An NMR approach to structural proteomics

Adelinda Yee*, Xiaoqing Chang*, Antonio Pineda-Lucena*, Bin Wu*, Anthony Semesi*, Brian Le*, Theresa Ramelot†, Gregory M. Lee‡, Sudeepa Bhattacharyya§, Pablo Gutierrez¶, Aleksej Denisov¶, Chang-Hun Lee‖, John R. Cort†, Guennadi Kozlov¶, Jack Liao*, Grzegorz Finak¶, Limin Chen*, David Wishart§, Weontae Lee‖, Lawrence P. McIntosh‡, Kalle Gehring¶, Michael A. Kennedy†, Aled M. Edwards*, and Cheryl H. Arrowsmith*,**

*Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, 101 College Street, Toronto, ON, Canada M5G 1L7; †Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, K8-98, Richland, WA 99352; ‡Departments of Biochemistry and Molecular Biology, Chemistry, and the Biotechnology Laboratory, 2146 Health Sciences Mall, University of British Columbia, Vancouver, BC, Canada V6T 1Z3; §Faculty of Pharmacy and Pharmaceutical Sciences, 3118 Dentistry/Pharmacy Centre, University of Alberta, Edmonton, AB, Canada T6G 2N8; ¶Department of Biochemistry and Montreal Joint Centre for Structural Biology, McGill University, 3655 Promenade Sir William Osler, Montreal, QC, Canada H3G 1Y6; and ‖Department of Biochemistry and Protein Network Research Center, Yonsei University, 134 Shinchon-Dong Seodaemoon-Gu, Seoul, Korea 120-749

Communicated by Louis Siminovitch, Mount Sinai Hospital, Toronto, Canada, December 19, 2001 (received for review September 11, 2001)

The influx of genomic sequence information has led to the concept of structural proteomics, the determination of protein structures on a genome-wide scale. Here we describe an approach to structural proteomics of small proteins using NMR spectroscopy. Over 500 small proteins from several organisms were cloned, expressed, purified, and evaluated by NMR. Although there was variability among proteomes, overall 20% of these proteins were found to be readily amenable to NMR structure determination. NMR sample preparation was centralized in one facility, and a distributive approach was used for NMR data collection and analysis. Twelve structures are reported here as part of this approach, which allowed us to infer putative functions for several conserved hypothetical proteins.

Structural proteomics, which aims to determine the three-dimensional (3D) structures of all proteins, has become a major initiative within the biomedical community (see ref. 1 and other articles in the same issue). The large number of protein structures expected from these projects will yield valuable clues to the rules for predicting protein folding and understanding biochemical function. In these early stages of the structural proteomics effort, one of the main goals is to identify the best technologies and the most efficient processes to convert gene sequence into 3D structural information. One of the decisions will be to determine the optimal use of x-ray crystallography and NMR spectroscopy, which are the two techniques that will provide the majority of experimental data for these initiatives.

X-ray crystallography currently is perceived as the potential workhorse for structural proteomics, because if provided with a well diffracting crystal it is possible to determine a 3D structure in hours. However, the throughput of structure determination using x-ray crystallography remains unclear, because the rate-determining step continues to be the production of well diffracting crystals, a process that is unpredictable and can take between hours and months.

NMR structure determination is limited currently by size constraints and lengthy data collection and analysis times (often months), and the method is best applied to proteins smaller than 250 amino acids. On the other hand, NMR experiments do not require crystals, and samples appropriate for structure determination can be identified within minutes of the protein being purified. In summary, x-ray crystallography and NMR spectroscopy seem to have complementary deficiencies, and the relative success of these methods in structural proteomics remains to be determined.

We have shown previously that NMR spectroscopy can play a significant role in structural proteomics even with its current limitations (2). The initial pilot project, based on a limited number of proteins from the thermophilic archaebacterium *Methanobacterium thermoautotrophicum* (Mth) suggested that smaller proteins may be more amenable to structure analysis, because in this genome a higher proportion of smaller proteins were soluble compared with larger proteins. However, this study was performed on a single proteome, and it was unclear how general these conclusions were.

Here we outline a strategy for the use of NMR spectroscopy for structural proteomics of small proteins based on data from 513 proteins from five microorganisms. These microorganisms include both thermophilic and mesophilic species and representatives from the prokaryotes, archaea, and eukaryotes [*Escherichia coli* (ecoli), *M. thermoautotrophicum*, *Thermotoga maritima* (TM), *Saccharomyces cerevisiae*, and the myxoma virus (Myx)]. We used an approach in which all proteins were cloned, expressed, and screened for suitability for NMR analysis in a single laboratory, and NMR data collection and structure determination were distributed among several NMR laboratories. Here we report the 3D structures of 12 proteins fully analyzed in this manner. These proteins are conserved but mostly unannotated from four species ranging in size from 8.4 to 22.6 kDa. The solution structures revealed several unusual folds as well as hint about the type of (and frequency with which) functional inferences that can be expected from larger scale structural proteomics projects focused on small proteins.

