A TOUR OF STRUCTURAL GENOMICS

Steven E. Brenner

Structural genomics projects aim to provide an experimental or computational three-dimensional model structure for all of the tractable macromolecules that are encoded by complete genomes. To this end, pilot centres worldwide are now exploring the feasibility of large-scale structure determination. Their experimental structures and computational models are expected to yield insight into the molecular function and mechanism of thousands of proteins. The pervasiveness of this information is likely to change the use of structure in molecular biology and biochemistry.

Department of Plant and Microbial Biology, University of California, 461A Koshland Hall, Berkeley, California 94720-3102, USA.
e-mail: brenner@compbio.berkeley.edu

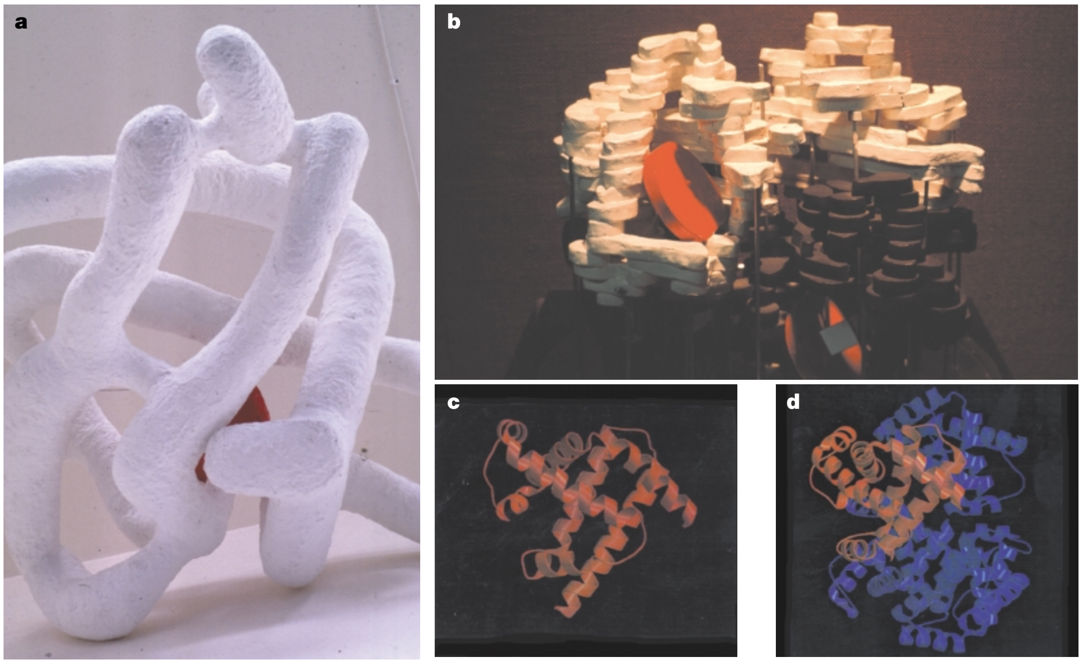
The explosive growth of genetic sequence information has offered us comprehensive collections of the protein sequences found in many living organisms. Most of these are not experimentally characterized. Although half of the proteins that are encoded in sequenced eukaryotic genomes have computationally recognized homology to at least one well-characterized domain[1,2], functional interpretation of these matches is fraught with difficulty. Functional changes over evolutionary time[3,4] and database errors[5] confound reliable computational prediction of the precise roles of newly discovered genes. Even proteins with recognized domains are often scattered with regions of unmatched sequence. So, most of the residues in putative gene products lack any computational annotation, and there exists no general experimental approach to directly ascertain their molecular role.

The challenge of understanding these gene products has led to the development of functional genomics methods, which collectively aim to imbue the raw sequence with biological understanding. Structural genomics is one such approach, with unique promise to reveal the molecular function[6] of protein domains.

Protein structure represents a powerful means of discovering function, because structure is well conserved over evolutionary time, and it therefore provides the opportunity to recognize homology that is undetectable by sequence comparison. This became apparent with the first two protein structures that were determined, because their common ancestry was clear from the three-dimensional fold[7] (FIG. 1), although their sequences did not contain recognizable similarity[8]. (Modern sequence analysis, however, would now detect their similarity.) Today, the literature is rich with celebrated cases of homology inferred from structure, including the unexpected similarity between actin and the 70-kDa heat-shock cognate protein[9], the TopRim domain shared between some topoisomerases, primases and nucleases[10,11], and the highly similar constant and variable domains of immunoglobulins. Indeed, most evolutionary relationships cannot be detected from sequence[12].

In addition, the three-dimensional structure of a protein can yield direct insight into its molecular mechanism. For example, the structure of the TATA-box-binding protein (TBP) when it is bound to DNA provides not only a sense of how these molecules interact in general, but also some fascinating clues about DNA-binding specificity. Furthermore, structural understanding of recognition mechanisms in major histocompatibility complex molecules and T-cell receptors helped to make immunology comprehensible at a molecular level[13,14]. Structural genomics efforts plan to extend structural insight to a broad repertoire of proteins, using large-scale high-throughput techniques[15–26].

While the term 'structural genomics' is sometimes loosely used to encompass disparate large-scale efforts to determine protein structure, by international agreement it has come to have a relatively specific meaning (see link to the Airlie Agreement for 'Agreed Principles



e

```
            4hhba  VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.DLS.....HGSAQVKGHGKKVA
            1mbd_  VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVL

4hhba DALTNAVAHVD..DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR......
1mbd_ TALGAILKK.K.GHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
```

Figure 1 | **Structure similarity without sequence similarity.** The first two protein structures that were solved — sperm-whale myoglobin and horse haemoglobin — were recognizable as homologues even at low resolution, even though their sequences were more different than similar. **a** | Papier mâché model of sperm-whale myoglobin. **b** | Baked and painted foam model of horse haemoglobin. Modern representations of these structures clearly show the areas of structural similarity (highlighted in red in **c** and **d**). **c** | Myoglobin (Protein Data Bank (PDB) code 1mbd)[117]. **d** | Human haemoglobin (PDB code 4hhb)[118]. **e** | Alignment of horse myoglobin and human α-haemoglobin sequences[119] shows little sequence similarity. Photos taken of the structures at the MRC Laboratory of Molecular Biology by S.E.B. Computer images were generated using Rasmol[120], Molscript[121] and Raster3D[122].

and Procedures'). In this more purist sense, structural genomics is an effort to create a representative set of experimental macromolecular structures, which will be augmented by computational methods to provide model structures for most tractable macromolecules. Although this reflects a primary focus on surveying the structures of different families, agreed goals of structural genomics include the study of biologically interesting molecules, such as those from model organisms and those with medical importance. In addition, structural genomics specifically aims to derive function from the structures.

