Acc. Chem. Res. **2003,** *36,* 183−189

Structural Proteomics: Toward High-Throughput Structural Biology as a Tool in Functional Genomics

ADELINDA YEE,[†] KEITH PARDEE,[‡] DINESH CHRISTENDAT,[†,‡,§] ALEXEI SAVCHENKO,[†] ALED M. EDWARDS,[†,‡] AND CHERYL H. ARROWSMITH*[,†]

Ontario Cancer Institute and Department of Medical Biophysics, and Department of Medical Genetics, University of Toronto, 200 Elizabeth Street, Toronto, ON, Canada M5G 2C4

Received May 2, 2002

ABSTRACT

Structural proteomics is the determination of atomic resolution three-dimensional protein structures on a genome-wide scale in order to better understand the relationship between protein sequence, structure, and function. Here we describe our ongoing structural proteomics project on the nonmembrane proteins of the archeaon, *Methanobacterium thermoautotrophicum*. This article provides a snapshot of an ongoing pilot project in an emerging area of multidisciplinary research that involves bioinformatics, molecular biology, biochemistry, and instrumental methods such as NMR spectroscopy and X-ray crystallography. An assessment of the technical challenges in this type of large-scale project along with a comparison of the efficiency of sample production for both X-ray crystallography and NMR spectroscopy will be discussed. Examples of new insights into protein function and the relationship between structure and sequence will also be presented.

Introduction

Recent completion of several genome projects has provided scientists with a wealth of information in the form of gene sequence data. For realization of its true value, however, these sequences must be related to the proteins they encode and in-turn their biological and biochemical importance in the organism. Since the three-dimensional structure of a protein polypeptide chain determines its biochemical function, the building of structure−function correlations for novel and diverse protein conformations is a critical next step in genomics research. For these reasons, many scientists consider functional genomics or proteomics, including the determination of 3D structures of proteins, to be the natural progression in the characterization of the genome. Because computational methods are not yet capable of accurately predicting 3D structures of native proteins from amino acid sequence alone, it is necessary to use experimental methods to determine the configurations of atoms that confer biochemical activity (for example enzymatic activity).

With recent technical advances in the fields of X-ray crystallography[1,2] and nuclear magnetic resonance (NMR) spectroscopy,[3] structural biologists can now contemplate applying these technologies to help annotate the structures and biochemical functions of proteins on a genome-wide scale. The genome-wide approach to protein structure determination, termed structural proteomics, provides a new rationale for structural biology. Traditionally, structural biologists attacked a problem only after it had been firmly characterized using biochemical and/or genetic methods. However, relying on structure−function relationships, it will now be possible to suggest a biochemical function of uncharacterized proteins based solely on structural homology to another protein with a known function. Such a predicted function could then provide the foundation for a hypothesis that could be tested with additional biochemical experiments. For proteins with functional annotations derived solely from sequence homology with proteins of known function, the structure can be used to understand in more detail the putative activity or function. Traditional applications of protein structure remain important, particularly for understanding at atomic resolution the details of biochemical and enzymatic mechanisms.

Pilot Project on *Methanobacterium thermoautotrophicum*

Strategy. Several years ago, we launched a project to determine the feasibility of large-scale structural biology. We selected several hundred proteins from the archaeon, *Methanobacterium thermoautotrophicum*, also known as

* Corresponding author. Tel: (416) 946-2017. Fax: (416) 946-6529. E-mail: carrow@uhnres.utoronto.ca.

[†] Ontario Cancer Institute and Department of Medical Biophysics.

[‡] Department of Medical Genetics.

[§] Present address: Department of Botany, University of Toronto.

Adelinda Yee received her B.Sc. in Chemical Engineering from Mapua Institute of Technology (Philippines) and M.Sc. and Ph.D. in Chemistry from University of Manitoba. She carried out postdoctoral training with Dr. Arrowsmith studying DNA-binding proteins by NMR spectroscopy and is currently working with the structural proteomics project.

Keith Pardee received a B.Sc. in Plant Physiology from University of Alberta and an M.Sc. in Botany from University of British Columbia and is currently a Ph.D. candidate at the Department of Molecular and Medical Genetics at the University of Toronto.

Dinesh Christendat received his B.Sc. and Ph.D. from Concordia University. He carried out his postdoctoral training as part of this project and has now moved on to a faculty position in the Botany Department of the University of Toronto.

Alexei Savchenko received his B.Sc. and M.Sc. from Yerevan State University (Armenia) and Ph.D. in Molecular Biology/Microbiology from University of Nantes (France). He carried out his postdoctoral training at Michigan State University and is currently working with the structural proteomics project.

Aled Edwards received his B.Sc. and Ph.D. in Biochemistry from McGill University, Montreal. He is also carried out postdoctoral training at Stanford University and was an assistant professor at McMaster University before taking his current positions at the University of Toronto and the Ontario Cancer Institute in 1997.