BIOCHEMISTRY

Materials and Methods

Expression and Purification. The targets selected for NMR screening had the following properties: single chain polypeptide with molecular mass under 23 kDa, no predicted transmembrane helix using TMHMM (www.cbs.dtu.dk), and no sequence homologue in the PDB database as identified by a BLAST search with an e-value cut-off of 10^{-4}. Targets were PCR-amplified from genomic DNA and subcloned in 96-well format into a pET15b (Novagen) or a modified pET15b vector with the thrombin cleavage site replaced with a TEV protease cleavage site. These vectors express proteins with an N-terminal hexahistidine tag followed by a thrombin or TEV protease cleavage site.

For screening, all proteins were expressed in *E. coli* strain BL21- (Gold λDE3), and in the case of the archaeal and eukaryote proteins the cells were cotransformed with a plasmid encoding three transfer RNAs for rare *E. coli* codons. Cells were

Abbreviations: 3D, three-dimensional; Myx, myxoma virus; HSQC, heteronuclear single quantum coherence; COG, cluster(s) of orthologous groups.

**To whom reprint requests should be addressed. E-mail: carrow@uhnres.utoronto.ca.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data bank (PDB), www. rcsb.org, and chemical shifts have been deposited in the BioMagResBank (BMRB), www.bmrb.wisc.edu: Mth0637 (PDB, 1JRM; BMRB, 5104), Mth0695 (PDB, 1IIO; BMRB, 4996), Mth0895 (PDB, 1IL0; BMRB, 4991), Mth1598 (PDB, 1JW3; BMRB, 5165), Mth1743 (PDB, 1JSB; BMRB, 5106), Mth1880 (PDB, 1IQO; BMRB, 5129), TM0983 (PDB, 1JDQ; BMRB, 5060), YedF_ecoli (PDB, 1JE3; BMRB, 5059), Yjbj_ecoli (PDB, 1JYG; BMRB, 5106), Hha_ecoli (PDB, 1JW2; BMRB, 5166), and Myxv156r (PDB, 1JJG; BMRB, 5077).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

grown in 1 liter of M9 minimal medium containing $^{15}NH_4Cl$ as the sole nitrogen source and supplemented with $ZnCl_2$, thiamine, and biotin. The cells were grown at 37°C to an OD_{600} of 0.6 and induced with 1 mM isopropyl β-D-thiogalactoside. Afterward, the temperature was reduced to 15°C, and the cells were allowed to grow overnight before harvesting. Frozen cell pellets were thawed in 500 mM NaCl/20 mM Tris/5 mM imidazole (pH 8) and lysed by sonication. The proteins were extracted from the lysates by batch Ni^{2+} affinity chromatography (Qiagen). The Ni^{2+} affinity beads were washed three times with 5 column volumes of 500 mM NaCl/20 mM Tris/30 mM imidazole (pH 8), and the proteins were eluted with 5 column volumes of 500 mM imidazole in this same buffer. The purified proteins were concentrated, and buffer was exchanged by ultrafiltration and dilution/reconcentration. The final "generic" NMR buffer comprised 450 mM NaCl/25 mM Na_2PO4/10 mM DTT/20 μM Zn^{2+}/1 mM benzamidine/1× inhibitor mixture (Roche Molecular Biochemicals)/0.01% NaN_3 (pH 6.5).

The proteins selected for structure determination generally were expressed in M9 minimal medium containing both $^{15}NH_4Cl$ and [$^{13}C_6$]glucose purified as described above plus a further chromatography step using either SP Sepharose or DEAE Sepharose (Amersham Pharmacia) ion-exchange columns. In some cases, the $(His)_6$ tag was cleaved by incubation with thrombin or TEV and removed by using a Ni^{2+} affinity column. When necessary, proteins were passed through an additional benzamidine Sepharose (Amersham Pharmacia) column to remove the thrombin.

NMR Spectroscopic Screening. All ^1H–^{15}N heteronuclear single quantum coherence (HSQC) spectra were acquired at 25°C by using a Varian INOVA 500- or 600-MHz spectrometer equipped with a pulse-field gradient unit and actively shielded z-gradient triple resonance probes. The total number of t1 increments was 64, with 8–64 scans per increment depending on the concentration of the sample being screened. The data were processed by using the NMRPIPE software package (3).

NMR Structure Determination. Twelve of the proteins that were deemed suitable for structure determination were distributed among six NMR laboratories. The proteins were labeled uniformly with ^{13}C and ^{15}N, resonances were assigned by using conventional triple resonance techniques, and structures were calculated by using distance and dihedral angle constraints derived from nuclear Overhauser effects and coupling constants. The choice of NMR experiments varied between laboratories and is documented in Figs. 4 and 5, which are published as supporting information on the PNAS web site, www.pnas.org. The total acquisition time varied between 204 and 1,390 h. On average, the acquisition times were 200 h for a suite of experiments for backbone resonance assignments, 202 h for side chain resonance assignments, and 261 h for distance and dihedral angle restraints.

