# Structural proteomics of an archaeon

Dinesh Christendat[1,2], Adelinda Yee[1,2], Akil Dharamsi[1,3], Yuval Kluger[4], Alexei Savchenko[1], John R. Cort[5], Valerie Booth[1], Cameron D. Mackereth[6], Vivian Saridakis[1], Irena Ekiel[7], Guennadi Kozlov[8], Karen L. Maxwell[9], Ning Wu[1], Lawrence P. McIntosh[6], Kalle Gehring[8], Michael A. Kennedy[5], Alan R. Davidson[9,10], Emil F. Pai[1,9,10], Mark Gerstein[4], Aled M. Edwards[1,11] and Cheryl H. Arrowsmith[1]

**A set of 424 nonmembrane proteins from *Methanobacterium thermoautotrophicum* were cloned, expressed and purified for structural studies. Of these, ~20% were found to be suitable candidates for X-ray crystallographic or NMR spectroscopic analysis without further optimization of conditions, providing an estimate of the number of the most accessible structural targets in the proteome. A retrospective analysis of the experimental behavior of these proteins suggested some simple relations between sequence and solubility, implying that data bases of protein properties will be useful in optimizing high throughput strategies. Of the first 10 structures determined, several provided clues to biochemical functions that were not detectable from sequence analysis, and in many cases these putative functions could be readily confirmed by biochemical methods. This demonstrates that structural proteomics is feasible and can play a central role in functional genomics.**

The completion and near completion of the sequencing phase of genome projects has ushered in the age of proteomics, the study of all gene products in an organism. This flood of sequence information coupled with recent advances in molecular and structural biology have led to the concept of 'structural proteomics' or 'structural genomics', the determination of three-dimensional protein structures on a genome-wide scale. An important use of three-dimensional structural information of proteins is to uncover clues as to a protein's function that are not detectable from sequence analysis[1,2]. This application of structural proteomics is driven by the realization that <30% of all predicted eukaryotic proteins have a known function. A related use of structural proteomics information is to determine a sufficient number of three-dimensional structures necessary to define a 'basic parts list' of protein folds[3,4]. Most other structures could then be modeled from this basis set using computational techniques[3,5]. The long term goal is to determine experimental structures for all proteins because it is the subtle differences in protein structure that contribute to the diversity and complexity of life, and current modeling techniques are not yet accurate enough to reveal these subtleties[6].

As reported in this manuscript, we initiated a prototype structural proteomics study of 424 nonmembrane proteins from the proteome of *Methanobacterium thermoautotrophicum* Δ*H* (*M.th.*). The primary goals of this research are to evaluate the technical hurdles involved in such a high throughput project, to estimate the percentage of proteins encoded by a genome that are immediately amenable to structure analysis, and to assess the extent to which function can be inferred from structure.

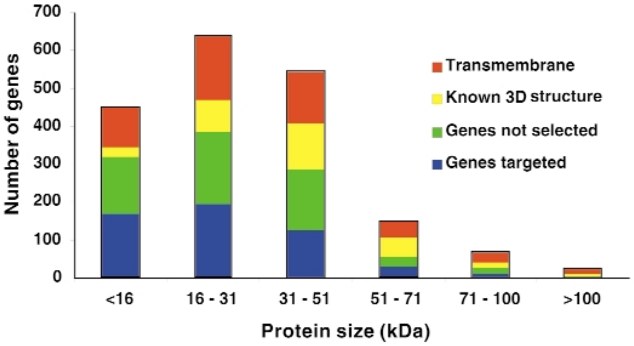

**Fig. 1** *M.th.* target ORFs. A histogram representing the numbers of different classes of *M.th.* ORFs according to predicted protein size showing unbiased sampling of nonmembrane proteins of unknown structure.

## Target selection

*M.th.* is a thermophilic archaeon whose genome comprises 1,871 open reading frames (ORFs)[7]. Archaeal proteins share many sequence and functional features with eukaryotic proteins, but are often smaller and more robust, and thus serve as excellent model systems for complex processes. Only two exclusionary criteria were implemented in our target selection scheme (Fig. 1). First, membrane associated proteins, which comprise ~30% (267–422 of 1,871 ORFs) of the *M.th.* proteome, were excluded. Although this class of proteins is of great biological significance, the science of membrane protein structure determination has not yet progressed to the point at which one would consider high throughput approaches. Second, proteins that had clear homologs in the Protein Data Bank (PDB) were excluded (~27%

[1] Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto 610 University Ave, Toronto, Ontario, Canada M5G 2M9. [2] These two authors contributed equally to this work. [3] Present address: Integrative Proteomics Inc., Toronto, Ontario, Canada. [4] Department of Molecular Biophysics and Biochemistry and Computer Science, PO Box 208114, Yale University, New Haven, Connecticut 06520, USA. [5] Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratories, EMSL 2569 K8-98, Richland, Washington 99352, USA. [6] Departments of Biochemistry and Molecular Biology, Chemistry, and the Biotechnology Laboratory, 2146 Health Sciences Mall, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3. [7] Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Ave., Montreal, Quebec, Canada H4P 2R2. [8] Department of Biochemistry and Montreal Joint Centre for Structural Biology, McGill University, 3655 Promenade Sir William Osler, Montreal, Quebec, Canada, H3G 1Y6. [9] Department of Biochemistry and [10] Department of Molecular and Medical Genetics, University of Toronto, 1 Kings College Circle, Toronto, Ontario, Canada M5S 1A8. [11] Banting and Best Department of Medical Research, C.H. Best Institute, University of Toronto, 112 College St., Toronto, Ontario, Canada M5G 1L6.

