Extremophiles (2008) 12:39–50
DOI 10.1007/s00792-007-0087-9

REVIEW

The impact of extremophiles on structural genomics (and vice versa)

Francis E. Jenney Jr · Michael W. W. Adams

Received: 22 December 2006 / Accepted: 19 April 2007 / Published online: 13 June 2007

Abstract The advent of the complete genome sequences of various organisms in the mid-1990s raised the issue of how one could determine the function of hypothetical proteins. While insight might be obtained from a 3D structure, the chances of being able to predict such a structure is limited for the deduced amino acid sequence of any uncharacterized gene. A template for modeling is required, but there was only a low probability of finding a protein closely-related in sequence with an available structure. Thus, in the late 1990s, an international effort known as structural genomics (SG) was initiated, its primary goal to ''fill sequence-structure space'' by determining the 3D structures of representatives of all known protein families. This was to be achieved mainly by X-ray crystallography and it was estimated that at least 5,000 new structures would be required. While the proteins (genes) for SG have subsequently been derived from hundreds of different organisms, extremophiles and particularly thermophiles have been specifically targeted due to the increased stability and ease of handling of their proteins, relative to those from mesophiles. This review summarizes the significant impact that extremophiles and proteins derived from them have had on SG projects worldwide. To what extent SG has influenced the field of extremophile research is also discussed.

Communicated by D.A. Cowan.

F. E. Jenney Jr · M. W. W. Adams (✉)
Department of Biochemistry and Molecular Biology,
University of Georgia, Davison Life Sciences Complex,
Green Street, Athens, GA 30602-7229, USA
e-mail: adams@bmb.uga.edu

Keywords Structural genomics · Open reading frame · Extremophiles · Thermophiles · Hyperthermophiles · X-ray crystallography · NMR spectroscopy

Abbreviations

ORF	Open reading frame
SG	Structural genomics

Introduction

The release of the first complete genome sequence of a microorganism in 1995 (*Haemophilus influenzae*) and of the first draft of the human genome 6 years later (Fleischmann et al. 1995; Venter et al. 2001) heralded the new age of genomics. Currently the complete genome sequences of almost 500 organisms are available, and the genome sequences of four times this number of organisms are in progress (http://www.genomesonline.org). One of the major revelations of this revolution has been the discovery of a large number of conserved/hypothetical genes, the function of which is essentially unknown. At least some insight into the function of the proteins that such genes encode might be available from their 3D structures, which could be predicted if the structure of a protein closely-related in sequence was available. However, in the late 1990s it was realized that for any uncharacterized protein (deduced amino acid sequence), the chance of finding a protein closely-related in sequence for which a crystal structure was available (to serve as a template for modeling) was very limited (Holm and Sander 1994, 1997; Orengo et al. 1999). Thus an international effort known as ''structural genomics'' (SG) was initiated to fill ''sequence-structure space''. This was to be achieved by

determining the 3D structures, mainly by X-ray crystallography, of representatives of all known protein families. It was anticipated that achieving this would require more than 5,000 new structures (Brenner 2000; Brenner et al. 1997; Burley 2000; Liu et al. 2004).

The SG concept was therefore a rational post-genomic goal to follow the success of the genome sequencing projects. In the United States the SG effort took the form of the Protein Structure Initiative spearheaded by the National Institutes of General Medical Sciences (NIGMS/NIH). It began in 2000 with the goal of developing high-throughput (HTP) protocols to increase the rate, and decrease the cost, of determining the 3D structures of proteins by X-ray crystallography (X-ray), and to a lesser extent by nuclear magnetic resonance (NMR) spectroscopy, in order to find representative structures of all possible protein folds in the biological world (Brenner and Levitt 2000; Gaasterland 1998). While the individual tasks of selecting genes to be expressed, obtaining the recombinant proteins in a stable purified form, crystallizing them, and determining their structures, were commonplace, developing HTP methods for all aspects, with the goal of generating thousands of structures from a single center, was a daunting challenge. In the first 5-year pilot phase of the NIH-funded project, nine large, multi-disciplinary groups were formed between multiple university, government, and industrial laboratories (for example see Bonanno et al. (2005)] to develop the necessary technologies (see http://www.nigms.nih.gov/Initiatives/PSI for a complete list). SG efforts were also initiated around the globe with essentially the same goals and with significant international coordination and communication (http://www.isgo.org).

Initially, most of the SG groups targeted the proteins (genes) encoded by the sequenced genomes of one or more model organisms, typically including both prokaryotic and eukaryotic representatives. It was at this early stage that extremophilic microorganisms had the most impact on the SG phenomenon. In particular, among the first organisms to be targeted were the thermophilic bacteria *Thermus thermophilus* and *Thermotoga maritima* (DiDonato et al. 2004; Ito et al. 2006), and thermophilic archaea *Methanobacterium thermoautotrophicum* [now *Methanothermobacter thermoautotrophicus* (Christendat et al. 2000)] *Pyrobaculum aerophilum* (Mallick et al. 2000), and *Pyrococcus furiosus* (Adams et al. 2003). It is important to note that one of the original SG initiatives began at RIKEN in Japan (Yokoyama 2005; Yokoyama et al. 2000) where the initial focus was exclusively on the extremophiles, *Thermus* and *Pyrococcus*. Proteins from thermophiles were prime targets in the initial stages of the SG projects mainly because of the anecdotal belief that such proteins crystallize more easily, and would therefore increase the overall success rates of transforming the gene of a hypothetical protein into a three dimensional protein structure. Of course, there is no doubt that proteins from thermophilic organisms are more stable (Sadeghi et al. 2006; Szilagyi and Zavodszky 2000) than those from mesophiles, and thus more easily purified, manipulated, and more stable during the long time periods needed for crystallization. However, whether they really do crystallize more readily is an open question (Rees 2001). In any event, the purpose of this paper is to briefly review the interaction between the world of SG and that of extremophiles, with a specific focus on protein targets from thermophilic organisms.

Structural genomics: from the early years to the present day

The immediate goals of the SG centers were to develop new technologies for bioinformatics analyses of multiple genomes, HTP gene cloning, protein expression, purification, crystallization, and structure determination protocols. Traditionally in the biochemistry field, for any one protein under investigation by a single group, it was often the case that the process from gene cloning to the structure of the encoded protein would take many years. In contrast, early estimates were as high as 15,000 for the number of protein structures that needed to be detemined in order to cover the perhaps 3,000 or more unique proteins folds that may be represented in living organisms (Gaasterland 1998) and this might increase as the database of available sequences increased (Marsden et al. 2006). Consequently, by conventional technologies during the 1990s, it would have taken as many as 50,000 person-years worth of work for full coverage of all possible protein fold families, at an astronomical cost. Thus the primary goal of the SG effort was to reduce this time and cost by many orders of magnitude, by automating and streamlining as many steps as possible in the process.