Results and Discussion

Protein Production and Screening by ^{15}N HSQC. To screen as many proteins as possible, a total of 513 ORFs were expressed, purified, and examined by NMR under identical conditions. Most proteins (85%) expressed well in *E. coli* grown in minimal medium, and 68% of the expressed proteins remained in the soluble fraction of the cell lysate (Fig. 1). Thus, over half the ORFs chosen in this study could be expressed in soluble form in *E. coli*.

We selected the proteins best suited for NMR structure determination by employing a rapid batch purification of polyhistidine-tagged ^{15}N-labeled protein followed by a rapid "screening" of labeled proteins by ^1H–^{15}N HSQC spectroscopy. The HSQC spectrum provides a diagnostic fingerprint of a protein.

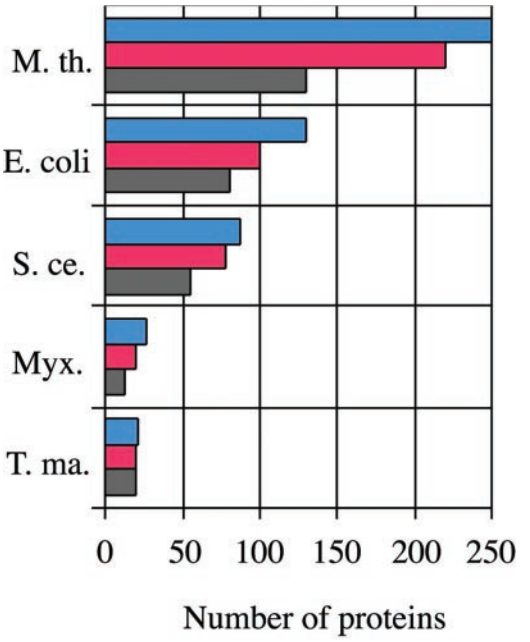

Fig. 1. Histogram of the number of proteins cloned (blue), expressed (red), and soluble (gray) from each organism.

One peak is expected for each nonproline residue. ^1H–^{15}N HSQC spectra of the soluble proteins could be classified as good, promising, poor, or mostly unfolded. The "good" spectra showed dispersion of peaks with roughly equal intensity and in the number expected from the sequence of the protein. These spectra indicated that the protein was amenable to structure determination by NMR methods. Approximately 33% of the soluble proteins, or 19% of the total, gave HSQC spectra that could be classified as good (Fig. 2). "Promising" spectra showed well dispersed peaks but were either too few or too many in number and were often of differing intensities. Such spectral features are indicative of conformational heterogeneity with slow or nonexistent interconversion between states (too many peaks) or the presence of dynamic processes on an intermediate time scale that can broaden and obscure the NMR signals. The behavior of these proteins sometimes can be optimized by changing either the protein construct or the solution conditions. Between 11 and 33% of soluble proteins in each proteome gave promising HSQCs.

HSQC spectra that showed very little peak dispersion were classified as either "mostly unfolded" or "poor." Mostly unfolded proteins showed many very sharp, intense peaks with random coil chemical shifts. Less than 10% of the soluble proteins fall under this classification. The poor HSQC spectra were characterized by a cluster of broad peaks with little dispersion in the center of the spectrum. Typically these spectra did not have the requisite number of peaks for the size of protein examined and likely reflect aggregated and/or conformationally unstable proteins that may be partially unfolded or that interconvert between multiple conformations. Approximately 27–55% of the soluble proteins exhibited these properties. Large, stable oligomers of the small proteins also could generate spectra that could be classified as poor. The remaining 10–25% of the soluble proteins precipitated during concentration and could not be subjected to further NMR analysis.

Comparison Between Species. One of the goals of this work was to determine whether there are trends among species for overall protein "behavior" to facilitate target selection for structural