Cheryl Arrowsmith received a B.Sc. in Chemistry from Allegheny College and a Ph.D. in chemistry from University of Toronto. She carried out postdoctoral training at Stanford University. Since 1992 she has been a Senior Scientist at the Ontario Cancer Institute and member of the Department of Medical Biophysics, University of Toronto.

10.1021/ar010126g CCC: $25.00
Published on Web 02/06/2003

Methanothermobacter thermoautotrophicus,[4] for our study. The selection of this thermophile, which grows optimally at 65 °C, was made in a bid to overcome difficulties in protein stability commonly experienced during purification. Using standard genetic engineering techniques, the genes of interest were "subcloned" into plasmid DNA, which when incorporated into bacteria, result in vast over production of the protein of interest, in some cases up to 50% of the cellular protein. Recombinant proteins were purified from host bacteria using affinity chromatography and evaluated for suitability for 3D structure determination by NMR and/or X-ray crystallography. Management and coordination of the above workflow on a genome-wide scale was an unprecedented undertaking that permitted the identification of bottlenecks in the structure determination process and allowed evaluation of the relative merits of NMR and X-ray crystallography for protein structure determination. The preparation of protein samples to yield good quality structural data was anticipated to be the most time-consuming phase of the structural proteomics program. Structure determination demands crystals which diffract to better than 3 Å resolution or proteins which remain stable and nonaggregated at high concentrations to yield high quality ^{15}N-heteronuclear single quantum coherence (HSQC) NMR spectra.

Our pilot project started in 1998 and covered over 700 proteins from archeaebacterium *M. thermoautotrophicum ΔH (Mth)*. Membrane proteins, which comprise about 30% of the *M. thermoautotrophicum* genome, were excluded from our target list because of low probability for success in our single "generic" sample preparation and crystallization protocol. Excluding membrane proteins avoided the complicating factor of working with structures whose conformation must span the lipid-rich cell membrane with alternating hydrophobic−hydrophilic domains. Furthermore, because our goal focused on unique structures, proteins with clear sequence similarity (BLAST[5] search with an *e*-value cutoff of 10^{-4}) to proteins in the Protein Data Bank (PDB) were excluded. Computational biology can be used to compare and assign similarity scores to proteins based on sequence and predicted protein topology. Because common interactions govern protein-folding and stability, proteins with comparable amino acid sequences often assume similar stable conformations.[6] Consequently, by removing redundant sequences during target selection, bioinformatics allowed us to maximize our survey of structural diversity. The remaining proteins were not prioritized as we set out to compile a broad, unbiased list of targets. The analysis of this set of proteins enabled us to study whether proteins with certain biophysical properties such as amino acid sequence, fold class, etc., were more amenable to the proteome-wide approach. Likewise, as structure−function relationships become more defined and annotated, it will become clear whether particular function categories (for example enzymes) are more amenable to this approach.

At the onset of the project, we decided to use both X-ray crystallography and NMR spectroscopy as structure determination tools. The quality of the NMR signal is highly dependent on the tumbling rate of the molecule being studied. The slower rotational correlation time and faster relaxation of the NMR signals associated with larger proteins complicates the NMR analysis. Consequently, we have chosen an arbitrary cutoff of around 20 kDa since the majority of the protein structures in the PDB solved by NMR at that time were of this size.[7] The development of a generic sample preparation protocol took on several stages. We decided to clone our targets with a fusion tag for ease of purification, and we choose the hexahistidine tag over other tags because it is small enough that we can "screen" the proteins by NMR with or without the tag. Thus, small proteins destined for NMR spectroscopy were labeled isotopically with ^{15}N and screened for suitability for NMR analysis using the ^{15}N-HSQC NMR experiment. Crystallization trials were initially used to screen only proteins larger than 20 kDa. However, more recently we have also included small proteins that failed to yield good quality NMR spectra. Figure 1 summarizes the workflow employed in this pilot project as well as the results achieved for each step. Experimental procedures are detailed in Christendat et al. 2000,[8] and Yee et al. 2002[9] and references are cited in Table 1.

The Sample Pipeline. Over 94% of the targets were successfully cloned in *Escherichia coli* expression vectors (Figure 1). A total of 70% of these clones yielded overexpressed recombinant protein. Only a fraction of these, 67% and 57% of the small and large proteins, respectively, were soluble in the *E. coli* lysate. A further 40% and 45% of the small and large proteins, respectively, were unable to be purified, largely due to protein precipitation either during nickel nitrilotriacetic acid (Ni−NTA) agarose (Qiagen) affinity chromatography or upon concentration in the final sample buffer. Such problems may be remedied in future proteomic screens with changes to buffering conditions, including the addition of ligands, detergent, salt, or glycerol, to promote solubility. A smaller fraction of proteins were lost because of physical or enzymatic degradation during the purification process or their inability to bind the Ni−NTA beads. Ni−NTA chromatography exploits the affinity of the imidazole moiety of histidine residues for divalent metal cations such as nickel. By expression of recombinant proteins with a hexa-histidine "tag" appended to the N-terminus of the polypeptide chain, target proteins were affinity purified efficiently. However, in rare cases the "tag" may be buried within the protein confounding purification efforts. A target protein was finally considered successfully purified if concentration for NMR or crystal trials (typically, >0.2 mM) was achieved without precipitation.

