court, Electra Way of the United Kingdom, Key Drug Prototyping BV of Netherlands, University of ZU LüBECK, PLS Design GMBH and Hamburg University of Germany, Triana Science & Technology and Conejo Superior De Investigaciones Cientificas of Spain, and so on[22]. The multi-national programs are primarily concerned about the development of computing methods for the fast structure analysis of the proteins based on NMR and X-ray crystallization data. University of Utrecht aims to develop tools for the fast interpretation of NMR data for structural and functional analysis of proteins and for protein function prediction, the University of York is engaged in an effort to interface major software for macromolecular crystallography in order to create an integrated, partly automated and user-friendly environment for structure analysis. Berlin has set up a Center for Molecular Medicine focusing on human proteins and their domains analysis by NMR and X-ray method and Max Planck Institute of Germany is embarking on facilities and methods for the high-throughput crystal structural analysis of human proteins[23]. The funding from French Genome Project is expected to focus on a small number of targeted protein classes from *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae* as well as nuclear receptors and transcription factors from multicellular eukaryotes[21]. SLS (Switzerland), SOLELL (France) and DIAMOND (UK) are improving X-ray diffraction by increasing in the number of beamlines, University of Utrecht of Netherlands used a fast microtiterplate-based expression and solubility screening procedure to purify human protein domains for NMR analysis[24]. The Royal Institute of Technology Center for Physics, Astronomy and Biotechnology of Sweden has tested a methodology to rapidly subclone a large number of human genes and screen these for expression and protein solubility in E. coli[25] and used AKTA 3D

system for automated multi-dimensional purification of His6-and GST-tagged proteins[26] as well as developed and tested a simple and efficient protein purification method for biophysical screening of proteins and protein fragments by NMR and optical methods such as circular dichroism spectroscopy[27]. Ontario Institute for Cancer Research of Canada and University of Cambridge of the United Kingdom organizes the International Cancer Genome Consortium (ICGE) to collaboratively characterize genomic abnormalities in 50 different cancer types and has created the ICGC DATA portal to make the data available[28].

3 Structural Genomics Projects in Japan

Japan was one of the first countries to embrace structural genomics, Japanese-led projects focused on this effort were conceptualized as early as 1995, with the Protein Folds Project with Yokoyama as the project director at the RIKEN Institute in April, 1997 and then being transferred to the newly established RIKEN Genomic Sciences Center (GSC: http://www.riken.go.jp) in October, 1998[29]. Another project, the Structurome Project, with Kuramitsu as the project leader, began in October, 1999, at the Riken Harima Institute at the Super Photon Ring-8 (Spring-8) synchrotron, focusing on *Thermus thermophilus* proteins. The two projects have been combined as the Riken Structural Genomics Initiative later. The Science and Technology (STA) of Japan funds both the protein Folds Project and the Structurome Project, meanwhile, the Ministry of International Trade and Industry (MITI) of Japan funds a structural genomics project on human membrane proteins with Kyogoku as project director at newly established Biological Information Research Center (BIRC). The Riken Structural Genomics Initiative contributes to the international structural genomics effort, uses a combination of target selection from bacteria (including *T. thermophilus*), animals and plants, and prioritizes proteins according to potential biological/medical importance or expectation of structural novelty. Sequencing of the 1.8 Mb genome encoding approximate 1,000-2,000 proteins of *T. thermophilus* HBB has been completed, and about its 600 ORFs have been amplified by PCR and cloned into pET expression vectors. Then its proteins are overproduced either *in vivo* or *in vitro* in a cell free system, purified and crystallized, various universities and research institutes in Japan are involved. The Crystallographic Society of Japan began to organize crystallography laboratories for structural genomics initiative with I. Tanaka of Hokkaido University as the chair since 2000[29]. The mammalian and plant genomes groups led by Hayashizaki and Shinozaki respectively of the Riken GSC are collecting and sequencing tens of thousands of full-length cDNAs from mouse and *Arabidopsis thaliana* respectively. As the success rates of overproducing mammalian and plant proteins are not high, the cell free protein synthesis method is explored by the Riken Structural Genomics Initiative[29]. The MITI-funded structural genomics project at the BIRC focuses on membrane protein structures and aims to pursue basic research on post-genome sequencing science and to provide the results mainly for industry. BIRC is composed of the Structural Genomics group, the Functional Genomics group and the Integrated Database group[29]. The goal of Japan for structural genomics research is to use the information of structural biology to develop medicines. Hokkaido University of Japan is exploring functionally related enzymes using radially distributed properties of active sites around the reacting points of bound ligands[30].