Because structural genomics is in its infancy, its course might change over the next several years; indeed, the experiences of the current pilot centres will inform future directions. However, the relatively precise definition of structural genomics includes several hints about the limitations and scope of the field. For example, structural genomics efforts often study individual protein domains, rather than whole proteins or complexes, because domains are the fundamental units of protein structure and evolution. For the time being, proteins and other macromolecules that are not tractable for high-throughput characterization will largely be left unconsidered by structural genomics efforts. Over time, however, the gamut of molecules suitable for large-scale studies is likely to increase; one can already imagine what structural genomics of RNA might involve[27], although no such projects are underway at present. Moreover, rather than solving the structures of all domains, the general intent at present is to solve experimentally the structure of one representative domain from each family, and use computational comparative modelling to provide the COORDINATES for related proteins. In this way, current structural genomics is a conjoined experimental and computational effort, which expects to provide a comprehensive repertoire of models of soluble globular-protein domains. This review outlines how proteins are selected for structural genomics and how they are experimentally characterized in a typical pilot centre, discusses some early results, and suggests what they might mean for the future of the field.

The process

The principles of experimental structural genomics are largely the same as those for traditional structural biology, but differ in motivation, automation and scale. The key to the success of this scientific venture is the ability to optimize the structure-determination process, so as to reap economies of scale as centres increase their throughput.

COORDINATES
A set of numbers that specify the *X*, *Y* and *Z* positions for each atom in a protein. Together, they describe the molecular structure.

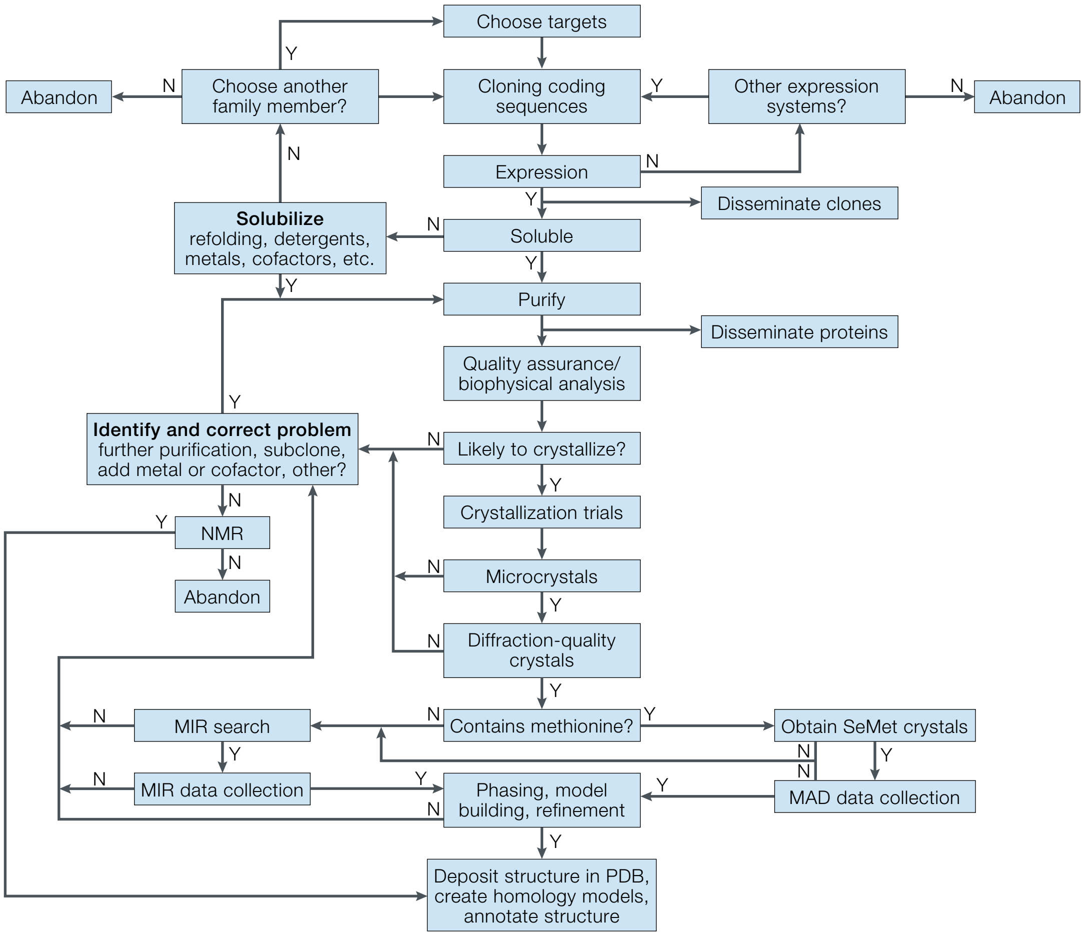

Choose targets
Y
N
Abandon
Choose another family member?
Cloning coding sequences
Y
Other expression systems?
N
Abandon
N
Expression
N
Y
Disseminate clones
Solubilize refolding, detergents, metals, cofactors, etc.
N
Soluble
Y
Y
Purify
Disseminate proteins
Quality assurance/ biophysical analysis
Y
Identify and correct problem further purification, subclone, add metal or cofactor, other?
N
Likely to crystallize?
Y
N
Y
NMR
Crystallization trials
N
Abandon
N
Microcrystals
Y
N
Diffraction-quality crystals
Y
N
MIR search
N
Contains methionine?
Y
Obtain SeMet crystals
Y
N
Y
N
N
MIR data collection
Y
Phasing, model building, refinement
N
Y
MAD data collection
Y
Deposit structure in PDB, create homology models, annotate structure

Figure 2 | **Processes involved in high-throughput structural genomics using X-ray crystallography.** N indicates that a process has failed and Y that it has succeeded. (MIR, multiple isomorphous replacement; an alternative to multiple anomolous dispersion (MAD) phasing for structure determination; NMR, nuclear magnetic resonance; SeMet, selenomethionine.) (Modified with permission from REF. 16.)

Experimental structural genomics faces no single bottleneck to overcome: nearly every stage of the process needs to be refined and optimized. Moreover, many individual proteins are expected to be intractable without specialized extensive effort. Therefore, parallel studies on related proteins are being relied on to increase the likelihood of readily solving a structure for a family of proteins. The progress of individual protein targets through the experimental process will be like a funnel, with many targets starting at the same time, and a fraction failing at each stage of the process. The slope of the funnel is dependent on the effort devoted at each step, which is, in turn, a consequence of the specific motivations of the particular structural genomics centre.