Of the samples that could be purified in sufficient quantity, 41 out of the 97 large proteins gave crystals, while 48 out of 115 small proteins gave a good HSQC NMR spectrum. A subset of the smaller proteins that exhibited poor NMR spectra was sent to crystal trials in order to improve the likelihood of structure determination. Strikingly, 22 of these 59 targets were "recovered" in crystallization. Overall, 9% of the initial large protein targets and 19% of the initial small proteins gave crystals or good

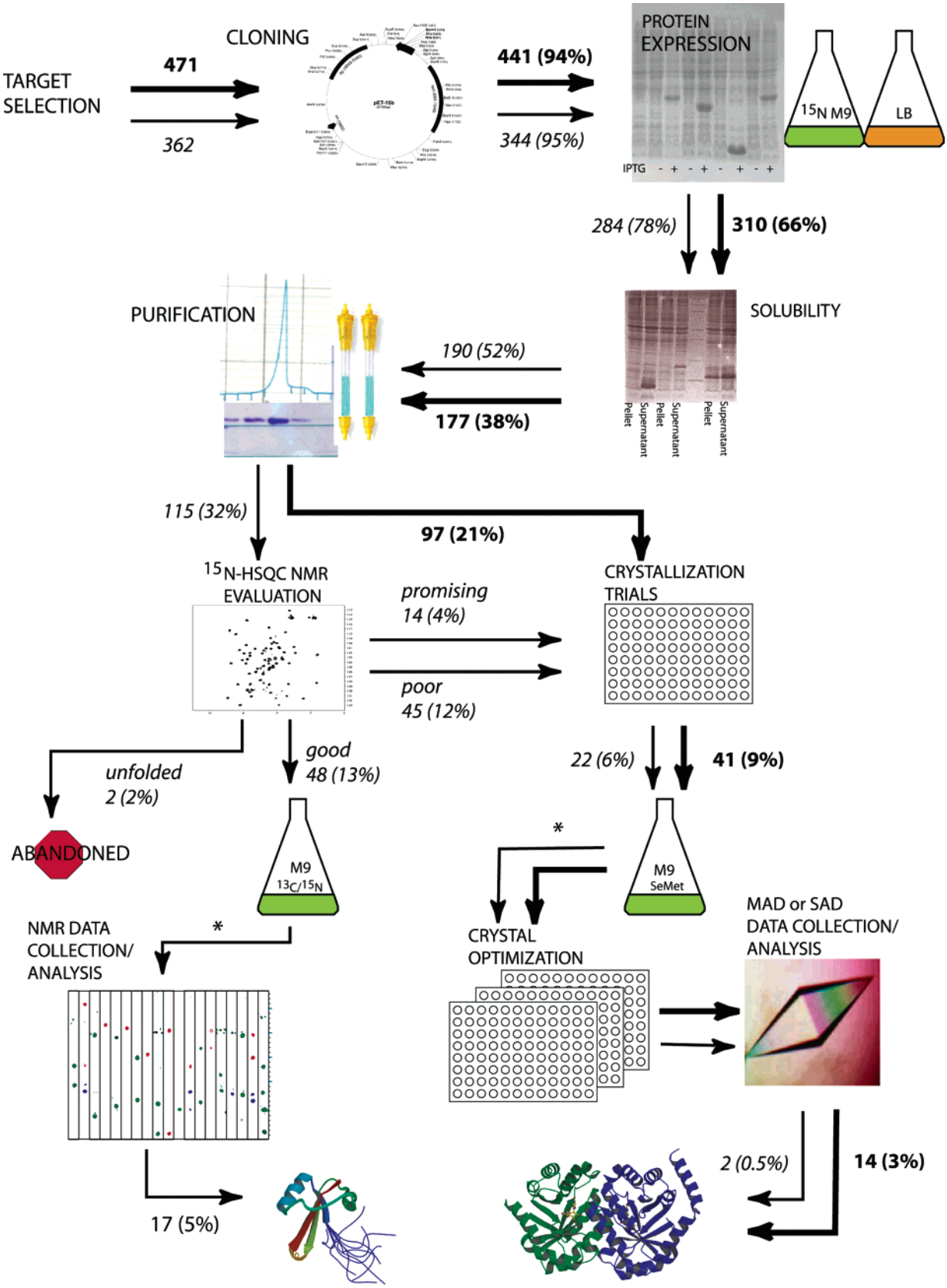

FIGURE 1. Schematic flow diagram of the strategy used in the *M. thermoautotrophicum* structural proteomics project. The number of targets after each step and the percentage relative to the number of starting targets are indicated in brackets. Thin arrows and italicized numbers are for smaller molecular weight proteins, and wide arrows and bold numbers are for larger molecular weight proteins. *Note that for NMR structures, not all proteins with good HSQC spectra were pursued for structure determination, and not all crystals were exhaustively screened for optimal crystal conditions, resulting in somewhat lower yield of structures than had we targeted more resources to these activities.

