REVIEW

Class-directed structure determination: Foundation for a Protein Structure Initiative

THOMAS C. TERWILLIGER, GEOFFREY WALDO, THOMAS S. PEAT, JANET M. NEWMAN,
KELVIN CHU, AND JOEL BERENDZEN

Structural Biology and Biophysics Groups, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(RECEIVED February 25, 1998; ACCEPTED May 27, 1998)

Abstract

The recent sequencing of many complete genomes, combined with the development of methods that allow rapid structure determination for many proteins, has changed the way in which protein structure determinations can be approached. One-by-one determinations of individual protein structures will soon be augmented by class-directed structure analyses in which a group of proteins is targeted and structures of representative members are determined and used to represent the entire group. Such a shift in approach would be the foundation for a broad protein structure initiative targeting classes of proteins important for biotechnology and for a fundamental understanding of protein function.

Keywords: biotechnology; class-directed structures; genome projects; microorganisms; protein structure determination

The genome project will be generating an enormous number of protein sequences over the next few years. Already the complete genomes of over a dozen microbial species have been reported; indications are that the rate of obtaining them is doubling every year.¹ Thousands of human genes have been sequenced already and the various genome projects are attempting to obtain sequences of all 80,000 or so human genes, and of many genes from other higher organisms within the next decade (Collins et al., 1997; Rowen et al., 1997). It has been recognized for some time that the vast numbers of sequences resulting from the genome project could transform many areas of biological research, allowing both more detailed and more global views of entire organisms than was possible before (Gilbert, 1991; Dev, 1997). This view of all the genes in an organism is already evident in the analyses of genomes of microbial species that have recently been presented (Bult et al., 1996; Fraser et al., 1997; Tomb et al., 1997).

It is now becoming clearer how this wealth of sequence data will affect biology. It seems likely that studies of protein structure and function will shift from the properties of individual proteins to representative members of entire classes of proteins. We call this collective analysis of groups of protein structures “class-directed” structure determination and analysis. Here we discuss why a change

in strategy of this kind is likely to occur in the area of protein structure determination, what form this approach to structure determination might take, scientific and technological questions that could be addressed with class-directed structure analyses and why a coordinated initiative to identify and determine structures of important classes of proteins should begin now.

Why class-directed structure determinations?

Protein structures are currently determined in a one-by-one fashion using either NMR or X-ray crystallographic techniques. The process is often lengthy, requiring years to complete. Considering the large time investment required to solve the structure of a particular protein molecule, it is not surprising that the targets of protein structure determinations are proteins of high intrinsic interest to the biological community. There are two reasons why structure determination is likely to shift toward class-directed analyses. First, classification and associated grouped analyses are now possible on a large scale and can yield information about protein structure and function that is not as readily obtained by individual consideration of proteins. Second, structural information can be obtained far more readily if the structure of any member of a class is sufficient.

Classification of proteins can be carried out in sophisticated ways because of the ready access to sequences of proteins from entire organisms and from many different organisms. For example, searching for a new sequence is much easier if one knows a large group of related sequences than if one only one member of the cluster is known (Rost & Sander, 1995). Classification by protein

Reprint requests to: Thomas C. Terwilliger, Structural Biology and Biophysics Groups, Los Alamos National Laboratory, Los Alamos, New Mexico 87545; e-mail: terwilliger@lanl.gov.

¹See, for example, the TIGR microbial genome database web site at www.tigr.org.

structure not only allows comparisons within a class to infer shared properties of other members (Firestone et al., 1996), but also comparisons between classes that can yield information of specific structure-function relationships. An example of the latter can be found in the analyses of known antibody combining site structures and associated protein sequences to suggest which parts of the combining site can be most readily replaced to change binding specificity (Rapley, 1995).