Correspondence should be addressed to C.H.A. *email: carrow@oci.utoronto.ca* or A.M.E. *email: aled.edwards@utoronto.ca*

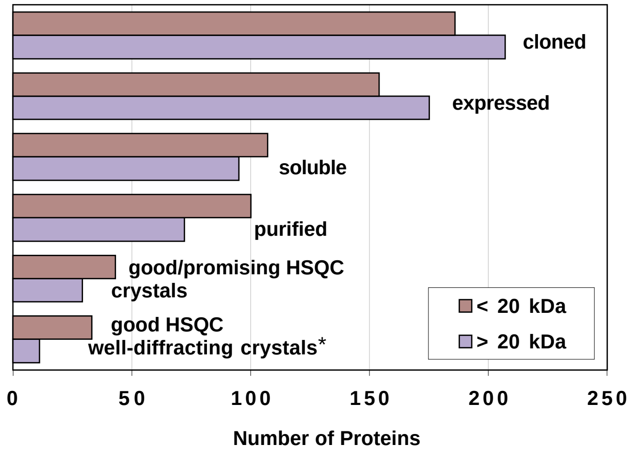

**Fig. 2** Histogram of the number of *M.th.* proteins at the end of each step of the cloning, expression, purification and sample screening process. *Note that not all proteins that crystallized in the initial screen were put through an optimization screen. Therefore the number of well-diffracting crystals is an underestimate of the total number that could readily be achieved. See text for further discussion.

of *M.th.* proteins). The remaining proteins (~900) were not prioritized based on their probability of having a new fold, nor in terms of 'biological relevance'. We chose to invest our effort in developing high throughput methods to generate a large collection of proteins to test as candidates for structural analysis, rather than concentrating on a small set of 'high priority' targets, a large proportion of which may not be amenable for immediate structural analysis. Thus, 424 of the 900 final target *M.th.* proteins (almost a quarter of the entire proteome and a third of the nonmembrane proteins) were chosen for cloning, expression and structural studies. These represent an unbiased sampling of nonmembrane proteins from a single proteome with 34% having a functional annotation, 54% classified as 'conserved' and 12% as 'unknown'. This diverse collection of proteins is also particularly valuable for retrospective bioinformatics surveys aimed at identifying sequence features that are predictive of protein biophysical behavior.

## Cloning and expression strategy

*M.th.* ORFs were cloned into an *Escherichia coli* expression vector containing an N-terminal hexa-His tag followed by a thrombin cleavage site. In the interest of throughput, no other expression vectors or organisms were used. A single PCR protocol and set of cloning conditions were optimized for *M.th.* based on an initial set of 50 genes. Positive clones were confirmed by colony PCR screening using Taq DNA polymerase. The generic nature of the procedure resulted in some PCR and subcloning failures, leading to a cumulative attrition rate of ~6%. At the end of the optimization process, the throughput (without automation) was 20 clones per person every three days. This protocol is readily scalable to 96-well format and can be extended to alternative vectors and expression organisms.

The *M.th.* ORFs were divided arbitrarily into two groups, 'large' (>20 kDa monomer size) and 'small' (<20 kDa monomer size). Large proteins were processed for crystallization trials and small proteins for NMR feasibility studies. Most (~80%) successfully cloned *M.th.* proteins could be expressed in *E. coli* BL21-Gold (DE3) cells (Stratagene), although efficient expression often required the presence of a second plasmid encoding three tRNAs that are frequently used by archeons and eukaryotes but are rare in *E. coli*. While most proteins could be expressed to reasonable levels, many were not expressed in soluble form (<0.5 mg $l^{-1}$ soluble protein), especially in the case of the larger proteins (Fig. 2). It may be possible to reduce the attrition rate due to poor solubility by optimizing the expression conditions for each clone. However, in the interests of throughput we used a single set of growth conditions optimized for the majority of *M.th.* proteins.

## Preparation and screening of structural samples

Large proteins were purified for crystal trials using a commercial screen as described in Methods. For each protein that crystallized in the initial screen, conditions were further optimized using an expansion of related solution conditions (typically 18–20 screens of 24 conditions for each protein). This process of optimizing the crystallization conditions proved to be one of the most labor intensive steps in the entire project. For every person-hour (h) spent in crystallization screens, more than 10 h were spent in implementing the expanded crystallization trials. Consequently, we chose 24 of the proteins that formed crystals in the primary screen to follow up with optimization screens. Of these, 11 formed well-diffracting crystals (<3.0 Å). The implementation of automated methods for setting up and monitoring crystal screens will greatly improve this process.