Table 1: Protein Structures from the *M. thermoautotrophicum* Proteome that Were Solved as Part of This Structural Proteomics Project[a]

gene number	original functional annotation[b]	structure-based annotation	PDB accession number	fold classification	number of sequence homologues[c]	structure determination method[d]	ref
Mth0001	CHP[e]	unknown	1K3R	α/β	19	X-ray[f]	–
Mth0040	RNA polymerase subunit 10	RNA polymerase subunit 10	1EF4	all α	30	NMR[g]	22
Mth0129	orotidine decarboxylase	orotidine decarboxylase	1DV7	α/β	63	X-ray[h]	23
Mth0146	precorrin 8w decarboxylase	precorrin 8w decarboxylase	1F38	α/β	39	X-ray[i]	32
Mth0150	CHP	NMNATase	1EJ2	α/β	30	X-ray	24
Mth0152	CHP	FMN-binding protein	1EJE	all β	40	X-ray	–
Mth0169	CHP	nucleotide biosynthesis	1GTD	α/β	7	X-ray[j]	–
Mth0256	CHP	unknown	1NE3	allβ	0	NMR	–
Mth0538	CHP	response regulatory system	1EIW	α/β	0	NMR[k]	25
Mth0637	CHP	unknown	1JRM	α/β	6	NMR	26
Mth0677	CHP	unknown	n.a.[l]	α/β		NMR[m]	
Mth0777	CHP	unknown	1KJN	α/β		X-ray	–
Mth0863	CHP	n.a.	n.a.	α + β		X-ray	–
Mth0865	CHP	unknown	1IIO	all α	4	NMR[g]	27
Mth0895	CHP	thioredoxin-like	1ILO	α + β	6	NMR[n]	28
Mth0938	CHP	unknown	1IHN	α/β	8	X-ray[o]	29
Mth1020	CHP	unknown	1KUU	α/β	3	X-ray	–
Mth1048	RNA polymerase subunit H	RNA polymerase subunit H	1EIK	α + β	44	NMR	30
Mth1175	CHP	unknown	1EO1	α/β	15	NMR[k]	31
Mth1184	CHP	unknown	1GH9	small protein	0	NMR[p]	8
Mth1187	CHP	unknown	1LXN	α/β	10	X-ray[j]	–
Mth1491	CHP	possible oxido-reductase	1L1S	α/β	8	X-ray	–
Mth1598	CHP	unknown	1JW3	α/β	21	NMR	9
Mth1615	CHP	nucleic acid binding	1EIJ	all α	13	NMR	8
Mth1675	CHP	unknown	n.a.	α/β	4	X-ray	–
Mth1692	CHP	RNA binding	1JCU	α/β	93	NMR°	–
Mth1699	CHP	translation elongation factor 1b	1GH8	α + β	14	NMR°	8
Mth1743	CHP	ubiquitin-like C-terminal conjugation protein	1JSB	α/β	0	NMR	9
Mth1747	CHP	dihydroxyacid dehydrogenase	1I36	α/β	5	X-ray	–
Mth1790	epimerase	epimerase	1EPZ	all β	99	X-ray	33
Mth1791	glucose-1-phosphate thymidylyl transferase	glucose-1-phosphate thymidylyl transferase	1LVW	α/β	99	X-ray[h]	–
Mth1821	CHP	unknown	n.a.	α + β	0	NMR[q]	–
Mth1880	CHP	Ca^{2+} binding protein	1IQO	α +β	1	NMR[q]	9

[a] Proteins were cloned, expressed, purified, and identified as samples that either form well-diffracting crystals or give good NMR spectra, in our laboratory. [b] Annotation as indicated in www.biosci.ohio-state.edu/~genomes/mthermo. [c] Based on BLAST search of nonredundant database using an *e*-value cutoff of 10^{-4}. [d] Unless otherwise indicated, structures were determined in the laboratory of C. Arrowsmith by NMR spectroscopy or A. Edwards by X-ray crystallography. [e] Conserved hypothetical protein. [f] A. Joachimiak, Angonne National Laboratory. [g] L. McIntosh, University of British Columbia. [h] E. Pai, University of Toronto. [i] J. Hunt, Columbia University. [j] L. Tong, Columbia University. [k] M. Kennedy, Pacific Northwest National Laboratory. [l] Final stages of structural refinement, PDB submissions in progress. [m] M. Rico, CSIC, Spain. [n] D. Wishart, University of Alberta. [o] E. Arnold, Rutgers University. [p] K. Gehring, McGill University. [q] W. Lee, Yonsei University.