One critical issue for the SG efforts was target selection, both from the perspective of limiting overlap between groups targeting homologous proteins in different organisms, and to limit protein targets to those expected to yield novel folds. The targets initially chosen by particular SG groups typically represented the research interests of individual members. Some SG groups had a common theme, for example, the SG of Pathogenic Protozoa Consortium (http://www.sgpp.org) and the TB SG Consortium (TBSGC http://www.doe-mbi.ucla.edu/TB/) focused on pathogenic protozoa and the tuberculosis-causing bacterium *Mycobacterium tuberculosis*, respectively. Similarly, the Center for Eukaryotic Structural Genomics (CESG http://www.uwstructuralgenomics.org) focused on the plant model organism *Arabidopsis thaliana*. In addition to the NIH projects in the US and RIKEN in Japan, there were also a large number of projects in Europe, including the

Structural Proteomics in Europe (SPINE) project (Berry et al. 2006), as well as efforts in Germany (Banci et al. 2006), France (Abergel et al. 2003), and England (Cianci et al. 2005). Links to all relevant sites can be found at the International Structural Genomics Organization (http://www.isgo.org).

As noted above, however, a number of centers targeted proteins from thermophilic organisms, both from interest and from the perspective that their more stable proteins would be more tractable. For example, *P. aerophilum* was targeted by the TBSGC as an early validation test for target selection by assigning protein fold predictions to open reading frames (ORFs), in order to eliminate targets for which homologous structures may already exist (Mallick et al. 2000). Some of the earliest pioneering work, and the earliest success story, was with *Methanobacterium thermoautotrophicum* (Mt) carried out at the University of Toronto (Christendat et al. 2000). This study is particularly useful as it demonstrated all the strengths and weaknesses of the SG approach (*vide infra*). For example, one of the first problems that became evident is the high attrition rate as targets pass through the SG ''pipeline'' from gene to structure, and this aspect is discussed further below.

All SG groups soon realized the difficulties in obtaining recombinant forms of proteins in a HTP-mode. Consequently, many ORFs were automatically removed from target lists, such as those predicted to encode membrane proteins and others expected to be recalcitrant, such as very large proteins. Even so, success in recombinant protein production was much lower than anticipated. We and others (Adams et al. 2003) proposed that the high attrition rate could be due in part to proteins that are very unstable and perhaps rapidly degraded by the expression host. This instability was proposed to arise because of the lack of either simple (e.g., Fe or Zn) or complex (e.g., flavin) cofactors that are not properly ''inserted'' (and in some cases even synthesized) by the expression host, and/or to proteins which may only be stable when coexpressed with their partners to form a multiprotein complex (Adams et al. 2003). It was estimated that less than 20% of the ORFs in any genome would likely be expressed as a stable, properly folded protein, commonly named the ''low-hanging fruit''. This prediction is consistent with the data from this first comprehensive SG report [<10 structures out of 424 targets from Mt (Christendat et al. 2000)], and with data available from many groups [Protein Data Bank (PDB), http://www.rcsb.org/pdb/, 2004]. While cloning is virtually 100% successful even at the HTP level, the loss at each step from gene target to structure can be as high as 50%, and is not correlated with research group, protocols used, or ORF targets (Acton et al. 2005), even when the more difficult membrane proteins have been removed from the target list (Bonanno et al. 2005).

Over the 5 years of the first phase of the NIH-funded SG effort, a number of key advances were made in bioinformatics (Bravo and Aloy 2006; Ginalski et al. 2005; Wolfson et al. 2005), recombinant protein production (Dieckman et al. 2006; Esposito and Chatterjee 2006; Hart and Tarendeau 2006; Marsischky and LaBaer 2004; Zhou and Chen 2004), and structure determination techniques (Arzt et al. 2005; Atreya and Szyperski 2005; McPherson 2004; Pusey et al. 2005; Wang et al. 2005). The results from the product-driven SG centers were recently compared and contrasted with those of conventional hypothesis-driven laboratories of individual investigators carrying out traditional (non HTP) approaches (Chandonia and Brenner 2006). An attempt was made to analyze quantitatively the cost and impact of protein structure determination between the two types of groups. It was found that about half of all novel protein structures are now solved at SG centers, and very significantly, the cost of solving a structure at these centers has been reduced to 25% of the estimated cost at a traditional laboratory (Chandonia and Brenner 2006). Nonetheless, while traditional hypothesis-driven structure laboratories may work on more difficult targets (such as protein complexes), their efficiency is similar to the HTP SG centers. In addition, publications from the traditional non-SG laboratories are cited more frequently, indicating that structures from the HTP SG laboratories are having a significantly lower impact. In fact, as discussed below, one of the major limitations of protein structures from the SG laboratories is that they are only deposited electronically and are not accompanied by a publication describing the biological consequences of the new structure.

The second and so-called production phase of the NIH-funded SG initiative in the US began in 2005 and this time two types of centers were created (Service 2005). They included large-scale centers dedicated to HTP protein production of novel targets, targeting orthologs of a particular protein from multiple species as at least one ortholog will typically be successful (Savchenko et al. 2003), and smaller, specialty centers dedicated to research on the more difficult problems, such as membrane protein and multiprotein complex expression (http://www.nigms.nih.gov/Initiatives/PSI), the so-called ''high-hanging fruit''. There are currently four large production centers, and their common theme is HTP production to meet the project goal of 4,000 structures from different protein families that currently have no representative structure. In all cases technology development is a major focus, as well as dissemination of these technologies. The primary focus of the 'Big4' centers is to efficiently cover structure space. There has been a major philosophical shift from the first 5 years of the NIH initiative, as the goal is now to reach maximum efficiency of protein structure production

and maximum coverage of sequence-structure space (http://www.psi-big4.org/index.php). As of December 2006, the 'Big4' have already produced almost 80% as many structures (671 vs. 854) in the first year of the second SG phase as they did in the 5 years of the first initiative. The six smaller, specialized centers will take the technologies developed in the first phase of the NIH-funded SG initiative and use these to develop more HTP protocols for the much more difficult proteins. While proteins from extremophiles and particularly thermophiles enjoyed premiere status in the first 5 years of the SG initiative, there is no longer any emphasis on these proteins. Their genes are still selected, but only as one of a large group of orthologous sequences for a particular target protein of interest, and not because they are like to generate highly stable proteins that are more amenable to crystallization.

Consequently, the protein (gene) target list has expanded dramatically in the recently initiated second phase of SG and contains many new species as sources of new orthologs of protein targets. There are currently over 600 species and strains from all three kingdoms of life and viruses represented with, in some cases, thousands of genes. While this incorporates genes from dozens of extremophiles including psychrophiles, (hyper) thermophiles, halophiles, acidophiles, etc. (see http://targetdb.pdb.org for a complete list of registered targets), these organisms no longer receive any ''special'' status and have no higher or lower priority than the genes from any other organism. In the second phase of SG, organisms per se are not the issue, protein families are. Nevertheless, as noted above, a number of the extremophiles were specifically targeted in the first 5 years of the SG initiative, and these projects have generated an enormous amount of information on their ''extremophilic'' proteins.