The smaller proteins (<20 kDa predicted monomer size) destined for NMR analysis were directly expressed in $^{15}$N-labeled media, purified and the $^{15}$N-HSQC NMR spectrum taken at 25 °C at 500 or 600 MHz. The HSQC spectra were classified into one of three categories. The first, termed 'excellent' and indicative of soluble, globular proteins, contained the predicted number of dispersed peaks of roughly equal intensity. These excellent spectra suggested that the process of determining their three-dimensional structures should be relatively straightforward. The second type of spectrum, termed 'promising', had features such as too few or too many peaks and/or broad but dispersed signals. This suggested that optimization of either the protein construct or the solution conditions would be needed to yield an excellent sample. The last category, termed 'poor', comprised two kinds of spectra. The first, which had intense peaks but with little dispersion in the $^{15}$N dimension, most likely reflected proteins that were soluble yet largely unfolded. The second class had very low signal-to-noise and/or a single cluster of very broad peaks in the center of the spectrum. This class probably represents proteins that aggregate nonspecifically at concentrations required for NMR spectroscopy and thus are not readily amenable to structural analysis. For the 100 soluble proteins tested, the ratio of excellent/promising/poor spectra was 33/10/57, respectively.

Of the 33 proteins showing excellent spectra, seven were initially chosen for more detailed structure determination using NMR spectroscopy. For these samples the hexa-His tag was removed by proteolytic cleavage; this did not markedly change the spectral properties of the proteins, suggesting that this step may be omitted in the interest of saving time and maximizing protein yield. In one case (MTH40) it was necessary to further optimize solution conditions in order to prepare a sample that was stable for the time period (several weeks) necessary for NMR data collection.

## Analysis of protein folding and stability

To explore how other spectroscopic techniques might aid in the identification of proteins suitable for detailed structural analysis, circular dichroism (CD) spectroscopy was performed on 100 of the small, soluble *M.th.* proteins. Of the 28 proteins with excellent

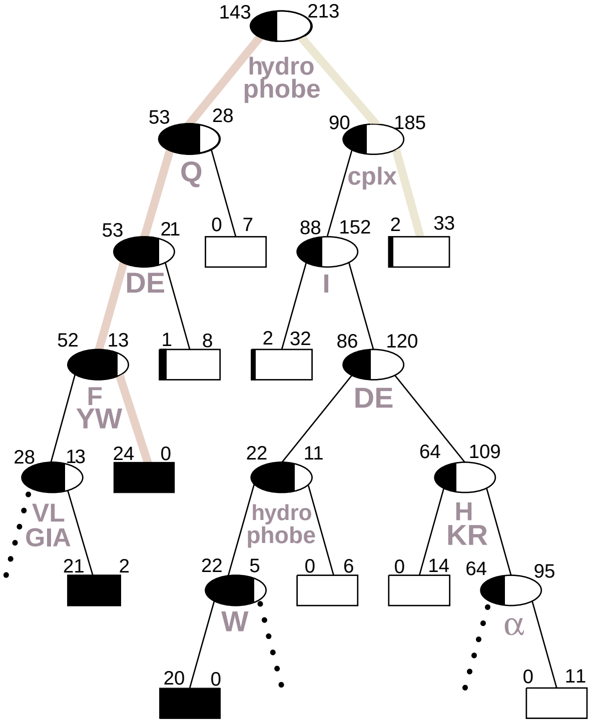

**Fig. 3** A decision tree discriminating between soluble and insoluble proteins. Ellipses represent intermediate nodes, and rectangles final nodes (leaves). The numbers of soluble and insoluble proteins are indicated above each node (right and left, respectively). The fraction of insoluble proteins is proportional to the shaded area. The highlighted decision pathways (in color) terminate in highly homogeneous nodes. Definitions of variables are in the Methods.

NMR spectra that were examined, all but six displayed CD spectra that were typical of folded proteins containing a significant fraction of α-helical and/or β-sheet secondary structure. The six atypical spectra may have resulted from unusual structural features of the proteins in question (for example, small β-sheet proteins like SH3 domains possess very unusual CD spectra). Interestingly, 24 out of 32 proteins classified as 'aggregated' by NMR spectroscopy displayed CD spectra consistent with stable, folded proteins. This suggests that the aggregation mechanism for many of the NMR samples may be due to surface interactions in the folded state, as opposed to aggregation of the exposed hydrophobic cores of unfolded proteins. Knowledge of the aggregation mechanism will be useful for optimizing solution conditions that disfavor aggregation, and, therefore, CD may provide a useful secondary screen in structural proteomics projects.

To better understand the contribution of protein stability to sample behavior, the thermal unfolding of 60 folded *M.th.* proteins was analyzed. Of these, 22 could be unfolded and refolded in a fully reversible manner. However, among the 19 proteins with 'excellent' NMR spectra that were tested in this manner, only nine refolded reversibly. The others precipitated at high temperatures, demonstrating that even among well-folded, small, soluble proteins, reversible thermal unfolding *in vitro* is not a ubiquitous property. Surprisingly, eight proteins classified as 'aggregated' by NMR were well behaved in thermal unfolding experiments, indicating that these proteins are probably large discrete oligomers rather than nonspecific aggregates.