HSQC spectra. These percentages represent the proteins that are amenable to structure analysis with a single expression/purification procedure without extra manipulation of conditions.

The quality of the HSQC spectra was found to be a good indicator of whether a solution structure could ultimately be determined for a protein (see spectra at: www.uhnres.utoronto.ca/proteomics). NMR structure determination requires weeks of data acquisition and significant manual analysis of the data. It was therefore evident that our ability to produce excellent NMR samples would exceed our capacity to determine the structures. To increase the

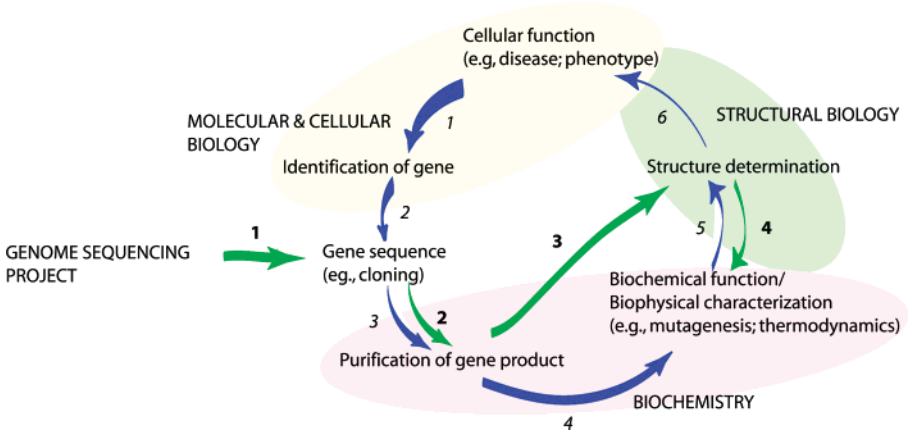

FIGURE 2. Schematic diagram showing the traditional path (blue arrows) and structural proteomics (green arrows) followed in the study of a particular gene and its product from the molecular and cellular biology field to biochemistry field and to structural biology.

throughput of structures, we elected to distribute the protein samples to collaborating NMR laboratories for analysis. Using this strategy, to date a total of 17 NMR structures had been determined from *M. thermoautotrophicum*, with several more in various stages of resonance assignment and structure refinement.

For well-diffracting crystals, in favorable instances, the protein structure can be solved within hours after acquiring the diffraction data. The bottleneck in crystallography is obtaining a well-diffracting crystal that produces high-quality data. Crystal formation relies on the uniform deposition of protein molecules. Sample inhomogeneity or disordered regions of the protein such as distal loops or unfolded termini can impair crystallization. Ideal crystals are regular arrays of closely packed protein molecules that diffract X-rays to high resolution (<2.8 Å). Only 42% of the purified proteins that went into the initial crystallization trials crystallized. So far we have optimized the crystallization conditions for 34% of initially crystallized proteins, to give well-diffracting crystals from which structures can be solved.

Of the 51 protein entries from *M. thermoautotrophicum* currently found in the PDB representing 34 genes, 36 of these, representing 29 genes, are the product of this structural proteomics pilot project (Table 1). The multiple PDB entries for several proteins reflect different crystallization conditions and several structures with bound cofactors or ligands. At the time of submission of this article, four additional structures were in the final stages of refinement and are also listed in Table 1. Sixteen structures are solved in six different X-ray crystallography laboratories, and 17 were solved in seven different NMR laboratories. This is a marked improvement in terms of the number of structures solved per structural biology laboratory. In a traditional structural biology research setting, scientists start from a well-characterized gene product and use the protein structure to explain the observed biochemical function or cellular function (blue arrows in Figure 2). The major portion of time spent by structural biologists is usually the preparation and screening of the proper protein construct and conditions to yield a good NMR sample or well-diffracting protein crystal. The rapid, parallel sample preparation and screening employed in structural proteomics has allowed structural biologists (in this case, our laboratory and our collaborators) to solved more structures than would otherwise be the case for a given amount of funding (green arrows in Figure 2).

