Biological function made crystal clear – annotation of hypothetical proteins via structural genomics

Edward Eisenstein*, Gary L Gilliland*, Osnat Herzberg*, John Moult*, John Orban*, Roberto J Poljak*, Linda Banerjei†, Delwood Richardson† and Andrew J Howard‡

Many of the gene products of completely sequenced organisms are 'hypothetical' – they cannot be related to any previously characterized proteins – and so are of completely unknown function. Structural studies provide one means of obtaining functional information in these cases. A 'structural genomics' project has been initiated aimed at determining the structures of 50 hypothetical proteins from *Haemophilus influenzae* to gain an understanding of their function. Each stage of the project – target selection, protein production, crystallization, structure determination, and structure analysis – makes use of recent advances to streamline procedures. Early results from this and similar projects are encouraging in that some level of functional understanding can be deduced from experimentally solved structures.

Addresses
*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850, USA
†The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA
‡Center for Synchrotron Radiation Research and Instrumentation, Biological, Chemical and Physical Sciences Department, Illinois Institute of Technology, Chicago, IL 60616, USA
Correspondence: Edward Eisenstein

Current Opinion in Biotechnology 2000, **11**:25–30

0958-1669/00/$ – see front matter

Abbreviations

HI	*Haemophilus influenzae*
MAD	multiple wavelength anomalous diffraction
ORF	open reading frame

Introduction

Recent developments in automated techniques for DNA sequencing have led to an explosion of information on the sequences of the genomes of several organisms. Complete genomic sequences of over two-dozen microorganisms and two eukaryotes are available now, and soon the genomes of several dozen additional organisms will be completed [1•]. A striking observation that has been made as each organism's genome is analyzed is that about one third of the observed open reading frames (ORFs), although conserved among several organisms, encode for 'hypothetical' proteins that cannot be related to other proteins of known function or structure. Understanding the physiological function of the protein products of these so-called 'orphan' genes has emerged as a major challenge. Knowledge of the complete complement of genetic information needed for the viability of any free-living organism is required to realize the full potential of utilizing genomic information for applications in biotechnology.

A broad spectrum of genetic, biochemical and computational approaches is being employed for annotating the physiological function of 'hypothetical' proteins encoded by orphan genes. Advances in experimental and computational approaches to structure determination and analysis have recently spawned several initiatives that aim to annotate genomes via a structural approach. The idea of annotating the biological function of a macromolecule from its high-resolution structure — as determined by X-ray crystallography or NMR — stems from the fact that the structure of a protein is absolutely essential for an understanding of its function at the molecular level. In favorable cases, determining three-dimensional structures could lead to the detection and characterization of prosthetic groups, or metal ligands, and reveal catalytic or regulatory sites in enzymes. From these structural features, catalytic mechanisms, protein–protein associations or protein–nucleic acid interactions could be predicted or proposed. Thus, an analysis of the structural characteristics of new proteins might identify unique attributes that would provide key insight about their function. Additionally, this research has the potential of significantly increasing the number of single-domain three-dimensional folding patterns because it is directed towards previously undetected or poorly understood proteins, which could be useful for protein modeling.

Is 'structural genomics' a viable approach to generating useful hypotheses for the role of proteins of unknown function? A common theme among the structural genomics initiatives at Berkeley [2•], Department of Energy/Los Alamos [3•], New York [4•], Toronto [5•], and Maryland [6•] is to provide clues about biological function through structure determination. It is too early to judge the success of this initiative, although the results so far are intriguing and encouraging. Our intent in this review is to assess the capabilities and prospects for going from structure to function for *bona fide* hypothetical proteins on a large, genome-wide scale. The authors of the review are the Principal Investigators on a structural genomics project aimed at determining the structure of ~50 proteins of unknown function from *Haemophilus influenzae* (HI) over five years.

Selection of microbial targets

There are several advantages of focusing on microorganisms such as HI for structural genomics. Firstly, the HI genome is relatively small for a free-living organism. Secondly, because *Haemophilus* can be grown on defined media, the metabolism of mutant strains can be manipulated in cellular studies aimed at defining function.

Thirdly, the lack of unusual codons in HI should minimize obstacles to heterologous gene expression in *Escherichia coli*. Finally, the high level of sequence similarity from bacteria to man among many ORFs of unknown function suggests they play critical cellular functions and that the biological information gleaned from one target could fill gaps in the annotation of several genomes.