Targeted genomes from extremophiles

Methanobacterium thermoautotrophicum

M. thermoautotrophicum (Mt, now *Methanothermobacter thermoautotrophicus*) is a thermophilic (T_{opt} 65°C) lithoautotroph isolated from sewer sludge (Zeikus and Wolfe 1972) which uses energy from H_2 to reduce CO_2 to CH_4. Its 1.75 Mbp genome contains approximately 1,871 ORFs (Smith et al. 1997). This organism was initially the flagship organism of the SG world as a collection of ORFs from Mt was the first real test of the SG protocol (Christendat et al. 2000). Out of the 1,871 ORFs in the genome, 424 were selected (none were predicted to encode membrane proteins). Approximately 80% of these yielded protein when they were expressed in *E. coli*, but only about half gave rise to soluble protein. A total of 175 proteins were purified, and about half gave promising results in initial NMR and crystallization screens. Of the ones that formed crystals, 24 were chosen for optimization, and ten structures were solved (Christendat et al. 2000). In this first SG test case, the inherent problem of target attrition in the gene to protein structure pipeline was evident, but this project also demonstrated that the structure determination of proteins of unknown function could, at least in some cases, give strong indications as to their in vivo function. Of the ten structures reported in this work, five co-crystallized with ligands. For example, the structure of the uncharacterized protein MTH150 showed that NAD was bound to it. The structure also indicated a nucleotide binding fold, and biochemical assays demonstrated that the protein has nicotinamide mononucleotide adenylyltransferase activity (Saridakis et al. 2001). Other structures were for proteins of known function for which there was no existing structure. These included MTH129, which is an orotodine 5′ monophosphate decarboxylase, and the NMR structure of MTH40 which is homologous to a subunit of RNA polymerase II revealed a novel Zn-binding motif (Christendat et al. 2000). Mt proteins continue to be targets of various SG centers and to yield novel structural information (Lee et al. 2004).

Thermotoga maritima

T. maritima (Tm) is a hyperthermophilic heterotrophic bacterium (T_{opt} 80°C) isolated from a hot marine sediment in Vulcano, Italy (Huber et al. 1986). It ferments sugars and produces H_2, making it of particular interest for the production of biofuels. The genome (1.86 Mbp) is predicted to contain 1,877 ORFs (Nelson et al. 1999), and it was also an early target of one of the most successful SG projects (Lesley et al. 2002) at the Joint Center for Structural Genomics (http://www.jcsg.org). The JSCG alone has deposited 162 structures of Tm proteins in the PDB (http://www.rcsb.org/pdb/, 2004). Out of 21 novel protein folds discovered by this SG group to date, 15 are in Tm proteins. Overall, various SG centers have produced 220 structures of Tm proteins (including 10 determined by NMR). This represents 11% of the total number of ORFs in this organism, which is a remarkable feat. Almost 1,000 recombinant Tm proteins have been purified successfully, 770 of them at the JCSG. An important point to note here is that this is a very valuable potential resource for the community of thermophile researchers. Many of these proteins may only be expressed at low levels, but there are now clones and data available on how to express and purify proteins representing at least half of the Tm genome. As discussed below, such resources are also available for several other extremophiles.

A number of structures of Tm proteins have yielded novel insights into the function of unknown proteins. For example, TM1662 encodes a surE homolog by sequence analysis and the structure of the Tm protein was determined (Zhang et al. 2001). The surE protein is conserved across all domains of life. However, its function is not clear, although it is expressed in stationary phase growth of *E. coli*. Through the SG effort, TM1662 was shown to be an acid phosphatase, despite having no sequence similarity to any other acid phosphatases. Another example is TM0654, which represented the first structure of an aminopropyltransferase (Korolev et al. 2002). This protein is involved in biosynthesis of common polyamines such as spermidine. The structure indicated that the active site was highly conserved in bacteria and eukaryotes, thus suggesting a universal catalytic mechanism and the specific residues likely to be involved (Korolev et al. 2002). Other structures of previously unknown ORFs have led to new, relevant insights into protein function. These include TM1643, which represents a completely novel family of enzymes, aspartate dehydrogenase, that catalyzes the first step of NAD biosynthesis (Yang et al. 2003). In addition, structures of unknown proteins can illuminate entire families of unknown genes. A good example is the NMR structure of TM0487, for which there are more than 200 homologs in the database. The structure of the Tm protein indicates a possible active site with a buried Asp residue (Almeida et al. 2005). Other structures of Tm proteins produced by SG groups have indicated unique covalent protein dimers and a novel DNA binding protein (Liu et al. 2005; Zhang et al. 2006).

The extensive library of structures of Tm proteins produced by SG efforts has also allowed for some initial attempts at correlating their high thermal stability with structural elements such as contact order (Robinson-Rechavi and Godzik 2005), density of salt bridges, and compactness (Robinson-Rechavi et al. 2006; Robinson-Rechavi and Godzik 2005). These data indicate a clear correlation between an increase in contact order between residues in the thermophilic proteins relative to mesophilic ones (Robinson-Rechavi et al. 2006). This is a particularly significant contribution to the understanding of protein stability, as there are many different proposals for the basis of the extreme stability of proteins from hyperthermophiles (Chakravarty and Varadarajan 2002; Sadeghi et al. 2006). The Tm protein collection has also been used for extensive screening of NMR structure candidates (Peti et al. 2004), protein solubility screening for crystallization optimization (Collins et al. 2005), and to design a HTP pipeline from cloning to structure determination (DiDonato et al. 2004). Clearly, the work with Tm has had a significant impact on the SG world in general and it remains one of the most studied organisms in this regard.

Thermus thermophilus

Thermus thermophilus (Tt) is an aerobic, thermophilic (T_{opt} 68°C), gram negative bacterium originally isolated from a thermal environment in Japan (Oshima and Imahori 1971). This organism is of significant biotechnological interest as it is tolerant to a number of stress conditions (Koyama et al. 1986). It is amenable to genetic manipulation (Hashimoto et al. 2001) and is closely related to the mesophilic, radiation-resistant *Deinococcus radiodurans* (Henne et al. 2004). Its 2.12 Mbp genome is predicted to contain 2,238 ORFs. The Tt SG effort is being carried out by groups at Osaka University and RIKEN [see http://www.thermus.org/e_index.htm (Yokoyama et al. 2000)]. So far 1,450 Tt ORFs have been heterologously expressed, 930 recombinant proteins have been purified, 632 have been crystallized. These have yielded 438 structures to date by these groups deposited in the PDB (http://www.pdb.org) although very few have been formally described in publications, and hence few have any degree of biochemical characterization. Unfortunately, this is one of the drawbacks of the SG approach, where the primary goal is structure determination. The interpretation of a structure, particularly if it is not novel (in structural terms), is typically not a priority and is left to those outside of the SG projects.