As expected for proteins from a thermophilic organism, those from *M.th.* all possessed high thermostability with transition midpoint temperature ($T_m$) values between 68 and 98 °C. Due to their low change in heat capacity ($\Delta C_p$) upon unfolding, small proteins are generally expected to have higher $T_m$ values compared to larger proteins[8]. Here, however, we observed no correlation between the length of the *M.th.* proteins and their $T_m$ values. The $\Delta C_p$ values of small *M.th.* proteins were within the expected range compared to a large number of other proteins that have been investigated[9]. These data suggest that except for their high thermal stability, the overall thermodynamic behavior of *M.th.* proteins studied here may be representative of other mesophilic organisms.

## Retrospective analysis of biophysical data

Our studies revealed that poor expression and solubility accounts for almost 60% of the recalcitrant proteins. To identify the parameters that contribute to this poor sample behavior (and other factors related to suitability for NMR spectroscopy and crystallography), we adopted a retrospective data mining approach. We tried a number of data mining schemes (such as decision trees, Bayesian classifiers, logistic regression, and support vector machines) and found that most had comparable prediction performance, overall. However, for certain subsets of the data, decision trees often performed better. We therefore focussed on this method as it was the most useful for comprehensibly summarizing multivariate data and developing simple prediction rules. Growing the trees requires devising strategies regarding which variables (or combination of variables) to divide on, and what threshold to use to achieve the split. The 53 'splitting variables' we used were derived from attributes of each sequence (for example, amino acid composition, similarity to other proteins, measures of hydrophobicity, regions of low sequence complexity).

The full tree classifying the proteins according to their solubility (yes/no) had 35 final nodes and 65% overall accuracy in cross-validated tests. However, a number of the rules encoded within the tree were of better predictive value (these are highlighted in Fig. 3). For example, proteins that fulfill the following sequence of four conditions are likely to be insoluble: (i) have a hydrophobic stretch — a long region (>20 residues) with average GES-scale hydrophobicity < -0.85 kcal mole$^{-1}$; (ii) Gln composition <4%; (iii) Asp + Glu composition <17%; and (iv) aromatic composition >7.5%. This rule has a 14% error rate in comparison to the default error rate of 39% for choosing a soluble protein without the aid of the tree. The probability that it could arise by chance is 1%, assuming one randomly chose the 24 insoluble proteins from the initial pool of 143 insoluble and 213 soluble proteins. These calculations are based on a 'pessimistic estimate for errors'[10], taking the upper bound of the 95% confidence interval (see Fig. 3 for details). Conversely, proteins that do not have a hydrophobic stretch and have more than 27% of their residues in (hydrophilic) 'low complexity' regions are likely to be soluble. This rule has a 'pessimistic' error rate of 20% in contrast to 39% without the tree and a 1% probability of occurring by chance.

We also derived similar trees for expressibility and crystallizability (available from http://bioinfo.mbb.yale.edu/labdb/datamine). The statistics for these were less reliable due to their smaller size. However, we did find that composition of Asn appeared to be relevant to crystallizability. In particular, an Asn threshold of 3.5% was able to select a set of 18 crystallizable and only one noncrystallizable protein from our initial set of 25 crystallizable and 39 noncrystallizable proteins.

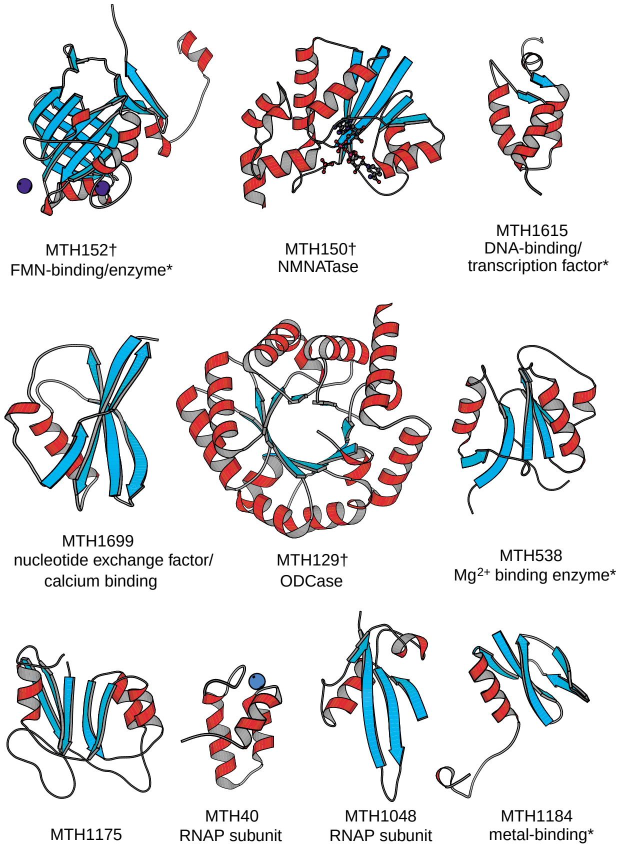

**Fig. 4** Backbone ribbon representations of the first 10 protein structures. β-sheets are shown in cyan and α-helices in red. Bound cofactors and ligands are shown as ball-and-stick models and metal ions as spheres. The *M.th.* gene number is given along with the confirmed and/or putative (asterisks) biochemical function of the protein. †Note that MTH150 is a homohexamer, and MTH152 and MTH129 are homodimers, although only a single subunit is displayed here.