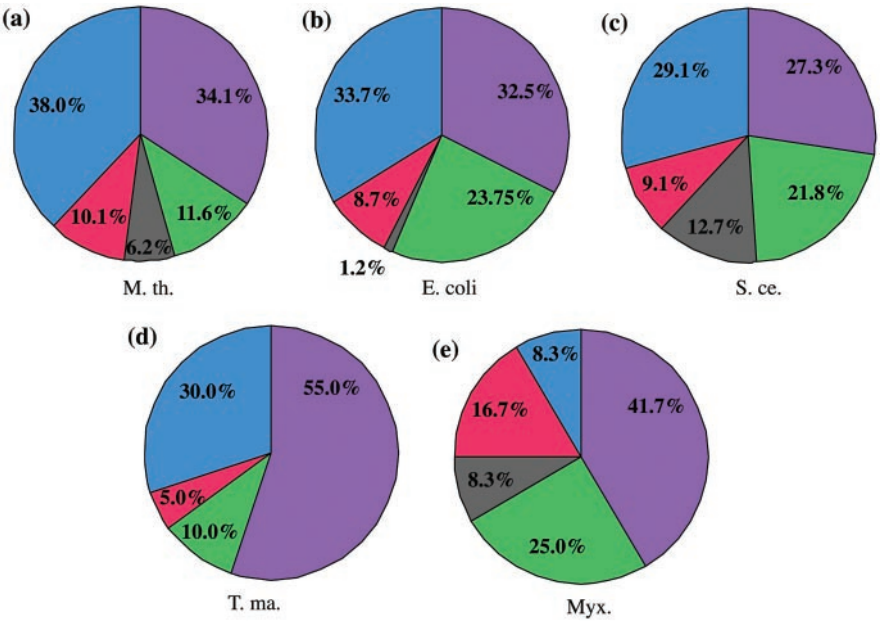

Fig. 2. Distribution chart of the HSQC classifications for soluble, purifiable proteins. (*a*) *M. thermoautotrophicum* (M. Th.); (*b*) *E. coli*; (*c*) *S. cerevisiae* (S. ce.); (*d*) *T. maritima* (T. ma.); (e) Myx. Spectra are classified as good (blue), promising (red), mostly unfolded (gray), poor (purple), and those for which no HSQC could be obtained because of a loss of protein during the concentration procedure (green).

proteomics projects. We therefore compared the expression, solubility, and HSQC results for the five species studied here. Among the proteomes, *T. maritima* proteins showed the largest proportion of soluble proteins (95%), followed by *S. cerevisiae* (63%), *E. coli* (61%), *M. thermoautotrophicum* (51%), and Myx (46%), respectively. Similar trends between organisms were observed for classification of HSQC spectra; 33% of *T. maritima* proteins had good or promising spectra followed by *E. coli* (26%), *M. thermoautotrophicum* (25%), *S. cerevisiae* (24%), and Myx (11.5%). In an effort to better understand whether the relative amino acid content of proteins in these organisms was correlated with their suitablilty for NMR analysis, we compared the percentages of each amino acid content within four broad classes of protein behaviors: insoluble proteins, good/promising HSQC, poor or no HSQC, and unfolded proteins. These percentages were compared also with those for proteins deposited in the PDB to see whether, for example, the amino acid content of proteins with good/promising HSQCs more closely matched that of proteins in the PDB compared with insoluble proteins (Fig. 5). We were unable to detect any significant trends in amino acid composition either among genomes or between different classes of protein behavior. The most significant differences in amino acid content correlated with the difference between thermophilic and mesophilic organisms and not necessarily with protein solubility or HSQC classification.

Among the prokaryotes, it appears that proteins from the thermophilic organism *T. maritima* had better-behaved proteins under our growth and lysis conditions. However, this conclusion could be biased by the relatively small number of proteins that we sampled from *T. maritima* (only 21 compared with 130 from *E. coli*). Surprisingly, compared with *T. maritima*, a similar proportion of soluble proteins from *E. coli*, *M. thermoautotrophicum*, and *S. cerevisiae* gave good/promising HSQCs. This result may suggest that the thermophilic properties of *M. thermoautotrophicum* proteins do not provide a significant advantage for structural proteomics of small proteins by using NMR. Proteins from Myx showed a particularly low percentage of well behaved proteins under our generic conditions. This poor behavior may reflect the fact that most viral proteins have evolved to interact with proteins in the host cell and may not be stable when expressed in isolation in *E. coli*.

Selection of Screening Conditions. There are several published methods to assess the suitability of a protein for NMR structure determination rapidly. In these methods, for example the ^{15}N-pulse labeling (4) and *in vivo* ^{15}N HSQC (5), the suitability for structure determination is estimated by the behavior of the impure protein in extracts or cells. Our method of using generic expression, purification, and final NMR solution condition also allowed us to screen a large number of proteins in a cost-effective and efficient manner. We elected to pursue this generic method, because our earlier studies indicate that only a small portion of the proteins that are soluble at lower concentration in the lysate or cells can be purified and concentrated. Therefore, many proteins that may have given good HSQC spectra in an *in vivo* screening protocol may not necessarily have been amenable to *in vitro* structure determination. However, in the future, with the advent of improved sensitivity from cryogenic probes, it may not be necessary to concentrate proteins as much in order collect NMR data for structure determination, and the *in vivo* screening strategy may be more appropriate.