A number of selection criteria were used to construct a set of 65 target proteins from the 1743 predicted coding regions in *Haemophilus* [7]. Firstly, only proteins of unknown function, originally annotated as 'hypothetical', were considered. Secondly, soluble proteins (those containing less than three transmembrane segments according to TopPhred) [8] were selected. Next, to eliminate any falsely predicted ORFs, there had to be at least three members in the sequence family for the targets, either in HI itself, or in one or more of the other seven microbial genomes completed at that time. The resulting 124 targets were further reduced to 104 by visually eliminating possible membrane proteins and poorly aligned families. The final selection process was facilitated by making all of the information, including BLAST/BEAUTY [9] and FASTA [10] sequence searches, PROSITE [11], Swiss Prot annotation [12], GeneQuiz [13], a set of threading predictions for the HI genome [14], and TopPred results, available via Internet pages. Project members then critically assessed the functional annotation for each potential target and assigned top priority to an initial set of 65 targets.

The rapidly increasing size of the sequence databases and continued improvements in sequence search techniques make it necessary to continually review the status of each target protein. New searches for functional and structural relatives are conducted once a month, using psi-BLAST [15] with the 'non-redundant database', and a local version of psi-PDB [16]. New genome-oriented techniques for identifying protein function [17] have also been implemented.

A project of this size requires careful tracking of progress and bottlenecks. A relational database with an Internet front end is used to record results and comments on each target. Details of the current status of each target and summaries of the experimental work are available on the project website [6•].

Cloning, expression and purification of hypothetical proteins

A key factor in any structural genomics initiative is maintaining an adequate supply of highly purified, native proteins for structure determination. Heterologous expression of HI targets in *E. coli* has been achieved by amplifying selected ORFs using PCR and cloning them as 'ATG-to-TAA cassettes' into plasmid vectors containing either *trc* or T7 promoter systems for high-level, regulated expression [18,19]. Additional features of the commercially available or modified vectors included the option of expressing the target either as the native polypeptide or as a fusion protein containing purification tags, such as a thrombin-cleavable polyhistidine sequence or a chitin-binding domain adjacent to a self-cleaving intein sequence. A qualitative comparison of expression was assessed by inspection of appropriate zones on SDS-PAGE as a function of induction time and growth temperature to identify optimum conditions for each protein. Expression must be assessed not only in rich medium, but also minimal medium containing anologs such as selenomethionine or ^{15}N and ^{13}C-labeled nutrients for structural studies. An interesting pattern has emerged from a preliminary analysis of the first 53 ORFs cloned in three different vectors. About 25% of the 53 targets show little or no expression as either native or histidine-tagged proteins from either the trc or T7 promoters. About half of the targets readily express as a histidine-tagged fusion protein, and yield between 1–100 mg of purified, native-like protein that is worthy of further study. The final 25% are candidates for native protein expression and purification because their expression as fusion proteins, affinity purification or cleavage of the histidine-tagged affinity peptide presents a significant obstacle to high throughput.

Because of the challenge of purifying and crystallizing proteins that one knows little or nothing about, an analysis of several physical and chemical properties of the target polypeptides can be quite useful in establishing conditions where polypeptides remain 'folded' or native-like, monodisperse, and soluble. The time- and temperature-dependence of spectroscopic properties such as circular dichroism or fluorescence can yield the apparent stability and a useful shelf life for a new protein. An analysis of molecular weight distributions and hydrodynamic properties for a limited set of solution conditions by analytical ultracentrifugation [20,21] or light scattering [22] can provide valuable clues about the optimal solution conditions for crystallization trials or solution structure assignment by NMR.

Crystallization

The development of high-throughput crystallization techniques is a critical aspect of structural genomics. Fast screen approaches that quickly sample a large number of solution conditions for their ability to induce crystallization of a protein are already used broadly in structural biology laboratories [23]. If fast screen experiments are unsuccessful, a more systematic approach must be undertaken [24,25], which involves experiments using reagents over broad ranges of concentration, pH 2–10 and temperature 6–35°C. Techniques such as vapor diffusion or microbatch for both fast and systematic screening are easily automated using robotics [26]. Robotic automation will play a key role in reagent preparation and monitoring crystallization experiments. Once crystallization conditions are discovered, automated approaches can also be used to optimize the production of crystals of a size and quality for diffraction experiments. This may be important for selenomethionine-containing protein or other variants that would be useful for improving diffraction quality and structure determination.