Although the detailed processes of scaling up the procedures involved in structure determination are unique to each centre for structural genomics, several characteristics are shared among most centres (FIG. 2). The experimental process begins with the cloning of selected target sequences, frequently with recombination-based vectors that allow the creation of many different constructs. These vectors incorporate different affinity tags, such as HIS-TAGS and glutathione-*S*-transferase (GST), to aid purification, as well as promoters that allow trials in different expression systems[28]. Expressing high levels of soluble protein is a particular challenge, so there is considerable interest in fusions between the target protein and green fluorescent protein that fluoresce only when soluble and folded, therefore indicating folded proteins in solution[29]. Cell-free expression systems hold great promise for improving yields and allowing the production of toxic proteins[30]. Another optimization is the use of hyperthermophilic proteins, which are easier to purify when expressed in MESOPHILIC hosts, as they are resistant to heat that will denature most of the proteins of the host.

The expressed proteins might have their domain boundaries identified by proteolysis and mass spectrometry, and several groups subject samples to DYNAMIC LIGHT SCATTERING to detect when proteins have formed heterogeneously sized oligomers that are unlikely to crystallize. In some centres, the proteins are studied by a heteronuclear single-quantum coherence nuclear magnetic resonance (HSQC NMR) experiment, because this technique gives insight into the 'foldedness' of a protein[31,32]. Any promising purified soluble proteins are then subjected to crystallization trials or NMR experiments.

HIS-TAG
A series of histidine residues fused to a protein that aids protein purification because of its strong binding to nickel columns.

MESOPHILE
An organism that grows at moderate temperature.

DYNAMIC LIGHT SCATTERING
A technique for determining apparent molecular size, in which laser light is shone on a solution. Its scatter corresponds to the diffusion rate and, therefore, the size of the molecules in solution.

Table 1 | **Centres that are undertaking structural genomics projects**

Centre	Leader	Key ideas	Website	Reference
Berkeley Structural Genomics Center	Sung-Hou Kim	Complete structural genomics of *M. genitalium* and *M. pneumoniae*	http://www.strgen.org/	–
Joint Center for Structural Genomics	Ian Wilson	Large-scale automation; proteins from *T. maritima* and *C. elegans*	http://www.jcsg.org/	–
Midwest Center for Structural Genomics	Andrzej Joachimiak	Novel protein folds and technology development	http://www.mcsg.anl.gov/	
New York Structural Genomics Research Consortium	Stephen Burley	Yeast proteins with novel folds; technology development	http://www.nysgrc.org/	–
Northeast Structural Genomics Consortium	Gaetano Montelione	Complementarity of NMR and crystallography; coverage of structure space	http://www.nesg.org/	–
Southeast Collaboratory for Structural Genomics	Bi-Cheng Wang	Development of SAD technology; *P. furiosis*, *H. sapiens* and *C. elegans* proteins	http://www.secsg.org/	–
TB Structural Genomics Consortium	Thomas Terwilliger	*M. tuberculosis* proteins; new folds; large-scale collaboration	http://www.doe-mbi.ucla.edu/TB/	–
Structure to Function	Roberto J. Poljak	Functional characterization of *H. influenzae* proteins	http://s2f.carb.nist.gov/	123
Ontario Structural Proteomics Group	Aled Edwards	High-throughput; experimental target selection	http://www.uhnres.utoronto.ca/proteomics/	31
Protein Folds Project	Shigeyuki Yokoyama	NMR of proteins from mouse full-length cDNAs	http://www.rsgi.riken.go.jp/	30
Structurome Project	Seiki Kuramitsu	Complete structural genomics of *T. thermophilus* HB8	http://www.rsgi.riken.go.jp/	39
Protein Structure Factory	Udo Heinemann	Technology development; human proteins	http://userpage.chemie.fu-berlin.de/~psf/	124
StructuralGenomiX	Tim Harris	Company: structures relevant to medicine	http://www.stromix.com/	125
Syrrx	Wendell Wierenga	Company: structures relevant to medicine	http://www.syrrx.com/	–

C. elegans, *Caenorhabditis elegans*; *H. influenzae*, *Haemophilus influenzae*; *H. sapiens, Homo sapiens*; *M. genitalium*, *Mycoplasma genitalium*; *M. pneumoniae*, *Mycoplasma pneumoniae*; *M. tuberculosis*, *Mycobacterium tuberculosis*; *P. furiosus*, *Pyrococcus furiosus; T. maritima*, *Thermotoga maritima*; *T. thermophilus*, *Thermus thermophilus*; NMR, nuclear magnetic resonance; TB, tuberculosis; SAD, single wavelength anomalous diffraction.

Several centres are investing in considerable automation to allow parallel large-scale expression trials and parallel crystallization trials (TABLE 1); for example, the Joint Center for Structural Genomics hopes to be able to analyse up to 130,000 crystallization experiments per day[33]. To ensure optimal use of precious SYNCHROTRON time, BEAMLINE AUTOMATION is crucial[34]. In addition, careful tracking of laboratory results and analyses can be used to predict better which proteins will be most successful[35]; this information might then be fed into the target-selection process to improve future results.

Crystallography has benefited from many technologies, including the brilliance of synchrotron radiation and its tunability for multiple anomalous dispersion (MAD) PHASING[36]. Other improvements include charged coupled device detectors, as well as the enhanced stability provided by cryocrystallography. NMR has seen similar advances, including cryogenic probes and higher-field magnets, as well as new techniques such as transverse relaxation-optimized spectroscopy (TROSY)[32,37]. Consequently, although early plans for structural genomics focused primarily on crystallography, NMR has already proved to have great value for the field[32,38]. At this time, most centres in the United States have NMR spectroscopists, and roughly half of the structural genomics effort in Japan use NMR[39,30].

The refinement of crystallographic structures has been reported to be the slowest step in structure determination (S.-H. Kim, personal communication), and the advent of highly automated structure-determination software for both crystallography[40,41] and NMR[42,43] is therefore likely to have a marked effect on increasing the speed of solution of structures.

Target selection: which proteins and how many?

It would be desirable to have an experimental molecular structure for every known protein, such as the ~600,000 in the protein sequence databases SWISS-PROT and TrEMBL[44]. However, practicalities dictate a compromise, whereby a more modest number of structures are solved, and these are used as templates for the comparative modelling of most soluble protein domains. A rough consensus indicates that it could be feasible for 10,000 structures to be experimentally solved over the next decade[45].

Dennis Vitkup and colleagues have shown that this number of experimental structures is insufficient to provide templates for high-quality models of all protein domains[46]. To determine how many structure

SYNCHROTRON
A device that accelerates particles of atomic size through an electric field; it is used to produce synchronous packets of particles.

BEAMLINE AUTOMATION
Technologies to reduce human intervention on synchrotron beamlines, such as robots for mounting and centring crystals in the X-ray beam.

MAD PHASING
(Multiple anomolous dispersion). An approach to determining the phases of a crystal structure by relying on the anomalous scattering of X-rays near the absorption edge of the atom (such as selenium). It allows determination of phase from several sets of data collected from a single crystal.

Box 1 | Who is doing structural genomics?