Structural Results. The solution structures for the 12 proteins we report here provided insight into the functional information that will be expected from NMR-based structural proteomics and also enabled us to assess and optimize our distributive data collection strategy (Fig. 3 and Table 1). The majority of these proteins (10 of 12) were "conserved" or "hypothetical" proteins with no functional annotation. Most (8 of 12) also were members of uncharacterized protein superfamilies or clusters of orthologous groups (COG). Although several of the superfamilies/COG had members that were distantly related to proteins of known structure or function (for example ThiS (6) in COG2104), each of the proteins chosen here had a sufficiently unique sequence that a good structure alignment could not be established with a protein of known 3D structure. The sequences were submitted to SwissModeler (7, 8) to determine whether a 3D structure could be predicted based on sequence similarity. In

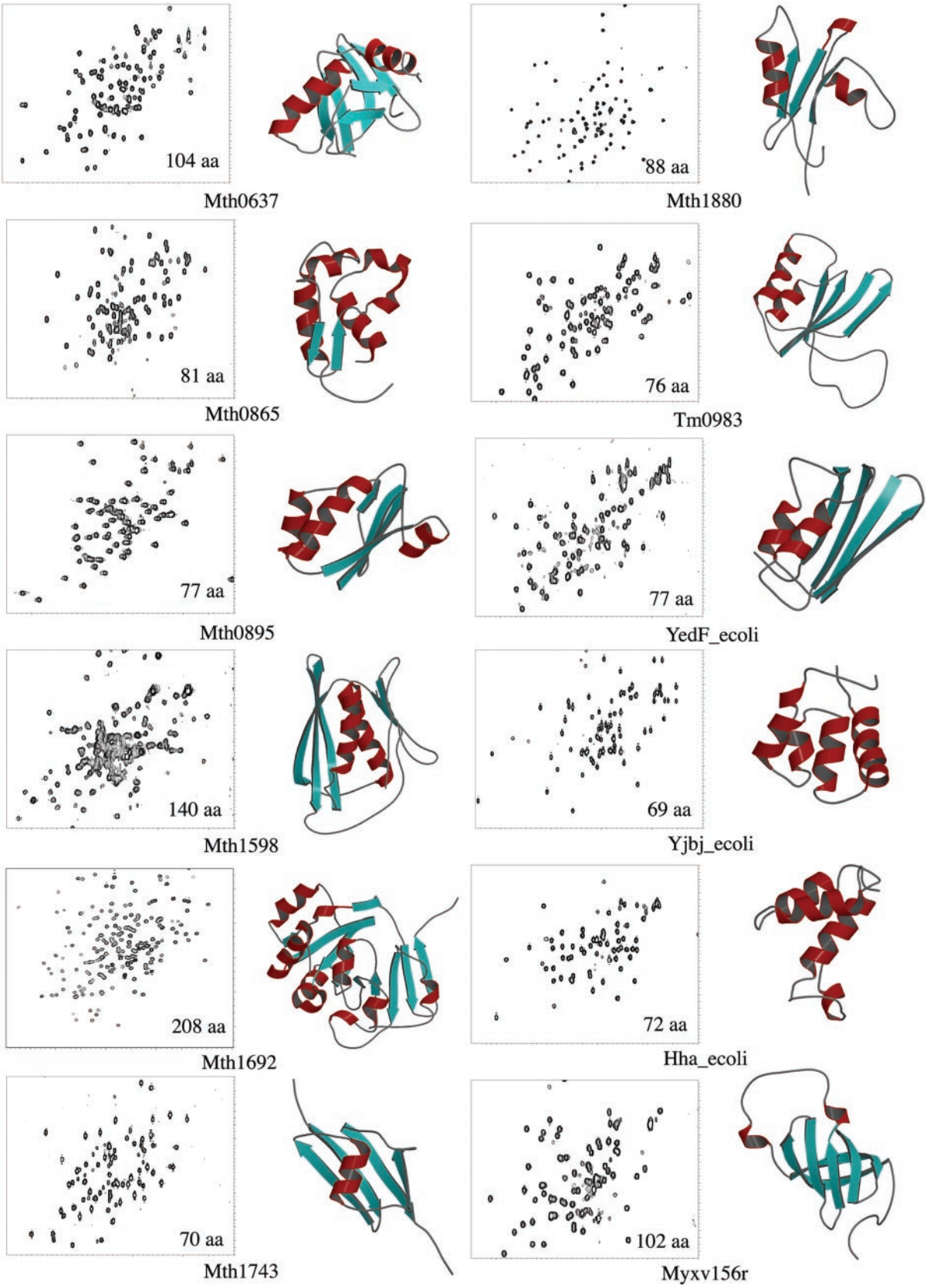

Fig. 3. ^{15}N HSQC spectra and the backbone ribbon representations of the 12 structures presented in this paper. All HSQCs are plotted from 6.0–10.5 ppm in the ^{1}H dimension (*x* axis) and from 107 to 133 ppm in the ^{15}N dimension (*y* axis). The number of residues for each protein is indicated on the HSQC spectrum. β-sheets are shown in cyan, and α-helices are shown in red. N-terminal residues 1–20 of yedF_ecoli and Myxv156r are unstructured and not shown. C-terminal residues 198–208 of Mth1692 are unstructured and not shown. All structure diagrams were created by using the MOLAUTO program within MOLSCRIPT (17).