If protein structure determination were completely straightforward, then, after classifying proteins in an appropriate way, the structures of the protein targets that promised the greatest amount of relevant structural information for the problem at hand would simply be determined. Unfortunately, structure determination is not this straightforward for macromolecules and, as a consequence, the ability to extract structural information from any one of the members of a class of proteins and apply it to the rest is very attractive. For both X-ray crystallography and NMR, the most severe bottlenecks in structure determination involve the properties of the protein molecules themselves. For crystallography, obtaining crystals is usually the most crucial step. Some proteins crystallize readily and under a variety of conditions, while others crystallize slowly or under very few conditions. Finding the right conditions is a time-consuming, laborious process with an uncertain outcome. For NMR, high solubility and well-defined NMR spectra are a crucial feature of proteins for structure determination. For both techniques, however, it is often found that relatively small differences in protein sequence can make the difference between a protein that is recalcitrant and one for which a structure can be obtained readily.

Given crystals, protein structure determination using X-ray diffraction can be rapid. Using the multiwavelength (MAD) approach to macromolecular structure determination of protein crystals containing selenium or other atoms with accessible X-ray absorption edges (Hendrickson & Ogata, 1997), the necessary data can be collected at a synchrotron in just hours to days (depending on the quality of the crystals, the intensity of the X-ray beam, and the size of the protein). With recent developments in analysis, diffraction data of this kind can now be converted into a three-dimensional picture of the protein molecule within a few hours or days. Building an atomic model based on this three-dimensional picture, and refining the model to agree as closely as possible with the data, currently takes considerably more time, but this process could be automated as well and considerable effort is being devoted to this goal. Data collection for NMR generally takes much longer than for X-ray diffraction, and the somewhat difficult data analysis is not yet fully automated. Additionally, NMR structure determinations are currently limited to proteins under about 30 kD, which includes fewer than half of the proteins of a typical microorganism such as *Methanobacterium thermoautotrophicum* (Smith et al., 1997). It seems likely, however, that efforts to automate NMR data analysis will speed up structure determination by this method and higher-field NMR and structure determinations of thermophilic proteins at higher temperatures will raise the size limit considerably.

The potential for very rapid X-ray diffraction and NMR structure determinations of a subset of protein molecules that are particularly well-suited for structural analyses means that protein structure determination might become a high-throughput technique for a well-suited subset of proteins. An important reason why a shift toward class-directed determination of protein structures is likely to occur is simply because class-directed structure determination can be based on solving just the structures of those well-

suited proteins. Class-directed structure determinations can be designed to make optimal use of the structural information that is most readily obtainable.

An important, but more subtle, aspect to the trade-off between individual and class-directed structure determinations is the investment required to identify the members of a class that are best-suited for structure determination. The following is a very rough estimate of this trade-off (efforts are being made to determine this empirically as well). Consider the case of determining the structure of a particular protein motif, represented by a cluster of, perhaps, 20 protein sequences. Using current approaches, if a particular protein sequence were picked for one-by-one structure determination the effort required to obtain that structure would typically be one to two person-years of work. In some cases it might be less, in other cases the structure might never be obtained. However, using current technologies, but with a high-throughput approach, we estimate that all 20 of those proteins could be tested for expression, the highest-expressing 10 or 15 of these tested for purification, and the most readily-purified 5 or 10 of these tested for rapid crystallization or for NMR spectral quality in parallel by one person in just a few weeks. The one or two proteins best-suited for structure determination could then be chosen for structure determination and structures obtained with a high likelihood of success shortly after this. The investment required to identify which structures are most readily solvable, and to determine the structures of these appears, on this basis, to be small compared to the time currently spent on obtaining a structure. It seems likely that this approach could increase the throughput of protein structures significantly. If many protein motifs are of interest, a more sophisticated experimental design, in which no more members of a class are considered once one structure in the class has been determined, could increase the throughput even further.

Steps in class-directed structure determination

A class-directed structure determination and analysis is likely to contain several steps:

1. Identification of important biological or technological targets; selection of a target that depends on a knowledge of protein structure
2. Classification of protein sequences based on that target
3. Structure determinations of appropriate representative proteins
4. Analysis of the structural, functional, and sequence information available after structure determinations.