Together these data suggest that, given a large enough data set, it may be possible to derive sets of 'rules' from primary sequence that are predictive of a given protein's biophysical properties.

## Structural and functional analysis

Proteins that formed well diffracting (<3 Å resolution) crystals were prepared with selenomethionine (SeMet) incorporation and their structures solved using multiwavelength anomalous dispersion (MAD) phasing[11]. Small proteins that gave excellent $^{15}$N-HSQC NMR spectra were prepared with uniform $^{13}$C,$^{15}$N incorporation and their solution structures solved using multinuclear, multidimensional NMR spectroscopy. Here we report the first 10 structures that were determined as part of this project (Fig. 4). At the start of this project, these 10 targets had no sequence homologs with a known three-dimensional structure. After determining their structures, we found that none had a completely new fold as determined using the FSSP classification and DALI server or the SCOP classification with automatic alignment[12–14]. Although this set is limited in number, the observation that some degree of similarity to members of the known structural data base exists for each *M.th.* protein indicates that the discovery of entirely new folds will be a relatively rare event. However, this argues favorably for the goal of determining (directly or subsequently) function from structure, as these structural similarities yielded a spectrum of clues to biochemical function that were not evident from sequence alone. In several cases the structures provided an atomic level model for interpretation of existing functional data, in others the structures and the presence of ligands generated a hypothesis for biochemical function that could be readily tested experimentally. Here we summarize the information that was gleaned from the structural data in order to illustrate the type of information that will flow from a larger structural genomics project. Due to space limitations experimental details of the structure determinations and full functional descriptions will be reported elsewhere.

Five of the 10 structures either contained a bound ligand (providing an immediate function) or a ligand binding site that could be inferred from structural homology. MTH150 was originally annotated as 'conserved', being highly homologous to a family of archaeal proteins of unknown function. This protein copurified and cocrystallized with NAD$^+$, immediately revealing at least one biochemical function. MTH150 has a nucleotide binding fold and structural similarity to a number of nucleotidyltransferases. Furthermore, MTH150 contains an HXGH motif that is similarly positioned to the acitive site HXGH motif found in these enzymes, suggesting a similar activity. A literature search for adenylyl transferases revealed that subsequent to our structure determination, a sequence homolog from *M. jannaschii* was reported to catalyze the condensation of nicotinaminde mononucleotide with ATP[15], suggesting that the MTH150 cocrystal contains the product-bound form of the *M.th.* enzyme. Additional biochemical studies have confirmed that MTH150 indeed has nicotinamide mononucleotide adenylyltransferase (NMNATase) activity.

MTH152 also shares sequence homology with several other archaeal proteins of unknown function. The purified protein was yellow, indicative of a flavin-like ligand bound to the protein, and crystallization required the presence of $Ni^{+2}$. Anomalous dispersion from the $Ni^{2+}$ ions and MAD phasing was used to solve the structure of this cocrystal. The structure showed that $Ni^{2+}$ is octahedrally coordinated with both the protein and the phosphate moiety of bound flavin mononucleotide (FMN), suggesting that $Ni^{2+}$ may play an integral role in cofactor binding and/or participate in a catalytic mechanism.

Although MTH538 is annotated as 'unknown', its NMR solution structure[16] uncovered a strong structural similarity with two protein classes, flavodoxin and the CheY family of bacterial response regulator proteins and domains. NMR chemical shift perturbation studies of MTH538 with either FMN or $F_{420}$, a related flavin-like compound found in methanogens, showed no evidence for binding of these cofactors. In contrast, titration of MTH538 with $Mg^{2+}$, a cofactor required for phosphorylation of CheY and related proteins, caused specific chemical shift perturbations for those residues that are predicted be affected by $Mg^{2+}$ based on structural homology[17]. However, MTH538 lacks a critical Asp, which is phosphorylated in CheY, and, unlike CheY, is not affected by treatment with acetyl phosphate. However, the substantial structural similarity between these proteins and the phosphorylation independent receiver domain AmiC, together with a small level of sequence similarity between MTH538 and a family of putative ATPases or kinases (COG1618), suggests MTH538 may have a role in a phosphorylation independent two component system, such as the AmiR-AmiC system[18].

MTH129 is a known ortholog of orotidine 5′ monophosphate decarboxylase, which catalyzes the exchange of $CO_2$ for a proton at the $C_6$ position of uridine 5′ monophosphate (UMP) at a rate 17 orders of magnitude faster than that of the uncatalyzed reaction in aqueous solution. Prior to this project, neither the structure nor the catalytic mechanism of any member of this enzyme family was known. The three-dimensional crystal structure of MTH129 in both the free and inhibitor bound forms allowed a detailed understanding of the remarkable catalytic power of this enzyme[19].