There are seven comprehensive pilot centres and one programme project that are funded by the National Institute of General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH)[33] (TABLE 1), and three more centres might be funded from pending applications. Large centres for structural genomics have also been established in Japan, Germany and Canada. In addition to these main funded centres, there are smaller programmes underway in the above countries, as well as France, Sweden, Australia, Israel and China. Funding has also been approved for programmes in Switzerland and Italy[107]. In the United Kingdom, the Wellcome Trust has proposed the formation of an industry-funded organization comparable with the SNP Consortium (Single Nucleotide Polymorphism Consortium), which would promptly release structures to the public[108,109]. Several companies are also involved in structural genomics, and two in particular — StructuralGenomiX and Syrrx — aim to solve numerous structures.

determinations are necessary to provide good three-dimensional models for all of the 1,626 non-membrane families in the Pfam database[47] (a collection of well-characterized protein-domain sequences), Vitkup clustered the sequences into groups with more than 30% identity. This produced 13,000 clusters, each requiring a structure determination. Even under the optimistic assumption that sequences outside Pfam belong to similarly large families, extrapolation shows that 64,000 structure determinations would be needed to provide structures for all soluble domains. However, if the goal is relaxed to provide models for 90% of all protein domains, then ~16,000 structure determinations might suffice. Vitkup points out that this reduction only holds if structural genomics efforts are optimally coordinated to solve structures from the largest families. In practice, smaller families will often be targeted because of their identified biological or medical importance, considerably increasing the number of structures required.

The number of requisite structures can be reduced greatly by relaxing the requirements for the quality of the model; for example, only one structure would be needed for each Pfam family if it were sufficient to know the fold-type of each protein without building a detailed coordinate model. The importance of cooperation between structural genomics centres is also evident, as in two instances already, independent groups have inadvertently solved the structures of homologous proteins[48–51]. Although determining the structures of highly homologous proteins often has great value to structural biology, it runs counter to the goals of structural genomics. To avoid future duplicated efforts, the structural genomics community has agreed to a set of principles and procedures for coordination[52] (see link to the Airlie Conference on Structural Genomics), which includes the sharing of lists of target proteins.

At present, each structural genomics centre (BOX 1) chooses protein targets using its own distinct criteria. The Ontario Structural Proteomics project, for example, aims to pursue those proteins that are experimentally most tractable, whereas the Berkeley Structural Genomics Center is pursuing a nearly complete repertoire of proteins from two *Mycoplasma* spp. Several centres focus on finding new topological protein folds. Nonetheless, general features of the target selection are common to all centres[53] (FIG. 3). First, proteins of interest are defined, and — to the extent possible — these proteins are typically divided into their constituent domains, as individual structural modules are more conducive to high-throughput studies[54]. At this point, domains are identified primarily on the basis of visual inspection of multiple-sequence alignments and comparison with well-described domains[47,55–58], but many automated approaches are being developed to incorporate alignment and other information. It has also been suggested that pairs of domains will be good targets[26]. This is because, out of the huge number of possible domain combinations, only a limited number are found to exist in individual proteins[59,60]. As these pairs often adopt regular conformations[61], solving the structures of domain pairs should provide an understanding of typical domain interaction, and give clues about overall protein structures.

All of the prospective target domains are put through a battery of computational tools; those proteins predicted to be membranous, unstructured or otherwise unsuitable are immediately removed from the pool as being intractable. Next, database searches are used, and proteins that can be computationally modelled by homology to known structures are also set aside. The remaining candidates are all valid 'structural genomics proteins,' as they are thought to be tractable, and their experimental characterization will provide structural information that could not have been predicted. Priority is assigned to families of structural genomics proteins according to their desirable characteristics[62], such as phylogenetic distribution[63,64], family size[46], likelihood of producing a new fold[65,66] and functional relevance[67].

The selected families contain the original candidate target, but often that protein will not be among those chosen for experimental characterization. Instead, in the selected families, individual proteins are chosen for study on the basis of their suitability for experimental characterization, including features such as length, thermostability, codon usage, ISOELECTRIC POINT (pI), ability to model other structures[42] and suitability for MAD phasing. This is deliberate, with the goal of reducing experimental effort. Indeed, following the 'class-directed' approach, in most cases, several homologous targets will be studied experimentally in parallel[68,69]. This is motivated by the expectation that one protein will fortuitously prove far more tractable than the others, therefore justifying the replicated effort at the early stages of the pipeline.

Function from structure

Elucidation of function from molecular structure is perhaps the most exciting, but also probably the least understood aspect of structural genomics[70–73]. Until recently, only proteins with well-characterized functions were candidates for structure determination. Structural genomics turns that logic on its head by using the structure to infer function. Although some basic principles for this process have been shown to be successful, the extent to which different approaches will prove valuable remains to be seen.

TROSY
(Transverse relaxation-optimized spectroscopy). A nuclear magnetic resonance technique that reduces the deterioration of signal from large proteins. It allows large proteins to be studied in high-field magnets.

ISOELECTRIC POINT
The pH at which a protein has zero net charge.

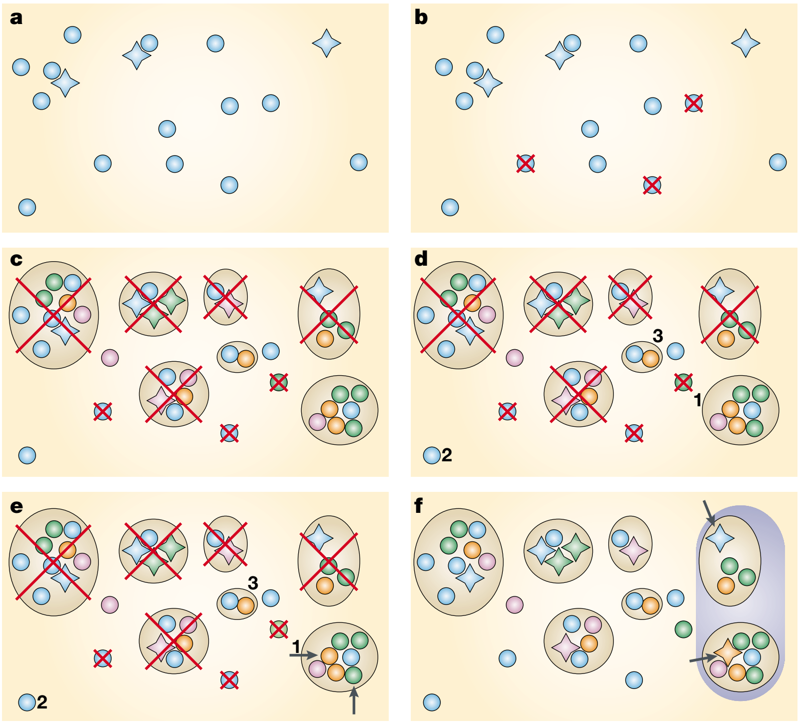


Figure 3 | **Target selection for structural genomics. a** | Proteins in the realm of interest (in this case a genome) are plotted as blue shapes in an arbitrary sequence space. Proteins of known structure are shown as stars, others as circles. **b** | Transmembrane proteins and those with low complexity are excluded, as indicated by a red cross. **c** | Homologues from other organisms (different colours) are identified and family relationships are determined (ovals). Families with a member of known structure are excluded, as indicated by a red cross. **d** | Priority is assigned to families. In this case, a pervasive taxonomically diverse family is ranked highest. **e** | Two proteins in the highest-priority family are chosen (arrows); note that they are not one of the original proteins of interest (blue), but they are homologous to such a protein. **f** | The solved structure is similar to, and homologous to, another structure that was previously known (arrows). This means that all of the proteins in the two families are homologous (indicated by the blue enclosure), and it might therefore be possible to make useful functional inferences.