As with Tm, the large amount of structural information generated on Tt proteins is being used to make global predictions about thermal stability, the solubility, and the crystallization ability of recombinant proteins. For example, 108 Tt sequences were used to predict structural domains, and experimentally assess these structural predictions and the stability of the recombinant proteins using NMR spectroscopy (Hondoh et al. 2006). A major part of the SG efforts with *Thermus* species in Japan has been the very promising development of HTP cell-free in vitro expression systems. This can eliminate a number of problems associated with in vivo expression such as cell lysis and multiple purification steps, as well as reducing the cost of isotopic labeling of protein targets (Endo and Sawasaki 2006; Yokoyama 2003; Yokoyama et al. 2000).

Pyrococcus furiosus and *P. horikoshii*

Two species of these obligately anaerobic, heterotrophic, hyperthermophilic archaea, both growing optimally near 100°C, have been the targets of SG projects. *Pyrococcus furiosus* (Pf) was isolated from a shallow marine solfatara near Vulcano, Italy (Fiala and Stetter 1986) and its genome of 1.9 Mbp contains approximately 2,200 ORFs. Pf was one of the inital target organisms at the NIH SG center SECSG (Adams et al. 2003). The specific goal was to express as many of its proteins as possible in a fully-folded,

functional form. This involved developing expression protocols for recombinant proteins that contain cofactors and/or are part of multiprotein complexes, for example, by growth of the heterologous host in the presence of excess Fe or Zn for metal cofactors (Jenney et al. 2005), or by coexpression of multiple ORFs for multiprotein heteromeric complexes. It was predicted that at least 20% of the ORFs would encode membrane proteins (Holden et al. 2001), and that few of these would yield soluble proteins.

One critical issue in designing any experiment involving the entire proteome of an organism is how to precisely define that proteome, both in terms of the total number of ORFs, and their putative translation start sites. While the original annotation of the Pf genome contained 2,065 putative ORFs (Robb et al. 2001), there are two annotations currently in the major databases (http://www.ncbi.nlm.nih.gov and http://www.tigr.org) where up to 2,261 ORFs are predicted (Poole et al. 2005). One major issue in annotations that is particularly important for SG efforts is the correct start site for a given ORF. For example, 552 ORFS, or about 25% of the total proteome, in the two current annotations of the Pf genome differ in their start-sites, many by the equivalent of more than 20 amino acids (Poole et al. 2005). The addition or deletion of a few critical residues at the N terminus of a protein could have a dramatic effect on protein stability, solubility and its ability to crystallize. There are no bioinformatic tools available to address this problem, so for the Pf project at the SECSG the maximum possible start site was chosen (which, if incorrect, would generate extended rather than truncated proteins) for all 2,192 of the predicted ORFs. Of these, 1,909 were cloned into an expression vector containing an N-terminal His affinity tag (MAHHHHHGS-). This allows protein purification by immobilized metal affinity chromatography (IMAC), as well as detection using an enzyme-linked immunosorbant assay (ELISA) with a commercial antibody against the His affinity tag [see http://www.secsg.org and (Sugar et al. 2005)].

For the production of recombinant Pf proteins in *E. coli*, an automated screening was performed using a small scale (1 mL) expression system (SSE). The soluble and insoluble fractions of cell-free extracts were separated robotically and recombinant protein production was assessed using an antibody to the His tag (Adams et al. 2003; Sugar et al. 2005). All of the SG centers have developed and demonstrated similar types of HTP heterologous protein expression screens, for example (Acton et al. 2005; Alzari et al. 2006; Cornvik et al. 2006; Dieckman et al. 2006; Douris et al. 2006; Hart and Tarendeau 2006; Vincentelli et al. 2005). In the case of Pf, the expression screen data were used to scale production to (at least) 1-L cultures of *E. coli* for the purification of the milligram amounts of protein necessary for analyses by X-ray crystallography (and NMR spectroscopy). Clones that failed the expression step (either due to no or limited amounts of recombinant protein, or the production of insoluble, presumably unfolded protein) were subjected to protocols of increasing complexity, such as alternative *E. coli* expression strains, recloning with different expression vectors or affinity tags, different host organisms, etc. For Pf, a total 2,381 cultures representing 1,008 unique ORFs were grown at the 1-L scale. Of these 57% (578) produced sufficient protein to be detected after SDS-polyacrylamide gel electrophoresis (after the IMAC step) and 388 proteins representing unique ORFs have been purified. Of these, 259 (67%) gave the predicted mass when analyzed by mass spectrometry, i.e., they had not been degraded, or subjected to some unknown post-translational modification in *E. coli*, and 240 (62%) were submitted for X-ray crystallography screening (and 137 for NMR screening). This resulted in 108 crystals, 59 of which diffracted, and 29 structures were obtained. The results to date are indicated in Table 1.

For the structures of Pf proteins determined by the SG effort, half of them (15 of 29) represented conserved hypothetical proteins. Unfortunately, insights into their biological functions provided by the structures were limited. For example, in the case of the hypothetical protein PF1455, its structure indicated that the protein is involved in the binding, transport, or detoxification of heavy metals (Mayer et al. 2006). On the other hand, some of the proteins enabled advances to be made in protein structure analysis. For example, PF1455 was used to demonstrate that with a rapidly collected, limited amount of NMR data (traditionally a slow method for structure determination), a structure can be modeled with sufficient detail to both

Table 1 December 2006 production statistics for *Pyrococcus furiosus* proteins from gene to structure, and for all structural genomics groups worldwide registered in the TargetDB [see http://www.secsg.org and Protein Data Bank (http://www.rcsb.org/pdb/, 2004)]

Status	Pf (SECSG)	All SG targets worldwide (PDB)
Selected	2,192	119,506
Cloned	1,911	80,788
Expression attempted	1,008	46,064
Soluble protein	578	23,540
Purified	388	19,511
Crystallized	110/259	7,199
Crystals diffract	59	3,122
Crystal structure	36	2,767
HSQC[a]	112/137	2,234
NMR structure	2	1,181
In PDB[b]	35	3,647

[a] Heteronuclear single quantum coherence (HSQC) measured

[b] Structure deposited and publically available

render a prediction as to its possible function, and to classify it as a novel fold. Such information is extremely important in SG screening so that protein targets are not duplicated (Mayer et al. 2006). The structure of another Pf protein (rubrerythrin, PF1283) provided an example of domain swapping, an unusual observation in protein structure. In this case, domains from two different monomers in a dimer were intertwined to form a structure homologous to that of a previously characterized protein [in which the same structure is made up of domains from one monomer (Tempel et al. 2004)]. Research on Pf proteins at the SECSG has led to a number of methods developments for HTP protein expression and structure determination (Jenney et al. 2005; Sugar et al. 2005; Valafar et al. 2004; Wang et al. 2005).