One of our objectives in this project was to assess the extent to which new structural information could provide new functional insight, particularly for previously uncharacterized proteins. In Table 1 we have categorized the structures into three groups on the basis of their original functional annotation. First, six structures were for proteins with a preexisting functional annotation based on sequence homology. For example, the 3D structure of mth0129, an orotidine 5′ monophosphate decarboxylase, revealed the atomic-level details of its extraordinary catalytic activity.[10] The second category comprised 12 conserved hypothetical proteins for which the structure suggested a possible biochemical function. In several cases our structure-based annotation was a direct result of having (inadvertently) cocrystallized the protein with a cofactor (mth0152 and mth863), substrate, or product (mth0150).[11] Other proteins in this category include those that share structural similarity to a class of proteins with conserved residues important for a specific biochemical function. Examples of such functions include nucleic acid binding (mth1615 and mth1692), metal binding (mth1880), or C-terminal conjugation (mth1743). The third category of proteins comprises those that had no previous annotation and for which a functional annotation could not readily be derived from the 3D structure. Often this was because the structure was a member of a common fold class (e.g., a small helix bundle) or the protein itself was new or contained an unusual fold that could not be matched to any other proteins in the PDB (mth0637, mth1598).

Roughly equal numbers of structures were determined using both NMR and crystallography (17 and 16, respectively). This suggests that NMR spectroscopy can make significant contributions to structural proteomic efforts, which traditionally focused heavily on crystallography, if small to medium-sized proteins are included in the target

list. The complimentary nature of these two techniques in structure determination was well illustrated by the subset of smaller proteins examined using both methods. As mentioned, in a bid to optimize output, 59 proteins that expressed well, but gave poor ^{15}N HSQC spectra, were redirected to crystal trials. Ultimately, 22 of these proteins crystallized, and to date two crystal structures have been determined (mth0169 and mth1491).

Although not a primary focus of this project, an important goal of structural genomics/proteomics is the "filing out of fold space"—that is, determining the 3D structures for all classes in which protein folds so that most proteins in the universe can be computationally modeled on the basis of similarity of their amino acid sequence to that of proteins for which there are experimental structures. Unfortunately, because the relationship between sequence and 3D structure is not fully understood, it is currently impossible to predict *a priori* which gene sequences encode proteins with new 3D folds. The structures we have solved so far have yielded very few completely new folds (2−5 depending on how strictly "new" is defined), suggesting that discovery of new folds is a rare occurrence. Similar results have been obtained for other structural proteomics projects, even those that specifically seek to identify new protein folds. The variety of different structures from this project suggests that our strategy did not select for a particular protein fold or functional class. On the other hand, because our strategy does select for proteins amenable to a single, specific expression/purification protocol, the structures in Table 1 may comprise a set of folds that are particularly amenable to structural analysis or our procedure in particular. In this respect it is interesting to note that many of these proteins fall into the most common α/β fold classes.[12]

Outlook

We have expanded this study to survey the structural proteomics of several other organisms using the same strategy for target selection, protein production, and data collection/analysis. The generic protocols we employed for *M. thermoautotrophicum* yielded similar results for both thermophilic and mesophilic prokaryotes (*Thermotoga maritima* and *Escherichia coli*). Interestingly, we did not observe a clear advantage to targeting thermophilic proteins.[13] From other prokaryotes, we and our collaborators have determined over 29 additional structures of novel proteins.[6,14,15,16]

Preliminary production efforts for proteins from the eukaryotic proteomes of *Saccharomyces cerevisiae*, *Arabidopsis thaliana* (unpublished data) and myxoma virus[9] suggest that our generic protocol optimized for *M. thermoautotrophicum* results in a somewhat lower yields of soluble proteins. This prompted us to explore the development of a hierarchical expression and purification strategy in which "failures" at each step in Figure 1 are subjected to one or more alternative protocols so as to maximize the total throughput of soluble proteins, while minimizing labor and material costs. For example, alternative purification protocols using denaturants can often facilitate the recovery of proteins that are expressed into inclusion bodies,[17,18] expression in different cell strains or under different conditions can increase yields of poorly expressed proteins, or the use of different fusion tag might help improve protein expression and solubility.[19]

The key going forward will be to rigorously test the content and the order of a matrix of protocols that yield the most efficient production of the largest numbers of structural samples. Another important aspect of building this type of production strategy at a genomic scale relates to optimizing resources. Incorporating informatics tools into our production strategy will allow us to improve our structure determination rates by improving our experimental strategy on the basis of past experience. To date, mining of empirical databases to uncover trends in protein behavior and guide the development of the hierarchical protocols is largely unexploited. We have carried out some initial efforts in this area which suggest that careful tracking and mining of both successes and failures throughout the project can provide valuable information that will allow us to design more successful experimental protocols in the future.[8,20,21]

We thank our many colleagues in Toronto and worldwide for data collection and structure determinations for the proteins described here. In particular, we thank our protein production team in Toronto (http://www.uhnres.utoronto.ca/proteomics/); Drs. Guy Montelione, Michael Kennedy, Liang Tong, John Hunt, Hao Wu, Mark Gerstein, Burkhard Rost, and their groups in the Northeast Structural Genomics Consortium (NESG); Dr. Andrzej Joachimiak and the members of Structural Biology Center at Argonne National Laboratory; and Dr. Janet Thornton and members of the Midwest Center for Structural Genomics (MCSG). Part of the NMR work was performed at Environmental Molecular Sciences Laboratory (a national scientific user facility sponsored by Department of Energy Biological and Environmental Research) located at Pacific Northwest National Laboratory operated by Battelle and funded by the U.S. Department of Energy Office of Biological and Environmental Research, under Contract W-31-109-Eng-38. This research was supported by the Ontario Research and Development Challenge Fund and the NIH Protein Structure Initiative (P50-GM62414-01 and P50-GM62413-02). A.M.E. and C.H.A. are CIHR Investigators.