4 Structural Genomics in Australia

Australians are interested in proteome for interpreting and making use of protein-protein interaction by trying to discover the link between protein structures and protein functions, and University of Queensland of Australia[31] is

interested in the fusion of a targeted protein to a large-affinity tag, such as the maltose-binding protein, thioredoxin[32], or glutathione-S-transferase[33], for high-throughput production of the interested proteins.

5 Expectations

Dr. Montelione from Rutgers University advanced five future directions for structural genomics research, including protein structure information and the relationship between the structure and function of the proteins from metagenomics and microbiomes , individual organisms and organelles, systems biology, networks, pathways, and complexes, and individual protein domain families as well as protein engineering and design[15], which are wise, appropriate and practical in current research situation.

Nucleic acid-protein interaction is a central issue for biological researchers[34]. DNA- or RNA-protein interaction determines the replication of DNA or RNA, then the transcription of RNA and then the translation of proteins, the structures and functions of proteins are determined by DNA or RNA sequences, and so DNA- or RNA-binding proteins should be explored so that the mechanism of nucleic acid-protein interaction can be clarified, accordingly, the mechanism of the pathogenesis of some diseases will be exposed, if the mechanism is known by scientists, some diseases such as cancer, diabetes and so on will be conquered. Structural genomics provides an ideal method and opportunity for scientists to know what a protein function is determined by what a protein structure, then what a structure of protein is determined by what a DNA or RNA sequence and then a chemical compound is designed to change the conformation of abnormal protein which interacts with DNA or RNA as a normal protein does and the abnormal functions of abnormal proteins will be corrected to have proper functions because the abnormal functions of the abnormal protein structures cause some diseases of animals or human beings. Thus, the final aim of structural genomics is to expound nucleic acid-protein interaction mechanism for scientists to use the protein structure to design medicines to make abnormal human nucleic acid-protein interaction in order as normal proteins do.

References:

[1] BURLEY SK. An overview of structural genomics[J]. Nature Structural Biology, 2000, 7(Supp.): 932-934

[2] ALI A. Genomics: Functional links between proteins[J]. Nature, 1999, 402: 23-26

[3] VOORHIS WCV, HOL WGJ, MYLER PJ, et al. The role of medical structural genomics in discovering new drugs for infectious diseases[J]. PLOS Computational Biology, 2009, 5(10): 1-7

[4] ANDERSON WF. Structural genomics and drug discovery for infectious diseases[J]. Infect Disord Drug Target, 2009, 9(5): 507-517

[5] TERWILLIGER TC. Structural genomics in North America[J]. Nature Structural Biology, 2000, 7(Supp.): 935-939

[6] NORVELL JC, MACHALEK AE. Structural genomics program at the US National Institute of General Medical Sciences[J]. Nature Structural Biology, 2000, 7(Supp.): 931

[7] LORIMER D, RAYMOND A, MIXON M, et al. Gene Composer in a structural genomics environment[J]. Acta Cryst. (Structural Biology and Crystallization Communication), 2011, F67: 985-991

[8] RAYMOND A, HAFFNER T, NG N, et al. Gene-designing, cloning and protein-expression methods for high-value targets at the Seattle Structural Genomics Center for Infectious Diseases [J]. Acta Cryst. (Structural Biology and Crystallization Communication), 2011, F67: 992-997

[9] BRYAN CM, BHANDARI J, NAPULI AJ, et al. High-throughput protein production and purification at the Seattle Structural Genomics Center for Infectious Diseases[J]. Acta Cryst. (Structural Biology and Crystallization Communication), 2011, F67: 1010-1014

[10] CHOI R, KELLEY A, LEIBLY D, et al. Immobilized metal-affinity chromatography protein-recovery screening is predicative of crystallographic structure success[J]. Acta Cryst. (Structural Biology and Crystallization Communication), 2011, F67: 998-1005

[11] BRENNER SE, Target selection for structural genomics[J]. Nature Structural Biology, 2000, 7(Supp.): 967-969