The other *Pyrococcus* species that is the specific target of an SG effort is *P. horikoshii* (Ph). In contrast to Pf, Ph was isolated from a deep sea hydrothermal vent in the Pacific Ocean (Gonzalez et al. 1998) although the two organisms are closely related and have similar size genomes (Lecompte et al. 2001). The SG effort with Ph at RIKEN (http://www.riken.go.jp) led to the production of 472 recombinant proteins, 447 of which were purified. Remarkably, this effort has led to over 180 structures of Ph proteins. Unfortunately, as is characteristic of SG, very few of these structures have been published in peer reviewed journals and so the information is not widely disseminated to those in the field of extremophiles. Two other closely-related species (Cohen et al. 2003; Fukui et al. 2005), *P. abyssi* (T_{opt} 98°C), isolated from a deep sea vent in the Pacific, and *Thermococcus kodakaraensis* (formerly *Pyrococcus* T_{opt} 85°C), isolated from a surface solfatara in Japan, appear on the target lists of various SG centers. However, only a few of their ORFs (36 and 6, respectively) have been utilized to produce proteins, and the organisms themselves (or rather their complete genomes) have not been targets of any SG effort.

Other extremophile targets

Extremophiles such as Tm, Pf, Tt and Mt are therefore unique as they have been specific targets of the initial SG efforts, and a large number of crystal structures of their proteins have been generated. The hyperthermophile *Pyrobaculum aerophilum* (Pa) was also one of the first target organisms at the beginning of one of the NIH-funded SG projects (for the Integrated Center for Structure and Function Innovation, formerly the TB Structural Genomics Consortium). However, this effort was not sustained as the focus of the center moved to the disease-causing, mesophilic bacterium *Mycobacterium tuberculosis*, and more recently, to a technology-based approach that emphasizes producing correctly-folded proteins regardless of source (http://techcenter.mbi.ucla.edu) (Protein Structure Initiative 2005). In a similar fashion, *Methanococccus jannaschii* (Mj) was a specific target organism of another SG center (at UC Berkeley, http://www.strgen.org). One of its early successes was the assignment of a biochemical function to a hypothetical Mj protein (Zarembinski et al. 1998). However, this SG group has since shifted emphasis to proteins from species of the mesophilic bacterium *Mycoplasma* (Chandonia et al. 2006). Although not a specific organismal target, ORFs from Mj are still the subject of study, with 317 targets listed in the PDB and the structures of 20 Mj proteins have been determined, some of which have been characterized biochemically. For example, MJ0936 was shown to be a novel phosphodiesterase (Chen et al. 2004) (Martinez-Cruz et al. 2002). Table 2 is a select list of some example target organisms from the TargetDB in the PDB, and demonstrates that a number of extremophiles have been targeted by the various SG centers around the world. It also shows that, at least in the early days of SG, the emphasis was clearly on thermophilic and particularly hyperthermophilic organisms (usually defined as those with $T_{opt} \geq 80$°C).

Table 2 Extremophiles as targets of structural genomics projects

Type	Example organism	Number of SG gene targets in PDB
Acidophilic bacterium	*Acidothermus cellulolyticus*	1
Thermophilic, acidophilic euryarchaeon	*Thermoplasma acidophilum*	461
Alkaliphilic bacterium	*Alkaliphilus metalliredigenes*	7
Halophilic bacterium	*Halobacterium* sp.	155
Radiation resistant bacterium	*Deinococcus radiodurans*	384
Psychrophilic bacterium	*Psychrobacter cryohalolentis*	4
Hyperthermophilic bacterium	*Aquifex aeolicus*	664
Hyperthermophilic euryarchaeon	*Archaeoglobus fulgidus*	761
Hyperthermophilic crenarchaeon	*Sulfolobus solfataricus*	606
Hyperthermophilic crenarchaeon	*Aeropyrum pernix*	564

Data are derived from the Protein Data Bank (http://targetdb.pdb.org/) representing data voluntarily deposited by SG centers worldwide. These data do not include individual research projects outside of the SG groups

Phase II of SG

Now that the 5-year pilot phase I of the NIH-funded SG initiative that began in 2000 is complete, the second, production phase is well underway. As stated above, there is a truly significant shift in priorities in this new phase. Individual organisms are no longer targeted and while extremophiles had a major impact on the first phase, their proteins (genes) are now lost in the sea of orthologs that are chosen entirely by bioinformatic criteria. Nonetheless, proteins from (hyper)thermophiles and other extremophiles will certainly be included on these target lists, and should it hold true that these proteins are more stable and crystallize more easily than mesophilic proteins, then they will likely be over represented in the list of protein structures that are produced.

A summary of current statistics for all SG groups can be found at the PDB (http://www.rcsb.org/pdb/, 2004), but at the time of this writing (December 2006), 119,506 targets from more than 600 organisms/strains have resulted in 2,767 crystal and 1,181 NMR structures (Table 1). Note also in this table that while the attrition rate across all groups has improved at some steps (for example, now as many as 82% of soluble proteins are successfully purified) only about 8% (3,948 of 46,064 proteins where expression has been attempted) have yielded either X-ray crystal or NMR structures. These numbers represent a glimpse at a rapidly changing scene, and a more in-depth analysis of these statistics has been reported recently (Chandonia and Brenner 2006). Of course, there continue to be general critiques of the SG philosophy [for example, Cyranoski (2006)], as significant funds in many countries, which could be directed towards individual research laboratories, have been directed towards the SG efforts.

The most serious problem in SG is that the steps in the ''gene to structure'' pipeline remain empirical—few predictive rules have become apparent and these mainly concern properties of proteins such as thermostability and correlation of physical properties with crystallization success (Canaves et al. 2004; Robinson-Rechavi et al. 2006). The hope is that a more extensive data set will allow better prediction of success in heterologous expression systems to obtain stable recombinant proteins. This will have a tremendous impact and make many more proteins available in fully-folded, functional forms for complete structural and functional characterization. As of yet, such predictions are still hampered by the incredible variability inherent in proteins.

The impact of SG on extremophiles

The impact of genome sequencing on a particular organism or a group of organisms is clear cut and readily appreciated, with quantitative results, such as number of bases, number of predicted ORFs, etc. The world of SG, however, is far more qualitative and it is hard to measure how much impact has been made in a particular field, such as extremophiles. In general terms, it is clear that in a few short years the SG efforts around the world have contributed a large number of novel structures to the public databases, and many are of proteins from extremophiles. SG efforts have also yielded new HTP technologies that have accelerated bioinformatics analyses, cloning and protein expression screening, and much more rapid structure determination, and these tools and protocols are available to all researchers. Those groups who are also interested in the biology of extremophiles and are directly involved in such efforts have also directly benefited. However, it is more difficult to say that the SG efforts have made a very specific impact on the field of extremophiles in general. A large number of structures from particular organisms are now available, especially from *Thermotoga*, *Pyrococcus* and *Methanothermobacter* species, and these in turn allow structure modelling of homologous proteins from many other organisms (Todd et al. 2005). However, as of yet the available SG structures have had no groundbreaking effect on extremophile research. In general, the biological contribution of SG efforts so far has been in using novel structure information to direct functional biochemical analyses (Sanishvili et al. 2003; Yakunin et al. 2004), but this has not really affected extremophiles.