The first step in class-directed structure determination and analysis is identification of an important target or goal that can be addressed with a knowledge of protein structure. There is a continuum of scientific and technological questions involving protein structure, ranging from those where any of a large number of protein structures can yield useful information to those where one particular structure must be obtained to answer the question. Those questions that can be answered using any of a number of protein structures lend themselves to a class-directed structure determination, while those requiring a particular protein structure are suitable for one-by-one analyses. The target for class-directed structure determination might be a scientific goal, such as obtaining a set of protein structure motifs that collectively represents much of the universe of existing protein folds and could be used to predict the structures

of newly-obtained protein sequences. Alternatively, it might be more of a technological question such as, "how can we design a drug to block human infection by *Helicobacter pylori*, the microorganism responsible for stomach ulcers?"

In a second step, known protein sequences are sorted with the question in mind. If one wanted to obtain a representative set of protein motifs, the classification might consist of clustering, at some level of sequence similarity, all known protein sequences or segments of sequences thought to represent folds. For drug targeting against *H. pylori*, the classification might simply be all *H. pylori* proteins that are thought to be involved in required metabolic pathways or in infection.

In a third step of class-directed structure determination, structural information is obtained about a subset of proteins, or protein domains, that can represent the remainder of the classes. Alternatively, the structure of the member of each cluster that is most amenable to structure determination would be obtained as a representative of the cluster. In some cases obtaining the structure of a representative of a particular class might be straightforward, but in others it might still be difficult, or indeed impossible. For example, structures of integral membrane proteins have proven difficult to obtain, and until vastly improved methods for structure determination of these proteins become available, they are unlikely to be obtained in anything resembling a high-throughput manner.

In the last stage of class-directed structure determination and analysis, the representative structures that are obtained would be analyzed together with existing known structures, biochemical information and information obtained from the classification process itself to address the target question. Once a set of representative protein motifs were obtained, for example, a novel protein sequence could be classified into one of the identified clusters of sequences, and the representative fold for that cluster could be used as a starting point for modeling the structure of the new protein. If more than one member of the cluster existed, then comparisons between the structures could provide additional information about the fold of the new protein. In the case of the *H. pylori* proteins, the situation is even simpler because each protein structure that is obtained is itself suitable as a starting point for directed drug design.

Scientific and technological questions that can be addressed by class-directed structure determinations

There are a number of major scientific and technological questions that depend heavily on protein structure and can now be addressed by class-directed structure determinations. Table 1 lists some of these that can be readily envisioned, and it is certain that more will become evident in the future. One of the most important aspects of these class-directed structure determinations is that they will provide the necessary information for constructing extensive databases of known protein structure and function. In the future when structural information is required, the necessary databases will already be in place.

A database of protein structural motifs

One of the most obvious tasks for class-directed analysis is the determination of a set of protein structure motifs that is representative of the "universe" of protein structural elements. The concept of obtaining such a set of motifs is currently being widely discussed. Obtaining these structures would require an informatics

Table 1. Target classes for a Protein Structure Initiative

Class of proteins	Importance
Database of protein structure motifs	Prediction of protein structure
Proteins from a microorganism	
Proteins from a pathogen	Potential drug targets
Proteins from a thermophile	Robust enzymes
Large-scale targets	
Human proteins	Medical applications
Plant or animal proteins	Biotechnology
Protein assemblies	Protein interactions
Small-scale targets	
Groups of structurally-similar proteins	Predicting protein evolution
Proteins from a metabolic pathway	Biocatalysis

effort to define clusters of related protein sequences, and a structure determination effort to obtain structural information on the members of each cluster best-suited for structure determination. There are many important uses for a fairly complete set of protein structural motifs, the most obvious of which is the prediction of structures of other cluster members. The structural information from one member of a cluster could be applied to another member either using homology modeling or threading approaches (Bowie et al., 1991; Demaeyer et al., 1997). An extension of this application would be prediction of the function of a new member of a cluster. This would be applicable to those clusters in which most members share some common function, as is known to occur for many (but not all) structurally related proteins (Orengo et al., 1994). A further extension of this use would be in the design of mutagenesis studies of other members of a cluster. The predicted structure of the new member could be used to identify amino acids that are near to a presumed active site, for example, and predictions of their importance could then be tested by changing them to other amino acids, as has been done recently. Another important use of such a set of structural motifs would be in the development of methods for prediction of protein structure. In the process of obtaining these motifs, it is likely that in many cases the structure of more than one member of a cluster will be obtained, and these could be used to cross-validate prediction approaches.