The key idea behind deducing function from structure is that protein structure is better conserved than sequence, and structure therefore provides a way of homology database searching that is more sensitive than sequence comparison. Hence, the logical first step in analysing a newly solved structural genomics protein is a structure comparison with the Protein Data Bank (PDB)[74], a database of known structures, using any of various popular tools[55,75–78]. However, none of these methods is guaranteed to find true matches in the database, and any of them can report high scores for evolutionarily unrelated proteins. Moreover, structural similarity alone is insufficient to determine whether two proteins are homologous, because they could have evolved by convergence to have the same structure.

As a consequence, it is also necessary to inspect the structures visually, and to provide expert judgement on whether there is similarity indicative of common ancestry. The primary aids for this task are databases such as SCOP[56] (structural classification of proteins) and CATH[57] (class, architecture, topology and homologous superfamily). These provide comprehensive hierarchical classifications of all known protein domains, with considerable manual review and annotation. The SCOP classification in particular is founded on using structure, along with functional and mechanistic information, to organize proteins according to their distant evolutionary relationships.

It remains to be seen to what extent new experimental work from structural genomics reveals recognizably homologous proteins. Extrapolations from historical structure determinations of proteins that could have been candidates for structural genomics indicate that ~45% of structural genomics proteins would be homologous to known proteins[79], and that 25–28% would have a new fold[79,80]. This trend seems to be roughly followed: in a small sample of 32 such domains that were recently solved, Teichmann and colleagues report that 34% are homologous and that 37% adopt a new fold, whereas the remainder are structurally similar to those seen before, but are not evolutionarily related[26].

In many cases, the homology that is inferred from structure has allowed interesting functional assignments to be made. For example, a hypothetical *Saccharomyces cerevisiae* protein was found to be a triosephosphate isomerase (TIM) barrel, the active site of which looks like alanine racemase, and preliminary studies indicate that it does have that biochemical activity[16]. However, homology has not proved to be definitive; indeed, of the ten structures solved by Christendat and co-workers, in no case did structurally inferred homology alone provide a robust functional prediction[31]. In several cases, common ancestry inferred from structure has not reflected common function; for example, *Methanobacterium thermoautotrophicum* MTH538 closely resembles the *Escherichia coli* response regulator CheY, but could not be shown to have any related aspartate-kinase activity[81]. Furthermore, two close homologues of unknown molecular function — YjgF from *E. coli* and YabJ from *Bacillus subtilis* — were both found to be similar in structure to chorismate mutase. However, the completely different active sites precluded the possibility of these proteins sharing chorismate-mutase function with their structurally similar homologue[50,51]. So, although structural analysis failed to show the role of YjgF and YabJ, it was key in allowing the researchers to realize that their homology did not reflect similar activity. Structure determination of *M. thermoautotrophicum* MTH1175 likewise showed structural similarity to *E. coli* RNaseH, but did not support a shared function between the two[82]. Because active sites can occur in different contexts and can change in homologous proteins, several automated methods have been developed to seek similarity in active sites to predict function[83–85] or specificity[86,87]; however, the application of these methods has not yet been described for the handful of published structural genomics proteins.

One of the more startling findings of structural genomics is that structures can often be functionally interpreted even when their folds are novel. For example, the discovery of a long, positively charged groove on the surface of the mouse tubby protein allowed Boggon and colleagues to postulate that it is a DNA-binding protein[88]. The structure also showed that all but one of the tubby mutations responsible for retinitis pigmentosa

Box 2 | Where are the structural genomics data?

The Airlie Agreement on structural genomics[52] specifies that, within 6 months of its completion, each protein structure will be deposited in the Protein Data Bank[110], a repository of all publicly solved structures. Structural and evolutionary relationships between these proteins can be found in the SCOP[56] and CATH[111] databases, whereas Dali provides automated structure comparison[55].

Lists of targets can be found on the websites of individual centres, as can information about protein production. Compendiums of targets and searching facilities can also be found in the PRESAGE database[112] and on the structuralgenomics.org website. The PRESAGE database also provides information about structure predictions, such as those in ModBase[113] and other fold predictions[114,115,113,116].

type 14 are found in a small region of the groove, even though they are dispersed within the sequence. Some of these replace positive amino acids with neutral ones, strengthening the hypothesis that surface charge is important. So, not only did three-dimensional structure provide insight into the molecular function of tubby, but also it helped to explain disease-causing mutations.

The *E. coli* YrdC protein was similarly found to have a concave surface with positive electrostatic potential, which led to experiments showing preferential RNA binding[89]. In another instance of structure directly indicating function, the *Methanococcus jannaschii* MJ0226 and *B. subtilis* Maf proteins established new structural superfamilies, although their structure was reminiscent of nucleotide-binding folds. Further tests showed that MJ0226 hydrolyses non-standard nucleotides[90,91]. In these cases, the functional inference would have been missed by all tools available at present; only the expertise of the structural biologists allowed these functional interpretations.

With surprising frequency, unexpected ligands identified in the crystal structure have also indicated the function of structural genomics proteins. Clues about the molecular mechanism of the proteins MTH150 and MTH152 (from *M. thermoautotrophicum*), HI0139 (from *Haemophilus influenzae*), and MJ0577 (from *M. jannaschii*) were shown by their co-crystallization with NAD$^+$, FMN (flavin mononucleotide), a selenium version of *S*-adenosyl-*L*-homocysteine, and ATP, respectively[31,92,93]. In each case, binding was sufficiently strong that the protein apparently scavenged the cofactor from the original expression system. In further tests, MJ0577 was found to hydrolyse ATP to ADP only in the presence of extract from its source organism, *M. jannaschii*, indicating that it might be a molecular switch[93].

Comparative modelling allows each experimentally determined fold to provide structure information for a family of related proteins[94]. The quality of the model can range from extremely good to virtually worthless, depending on the intended use and the evolutionary distance between the template (solved structure) and the query, in large part because of problems with alignment[95,96]. Because the structural information of most proteins will be available only as a homology model, understanding the strengths and limitations of the comparative modelling methods will be crucial for making informed use of structural genomics data.