The NMR solution structure of MTH40 (ref. 20), which is homologous to the essential RPB10 subunit of RNA polymerase II, revealed a novel Zn binding motif that we term a Zn bundle. The protein folds as a three-helix bundle stabilized by a metal ion coordinated by a highly conserved but atypical $CX_2CX_nCC$ sequence motif (where X is any naturally occurring amino acids). This represents the first example of two adjacent zinc binding Cys residues within an α-helix, thus expanding the data base of known metal binding motifs. Based on the pattern of conserved and charged surface residues, as well as structural similarity to the N-terminal Zn binding domains of HIV-1 and HIV-2 integrases, insights were gained into the potential role of RPB10 as a scaffold protein within the multisubunit polymerase. These insights were confirmed when the NMR derived structure of *M.th.* RPB10 was used to help interpret the electron density map of the homologous subunit in yeast RNA polymerase II[21]. This demonstrates that generating high quality structures of individual proteins or domains in the context of structural genomics projects can facilitate the structure determination of larger biomolecular complexes.

Similarly, MTH1048 is homologous to the RPB5 subunit of RNA polymerase II. The role of this subunit within polymerase is poorly understood. The solution NMR structure of this protein[22] revealed a distinctive 'mushroom'-like shape with one half of the molecular surface composed of conserved hydrophobic residues suggestive of a site for protein–protein interactions. In contrast the opposite surface contains fewer conserved residues and a high density of charged residues suggestive of a region that is either solvent exposed or possibly one that interacts with nucleic acids. This interpretation was also confirmed by examination of the crystal structure of yeast RNA polymerase II, in which the 'stem' of the mushroom is buried within RPB1, and the 'cap' of the mushroom is more exposed[21].

The case of MTH1615 is particularly illustrative of both the reiterative uses of NMR and other biophysical methods in determining protein structures and in how knowledge of a structure can lead to simple experiments that provide immediate insights into function. MTH1615 is a member of a gene family whose human ortholog was identified in a screen for genes involved in apoptosis[23]. The NMR spectrum was promising, but not immediately amenable to rapid structure determination; the initial HSQC of this protein had many peaks in the central random coil region, indicative of an unfolded polypeptide, as well as a large number of dispersed peaks indicative of a folded structure. Using limited proteolysis followed by mass spectrometry and NMR analysis, we found that the N-terminal 31 residues of MTH1615 were unstructured in solution. A smaller construct lacking the first 31 residues was then prepared, and the structure of this protease resistant domain determined using NMR spectroscopy. From this structure we modeled that of the human protein, revealing a conserved basic cleft. Since the human protein has been shown to localize to the nucleus, we also tested MTH1615 for DNA binding activity. Using electrophoretic mobility shift assays (EMSA), we found that MTH1615 can interact nonspecifically with a randomly chosen 20-mer of double stranded DNA, suggesting that the human protein may be involved in nucleic acid binding or metabolism. It is important to note that it was difficult to identify the unstructured N-terminal region from sequence analysis. This region is strongly predicted to be an α-helix using PHD[24] and the alignment of seven highly conserved orthologs from *M.th.* to human.

MTH1699 was identified from sequence analysis as an archaebacterial translation elongation factor 1 β (aEF-1β), which acts as a guanine nucleotide exchange factor. It has an α/β sandwich fold typical of many RNA binding proteins. While this structure determination was underway[25] the structure of human EF-1β[26] was published, revealing structural homology to the functionally similar eubacterial EF-Ts protein. These three structures provide a dramatic example of functional and structural conservation that is not evident from sequence. In all three kingdoms nucleotide exchange occurs in an identical manner, via a conserved Phe or Tyr on an exposed loop. However, unlike either hEF-1β or EF-Ts, MTH1699 was found to bind calcium. This novel feature may play a functional role in archaeal protein translation or may simply serve to increase the protein's thermal stability.

MTH1184 is annotated as 'unknown', and consists of a β-sheet region followed by an α-helix and an unstructured C-terminus. The β-sheet region contains a CXCX...XCXC sequence with Cys residues located in two proximal loops and pointing towards each other. While this motif is potentially capable of metal binding, we were unable to detect zinc binding to the protein, suggesting specificity for another metal.

MTH1175 is a member of an uncharacterized COG (COG1433) that is predominantly represented by archaea. Therefore, from sequence homology it is currently impossible to gain insight into its function. SCOP analysis of MTH1175 reveals that it is most similar to structures within the ribonuclease H superfamily. While, MTH1175 lacks the key catalytic residues of RNase H, an RNA binding capability is suggested by a Gly and Arg-rich region in the flexible C-terminus. However the biochemical function of this protein remains to be determined.

## Conclusions

This study of the proteome of *M.th.* demonstrates that structural proteomics is a feasible concept, and that NMR spectroscopy can play a significant role. However, we note that the rate of structure determinations for this project was limited by access to expensive instrumentation, such as NMR spectrometers and synchrotron radiation sources. This, therefore, should be one of the major issues to be addressed in the pending large scale, internationally coordinated structural genomics/proteomics projects. Our data also indicate that considerable effort must be invested in improving the attrition rate due to proteins with poor expression levels and unfavorable biophysical properties. Our retrospective analysis of the biophysical behavior of *M.th.* proteins suggests that the creation and mining of an empirical data base of biophysical properties may allow investigators to predict which proteins will (or will not) be amenable to structural analysis under a given set of conditions. The structural results suggest that the discovery of completely new protein folds may not be a common occurrence, but that considerable biochemical insights can be gained either from the structures themselves or from subsequent biochemical experiments suggested by the structures.