After a protein's crystallization has been optimized, preliminary diffraction studies must be carried out to assess the potential for the crystals to be used in structure determination. Additionally, stabilizing cryoprotectants are needed for data collection at low temperatures [27]. Although many solutions that contain alcohols and low-molecular-weight polyethylene glycols flash freeze without problems, other solutions may require the addition of cryosolvents, such as glycerol, which may require further crystallization optimization to facilitate freezing [28]. These studies can be easily carried out at departmental X-ray sources. When crystals that behave well at low temperature have been produced, they can be stored and transported at liquid nitrogen temperatures to synchrotron X-ray sources for subsequent data collection.

X-ray diffraction and structure determination

Advances in methods for X-ray diffraction have made it possible to progress from X-ray data to a protein model in a few days, or even a few hours. The ability to acquire X-ray diffraction data at multiple wavelengths by exploiting the absorption edge of certain heavy atoms (the multiple wavelength anomalous diffraction [MAD] method) is crucial for speedy phase determination and for obtaining high-quality initial electron density maps [29,30]. With the advent of tunable synchrotron X-ray sources, the development of charged-coupled device (CCD) detectors, the application of crystal flash-freezing techniques that reduce X-ray radiation damage, and the ability to express selenomethionine-containing proteins, phase determination by MAD methods has become the method of choice for high-throughput structure determination. For the project described here, MAD experiments have been performed at beam line X12C of the National Synchrotron Light Source (Brookhaven National Laboratory, Upton, NY), and at the Industrial Macromolecular Crystallography Association Collaborative Access Team (IMCA-CAT) of the Advanced Photon Source (Argonne National Laboratory, Argonne, IL). In addition to exploiting the selenium absorption edge, mercury and platinum MAD experiments have also been performed, to either improve the phases for proteins that contain few methionines, or when the heavy atom derivative could be prepared easily.

Automated programs and direct method algorithms are now available for positioning the anomalous scatterers. Both SOLVE [31] and SHELEX [32] have been used in the current project, as well as heavy atom parameter refinement programs, such as MLPHARE [33]. Solvent flattening and, when applicable, non-crystallographic symmetry averaging can improve the quality of electron density maps [34,35]; the program DM [36] has been used successfully thus far.

Structure determination is accelerated by the high quality of phases derived from the MAD method, which in turn are suitable for computer programs that automatically trace polypeptide chains and produce more accurate models than those built manually. Consequently, refinement is no longer a rate-limiting step in structure determination. In the current project, two of the structures determined by the MAD method have been obtained at a resolution better than 2.3 Å, resulting in phase quality suitable for automatic polypeptide chain tracing. For both cases, approximately half of the polypeptide chain could be traced automatically, using the program ARP/wARP [37], and the remaining model was traced manually, with the program O [38]. Acceleration of the refinement process is further aided by automatic selection of water molecules, and both ARP/wARP and CNS [39] have been used for this purpose.

NMR spectroscopy

An important advantage of NMR spectroscopy in structural genomics is that protein structure determination can be performed in solution so that crystal growth is unnecessary. However, NMR is typically most useful for smaller proteins (<30 kDa) that are highly soluble (millimolar concentrations). Another issue is the time required for determining a fully refined NMR structure. The length of both data collection (45–60 days) and analysis (6–12 months) pose challenges for high-throughput structure determination by NMR for structural genomics initiatives. Recently developed cryoprobes that contain coils and preamplifiers that operate at low (~25 K) temperatures can yield a 3–4-fold improvement in signal-to-noise over conventional probes. The application of cryoprobes should have a major impact in NMR structure determinations by reducing the time for data acquisition (15–20 days), and permitting investigations of proteins that either have low solubility or yield poorly to purification.

The most time-consuming aspect of structure determination is the interpretation of the large number of NMR spectra from $^{13}C/^{15}N$-labeled samples. Heteronuclear multidimensional data, however, can reduce signal overlap for data sets enough for data to be analyzed automatically. New automation routines can assign the peptide backbone in less than 1 min with reasonable accuracy [40]. The most laborious part of automatically assigning the peptide backbone of two proteins, HI0719 and HI0257, with AUTOASSIGN [41] was peak picking, which could be completed in less than a day, yielding results that were in good agreement with manual assignments. Because chemical shift assignments yield valuable secondary structure information [42,43] and provide the basis for global fold determination using nuclear Overhauser effects (NOEs) and residual dipolar couplings [44], the global fold can often be determined quickly, which can yield clues about biochemical function. Further efforts on the automation of sidechain and NOESY assignments [40,45–47] and the integration of these procedures into an automated structure determination package will doubtless enhance the role of NMR in structural genomics.