When considering an effort to obtain a set of protein structure motifs that represent most protein folds, it is important to recognize that the level of sequence similarity shared among members of a cluster of proteins, the degree of structural similarity among these members and the number of clusters necessary to represent most protein folds are all highly related (Orengo et al., 1994). It was shown some time ago (Chothia & Lesk, 1986) that when structures of pairs of proteins with varying degrees of sequence identity ranging from 80% down to 20% are compared, the "core" portions of these proteins (essentially excluding loops) differ by RMS shifts ranging from 1 Å (for the proteins with sequence identity above about 50%) to about 2 Å (for those that have sequence identity of only 25%). Clearly if clusters are chosen such that members of a cluster share 50% sequence identity then there will be far more clusters than if a cutoff of 25% identity were chosen. The number of families of protein structure at about a 30% identity level was recently estimated to be about 20,000 (Orengo

et al., 1994). Correspondingly, the structural information that one member of a cluster provides about another member is far greater if the clusters share 50% sequence identity than if they share only 25%. For purposes of predicting general classes of protein function and for targeting mutagenesis experiments, clusters with identities of 25% or more are likely to be of at least some use. On the other hand, for purposes such as drug design, the structure of the actual protein of interest or a very near relative would be far more useful, and many clusters with sequence identities over at least 50% would be necessary before one member of a cluster could be used effectively for prediction of the structure of another.

Proteins from a microorganism

The class of all proteins from a microorganism is a very attractive target for class-directed structure determination. Many complete genomes of microorganisms are known, and expression and purification of proteins from microorganisms are considerably simpler than for proteins from higher organisms. High-throughput expression, purification and crystallization or NMR analyses of the most suitable 10% or so of proteins in a microorganism could probably be carried out in as little as a few years by several teams of 5 to 10 people. Recently it was proposed that structures of all proteins in a microorganism be determined (Frishman & Mewes, 1997). This would be a much more difficult undertaking, but the structures of even 10% of the proteins from a microorganism would yield a wealth of new information. Two particularly suitable types of microorganisms for such an effort would be pathogens and thermophiles, as the proteins from each have clear biotechnological uses.

An understanding of the details of the biology of pathogenic organisms is important for control of both human and animal diseases. A major effort is already underway to obtain complete genomic sequences of a large group of pathogenic organisms, and efforts to unravel the metabolic pathways of these organisms are certain to follow. Structural information on proteins that are necessary for growth of pathogenic organisms or for infection of their hosts is likely to play a central role in the development of therapeutic agents. These will surely include drugs that inhibit the action of key metabolic enzymes or that block binding to receptor molecules on their hosts. A promising set of targets for class-directed structure determinations would be all these key proteins from a pathogenic organism such as *Staphylococcus aureus* (cause of many antibiotic-resistant hospital infections), *Borrelia burgdorferi* (cause of Lyme disease), or *H. pylori*. For this project to be useful, it would not be necessary to obtain the structures of all these proteins, rather the structures of any subset of them from each targeted pathogen could provide a basis for rational drug design, and the more structures obtained the greater the likelihood of finding a useful therapeutic agent. An exciting extension of a project to determine structures of proteins from pathogens would be to obtain structures of the corresponding proteins from the host as well (Velinde & Hol, 1994). Comparisons of the structures of key metabolic enzymes from a human pathogen such as *S. aureus* and the corresponding protein from humans could be exceptionally useful in developing drugs that selectively inhibit the action of the pathogen protein without causing harm to humans.