## Methods

**Target selection.** Putative membrane proteins were identified using the program TMHMM 1.0 (www.cbs.dtu.dk)[27]. Proteins with sequence homologs in the PDB were identified using a BLAST search[28] against the PDB with an e-value cutoff of $10^{-4}$.

**Sample preparation.** All cloning and initial expression, purification and HSQC/crystal screens were performed at the Ontario Cancer Institute (OCI) over a 12 month period by A.D., D.C. & A.Y.

with the help of one FTE technician and (for 4 months) six summer students. In some cases clones were sent to individual labs where they were expressed with the appropriate isotopes or SeMet labels, data acquired and structures solved (MTH129 by N.W.; MTH1184 and MTH1699 by G.K.and I.E.; MT0040 by C.M.). In the remaining cases the NMR sample or SeMet crystal was prepared at the OCI and the NMR or diffraction data collected and structures solved by individual labs (MTH538 and MTH1175 by J.R.C. and M.A.K.; MTH150 by D.C.; MTH1615 by V.B.; MTH150 by V.S.; MTH1048 by A.Y.). CD data were collected and analyzed by A.S., K.L.M. and A.R.D.

Each target gene was PCR amplified from genomic DNA, with terminal incorporation of unique restriction sites, using high fidelity *Pfu* DNA polymerase (Stratagene). The PCR products were directionally cloned into the pET15b bacterial expression vector (NOVAGEN). A single PCR protocol and set of cloning conditions were optimized for *M.th.* based on an initial set of 50 genes. Positive clones were confirmed by colony PCR screening using Taq DNA polymerase.

For large proteins (>20 kDa per monomer), three colonies from each transformation were tested for protein expression on a small scale (50 ml). Proteins found to be soluble by SDS-PAGE analysis of the bacterial extract were prepared on a larger scale (2 l). These proteins were purified by a combination of heat treatment (55 °C) and nickel affinity chromatography, followed by thrombin cleavage and removal of the hexa-His tag. The heat treatment caused a significant enrichment of many, but not all, *M.th.* proteins. Proteins were judged to be 99% pure as judged by an overloaded coumassie blue stained SDS gel. Occasionally mass spectrometry was used to monitor the integrity of purified proteins. No evidence of frequent or systematic mutations was found. Proteins that 'survived' the purification process (~75%) were concentrated to 10 mg ml$^{-1}$ and subjected to a sparse-matrix crystallization screen of 48 conditions at room temperature (Msatrix screen 1; Hampton Research).

Proteins for initial NMR HSQC screens were expressed five at a time, each in 1 l of $^{15}$N-enriched minimal media and purified in parallel using metal affinity chromatography. The resulting $^{15}$N-labeled hexa-His fusion proteins were concentrated by ultrafiltration to ~5–20 mg ml$^{-1}$ and were typically ~90–95% pure as judged by coumassie blue stained SDS-PAGE. $^{15}$N,$^{13}$C-labeled proteins prepared for three-dimensional structure determination were further purified to >98% purity by either gel filtration or ion exchange chromatography after removal of the His tag.

**Decision tree analysis**. This analysis was carried out by Y.K. and M.B.G. Under each intermediate node, the decision tree algorithm calculates all possible splitting thresholds for each of 53 variables (hydrophobicity, amino acid composition, etc.). It picks the optimal splitting variable and its threshold, in order for at least one of the two daughter nodes to be as homogeneous as possible. When a variable, v, is split, v < threshold is the left branch, and v > threshold is the right branch. The specific parameters used at each node and their thresholds for the right branches shown in Fig. 3 are in descending order (from top root to bottom leaves): hydrophobe > 0.85 kcal mole$^{-1}$ (where 'hydrophobe' represents the average GES hydrophobicity of a sequence stretch; the higher this value, the lower the energy transfer); cplx > 0.28 (where 'cplx' is a measure of a local sequence complexity region based on the SEG program[10] ); Q > 4%; DE > 17%; I > 5.6%; FWY > 7.5%; DE > 13.6%; GAVLI > 42%; hydrophobe > 0.01 kcal mole$^{-1}$; HKR > 12%; W composition > 1.2%; and α-helical secondary structure composition > 58%. In the preceding pathway, Q represents Gln composition; DE, Asp + Glu composition, and other quantities are defined similarly. Note that two of the variables are conditioned on more than once (hydrophobe, Asp + Glu). The shorter the decision pathway and the larger the number of cases in the terminal node, the lower the risk of over-fitting the data. Heterogeneous leaves could be further split (dotted lines in Fig. 3) improving the error rate but risking overfitting of the training set. The predictive values of the pathways were evaluated using a 'pessimistic estimation' procedure that assumed that the error rate at each node is bionomially distributed, and then inflates the rate found on a tree based on all the data (by ~2 standard deviations) to arrive at a more realistic estimate[10,29]. Further details can be found at http://bioinfo.mbb.yale.edu/labdb/datamine.