Table 1. Sequence and structural characteristics of 12 structural proteomics targets

Protein	Annotation	Sequence homologues*	Fold class	Structural homologues†‡	Possible function
Mth0637	CHP§	SF004865 COG1872	Orthogonal β-sandwich with 2 helices	Translation initiation factor (2IF1)¶	Unknown
Mth0865	CHP	1 archeal protein	Helix bundle with β-hairpin	Bsobi restriction endonuclease (1DC1)	ATP binding
Mth0895	CHP	SF006424 COG0526	Thioredoxin	Thioredoxin (1THX)	Redox
Mth1598	CHP	SF006548 COG1371	α/β Sandwich	Heat-shock protein 33 (1I7F)	RNA binding
Mth1692	CHP	SF004931 COG0009	α/β Twisted Open-sheet	yrdC_ecoli (1HRU)	RNA binding
Mth1743	CHP	COG2104	Ubiquitin fold	Ubiquitin (1UBI)	C-terminal Conjugation
Mth1880	CHP	SF005648	$\alpha/\beta/\alpha$ Fold	None	Ca^{2+}-dependent Protein binding
TM983	CHP	SF006277 COG0425	Two-layered α/β sandwich	Translation initiation factor 3 (1TIG)	RNA binding
YedF_ecoli	CHP	SF006277 COG0425	Two-layered α/β sandwich	yhhp_ecoli (1DCJ)	RNA binding
Yjbj_ecoli	Hypothetical	None	4-Helix bundle	None	Unknown
Hha_ecoli	Hemolysin expression modulating	SF006255	Helix–turn–helix	Methane monooxygenase hydroxylase (1MTY-G)	DNA binding
Myxv156r	Interferon resistance, eIF2α homolog	C8, K3L viral proteins	β-Barrel	PNPase fragment (1SR0)	RNA/protein binding

*Superfamily and COG numbers were obtained from the Integrated Protein Classification database (pir.georgetown.edu/iproclass).
†Coordinates were submitted to DALI server (www2.ebi.ac.uk/dali) and structures with Z values larger than 2.0 were examined visually for similarity. Only a representative structural homologue is listed.
‡The PDB accession numbers for the structural homologues are indicated in parentheses.
§Conserved hypothetical protein.
¶Structural homology was observed from manual search of the CATH protein structure classification (www.biochem.ucl.ac.uk).

only two cases (Mth1692 and YedF_ecoli) was a structure prediction returned. In these cases the sequence alignment was based on a 27.4 and 30.6% sequence identity between Mth1692 and Yrdc_ecoli, and YedF_ecoli with YhhP_ecoli, respectively. The predicted models had an rms deviation between backbone atoms compared with the experimental NMR structures of 4.4 and 5.8 Å for Mth1692 and YedF_ecoli, respectively. Thus, these 12 structures represent 3D structures and form the basis for modeling of up to 87 additional homologous proteins from all three kingdoms.

Although none of the 12 proteins had a novel fold, several were novel variants of known folds. Most structures could be classified as being structurally similar to other structures in the PDB using tools such as DALI (9, 10). The most unique fold is that of Mth0637, which has a fold resembling that of translation initiation factor 1 (PDB ID code 2IF1), a two-layered α/β fold. However, Mth0637 has an insert of a two-stranded β-sheet in the middle of the sequence that packs orthogonally to the main β-sheet. Mth0637 also has a similar architecture (from the CATH classification system (11, 12) to that of a PDZ domain (PDB ID code 1Q1C) with the N-terminal β-strand of Mth0637 taking the place of the C-terminal strand in the PDZ domain. TM0983 and YedF_ecoli (which are members of the same COG) formed a two-layered α/β fold that is structurally homologous to translation initiation factor 3 (PDB ID code 1TIG). Myxv156r has a β-barrel fold found in S1 RNA-binding domains and cold-shock domain (13). Myxv156r was annotated as an IFN-resistance eIF2α homolog, and the structure is very similar to that predicted for the N-terminal domain of eukaryotic eIF2α. Hha_ecoli is annotated as a hemolysin modulating protein and belongs to a superfamily of transcriptional regulatory and DNA-binding proteins. The structure of Hha_ecoli reveals a helix–turn–helix motif typical of many other bacterial DNA-binding proteins. Mth1598 has structural homology to heat-shock protein 33, which acts as a chaperonin by inhibiting aggregation of partially denatured proteins (14). Mth1598 also has structural homology at the architecture level to S8 ribosomal protein (PDB ID code IFKA-H).