The most important ramification of SG efforts for those who study extremophiles is more technical than scientific. This concerns the large number of recombinant proteins produced from a variety of genes from numerous extremophile sources. More importantly, the procedures and protocols to produce these recombinant proteins, while typically not published in the formal literature, are available on web sites from the various SG centers (links to all these centers can be found at http://www.nigms.nih.gov/Initiatives/PSI). Similarly, a huge collection of clones is also available for an even larger variety of extremophilic organisms, and these may or may not have been analyzed for the production of recombinant protein. The complete (and searchable) list of all targets selected by all SG centers worldwide can be found at the PDB (http://targetdb.pdb.org/). Such resources have been created by the SG phenomenon and are available to be utilized by the extremophile community at large.

Acknowledgments Work reported here from the authors' laboratory was supported in part by grants from the National Science Foundation, the Department of Energy, the National Institutes of Health, the University of Georgia and the Georgia Research Alliance.

References

Abergel C, Coutard B, Byrne D, Chenivesse S, Claude JB, Deregnaucourt C, Fricaux T, Gianesini-Boutreux C, Jeudy S, Lebrun R, Maza C, Notredame C, Poirot O, Suhre K, Varagnol M, Claverie JM (2003) Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. J Struct Funct Genomics 4:141–157

Acton TB, Gunsalus KC, Xiao R, Ma LC, Aramini J, Baran MC, Chiang YW, Climent T, Cooper B, Denissova NG, Douglas SM, Everett JK, Ho CK, Macapagal D, Rajan PK, Shastry R, Shih LY, Swapna GV, Wilson M, Wu M, Gerstein M, Inouye M, Hunt JF, Montelione GT (2005) Robotic cloning and protein production platform of the northeast structural genomics consortium. Methods Enzymol 394:210–243

Adams MW, Dailey HA, DeLucas LJ, Luo M, Prestegard JH, Rose JP, Wang BC (2003) The southeast collaboratory for structural genomics: a high-throughput gene to structure factory. Acc Chem Res 36:191–198

Almeida MS, Herrmann T, Peti W, Wilson IA, Wuthrich K (2005) NMR structure of the conserved hypothetical protein TM0487 from *Thermotoga maritima*: implications for 216 homologous DUF59 proteins. Protein Sci 14:2880–2886

Alzari PM, Berglund H, Berrow NS, Blagova E, Busso D, Cambillau C, Campanacci V, Christodoulou E, Eiler S, Fogg MJ, Folkers G, Geerlof A, Hart D, Haouz A, Herman MD, Macieira S, Nordlund P, Perrakis A, Quevillon-Cheruel S, Tarandeau F, van Tilbeurgh H, Unger T, Luna-Vargas MP, Velarde M, Willmanns M, Owens RJ (2006) Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. Acta Crystallogr D Biol Crystallogr 62:1103–1113

Arzt S, Beteva A, Cipriani F, Delageniere S, Felisaz F, Forstner G, Gordon E, Launer L, Lavault B, Leonard G, Mairs T, McCarthy A, McCarthy J, McSweeney S, Meyer J, Mitchell E, Monaco S, Nurizzo D, Ravelli R, Rey V, Shepard W, Spruce D, Svensson O, Theveneau P (2005) Automation of macromolecular crystallography beamlines. Prog Biophys Mol Biol 89:124–152

Atreya HS, Szyperski T (2005) Rapid NMR data collection. Methods Enzymol 394:78–108

Banci L, Bertini I, Cusack S, de Jong RN, Heinemann U, Jones EY, Kozielski F, Maskos K, Messerschmidt A, Owens R, Perrakis A, Poterszman A, Schneider G, Siebold C, Silman I, Sixma T, Stewart-Jones G, Sussman JL, Thierry JC, Moras D (2006) First steps towards effective methods in exploiting high-throughput technologies for the determination of human protein structures of high biomedical value. Acta Crystallogr D Biol Crystallogr 62:1208–1217

Berry IM, Dym O, Esnouf RM, Harlos K, Meged R, Perrakis A, Sussman JL, Walter TS, Wilson J, Messerschmidt A (2006) SPINE high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. Acta Crystallogr D Biol Crystallogr 62:1137–1149

Bonanno JB, Almo SC, Bresnick A, Chance MR, Fiser A, Swaminathan S, Jiang J, Studier FW, Shapiro L, Lima CD, Gaasterland TM, Sali A, Bain K, Feil I, Gao X, Lorimer D, Ramos A, Sauder JM, Wasserman SR, Emtage S, D'Amico KL, Burley SK (2005) New York-structural genomix research consortium (NYSGXRC): a large scale center for the protein structure initiative. J Struct Funct Genomics 6:225–232

Bravo J, Aloy P (2006) Target selection for complex structural genomics. Curr Opin Struct Biol 16:385–392

Brenner SE (2000) Target selection for structural genomics. Nat Struct Biol 7 Suppl:967–969

Brenner SE, Chothia C, Hubbard TJ (1997) Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 7:369–376

Brenner SE, Levitt M (2000) Expectations from structural genomics. Protein Sci 9:197–200

Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7(Suppl):932–934

Canaves JM, Page R, Wilson IA, Stevens RC (2004) Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics. J Mol Biol 344:977–991

Chakravarty S, Varadarajan R (2002) Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 41:8152–8161

Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311:347–351

Chandonia JM, Kim SH, Brenner SE (2006) Target selection and deselection at the Berkeley structural genomics center. Proteins 62:356–370

Chen S, Yakunin AF, Kuznetsova E, Busso D, Pufan R, Proudfoot M, Kim R, Kim SH (2004) Structural and functional characterization of a novel phosphodiesterase from *Methanococcus jannaschii*. J Biol Chem 279:31854–31862

Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH (2000) Structural proteomics of an archaeon. Nat Struct Biol 7:903–909

Cianci M, Antonyuk S, Bliss N, Bailey MW, Buffey SG, Cheung KC, Clarke JA, Derbyshire GE, Ellis MJ, Enderby MJ, Grant AF, Holbourn MP, Laundy D, Nave C, Ryder R, Stephenson P, Helliwell JR, Hasnain SS (2005) A high-throughput structural biology/proteomics beamline at the SRS on a new multipole wiggler. J Synchrotron Radiat 12:455–466

Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Querellou J, Ripp R, Thierry JC, Van der Oost J, Weissenbach J, Zivanovic Y, Forterre P (2003) An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol Microbiol 47:1495–1512

Collins B, Stevens RC, Page R (2005) Crystallization optimum solubility screening: using crystallization results to identify the optimal buffer for protein crystal formation. Acta Crystallogr Sect F Struct Biol Cryst Commun 61:1035–1038