Deducing function from structure

Protein structures provide many direct and indirect clues about molecular function that can be utilized in structural genomics. Four specific approaches are being used to analyze

the structures of hypothetical proteins in this project, depending on the case.

Case 1: the protein has a fold that has been seen before

If the fold is associated with one, or a few, biological functions, then an assessment can be made as to whether the new structure is compatible with one of the biological functions as most folds have only one or a few functions [48].

Case 2: the protein is an enzyme

Many enzyme mechanisms have been seen multiple times in proteins of different folds. This suggests convergent evolution of enzyme function, and also that a large fraction of the most common mechanisms have already been annotated. Several groups are compiling libraries of known three-dimensional catalytic motifs, and one example, the Ser–His–Asp catalytic triad of serine proteinases and lipases, has been published [49]. When such a library has been established, each new structure can be searched to see if it contains one of the known motifs.

Case 3: the protein binds one or more small-molecule ligands

Ligand-binding sites are almost always associated with the largest depressions in protein surfaces [50], and can be identified automatically. Given a binding site, the docking tools developed for structure-based drug design [51–53] have the potential for identifying binding ligands from a library of naturally occurring compounds.

Case 4: the protein interacts with other macromolecules

Five methods are currently available for determining macromolecular interactions depending on the nature of the interaction. First, in cases where electrostatics plays a major role in binding, such as when a protein associates with RNA or DNA, mapping of the surface potential can be an effective technique [54]. Second, sites of tight association with other proteins may be identified by analyses of surface composition [55]. Third, where a large family of sequences are available, mapping the extent of conservation of surface residues provides a means of identifying interaction sites [56]. Fourth, three new genome-scale non-structure-based methods hold promise for providing clues to identifying interacting proteins [17]. Hypotheses generated by these methods may be tested structurally by protein–protein docking methods that search for specific binding sites [57]. Fifth, it has also been suggested that interaction sites can also be identified by analyses of the correlation of sequence changes between pairs of proteins across many species [58]. These cases can also be clarified using protein–protein docking methods.

Many of these methods provide hypotheses concerning function, which then require experimental verification. An important component of the project is outreach to appropriate members of the larger experimental community, supplying them with information on possible function, and material for further work.

Genetic approaches to identify essential genes and their function

The phenotypes of specific gene deletions under various growth conditions can yield important clues on the biological roles for ORFs of unknown function, especially for those genes that are essential for growth or viability under laboratory conditions. The identification of essential ORFs of unknown function also can provide a starting point for uncovering novel and important biological processes [59,60,61•]. Additionally, as all conventional antibiotics target the products of essential genes, the discovery of new essential ORFs will have a significant impact on antimicrobial drug discovery. In a pilot study aimed to identify essential genes among the conserved hypothetical ORFs in the Rd strain of *H. influenzae*, 10 targeted gene deletions were generated by homologous recombination [62]. Five of the null mutants are viable on rich medium (brain–heart infusion broth); however, five ORFs could not be deleted, and appear essential for viability, highlighting the need for structure and function determination.

Conclusions: early structural results and future prospects

So far, there have been only a few examples of proteins of unknown function whose structures have been determined as part of structural genomics projects. More cases are required in order to assess the prospects of assisting the assignment of function based on structure. Nevertheless, the results accumulated during the past year are encouraging. Interestingly, in two cases, co-purified ligands were found bound to the protein in the crystal structure, thus shedding light on the function. In the first case, the structure contained ATP, hinting that the protein was an ATPase or an ATP-mediated molecular switch [63••]. In the second case, the molecule belongs to the α/β barrel fold and binds a prosthetic group, pyridoxal 5′-phosphate (unpublished data of the Structural Genomics Research Consortium, see [4•]. The structure coordinates are available in the Protein Data Bank, PDB accession number 1b54.) The third structure that has been published revealed a fold that is homologous in part to some nucleotide-binding proteins, and biochemical analysis inspired by the structure confirmed that the protein catalyzes the hydrolysis of a number of nucleotides [64••].