One of the goals of this research is to use structural homology to help reveal functional clues to previously unannotated proteins. Of the 12 structures solved here, nine proteins have significant structural similarity to a protein of known biochemical function. For the four proteins with initiation factor-like folds mentioned above, a possible RNA-binding function may be suspected. However, none of the four structures determined here had the key surface features of their structurally homologous RNA-binding domains, namely a large basic surface combined with several surface-exposed aromatic residues. Interestingly, in the case of Myx156r, the structural similarity to the predicted structure of the eIF2α N-terminal domain combined with the positions of conserved surface residues suggests that Myx156r and its viral homologues may be structural mimics of eIF2α and compete with eIF2α for regulators such as IFN-induced protein kinase PKR. Mth1692 has sequence and structural homology to yrdC_ecoli, which was shown to have high affinity for double-stranded RNA (15). The structure of Mth1692 shows a large positively charged cavity, also found in yrdC_ecoli, that is suspected to be the RNA-binding site.

Mth0865 and Mth1743 are members of COG that are distantly related to the thioredoxin and the ThiS/Ubiquitin families of proteins, respectively. Although the level of sequence identity with thioredoxin or ThiS is insignificant, Mth0865 and Mth1743

indeed have folds similar to these two proteins, respectively, and likely have similar but not identical biochemical activities. The putative function of the hypothetical protein Mth1880 was deduced not from 3D structure but from sequence homology. The second helix of Mth1880 has sequence homology to a helical region in syntaxin 1A involved in the calcium-dependent binding of syntaxin 1A to synaptotagmin (16). Mth1880 has been shown to bind calcium (G.M.L. and W.L., unpublished data), although at this point the binding partner of Mth1880 is unknown. The scaffolds that hold the homologous helices of syntaxin 1A and Mth1880 are very different even though the sequence homology between the two proteins extends beyond the active helix.

Conclusions

Here we have demonstrated the feasibility of an NMR approach to structural proteomics of small proteins. Our strategy relies on a centralized site for protein preparation and initial characterization followed by a distributive mechanism for NMR data collection and analysis. During this year-long project we identified several key "bottlenecks" in the process which, if eliminated, would allow a much higher throughput of structures. First, we have confirmed protein insolubility and/or aggregation as a major hindrance in structural proteomics. Our generic purification protocol may need to be expanded to explore 2–3 additional buffer conditions to allow more proteins to be concentrated for structural analysis. This problem in sample preparation applies to both x-ray crystallography and NMR spectroscopy. The use of higher-sensitivity cryogenic NMR probes will allow data collection for less concentrated samples, which should also help diminish the aggregation problem.

A second bottleneck for NMR-based structural proteomics is the amount of time required for data collection, currently ≈3–4 weeks per protein. Cryogenic probes will certainly assist in reducing this bottleneck. The third bottleneck is the NMR data analysis and structure determination phase. Although it currently takes on the order of months to assign NMR resonances and determine a structure, progress is being made on several fronts. Automated resonance- and nuclear Overhauser effect-assignment programs such as AUTOASSIGN (18), ANSIG (19), TATAPRO (20), ARIA (21), NOAH (22), and SANE (23) hold great promise in reducing the amount of time in structure determination; however, all these algorithms rely on frequency lists that often need manual peak-picking editing of lists generated by automatic peak-picking routines. Therefore, developing a more robust peak-picking algorithm that can better distinguish noise and artifacts from real peaks is crucial. Such a program combined with better data, both in terms of resolution and signal-to-noise, will enhance the reliability of automated and semiautomated procedures greatly. The use of programs such as RFAC (24) and MADIGRAS (25, 26) to back calculate the nuclear Overhauser effect spectroscopy spectra for comparison with the experimental spectra will reduce the need for manual validation of the calculated structure by measuring an R factor akin to that used in x-ray crystallography. Similarly, incorporation of residual dipolar coupling data (27, 28) into structure calculations also will facilitate more rapid and more accurate structure determination, although the generation of such data requires additional effort in sample preparation to identify appropriate orienting media.

Thus, we believe that a feasible and economical strategy for NMR-based structural proteomics would consist of two key components. The first is one or more centralized sample-preparation facilities that can ridentify the most appropriate sample conditions for modern NMR analysis rapidly and economically (see for example www.uhnres.utoronto.ca/proteomics and www.nesg.org). The second component is a consortium of NMR laboratories to which previously validated isotope-labeled samples would be sent for rapid acquisition of high signal-to-noise NMR experiments appropriate for automated or semiautomated data analysis.