References

(1) Abola, E.; Kuhn, P.; Earnest, T.; Stevens, R. Automation of x-ray crystallography. *Nat. Struct. Biol.* **2000**, *7*, 973−977.
(2) Lamzin, V.; Perrakis, A. Current state of automated crystallographic data analysis. *Nat. Struct. Biol.* **2000**, *7*, 978−981.
(3) Montelione, G. T.; Zheng, D.; Huang, Y.; Szyperski, T. Protein NMR Spectroscopy in Structural Genomics. *Nat. Struct. Biol.* **2000**, *7*, 982−984.
(4) Wasserfallen, A.; Nolling, J.; Pfister, P.; Reeve, J.; Macario, E. Phylogenetic analysis of 18 thermophilic Methanobacterium isolates supports the proposals to create a new genus, Methanothermobacter gen. Nov., and to reclassify several isolates in three species, Methanothermobacter thermoautotrophicus comb. nov., Methanothermobacter wolfeii comb. nov., and Methanothermobacter marburgensis sp. nov. *Int. J. Syst. Evol. Microbiol.* **2000**, *50*, 43−53.
(5) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(6) Patthy, L. Protein Structure. In *Protein Evolution*; Blackwell Science: Oxford 1999; Chapter 2.

(7) Guntert, P. Structure calculation of biological macromolecules from NMR data. *Q. Rev. Biophysics* **1998**, *31*, 145−237.
(8) Christendat, D.; Yee, A.; Dharamsi, A.; Kluger, Y.; Savchenko, A.; Cort, J.; Booth, V.; Mackereth, C.; Saridakis, V.: Ekeil, I.; et al. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **2000**, *7*, 903−909.
(9) Yee, A.; Chang, X.; Pineda-Lucena, A.; Wu, B.; Semesi, A.; Le, B.; Ramelot, T.; Lee, G. M.; Bhattacharyya, S.; Gutierrez, P., et al. An NMR approach to structural proteomics. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1825−1830.
(10) Wu, N.; Mo, Y.; Gao, J.; Pai, E. F. Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2017−2022.
(11) Saridakis, V.; Christendat, D.; Kimber, M. S.; Dharamsi, A.; Edwards, A. M.; Pai, E. F. Insights into ligand binding and catalysis of a central step in NAD+ synthesis: structures of *Methanobacterium thermoautotrophicum* NMN adenylyltransferase complexes. *J. Biol. Chem.* **2001**, *276*, 7225−7232.
(12) Gerstein, M.; Hegyi, H. Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **1998**, *22*, 277−304.
(13) Savchenko, A.; Yee, A.; Khachatryan, A.; Skarina, T.; Evdokimova, E.; Pavlova, M.; Semesi, A.; Northey, J.; Beasley, S.; Lan, N.; Das, R.; Gerstein, M.; Arrowsmith, C. H.; Edwards, A. M. Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and x-ray crystallography approaches. *Proteins: Struct., Funct., Genet.*, in press.
(14) Zhang, R. G.; Skarina, T.; Katz, J. E.; Beasley, S.; Khachatryan, A.; Vyas, S.; Arrowsmith, C. H.; Clarke, S.; Edwards, A.; Joachimiak, A.; Savchenko, A. Structure of *Thermatoga maritima* stationary phase survival protein SurE: a novel acid phosphatase. *Structure* **2001**, *9*, 1095−1106.
(15) Korolev, S.; Ikeguchi, Y.; Skarina, T.; Beasley, S.; Arrowsmith, C. H.; Edwards, A.; Joachimiak, A.; Pegg, A. E.; Savchenko, A. The crystal structure of spermidine synthase with multisubstrate adduct inhibitor. *Nat. Struct. Biol.* **2002**, *9*, 27−31.
(16) Zhang, R.; Kim, Y.; Skarina, T.; Beasley, S.; Laskowski, R.; Arrowsmith, C.; Edwards, A.; Joachimiak, A.; Savchenko, A. Crystal structure of Thermotoga maritima 0065 − a member of the IclR transcriptional factor family. *J. Biol. Chem.* **2002**, in press.
(17) Altamirano, M. M.; Golbik, R.; Zahn, R.; Buckle, A. M.; Fersht, A. R. Refolding chromatography with immobilized mini-chaperones. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 3576−3578.
(18) Xie, Y.; Wetlaufer, D. B. Control of aggregation in protein refolding: The temperature-leap tactic. *Prot. Sci.* **1996**, *5*, 517−523.