The first structure that has emerged from this project is of the hypothetical protein HI1434, one of a nearly 30-membered microbial protein family labeled in SwissProt as YbaK and ebsC (H Zhang *et al.*, unpublished data). This structure illustrates some challenges in assigning function from structure as the fold of the protien is not sufficiently close to other known structures to imply function. The structure of HI1434 is only remotely related to the C-lectin fold and, in particular, to endostatin, an inhibitor of angiogenesis. The similarity is too weak, however, to imply that the YbaK is a saccharide-binding protein. Nevertheless, a crevice that may accommodate a small ligand is evident. The putative binding site contains only one invariant

residue, a lysine, whereas enzymes usually contain several conserved functional groups to comprise their catalytic apparatus, implying that YbaK is not an enzyme. As the sequence of more genomes were completed, it was possible to identify new YbaK family members. Analyses of these sequences revealed homology to an insertion domain in prokaryotic prolyl-tRNA synthetase, underscoring the need for continuously updating sequence searches. Although the function of this insertion domain is unknown, a comparative model based on HI1434 suggests that it too should contain a putative binding site, which may play a role in nucleotide binding by the synthetase. By analogy, YbaK may bind mononucleotides or oligonucleotides, the nature of which is yet to be determined.

These examples illustrate that it is indeed feasible for the range of activities involved in structural genomics initiatives to annotate the biological role of proteins of unknown function. Greater success will doubtless result from improved experimental approaches for high-throughput protein expression and purification, crystallization and structure determination by X-ray diffraction and NMR spectroscopy, as well as from faster and more accurate computational tools to analyze with possible functions from protein structures. It is becoming crystal clear that structural approaches will play a key role in realizing the full potential of the genomics revolution.

Acknowledgement

This work is supported by NIH grant P01 GM57890.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. TIGR Microbial Database on World Wide Web URL:
• http://www.tigr.org/tdb/mdb/mdb.html
This website at TIGR gives the status of the latest developments in microbial genome sequencing, including links to sequence and annotation information.

2. Professor Sung-Hou Kim's Homepage on World Wide Web URL:
• http://www.cchem.berkeley.edu/~shkgrp/index.html
The website for the laboratory of Sung-Ho Kim includes information about the structural genomics project using proteins from the hyperthermophilc archaebacteria *Methanococcus jannaschii*.

3. Los Alamos National Laboratory Life Sciences Division Research
• Projects on World Wide Web URL: http://lsdiv.lanl.gov/research.htm
A website describing the approach and results of the US Department of Energy/University of California Los Angeles collaboration on the structural genomics of the hyperthermophilic archaebacteria Pyrobaculum aerophilum.

4. Structural Genomics Pilot Project on World Wide Web URL:
• http://genome5.bio.bnl.gov/Proteome
The website for the structural genomics collaboration between Brookhaven National Laboratory, Rockefeller University, and Albert Einstein College of Medicine focused on proteins from the yeast *Saccharomyces cerevisiae*.

5. Structural Proteomics: Structural Biology on Genome-wide Scale on
• World Wide Web URL:
http://diana.oci.utoronto.ca/arrowsmith/proteomics/index.html
A website summarizing the approach and results of the Ontario Cancer Institute's structural genomics project on 500 proteins from *Methanobacterium thermoautotrophicum*.

6. From Gene Structure to Function Project on World Wide Web URL:
• \http://s2f.carb.nist.gov
The website for the CARB–TIGR collaborative project on the structure and function of hypothetical proteins from *Haemophilus influenzae*.

7. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Doughertry BA, Merrick JM *et al.*: **Whole-genome random sequencing and assmembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496-512.

8. Claros MG, von Heijn G: **TopPredII: an improved software for membrane protein structure predictions.** *Comput Appl Biosci* 1994, **10**:685-686.

9. Worley KC, Wiese BA, Smith RF: **BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results.** *Genome Res* 1995, **5**:173-184.

10. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85**:2444-2448.

11. Baitroch A, Bucher P, Hafmann K: **The PROSITE database, its status in 1995.** *Nucleic Acids Res* 1995, **24**:189-196.

12. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence databank and its new supplement TrEMBL.** *Nucleic Acids Res* 1996, **24**:21-25.

13. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C: **GeneQuiz: a workbench for sequence analysis.** *Ismb* 1994, **2**:348-353.