We thank Dr. Grant McFadden for a gift of Myx genomic DNA, and Anna Kachatryan, Akil Dharamsi, Jun Gu, Alexei Savchenko, and Dinesh Christendat for excellent technical assistance and PCR cloning. This research was supported by the Ontario Research and Development Challenge Fund (A.M.E.), Canadian Institutes for Health Research (C.H.A, A.M.E, and K.G.), Protein Engineering Network of Centres of Excellence (L.P.M.), National Institutes of Health Structural Genomics Center Grant P50 GM62413-02 (C.H.A., A.M.E., and M.A.K.), and US Department of Energy contracts DE-AC06-76RL01830 (to M.A.K) and DE-FG06-92RL-12451 (to J.R.C.), Pacific Northwest National Laboratory Director's Research and Development funds (to M.A.K.), and Korea Science and Engineering Foundation through Protein Network Research Centre (W.L.). 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 and operated by Battelle. C.H.A., A.M.E., and M.A.K. are members of the Northeast Structural Genomics Consortium. A.M.E. and C.H.A. are Canadian Institutes for Health Research Scientists.

1. Smith, T. (2000) *Nat. Struct. Biol. Suppl.* **7,** 927.
2. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., *et al.* (2000) *Nat. Struct. Biol.* **7,** 903–909.
3. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995) *J. Biomol. NMR* **6,** 277–293.
4. Gronenborn, A. M. & Clore, G. M. (1996) *Protein Sci.* **5,** 174–177.
5. Serber, Z., Keatinge-Clay, A. T., Ledwidge, R., Kelly, A. E., Miller, S. M. & Dotsch, V. (2001) *J. Am. Chem. Soc.* **123,** 2446–2447.
6. Begley, T., Xi, J., Kinsland, C., Taylor, S. & McLafferty, F. (1999) *Curr. Opin. Chem. Biol.* **3,** 623–629.
7. Peitsch, M. C. (1996) *Biochem. Soc. Trans.* **24,** 274–279.
8. Peitsch, M. C., Schwede, T. & Guex, N. (2000) *Pharmacogenomics* **1,** 257–266.
9. Holm, L. & Sander, C. (1993) *J. Mol. Biol.* **233,** 123–138.
10. Holm, L. & Sander, C. (1996) *Science* **273,** 595–602.
11. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997) *Structure (London)* **5,** 1093–1108.
12. Pearl, F. M. G., Lee, D., Bray, J. E., Sillitoe, I., Todd, A. E., Harrison, A. P., Thornton, J. M. & Orengo, C. A. (2000) *Nucleic Acids Res.* **28,** 277–282.
13. Bycroft, M., Hubbard, T., Proctor, M., Freund, S. & Murzin, A. (1997) *Cell* **88,** 235–242.
14. Kim, S., Jeong, D., Chi, S., Lee, J. & Ryu, S. (2001) *Nat. Struct. Biol.* **8,** 459–466.
15. Teplova, M., Tereshko, V., Sanishvili, R., Joachimiak, A., Bushueva, T., Anderson, W. & Egli, M. (2000) *Protein Sci.* **9,** 2557–2566.
16. Fernandez, I., Ubach, J., Dulubova, I., Zhang, X., Sudhof, T. & Rizo, J. (1998) *Cell* **94,** 841–849.
17. Kraulis, P. (1991) *J. Appl. Crystallogr.* **24,** 946–950.
18. Zimmerman, D., Kulikowski, C., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. & Montelione, G. (1997) *J. Mol. Biol* **269,** 592–610.
19. Helgstrand, M., Kraulis, P., Allard, P. & Hard, T. (2000) *J. Biomol. NMR* **18,** 329–336.
20. Atreya, H., Sahu, S., Chary, K. & Govil, G. (2000) *J. Biomol. NMR* **17,** 125–136.
21. Nilges, M. & O'Donoghue, S. (1998) *Prog. Nucl. Magn. Reson. Spectrosc.* **32,** 107–139.
22. Mumenthaler, C. & Braun, W. (1995) *J. Mol. Biol.* **254,** 465–480.
23. Duggan, B. M., Legge, G. B., Dyson, J. & Wright, P. E. (2001) *J. Biomol. NMR* **19,** 321–329.
24. Gronwald, W., Kirchhofer, R., Gorler, A., Kremer, W., Ganslmeier, B., Neidig, K. & Kalbitzer, H. (2000) *J. Biomol. NMR* **17,** 137–151.
25. James, T. L. (1991) *Curr. Opin. Struct. Biol.* **1,** 1042–1053.
26. Borgias, B. A., Gochin, M., Kerwood, D. J. & James, T. L. (1990) *Prog. Nucl. Magn. Reson. Spectrosc.* **22,** 83–100.
27. Brunner, E. (2001) *Concepts Magn. Reson.* 13238–259.
28. Fowler, C., Tian, F., Al-Hashimi, H. & Prestegard, J. (2000) *J. Mol. Biol.* **304,** 447–460.
29. Gerstein, M. (1998) *Folding Des.* **3,** 497–512.