Proteins from thermophilic organisms have been the subject of considerable interest recently, primarily because of their exceptionally robust nature. These proteins remain active at temperatures as high as 70 °C to over 100 °C and often have a similarly high resistance to chemical denaturation. Proteins that are highly

robust could be useful in industrial processes that are more efficient at high temperature or that require long-term protein stability. Structures of many of the proteins from a thermophilic organism could be useful in several ways. One is that a large number of comparisons of structures of corresponding proteins from mesophilic and thermophilic organisms may help identify features of proteins that yield high thermostability (Vogt et al., 1997). Another use of structures of proteins from a thermophilic organism would be as scaffolds for the design of novel proteins (Pinto et al., 1997). As the folded forms of mesophilic protein molecules are only marginally stabilized over the unfolded state, adjustments made to their amino acid sequences intended to affect catalytic activity often cause unintended decreases in the stability of the protein. It seems possible that the likelihood of obtaining a protein that has desirable novel activity and that is highly stable would be considerably improved by taking from a protein that is already exceptionally stable as a starting point. Effective protein engineering of this sort requires structural information, and a large-scale effort to obtain structures of many proteins from a thermophile could result in a database of protein scaffolds that would be ready for application to protein design as the needs arise.

Large-scale class-directed structure determination targets

As the genome projects progress into the arena of "functional genomics," an increasing number of human genes are being expressed in a variety of systems and the corresponding proteins purified and tested for function (Hieter & Boguski, 1997). Although the functions of the vast majority of human proteins are unknown, it seems likely that, at some point, many of them will be found to be involved in some way with human disease. Each of these would be an important potential protein structure target. As this class of proteins becomes identified, structures of as many members as possible will be useful in understanding the diseased and normal states and in controlling these diseases. A truly daring alternative would be to consider all human proteins as potentially involved in disease, or at least important in understanding how we function, and to attempt to determine the structures of any human proteins that can be obtained. In the short term, such a project with some 80,000 protein targets may not be feasible. In the longer term, with major innovations in structure determination methods and improvements in facilities, this project might well be the target of choice. The agricultural analog of the determination of all human protein structures would be a large-scale effort to obtain structures of as many proteins as possible from an important species such as rice or corn. These may allow us to improve properties such as disease and pest resistance, crop yield, and the nutritional profile of agricultural proteins.

The structure of an individual protein molecule often gives only a part of the structural information necessary to understand an important biological process. Replication of DNA and protein synthesis, for example, involve dozens of proteins (along with RNAs in the case of protein synthesis) that assemble together into large functional units. Each individual structure of a part of such a complex assembly gives some information about the whole, but crucial information about interactions among the building blocks is lacking. In other cases such as signal transduction pathways, protein-protein contacts are transiently made that are central to the biological pathway. One of the large-scale projects proposed in the area of functional genomics is to identify pairs of interacting proteins, first in microorganisms and later in humans using yeast

two-hybrid or other systems for detecting these interactions (Hieter & Boguski, 1997). A possible class-directed structure determination project would be to determine structures of complexes formed by many pairs of proteins that are known to interact with each other and are involved in important pathways such as DNA repair and signal transduction. This class of proteins is suitable for the class-directed approach because its members can be identified in advance by experimental methods and by knowledge of the biochemistry of the relevant pathways and because the structures of any of the complexes formed along these pathways would be helpful in understanding the process. In some cases, of course, a pathway might be of such interest that each protein-protein interaction would be the subject of an intensive one-by-one structure determination.

Small-scale class-directed structure determination targets

The space of protein structures can be explored from many perspectives. One way that is orthogonal to finding protein structure motifs would be the analysis of a group of structurally-similar proteins with the goal of predicting likely pathways of evolution of one of the members. The development of resistance to drugs, or to the immune system, by pathogens has been a major difficulty in treating many diseases from AIDS to influenza. It seems possible that an analysis of the structures of a large group of related proteins could be used to develop an understanding of which structural features are most rigid and which are most plastic. This structural information would allow the development of a database of structures that are attainable. This in turn could be used to develop predictions of other structures that are likely to be attainable with minor protein sequence changes and, thereby, to anticipate possible ways that resistance to particular therapeutic agents could evolve.