(19) Hammarstrom, M.; Hellgren, N.; Van der Berg, S.: Berglund, H.; Hard, T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* **2002**, *11*, 313−321.
(20) Bertone, P.; Kluger, Y.; Lan, N.; Zheng, D.; Christendat, D.; Yee, A.; Edwards, A. M.; Arrowsmith, C. H.; Montelione, G. T.; Gerstein, M. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics, *Nucleic Acid Res.* **2001**, *29*, 2884−98.
(21) Kimber, M. S.; Vallee, F.; Houston, S.; Necakov, S.; Vedadi, M.; Skarina, T.; Evdokimova, E.; Beasley, S.; Christendat, D.; Savchenko, A.; Arrowsmith, C. H.; Gerstein, M.; Edwards, A. M. Data mining crystallization databases: Knowledge-based approaches to optimize protein crystal screens. *Proteins*: *Struct., Funct., Genet.*, in press.
(22) Mackereth, C. D.; Arrowsmith, C. H.; Edwards, A. M.; McIntosh, L. P. Zinc bundle structure of the essential RNA polymerase subunit RPB10 from *Methanobacterium thermoautotrophicum*. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6316−6321.
(23) Wu, N.; Mo, Y.: Gao, J.; Pai, E. F. Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2017−2022.
(24) Saridakis, V.; Christendat, D.; Kimber, M. S.; Dharamsi, A.; Edwards, A. M.; Pai, E. F. Insights into ligand binding and catalysis of a central step in NAD+ synthesis: structures of *Methanobacterium thermoautotrophicum* NMN adenylyltransferase complexes. *J. Biol. Chem.* **2001**, *276*, 7225−7232.
(25) Cort, J. R.; Yee, A.; Edwards, A. M.; Arrowsmith, C. H.; Kennedy, M. A. Structure-based functional clasification of hypothetical protein Mth538 from *Methanobacterium thermoautotrophicum*. *J. Mol. Biol.* **2000**, *302*, 189−203.
(26) Pineda-Lucena, A.; Yi, G.; Cort, J.; Kennedy, M.: Edwards, A.; Arrowsmith, C. H. Solution structure of the hypothetical protein Mth0637 from *Methanobacterium thermoautotrophicum*. *J. Biomol. NMR*, submitted.
(27) Lee, G. M.; Edwards, A. M.; Arrowsmith, C. H.; McIntosh, L. P. NMR-based structure of the conserved protein Mth865 from the archeon *Methanobacterium thermoautotrophicum*. *J. Biomol. NMR* **2001**, *21*, 63−66.
(28) Bhattarcharyya, S.; Habibi-Nazhad, B.; Amegbey, G.; Slupsky, C.; Yee, A.; Arrowsmith, C.; Wishart, D. Identification of a novel archaebacterial thioredoxin: Determination of function through structure. *Biochemistry* **2002**, *41*, 4760−4770.
(29) Das, K.; Xiao, R.; Wahlberg, E.: Hsu, F.; Arrowsmith, C. H.; Montelione, G. T.; Arnold, E. X-ray crystal structure of MTH938 from *Methanobacterium thermoautotrophicum* at 2.2 Å resolution reveals a novel tertiary protein fold. *Proteins* **2001**, *45*, 486-488.
(30) Yee, A.; Booth, V.; Dharamsi, A.; Engel, A.; Edwards, A. M.; Arrowsmith, C. H. Solution structure of the RNA polymerase subunit RPB5 from *Methanobacterium thermoautotrophicum Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6311−6315.
(31) Cort, J. R.; Yee, A.; Edwards, A. M.; Arrowsmith, C. H.; Kennedy, M. A. NMR structure determination and structure-based functional characterization of conserved hypothetical protein Mth1175 from *Methanobacterium thermoautotrophicum*. *J. Struct. Funct. Genom.* **2000**, *1*, 15−25.
(32) Keller, J.; Smith, P.; Benach, J.; Christendat, D.; deTitta, G.; Hunt, J. The Crystal Structure of MT0146/CbiT suggests that the putative precorrin-8W decarboxylase is a methyl transferase. *Structure* **2002**, *10*, 1475−1487.
(33) Christendat, D.; Saridakis, V.; Dharamsi, A.; Bochkarev, A.; Pai, E. F.; Arrowsmith, C. H.; Edwards, A. M. Crystal structure of dTDP-4-keto-6-deoxy-D-hexulose 3,5-epimerase from *Methanobacterium thermoautotrophicum* complexed with dDTP. *J. Biol. Chem.* **2000**, *275*, 24608−24612.

AR010126G