A fundamental limitation of structural genomics is that it typically only provides clues about molecular function[6], such as what a protein binds to or reacts with. Understanding this molecular function gives only limited insight into the cellular role. This limitation is endemic to homology-based methods and is therefore shared with sequence comparison. Fortunately, many other functional genomics approaches, especially expression profiling, yield precisely complementary data: although they cannot indicate the molecular action of a protein, they provide clues about its role in a wider context, such as in a signalling pathway or a cellular state.

Beyond structural genomics

Structural genomics will revolutionize biochemistry and molecular biology, making pervasive the use of three-dimensional structure information. Just as one can expect to find sequences for most genes of interest in public databases, structural genomics promises to offer a comparably comprehensive library of experimental and computational models (BOX 2). These will reveal new functions, indicate molecular mechanisms and explicate mutations.

Despite its promise, current structural genomics will not provide a perfect resource. Most membrane proteins and RNA structures[27] will probably be left unsolved for the time being, as will proteins without a defined structure[97–99]. Moreover, although most important families will have representative structures, rare unusual families with no known functional import are unlikely to be characterized soon. Finally, although structural genomics focuses on a complete repertoire of static individual domains of proteins, it fails to capture their interactions, complexes and dynamics at present.

Even as structural genomics provides a solid foundation for the future of structural biology research, its limitations leave much exciting work to be done. Improvements in sequence analysis[100] and comparative modelling will yield disproportionate enhancements in the number and quality of modelled structures. Likewise, building from the repertoire of known structures, computational methods using limited experimental data[101–103] and *ab initio* approaches[104] should help to fill in knowledge of domains beyond the resources of fully experimental approaches. The technology developed for structural genomics is also expected to provide a watershed for studies of those macromolecules not suited for high-throughput studies, by providing the means to rapidly explore several expression constructs and screen through many purification and crystallization protocols. It will also allow for parallel studies of homologues, such as all human kinases, to understand their specificity. In addition, structural genomics will provide a platform for detailed studies on molecular dynamics and interactions[105], and for the elucidation of large macromolecular complexes by X-ray crystallography and electron microscopy[106]. In this way, even as structural genomics brings our knowledge of protein-domain structures near to completion, it is a prelude to a still richer knowledge of molecular structure and function.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Devos, D. & Valencia, A. Practical limits of function prediction. *Proteins* **41**, 98–107 (2000).
4. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
5. Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
6. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
7. Perutz, M. F. *et al.* Structure of hæmoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* **185**, 416–422 (1960).
8. Kendrew, J. C. & Watson, H. C. Comparison between amino-acid sequences of sperm whale myoglobin and of human haemoglobin. *Nature* **190**, 670 (1961).
9. Flaherty, K. M., McKay, D. B., Kabsch, W. & Holmes, K. C. Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. *Proc. Natl Acad. Sci. USA* **88**, 5041–5045 (1991).
10. Aravind, L., Leipe, D. D. & Koonin, E. V. Toprim — a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res.* **26**, 4205–4213 (1998).
11. Berger, J. M., Fass, D., Wang, J. C. & Harrison, S. C. Structural similarities between topoisomerases that cleave one or both DNA strands. *Proc. Natl Acad. Sci. USA* **95**, 7876–7881 (1998).
12. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA* **95**, 6073–6078 (1998).
13. Bjorkman, P. J. *et al.* Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **329**, 506–512 (1987).
14. Wilson, I. A. & Garcia, K. C. T-cell receptor structure and TCR complexes. *Curr. Opin. Struct. Biol.* **7**, 839–848 (1997).
15. Blundell, T. L. & Mizuguchi, K. Structural genomics: an overview. *Prog. Biophys. Mol. Biol.* **73**, 289–295 (2000).
16. Burley, S. K. *et al.* Structural genomics: beyond the human genome project. *Nature Genet.* **23**, 151–157 (1999).
17. Domingues, F. S., Koppensteiner, W. A. & Sippl, M. J. The role of protein structure in genomics. *FEBS Lett.* **476**, 98–102 (2000).
18. Gaasterland, T. Structural genomics: bioinformatics in the driver's seat. *Nature Biotechnol.* **16**, 625–627 (1998).
19. Kim, S. H. Shining a light on structural genomics. *Nature Struct. Biol.* **5**, 643–645 (1998).
20. Mittl, P. R. & Grutter, M. G. Structural genomics: opportunities and challenges. *Curr. Opin. Chem. Biol.* **5**, 402–408 (2001).
21. Montelione, G. T. & Anderson, S. Structural genomics: keystone for a Human Proteome Project. *Nature Struct. Biol.* **6**, 11–12 (1999).
22. Sali, A. 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032 (1998).
23. Shapiro, L. & Lima, C. D. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure* **6**, 265–267 (1998).
24. Smith, T. A new era. *Nature Struct. Biol.* **7**, 927 (2000).
    The introduction to a supplement to *Nature Structural Biology* devoted to structural genomics, which contains 20 articles that address different aspects of the field.
25. Teichmann, S. A., Chothia, C. & Gerstein, M. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390–399 (1999).
26. Teichmann, S. A., Murzin, A. G. & Chothia, C. Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* **11**, 354–363 (2001).
    This review includes an analysis of 32 structural genomics proteins and presents lessons learned in each case.
27. Doudna, J. A. Structural genomics of RNA. *Nature Struct. Biol.* **7**, 954–956 (2000).
28. Edwards, A. M. *et al.* Protein production: feeding the crystallographers and NMR spectroscopists. *Nature Struct. Biol.* **7**, 970–972 (2000).
29. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnol.* **17**, 691–695 (1999).
30. Yokoyama, S. *et al.* Structural genomics projects in Japan. *Prog. Biophys. Mol. Biol.* **73**, 363–376 (2000).
31. Christendat, D. *et al.* Structural proteomics of an archaeon. *Nature Struct. Biol.* **7**, 903–909 (2000).
    Describes the determination of ten protein structures from *M. thermoautotrophicum*, using the principle of finding proteins that are most amenable to structural characterization.
32. Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C. & Szyperski, T. Protein NMR spectroscopy in structural genomics. *Nature Struct. Biol.* **7**, 982–985 (2000).
33. Terwilliger, T. C. Structural genomics in North America. *Nature Struct. Biol.* **7**, 935–939 (2000).
34. Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. Automation of X-ray crystallography. *Nature Struct. Biol.* **7**, 973–977 (2000).
35. Bertone, P. *et al.* SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29**, 2884–2898 (2001).
36. Hendrickson, W. A. Synchrotron crystallography. *Trends Biochem. Sci.* **25**, 637–643 (2000).
37. Wider, G. & Wuthrich, K. NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr. Opin. Struct. Biol.* **9**, 594–601 (1999).
38. Prestegard, J. H., Valafar, H., Glushka, J. & Tian, F. Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* **40**, 8677–8685 (2001).
39. Yokoyama, S. *et al.* Structural genomics projects in Japan. *Nature Struct. Biol.* **7**, 943–945 (2000).
40. Adams, P. D. & Grosse-Kunstleve, R. W. Recent developments in software for the automation of crystallographic macromolecular structure determination. *Curr. Opin. Struct. Biol.* **10**, 564–568 (2000).
41. Lamzin, V. S. & Perrakis, A. Current state of automated crystallographic data analysis. *Nature Struct. Biol.* **7**, 978–981 (2000).
42. Helgstrand, M., Kraulis, P., Allard, P. & Hard, T. Ansig for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra. *J. Biomol. NMR* **18**, 329–336 (2000).
43. Zimmerman, D. E. *et al.* Automated analysis of protein NMR assignments using methods from artificial intelligence. *J. Mol. Biol.* **269**, 592–610 (1997).
44. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
45. Norvell, J. C. & Machalek, A. Z. Structural genomics programs at the US National Institute of General Medical Sciences. *Nature Struct. Biol.* **7**, 931 (2000).
46. Vitkup, D., Melamud, E., Moult, J. & Sander, C. Completeness in structural genomics. *Nature Struct. Biol.* **8**, 559–566 (2001).
    This paper predicts the number of structure determinations necessary to provide three-dimensional models of all (or most) families of proteins.
47. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).
48. Kim, K. K., Hung, L. W., Yokota, H., Kim, R. & Kim, S. H. Crystal structures of eukaryotic translation initiation factor 5A from *Methanococcus jannaschii* at 1.8 Å resolution. *Proc. Natl Acad. Sci. USA* **95**, 10419–10424 (1998).
    A report of one of the first structural genomics proteins solved; it represented inadvertent duplication of effort, as the same structure was independently solved in the next reference.
49. Peat, T. S., Newman, J., Waldo, G. S., Berendzen, J. & Terwilliger, T. C. Structure of translation initiation factor 5A from *Pyrobaculum aerophilum* at 1.75 Å resolution. *Structure* **6**, 1207–1214 (1998).
50. Sinha, S. *et al.* Crystal structure of *Bacillus subtilis* YabJ, a purine regulatory protein and member of the highly conserved YjgF family. *Proc. Natl Acad. Sci. USA* **96**, 13074–13079 (1999).
51. Volz, K. A test case for structure-based functional assignment: the 1.2 Å crystal structure of the YjgF gene product from *Escherichia coli*. *Protein Sci.* **8**, 2428–2437 (1999).
52. Smaglik, P. Protein structure groups seek to draft common ground rules. *Nature* **403**, 691 (2000).
53. Brenner, S. E. Target selection for structural genomics. *Nature Struct. Biol.* **7**, 967–969 (2000).
54. Kuroda, Y., Tani, K., Matsuo, Y. & Yokoyama, S. Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics. *Protein Sci.* **9**, 2313–2321 (2000).
55. Dietmann, S. *et al.* A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29**, 55–57 (2001).
    An introduction to one of the most popular systems for automatically comparing proteins of known structure.
56. 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).
    The SCOP database is a comprehensive expert-curated hierarchical evolutionary classification of protein domains using structural information.
57. Pearl, F. M. *et al.* A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res.* **29**, 223–227 (2001).
    An introduction to CATH, a largely automated hierarchical classification of protein domain structures.
58. Siddiqui, A. S., Dengler, U. & Barton, G. J. 3Dee: a database of protein structural domains. *Bioinformatics* **17**, 200–201 (2001).
59. Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001).
60. Apic, G., Gough, J. & Teichmann, S. A. An insight into domain combinations. *Bioinformatics* **17** (Suppl. 1), S83–S89 (2001).
61. Saha, S. *et al.* Solution structure of the LDL receptor EGF-AB pair. A paradigm for the assembly of tandem calcium binding EGF domains. *Structure* **9**, 451–456 (2001).
62. Gerstein, M. Integrative database analysis in structural genomics. *Nature Struct. Biol.* **7**, 960–963 (2000).
63. Fischer, D. Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. *Protein Eng.* **12**, 1029–1030 (1999).
    This paper describes interesting features of genes without homologues and the ability of structural genomics to elucidate their provenance.
64. Galperin, M. Y. Conserved 'hypothetical' proteins: new hints and new puzzles. *Comp. Funct. Genomics* **2**, 14–18 (2001).
65. Linial, M. & Yona, G. Methodologies for target selection in structural genomics. *Prog. Biophys. Mol. Biol.* **73**, 297–320 (2000).
66. Mallick, P., Goodwill, K. E., Fitz-Gibbon, S., Miller, J. H. & Eisenberg, D. 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 (2000).
67. Erlandsen, H., Abola, E. E. & Stevens, R. C. Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites. *Curr. Opin. Struct. Biol.* **10**, 719–730 (2000).
68. Lewis, H. A. *et al.* A structural genomics approach to the study of quorum sensing. Crystal structures of three LuxS orthologs. *Structure* **9**, 527–537 (2001).
69. Terwilliger, T. C. *et al.* Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci.* **7**, 1851–1856 (1998).
70. Shapiro, L. & Harris, T. Finding function through structural genomics. *Curr. Opin. Biotechnol.* **11**, 31–35 (2000).
71. Skolnick, J., Fetrow, J. S. & Kolinski, A. Structural genomics and its importance for gene function analysis. *Nature Biotechnol.* **18**, 283–287 (2000).
72. Thornton, J. M. From genome to function. *Science* **292**, 2095–2097 (2001).
73. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. From structure to function: approaches and limitations. *Nature Struct. Biol.* **7**, 991–994 (2000).
74. Berman, H. M. *et al.* The Protein Data Bank and the challenge of structural genomics. *Nature Struct. Biol.* **7**, 957–959 (2000).
75. Gibrat, J. F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385 (1996).
76. Orengo, C. A. & Taylor, W. R. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635 (1996).
77. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747 (1998).
78. Subbiah, S., Laurents, D. V. & Levitt, M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* **3**, 141–149 (1993).
79. Brenner, S. E. & Levitt, M. Expectations from structural genomics. *Protein Sci.* **9**, 197–200 (2000).
    Uses historical data to predict the fraction of new folds and new superfamilies to be discovered by structural genomics.
80. Koppensteiner, W. A., Lackner, P., Wiederstein, M. & Sippl, M. J. Characterization of novel proteins based on known protein structures. *J. Mol. Biol.* **296**, 1139–1152 (2000).
81. Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H. & Kennedy, M. A. Structure-based functional classification of hypothetical protein MTH538 from *Methanobacterium*

thermoautotrophicum. *J. Mol. Biol*. **302**, 189–203 (2000).

82. Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H. & Kennedy, M. A. NMR structure determination and structure-based functional characterization of conserved hypothetical protein MTH1175 from *Methanobacterium thermoautotrophicum*. *J. Struct. Funct. Genomics* **1**, 15–25 (2001).
83. Fetrow, J. S., Godzik, A. & Skolnick, J. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711 (1998).
84. Wallace, A. C., Borkakoti, N. & Thornton, J. M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323 (1997).
85. Wei, L. & Altman, R. B. Recognizing protein binding sites using statistical descriptions of their 3D environments. *Pac. Symp. Biocomput.* **4**, 497–508 (1998).
86. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
87. Sowa, M. E. *et al.* Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nature Struct. Biol.* **8**, 234–237 (2001).
88. Boggon, T. J., Shan, W. S., Santagata, S., Myers, S. C. & Shapiro, L. Implication of tubby proteins as transcription factors by structure-based functional analysis. *Science* **286**, 2119–2125 (1999).
    This paper predicts the DNA-binding function of tubby proteins on the basis of examination of the surface electrostatics of the structure.
89. Teplova, M. *et al.* The structure of the YrdC gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. *Protein Sci.* **9**, 2557–2566 (2000).
90. Hwang, K. Y., Chung, J. H., Kim, S. H., Han, Y. S. & Cho, Y. Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nature Struct. Biol.* **6**, 691–696 (1999).
91. Minasov, G. *et al.* Functional implications from crystal structures of the conserved *Bacillus subtilis* protein Maf with and without dUTP. *Proc. Natl Acad. Sci. USA* **97**, 6328–6333 (2000).
92. Lim, K. *et al.* Crystal structure of YecO from *Haemophilus influenzae* (HI0319) reveals a methyltransferase fold and a bound *S*-adenosylhomocysteine. *Proteins* (in the press).
93. 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).
    This paper reports that a bound ATP that was found in the solved structure indicated that this hypothetical protein is a molecular switch.
94. Sanchez, R. *et al.* Protein structure modeling for structural genomics. *Nature Struct. Biol.* **7**, 986–990 (2000).
95. Friedberg, I., Kaplan, T. & Margalit, H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci.* **9**, 2278–2284 (2000).
96. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L. Jr Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* **40**, 6–22 (2000).
97. Dunker, A. K. *et al.* Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput*. 473–484 (1998).
98. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).
99. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
100. Schaffer, A. A. *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).
101. Fowler, C. A., Tian, F., Al-Hashimi, H. M. & Prestegard, J. H. Rapid determination of protein folds using residual dipolar couplings. *J. Mol. Biol.* **304**, 447–460 (2000).
102. Potts, B. C. & Chazin, W. J. Chemical shift homology in proteins. *J. Biomol. NMR* **11**, 45–57 (1998).
103. Young, M. M. *et al.* High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl Acad. Sci. USA* **97**, 5802–5806 (2000).
    In this work, cross-linking and mass spectrometry were used to glean limited structural information, sufficient to predict a protein fold.
104. Simons, K. T., Strauss, C. & Baker, D. Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199 (2001).
105. Wuthrich, K. Protein recognition by NMR. *Nature Struct. Biol.* **7**, 188–189 (2000).
106. Baumeister, W. & Steven, A. C. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.* **25**, 624–631 (2000).
107. Heinemann, U. Structural genomics in Europe: slow start, strong finish? *Nature Struct. Biol.* **7**, 940–942 (2000).
108. Butler, D. Wellcome discusses structural genomics effort with industry. . . but data release remains an open question. *Nature* **406**, 923–924 (2000).
109. Williamson, A. R. Creating a structural genomics consortium. *Nature Struct. Biol.* **7**, 953 (2000).
110. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
111. Orengo, C. A. *et al.* The CATH database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **27**, 275–279 (1999).
112. Brenner, S. E., Barken, D. & Levitt, M. The PRESAGE database for structural genomics. *Nucleic Acids Res.* **27**, 251–253 (1999).
113. Sanchez, R. & Sali, A. ModBase: a database of comparative protein structure models. *Bioinformatics* **15**, 1060–1061 (1999).
114. Huynen, M. *et al.* Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323–326 (1998).
115. Rychlewski, L., Zhang, B. & Godzik, A. Functional insights from structural predictions: analysis of the *Escherichia coli* genome. *Protein Sci.* **8**, 614–624 (1999).
116. Teichmann, S. A., Park, J. & Chothia, C. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA* **95**, 14658–14663 (1998).
117. Phillips, S. E. & Schoenborn, B. P. Neutron diffraction reveals oxygen–histidine hydrogen bond in oxymyoglobin. *Nature* **292**, 81–82 (1981).
118. Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J. Mol. Biol.* **175**, 159–174 (1984).
119. Bashford, D., Chothia, C. & Lesk, A. M. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216 (1987).
120. Sayle, R. A. & Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374 (1995).
121. Kraulis, P. J. Molscript: a program to produce both detailed and schematic plots of protein structure. *J. Appl. Crystallography* **24**, 946–950 (1991).
122. Merritt, E. A. & Bacon, D. J. Raster3d: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
123. Eisenstein, E. *et al.* Biological function made crystal clear — annotation of hypothetical proteins via structural genomics. *Curr. Opin. Biotechnol.* **11**, 25–30 (2000).
124. Heinemann, U. *et al.* An integrated approach to structural genomics. *Prog. Biophys. Mol. Biol.* **73**, 347–362 (2000).
125. Dry, S., McCarthy, S. & Harris, T. Structural genomics in the biotechnology sector. *Nature Struct. Biol.* **7**, 946–949 (2000).

Acknowledgements

This work is supported by NIH grants and a Searle Scholarship. S.E.B. is grateful to J.-M. Chandonia, L. Lo Conte and R. Peters for critical review of the manuscript.

Online Links

DATABASES

The following terms in this article are linked online to:

InterPro: http://www.ebi.ac.uk/interpro/
TIM | TopRim
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/
TBP | tubby
OMIM: http://www.ncbi.nlm.nih.gov/Omim/
retinitis pigmentosa type 14

FURTHER INFORMATION

Airlie Agreement:
http://www.nigms.nih.gov/news/meetings/airlie.html#agree
Airlie Conference:
http://www.nigms.nih.gov/news/meetings/airlie.html
CATH: http://www.biochem.ucl.ac.uk/bsm/cath_new/
Dali: http://www.ebi.ac.uk/dali/
ModBase: http://pipe.rockefeller.edu/modbase/
National Institute of General Medical Sciences (NIGMS):
http://www.nigms.nih.gov
Pfam: http://www.sanger.ac.uk/Software/Pfam/
PRESAGE: http://presage.berkeley.edu
Protein Data Bank: http://www.rcsb.org/pdb/
SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/
SNP Consortium: http://snp.cshl.org
Structuralgenomics.org: http://www.structuralgenomics.org
SWISS-PROT and TrEMBL: http://www.expasy.ch/sprot/