Cornvik T, Dahlroth SL, Magnusdottir A, Flodin S, Engvall B, Lieu V, Ekberg M, Nordlund P (2006) An efficient and generic strategy for producing soluble human proteins and domains in *E. coli* by screening construct libraries. Proteins 65:266–273

Cyranoski D (2006) 'Big science' protein project under fire. Nature 443:382

DiDonato M, Deacon AM, Klock HE, McMullan D, Lesley SA (2004) A scaleable and integrated crystallization pipeline applied to mining the *Thermotoga maritima* proteome. J Struct Funct Genomics 5:133–146

Dieckman LJ, Hanly WC, Collart ER (2006) Strategies for high-throughput gene cloning and expression. Genet Eng (N Y) 27:179–190

Douris V, Swevers L, Labropoulou V, Andronopoulou E, Georgoussi Z, Iatrou K (2006) Stably transformed insect cell lines: tools for expression of secreted and membrane-anchored proteins and high-throughput screening platforms for drug and insecticide discovery. Adv Virus Res 68:113–156

Endo Y, Sawasaki T (2006) Cell-free expression systems for eukaryotic protein production. Curr Opin Biotechnol 17:373–380

Esposito D, Chatterjee DK (2006) Enhancement of soluble protein expression through the use of fusion tags. Curr Opin Biotechnol 17:353–358

Fiala G, Stetter KO (1986) *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch Microbio 145:56–60

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fukui T, Atomi H, Kanai T, Matsumi R, Fujiwara S, Imanaka T (2005) Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res 15:352–363

Gaasterland T (1998) Structural genomics: bioinformatics in the driver's seat. Nat Biotechnol 16:625–627

Ginalski K, Grishin NV, Godzik A, Rychlewski L (2005) Practical lessons from protein structure prediction. Nucleic Acids Res 33:1874–1891

Gonzalez JM, Masuchi Y, Robb FT, Ammerman JW, Maeder DL, Yanagibayashi M, Tamaoka J, Kato C (1998) *Pyrococcus horikoshii* sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. Extremophiles 2:123–130

Hart DJ, Tarendeau F (2006) Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. Acta Crystallogr D Biol Crystallogr 62:19–26

Hashimoto Y, Yano T, Kuramitsu S, Kagamiyama H (2001) Disruption of *Thermus thermophilus* genes by homologous recombination using a thermostable kanamycin-resistant marker. FEBS Lett 506:231–234

Henne A, Bruggemann H, Raasch C, Wiezer A, Hartsch T, Liesegang H, Johann A, Lienard T, Gohl O, Martinez-Arias R, Jacobi C, Starkuviene V, Schlenczeck S, Dencker S, Huber R, Klenk HP, Kramer W, Merkl R, Gottschalk G, Fritz HJ (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*. Nat Biotechnol 22:547–553

Holden JF, Poole FL, Tollaksen SL, Giometti CS, Lim H, Yates JR, Adams MWW (2001) Identification of membrane proteins in the hyperthermophilic archaeon *Pyrococcus furiosus* using proteomics and prediction programs. Comp Funct Genomics 2:275–288

Holm L, Sander C (1994) Searching protein structure databases has come of age. Proteins 19:165–173

Holm L, Sander C (1997) New structure—novel fold?. Structure 5:165–171

Hondoh T, Kato A, Yokoyama S, Kuroda Y (2006) Computer-aided NMR assay for detecting natively folded structural domains. Protein Sci 15:871–883

Huber R, Langworthy TA, Konig H, Thomm M, Woese CR, Sleytr UB, Stetter KO (1986) *Thermotoga-maritima* sp-nov represents a new genus of unique extremely thermophilic eubacteria growing up to 90-degrees-C. Arch Microbiol 144:324–333

Ito K, Arai R, Fusatomi E, Kamo-Uchikubo T, Kawaguchi SI, Akasaka R, Terada T, Kuramitsu S, Shirouzu M, Yokoyama S (2006) Crystal structure of the conserved protein TTHA0727 from *Thermus thermophilus* HB8 at 1.9 angstrom resolution: a CMD family member distinct from carboxymuconolactone decarboxylase (CMD) and AhpD. Protein Sci 15:1187–1192

Jenney FE, Brereton PS, Izumi M, Poole FL, Shah C, Sugar FJ, Lee HS, Adams MWW (2005) High-throughput production of *Pyrococcus furiosus* proteins: considerations for metalloproteins. J Synchrotron Radiat 12:8–12

Korolev S, Ikeguchi Y, Skarina T, Beasley S, Arrowsmith C, Edwards A, Joachimiak A, Pegg AE, Savchenko A (2002) The crystal structure of spermidine synthase with a multisubstrate adduct inhibitor. Nat Struct Biol 9:27–31

Koyama Y, Hoshino T, Tomizuka N, Furukawa K (1986) Genetic transformation of the extreme thermophile *Thermus thermophilus* and of other *Thermus* spp. J Bacteriol 166:338–340

Lecompte O, Ripp R, Puzos-Barbe V, Duprat S, Heilig R, Dietrich J, Thierry JC, Poch O (2001) Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res 11:981–993

Lee CH, Jung JW, Yee A, Arrowsmith CH, Lee W (2004) Solution structure of a novel calcium binding protein, MTH1880, from *Methanobacterium thermoautotrophicum*. Protein Sci 13:1148–1154

Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T, Vincent J, Robb A, Brinen LS, Miller MD, McPhillips TM, Miller MA, Scheibe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC (2002) Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. Proc Natl Acad Sci USA 99:11664–11669

Liu JY, Huang CD, Shin DH, Yokota H, Jancarik J, Kim JS, Adams PD, Kim R, Kim SH (2005) Crystal structure of a heat-inducible transcriptional repressor HrcA from *Thermotoga maritima*: structural insight into DNA binding and dimerization. J Mol Biol 350:987–996

Liu X, Fan K, Wang W (2004) The number of protein folds and their distribution over families in nature. Proteins 54:491–499

Mallick P, Goodwill KE, Fitz-Gibbon S, Miller JH, Eisenberg D (2000) Selecting protein targets for structural genomics of *Pyrobaculum aerophilum*: validating automated fold assignment methods by using binary hypothesis testing. Proc Natl Acad Sci USA 97:2450–2455

Marsden RL, Lee D, Maibaum M, Yeats C, Orengo CA (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. Nucleic Acids Res 34:1066–1080

Marsischky G, LaBaer J (2004) Many paths to many clones: a comparative look at high-throughput cloning methods. Genome Res 14:2020–2028

Martinez-Cruz LA, Dreyer MK, Boisvert DC, Yokota H, Martinez-Chantar ML, Kim R, Kim SH (2002) Crystal structure of MJ1247 protein from *M. jannaschii* at 2.0 A resolution infers a molecular function of 3-hexulose-6-phosphate isomerase. Structure (Camb) 10:195–204

Mayer KL, Qu Y, Bansal S, LeBlond PD, Jenney FE, Brereton PS, Adams MWW, Xu Y, Prestegard JH (2006) Structure determination of a new protein from backbone-centered NMR data and NMR-assisted structure prediction. Proteins Struct Funct Bioinform 65:480–489