**Coordinates.** The PDB accession codes for each structure (and the BioMagResBank accession numbers where applicable) are as follows: MTH150, 1ej2; MTH152, 1eje; MTH538, 1eiw (4793); MTH129, 1dv7; MTH40, 1ef4(4571); MTH1048, 1eik(4678); MTH1615, 1eij(4674); MTH1699, 1d5k(4385); MTH1184, 1dw7(4740); MTH1175, 1eo1(4796).

**Acknowledgments**

We are grateful to J. Reeve for providing *M.th. ΔH* chromosomal DNA. We thank technicians, A. Engel, S. Beasely, B. Le and summer students F. Hsu, A. Tuite, G. Minoo, S. Fung J. Loo, and H. Javidni for help with protein expression and purification. We thank L. Daniels for a gift of coenzyme $F_{420}$, and A. Ayed for performing the EMSA of MTH1615. We acknowledge funding support from the Canadian funding organizations, MRC/CIHR (C.H.A, A.M.E, V.B., K.G.), NSERC (C.D.M), PENCE (L.P.M., E.F.P.) and US DOE contracts (M.A.K., J.R.C.), NIH and the Keck Foundation (M.G.), PNNL Laboratory Director's Research and Development funds (M.A.K.) and a Sloan Foundation-DOE fellowship (Y.K.). Part of the NMR work was performed at EMSL (a national scientific users facility sponsored by DOE Biological and Environmental Research) located at PNNL and operated by Battelle. X-ray data were collected at the Advanced Photon Source supported by the U.S. DOE, Basic Energy Sciences, Office of Science, and BioCARS Sector 14 supported by the NIH.

Received 20 June, 2000; accepted 2 August, 2000.

1. Zarembinski, T.I., *et al.* Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl. Acad. Sci. USA* **95**, 15189–15193 (1998).
2. Montelione, G.T. & Anderson, S. Structural genomics: keystone for a human proteome project. *Nature Struct. Biol.* **6**, 11–12 (1999).
3. Gerstein, M. & Hegyi, H. Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277 (1998).
4. Sali, A. 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032 (1998).
5. Sanchez, R. & Sali, A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
6. Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K. & Pedersen, J.T. Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* **Suppl**, 2–6 (1997).
7. Smith, D.R., *et al.* Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155 (1997).
8. Alexander, P., Fahnestock, S., Lee, T., Orban, J. & Bryan, P. Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures. *Biochemistry* **31**, 3597–3603 (1992).
9. Myers, J.K., Pace, C.N. & Scholtz, J.M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148 (1995).
10. Quinlan, J.R. Decision trees and decision making. *IEEE Transaction on Systems, Man and Cybernetics* **20** (1992).
11. Hendrickson, W.A., Smith, J.L. & Sheriff, S. Direct phase determination based on anomalous scattering. *Methods Enzymol.* **115**, 41–55 (1985).
12. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
13. Gerstein, M. & Levitt, M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.* **7**, 445–456 (1998).
14. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
15. Raffaelli, N., *et al.* Identification of the archaeal NMN adenylytransferase gene. *Mol. Cell. Biochem.* **193**, 99–102 (1999).
16. Cort, J.R., Yee, A., Edwards, A.M., Arrowsmith, C.H. & Kennedy, M.A. Structure-based functional classification of hypothetical protein MT538 from *Methanobacterium thermoautotrophicum*. *J. Mol. Biol.* **in the press** (2000).
17. Feher, V.A. & Cavanagh, J. Millisecond-timescale motions contribute to the function of the bacterial response regulator protein Spo0F. *Nature* **400**, 289–293 (1999).
18. O'Hara, B.P., *et al.* Crystal structure and induction mechanism of AmiC-AmiR: a ligand-regulated transcription antitermination complex. *EMBO J.* **18**, 5175–5186 (1999).
19. 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. USA* **97,** 2017–2022 (2000).
20. 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. USA* **97**, 6316–6321 (2000).
21. Cramer, P., *et al.* Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* **288**, 640–649 (2000).
22. Yee, A., *et al.* Solution structure of the RNA polymerase subunit RPB5 from *Methanobacterium thermoautotrophicum*. *Proc. Natl. Acad. Sci. USA* **97**, 6311–6315 (2000).
23. Liu, H., *et al.* TFAR19, a novel apoptosis-related gene cloned from human leukemia cell line TF-1, could enhance apoptosis of some tumor cells induced by growth factor withdrawal. *Biochem. Biophys. Res. Commun.* **254**, 203–210 (1999).
24. Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599 (1993).
25. Kozlov, G., *et al.* Rapid fold and strcuture determination of the archaeal translation elongation factor 1b from *Methanobacterium thermoautotrophicum*. *J. Biomol. NMR* **17**, 187–194 (2000).
26. Perez, J.M. *et al.* The solution structure of the guanine nucleotide exchange domain of human elongation factor 1β reveals a striking resemblance to that of EF-Ts from *Escherichia coli*. *Structure Fold. Des.* **7**, 217–226 (1999).
27. Sonnhammer, E.L.L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (eds. Glasgow, J., *et al.*) 175–182 (AAAI Press, Menlo Park, California; 1998).
28. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Quinlan, J.R. *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, California; 1992).