[12] KUHN P, WILSON K, PATCH MG, et al. The genesis of high-throughput structure-based drug discovery using protein crytstallography[J]. Current Opinion in Chemical Biology, 2002, 6: 704-710

[13] BAKER D, SALI A. Protein structure prediction and structural genomics[J]. 2001, Science, 294: 93-96

[14] LESLEY SA, et al. Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline[J]. PNAS, 2002, 99(18): 11664-11669

[15] MONTELIONE GT, The protein structure initiative: achievements and vision for the future[EB/OL], http://f1000.com/reports/b/4/7, 2012-04-02

[16] GERSTEIN M. Integrative database analysis in structural genomics[J]. Nature Structural Biology, 2000, 7(11): 960-963

[17] MAO B, GUAN R, MONTELION GT. Improved technologies now routinely provide protein NMR structures useful for molecular replacement[J]. Structure, 2011, 19: 757-766

[18] RAMAN S, LANGE OF, ROSSI P, et al. NMR structure determination for larger proteins using backbone-only data[J]. Science, 2010, 327: 1014-1018

[19] JHA RK, WU YL, ZAWISTOWSKI JS, et al. Redesign of the PAKI autoinhibitory domain for enhanced stability and affinity in biosensor applications[J]. J. Mol. Biol., 2011, 413: 513-522

[20] GABANYI MJ, ADAMS PD, ARNOLD K, et al. The structural biology knowledgebase: a portal to protein structures, sequences, functions and methods[J]. J. Struct. FUNCT. Genomics, 2011, 12: 55-62

[21] HEINEMANN U. Structural genomics in Europe: Slow start, strong finish? [J]. Nature Structural Biology, 2000, 7(Supp.): 940-942

[22] Optimisation of protein crystallisation for European structural genomics[EB/OL], http://www.google.com.hk/2006-12-01/2010-08-31

[23] HEINEMANN U, BUSSOW K, MUELLER, U, et al. Facilities and methods for the high-throughput crystal structural analysis of human proteins[J]. Accounts of Chemical Research, 2003, 36(3): 157-163

[24] FOLKER GE, BUUREN BNMV, KAPTEIN R. Expression screening, protein purification and NMR analysis of human protein domains for structural genomics[J]. J. Struct. Funct. Genomics, 2003, 4: 1-13

[25] HAMMARTROM M, HELLGREN N, BERG SVD, et al. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli[ED/OL], http://www.proteinscience.org/2001-06-04/2001-09-02

[26] SIGRELL JA, EKLUND P, GALIN M, et al. Automated multi-dimensional purification of tagged proteins[J]. J. Struct. Funct. Genomics, 2003, 4: 109-114

[27] WOESTENENK EA, HAMMARSTROM M, HARD T, et al. Screening methods to determine biophysical properties of proteins in structural genomics[J]. Analytical Biochemistry, 2003, 318: 71-79

[28] ZHANG JJ, BARAN J, CROS A, et al. International cancer genome consortium data portal-a one-stop shop for cancer genomics data[ED/OL]. http://creativecommons.org/licenses/by-nc/2.5/

[29] Yokoyama S, HIROTA H, KIGAWA T, et al. Structural genomics projects in Japan[J]. Nature Structural Biology, 2000, 7(Supp.): 943-945

[30] UENO K, MINETA K, ITO K, et al. Exploring functionally related enzymes using radially distributed properties of active sites around the reacting points of bound ligands[J]. MBC Structural Biology, 2012, 12: 5-28, http://www.biomedicentral.com/1472-6807/12/5.

[31] SMYTH DR, MROZKIEWICZ MK, MCGRATH WJ, et al. Crystal structure of fusion proteins with large-affinity tags[J]. Protein Science, 2003, 12: 1313-1322

[32] SACHDEV D, CHIRGWIN JM. Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin[J]. Protein Expr. Purif., 1998, 12: 122-132

[33] SMITH DB. Generating fusions to glutathione-S-transferase for protein studies[M]. Methods in Enzymol., 326: 254-270

[34] JIANG HC. A gel electrophoresis method for studying nucleic acid-protein system[J]. Progress in Biochemistry and Biophysics (in Chinese), 1989, 16(3): 180-183

(下转第94页)