McPherson A (2004) Protein crystallization in the structural genomics era. J Struct Funct Genomics 5:3–12

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329

Orengo CA, Todd AE, Thornton JM (1999) From protein structure to function. Curr Opin Struct Biol 9:374–382

Oshima T, Imahori K (1971) Isolation of an extreme thermophile and thermostability of its transfer ribonucleic-acid and ribosomes. J Gen Appl Microbiol 17:513–517

Peti W, Etezady-Esfarjani T, Herrmann T, Klock HE, Lesley SA, Wuthrich K (2004) NMR for structural proteomics of *Thermo-*

toga maritima: screening and structure determination. J Struct Funct Genomics 5:205–215

Poole FL 2nd, Gerwe BA, Hopkins RC, Schut GJ, Weinberg MV, Jenney FE Jr, Adams MW (2005) Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. J Bacteriol 187:7325–7332

Protein Structure Initiative P (2005) http://www.nigms.nih.gov/psi/, vol 2005. NIGMS/NIH Protein Structure Initiative

Pusey ML, Liu ZJ, Tempel W, Praissman J, Lin D, Wang BC, Gavira JA, Ng JD (2005) Life in the fast lane for protein crystallization and X-ray crystallography. Prog Biophys Mol Biol 88:359–386

Rees DC (2001) Crystallographic analyses of hyperthermophilic proteins. Methods Enzymol 334:423–437

Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB, Dunn DM (2001) Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. Methods Enzymol 330:134–157

Robinson-Rechavi M, Alibes A, Godzik A (2006) Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima*. J Mol Biol 356:547–557

Robinson-Rechavi M, Godzik A (2005) Structural genomics of *Thermotoga maritima* proteins shows that contact order is a major determinant of protein thermostability. Structure (Camb) 13:857–860

Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B (2006) Effective factors in thermostability of thermophilic proteins. Biophys Chem 119:256–270

Sanishvili R, Yakunin AF, Laskowski RA, Skarina T, Evdokimova E, Doherty-Kirby A, Lajoie GA, Thornton JM, Arrowsmith CH, Savchenko A, Joachimiak A, Edwards AM (2003) Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from *Escherichia coli*. J Biol Chem 278:26039–26045

Saridakis V, Christendat D, Kimber MS, Dharamsi A, Edwards AM, Pai EF (2001) Insights into ligand binding and catalysis of a central step in NAD+ synthesis: structures of *Methanobacterium thermoautotrophicum* NMN adenylyltransferase complexes. J Biol Chem 276:7225–7232

Savchenko A, Yee A, Khachatryan A, Skarina T, Evdokimova E, Pavlova M, Semesi A, Northey J, Beasley S, Lan N, Das R, Gerstein M, Arrowmith CH, Edwards AM (2003) Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. Proteins 50:392–399

Service R (2005) Structural biology. Structural genomics, round 2. Science 307:1554–1558

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Reeve JN et al (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 179:7135–7155

Sugar FJ, Jenney FE Jr, Poole FL 2nd, Brereton PS, Izumi M, Shah C, Adams MW (2005) Comparison of small- and large-scale expression of selected *Pyrococcus furiosus* genes as an aid to high-throughput protein production. J Struct Funct Genomics 6:149–158

Szilagyi A, Zavodszky P (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure 8:493–504

Tempel W, Liu ZJ, Schubot FD, Shah A, Weinberg MV, Jenney FE Jr, Arendall WB 3rd, Adams MW, Richardson JS, Richardson DC, Rose JP, Wang BC (2004) Structural genomics of *Pyrococcus furiosus*: X-ray crystallography reveals 3D domain swapping in rubrerythrin. Proteins 57:878–882

Todd AE, Marsden RL, Thornton JM, Orengo CA (2005) Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 348:1235–1260

Valafar H, Mayer KL, Bougault CM, LeBlond PD, Jenney FE Jr, Brereton PS, Adams MW, Prestegard JH (2004) Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target. J Struct Funct Genomics 5:241–254

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, et al (2001) The sequence of the human genome. Science 291:1304–1351

Vincentelli R, Canaan S, Offant J, Cambillau C, Bignon C (2005) Automated expression and solubility screening of His-tagged proteins in 96-well format. Anal Biochem 346:77–84

Wang BC, Adams MW, Dailey H, DeLucas L, Luo M, Rose J, Bunzel R, Dailey T, Habel J, Horanyi P, Jenney FE Jr, Kataeva I, Lee HS, Li S, Li T, Lin D, Liu ZJ, Luan CH, Mayer M, Nagy L, Newton MG, Ng J, Poole FL 2nd, Shah A, Shah C, Sugar FJ, Xu H (2005) Protein production and crystallization at SECSG—an overview. J Struct Funct Genomics 6:233–243

Wolfson HJ, Shatsky M, Schneidman-Duhovny D, Dror O, Shulman-Peleg A, Ma BY, Nussinov R (2005) From structure to function: methods and applications. Curr Protein Pept Sci 6:171–183

Yakunin AF, Yee AA, Savchenko A, Edwards AM, Arrowsmith CH (2004) Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol 8:42–48

Yang Z, Savchenko A, Yakunin A, Zhang R, Edwards A, Arrowsmith C, Tong L (2003) Aspartate dehydrogenase, a novel enzyme identified from structural and functional studies of TM1643. J Biol Chem 278:8804–8808

Yokoyama S (2003) Protein expression systems for structural genomics and proteomics. Curr Opin Chem Biol 7:39–43

Yokoyama S (2005) [Large-scale structural proteomics project at RIKEN: present and future]. Tanpakushitsu Kakusan Koso 50:836–845

Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, Kyogoku Y, Miki K, Masui R, Kuramitsu S (2000) Structural genomics projects in Japan. Nat Struct Biol 7(Suppl):943–945

Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci USA 95:15189–193

Zeikus JG, Wolfe RS (1972) *Methanobacterium thermoautotrophicus* sp. n., an anaerobic, autotrophic, extreme thermophile. J Bacteriol 109:707–715

Zhang R, Skarina T, Evdokimova E, Edwards A, Savchenko A, Laskowski R, Cuff ME, Joachimiak A (2006) Structure of SAICAR synthase from *Thermotoga maritima* at 2.2 angstroms reveals an unusual covalent dimer. Acta Crystallogr Sect F Struct Biol Cryst Commun 62:335–339

Zhang RG, Skarina T, Katz JE, Beasley S, Khachatryan A, Vyas S, Arrowsmith CH, Clarke S, Edwards A, Joachimiak A, Savchenko A (2001) Structure of *Thermotoga maritima* stationary phase survival protein SurE: a novel acid phosphatase. Structure (Camb) 9:1095–1106

Zhou CZ, Chen YX (2004) Developments in structural genomics: protein purification and function interpretation. Curr Genomics 5:37–48