14. Genomic Threading Database on World Wide Web URL:
http://globin.bio.warwick.ac.uk/genome/hi/

15. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

16. Salamov AA, Suwa M, Orengo CA, Swindells MB: **Genome analysis: assigning protein coding regions to three-dimensional structures.** *Protein Sci* 1999, **8**:771-777.

17. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome wide prediction of protein function.** *Nature* 1999, **402**:83-86.

18. Amann E, Brosius J: **'ATG vectors' for regulated high-level expression of cloned genes in *Escherichia coli*.** *Gene* 1985, **40**:183-190.

19. Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW: **Use of T7 RNA polymerase to direct expression of cloned genes.** *Methods Enzymol* 1990, **185**:60-89.

20. Hensley P: **Defining the structure and stability of macromolecular assemblies in solution: the re-emergence of analytical ultracentrifugation as a practical tool.** *Structure* 1996, **4**:367-373.

21. Schuster TM, Toedt JM: **New revolutions in the evolution of analytical ultracentrifugation.** *Curr Opin Struct Biol* 1996, **6**:650-658.

22. Frerre-D'Amare AR, Burley SK: **Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies.** *Structure* 1994, **2**:357-359.

23. Jancarik J, Kim S-H: **Sparse matrix sampling: a screening method for crystallization of proteins.** *J Appl Crystallog* 1991, **24**:409-411.

24. Gilliland GL, Tung M, Ladner J: **The biological macromolecule crystallization database and NASA protein crystal growth archive.** *J Res Natl Inst Stand Technol* 1996, **101**:309-320.

25. McPherson A: *Crystallization of Biological Macromolecules.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.

26. Shaw Stewart PD, Baldock PFM: **Practical experimental design techniques for automatic and manual protein crystallization.** *J Crystal Growth* 1999, **196**:665-673.

27. Garman EF, Schneider TR: **Macromolecular cryocrystallography.** *J Appl Crystallogr* 1997, **30**:211-237.

28. Garman E, Mitchell E: **Glycerol concentrations required for cryoprotection of 50 typical protein crystallization solutions.** *J Appl Crystallogr* 1996, **29**:584-587.

29. Karle J: **Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology.** *Int J Quantum Chem: Quantum Biol Symp* 1980, **7**:357-367.

30. Hendrickson WA: **Analysis of protein structure from diffraction measurement at multiple wavelengths.** *Trans Am Crystallogr Assoc* 1985, **25**:11-21.

31. Terwilliger TC, Berendzen J: **Automated structure solution for MIR and MAD.** *Acta Crystallogr* 1999, **55**:849-861.

32. Sheldrick GM: **SHELEX: applications to macromolecules**. In *Direct Methods for Solving Macromolecular Structures*. Edited by Forteir S. Dordrecht: Kluwer Academic Publishers; 1998:401-411.

33. Otwinowski Z: **Maximum likelihood refinement of heavy atom parameters.** In *Isomorphous Replacement and Anomalous Scattering, Proceedings of the CCP4 Study Weekend 25–26 January, 1991*. Edited by Wolf W, Evans PR, Leslie AGW. Warrington, England: Daresbury Laboratory; 1991:80-86.

34. Wang B-C: **Resolution of phase ambiguity in macromolecular crystallography.** *Methods Enzymol* 1985, **115**:90-112.

35. Bricogne G: **Geometric sources of redundancy in intensity data and their use for phase determination.** *Acta Crystallogr* 1974, **30**:395-405.

36. Cowtan K: **An automated procedure for phase improvement by density modification.** *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* 1994, **31**:34-38.

37. Perrakis A, Morris R, Lamzin VS: **Automated protein model building combined with iterative structure refinement.** *Nat Struct Biol* 1999, **6**:458-463.

38. Jones TA, Zou J-Y, Cowan SW, Kjelgaard M: **Improved methods for building protein models in electron density maps and the location of errors in these models.** *Acta Crystallogr* 1991, **47**:110-119.

39. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunsleve W, Jiang J-S, Kuszewski J, Nilges M, Pannu NS *et al.*: **Crystallography and NMR system: a new software suite for macromolecular structure determination.** *Acta Crystallogr* 1998, **54**:905-921.

40. Moseley HNB, Montelione GT: **Automated analysis of NMR assignments and structures for proteins.** *Curr Opin Struct Biol* 1999, **9**:635-642.

41. Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C-Y, Powers R, Montelione GT: **Automated analysis of protein NMR assignments using methods from artificial intelligence.** *J Mol Biol* 1997, **269**:592-610.

42. Wishart DS, Sykes BD: **The ^{13}C chemical shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical shift data.** *J Biomol NMR* 1994, **4**:171-180.

43. Cornilescu G, Delaglio F, Bax A: **Protein backbone angle restraints from searching a database for chemical shift and sequence homology.** *J Biomol NMR* 1999, **13**:289-302.

44. Tjandra N, Omichinski JG, Gronenborn AM, Clore GM, Bax A: **Use of dipolar ^{1}H-^{15}N and ^{1}H-^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution.** *Nat Struct Biol* 1997, **4**:732-738.

45. Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H: **Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin.** *J Mol Biol* 1997, **269**:408-422.

46. Mumenthaler C, Guntert P, Braun W, Wuthrich K: **Automated combined assignment of NOESY spectra and three-dimensional protein structure determination.** *J Biomol NMR* 1997, **10**:351-362.

47. Xu Y, Wu J, Gorenstein D, Braun W: **Automated 2D NOESY assignment and structure calculation of crambin (S22/I25) with the self-correcting distance geometry based NOAH/DIAMOD programs.** *J Magnet Resonance* 1999, **136**:76-85.

48. Murzin A, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classfication of proteins database for investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540.

49. Wallace AC, Laskowski RA, Thornton JM: **Derivation of 3D co-ordinate templates for searching structural databases: application to Ser–His–Asp catalytic triads in the serine proteinases and lipases.** *Protein Sci* 1996, **5**:1001-1013.

50. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function.** *Protein Sci* 1996, **5**:2438-2452.

51. Kuntz ID: **Structure-based strategies for drug design and discovery.** *Science* 1992, **257**:1078-1082.

52. Shoichet BK, Kuntz ID: **Matching chemistry and shape in molecular docking.** *Protein Eng* 1993, **6**:723-732.

53. Eisen MB, Karplus M: **HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecular binding site.** *Proteins* 1994, **19**:199-221.

54. Honig B, Nicholls A: **Classical electrostatics in biology and chemistry.** *Science* 1995, **268**:1144-1149.

55. Jones D, Thornton JM: **Protein–protein interactions: a review of protein dimer structures.** *Prog Biophys Mol Biol* 1995, **63**:31-65.

56. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342-358.

57. Strynadka NC, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F *et al.*: **Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase.** *Nat Struct Biol* 1996, **3**:233-239.

58. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein–protein interaction.** *J Mol Biol* 1997, **271**:511-523.

59. Arigoni F, Talabot FPM, Edgerton MD, Meldrum E, Allet E, Fish R, Jamotte T, Curchod ML, Loferer H: **A genome-based approach for the identification of essential bacterial genes.** *Nat Biotechnol* 1998, **16**:851-856.

60. Akerley BJ, Rubin EJ, Camilli A, Lampe DJ, Robertson HM, Mekalanos JJ: **Systematic identification of essential genes by *in vitro* mariner mutagenesis.** *Proc Natl Acad Sci USA* 1998, **95**:8927-8932.

61. • Reich KA, Chovan L, Hessler P: **Genome scanning in *Haemophilus influenzae* for identification of essential genes.** *J Bacteriol* 1999, **181**:4961-4968.

Transposon insertional mutagenesis is combined with analyses of growth rates of mutant strains to rapidly identify essential genes.

62. Barcak GJ, Chandler MS, Redfield RJ, Tomb J: **Genetic systems in *Haemophilus influenzae*.** *Methods Enzymol* 1991, **204**:321-342.

63. •• Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim S-H: **Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics.** *Proc Natl Acad Sci USA* 1998, **95**:15189-15193.

This is the first example of an hypothetical protein structure that has been determined as part of a structural genomics project. The biochemical studies that complimented the structural work followed the identification of bound ATP.

64. •• Hwang KY, Chung JH, Kim S-H, Han YS, Cho Y: **Structure-based identification of a novel NTPase from *Methanococcus jannaschii*.** *Nat Struct Biol* 1999, **6**:691-696.

In this case, fold recognition suggested testing for nucleotide hydrolysis, and it was confirmed that the protein could hydrolyze some nonstandard nucleotides.