Groups of structurally-similar proteins could provide an invaluable database for development of protein structure prediction methods as well. A set of many small-scale structure determination efforts of this kind, when combined together, would yield sets of closely related protein structures that could be predicted from each other in order to test structure prediction methods.

Another small-scale target for class-directed structure determination would be the proteins from an important metabolic or biochemical pathway. The pathway could be a relatively linear one for synthesis of a complex organic compound or a more complex one such as DNA replication. Although the structures of every member of such a pathway would be desirable, the structures of any of the members would be useful in understanding the pathway, and also in engineering the pathway for possible biotechnology applications. If some of the proteins in the pathway from one organism are not well-suited for rapid structure determination, homologous proteins from another organism could be included in the class and determined instead.

A Protein Structure Initiative: Is it time?

Class-directed structure determinations have the potential to make a rapid and dramatic impact on the amount of useful information about protein structures that is available. To fully realize this potential will require both a substantial investment in infrastructure and support, and coordination of efforts in class-directed structure determination and analysis. A protein structure initiative could provide a focus for this investment and help target large-scale protein structure determination efforts in a coordinated fashion.

A major development of infrastructure for analysis and determination of protein structures will be necessary for efficient class-directed structure determinations because the approach is otherwise likely to run into severe bottlenecks if it becomes widely used. For structure determination by X-ray crystallography using multiwavelength diffraction methods, access to tunable X-ray sources is essential. There are several dozen suitable synchrotron beamlines world-wide, devoted largely to macromolecular crystallography, but even now these facilities are heavily oversubscribed. With a large increase in the rate of obtaining crystals suitable for MAD structure determinations, this situation would become even more difficult unless both the number of beamlines and the throughput at each one could be increased substantially. In the case of NMR, more locations exist where macromolecular structures can be determined, but the time required for data collection is much longer and a large increase in the number of structural analyses does not appear feasible with existing facilities and technology. In the case of both NMR and X-ray crystallography, it seems apparent that not only enhancements in physical facilities, but also development of improved technologies for structure determination and analysis will be essential to obtain a large amount of protein structure information quickly.

Obtaining the large numbers of target protein structures listed in Table 1 will require coordination at a level that a protein structure initiative could provide. Although most of these targets would involve hundreds or even thousands of structure determinations, they could, in principle, be divided into many smaller projects, each of which still could be carried out in a high-throughput, class-directed fashion. For example, once known protein sequences have been clustered into groups based on sequence similarities, each of these groups can form a single class-directed structure determination that could be carried out on a relatively small scale. It might be feasible to have many small research groups working in parallel on a large-scale project to obtain a set of unique protein structure motifs in this way as long as there is a high degree of coordination in both the selection of clusters of similar sequences and in the structure determination efforts. Another aspect of coordination of a project of this kind is the construction and maintenance of databases of materials, such as genomic DNA, and expression libraries and databases of both work in progress and of completed work. Major efforts in informatics will be necessary to complement and target the structural work as well.

An additional and very important reason for a protein structure initiative would be the infrastructure itself that would be created to carry it out. At present, determining a protein structure is a major undertaking, requiring specialized knowledge, experience and equipment, and often taking years to complete. A protein structure initiative could change this completely. An infrastructure of technology for protein expression and purification, and for structure determination, is likely to be developed that will be high-throughput and available to anyone. After a protein structure initiative, protein structure would become a foundation of biochemical research rather than a desirable, but infrequent adjunct.

These are truly exciting times for structural biology and, indeed, all of biology. With the complete sequences of many genomes, carefully constructed databases of functional and structural information at the molecular level and comprehensive methods for analyzing and compiling this wealth of information, our view of life will be vastly broadened and more detailed than it is now. A protein structure initiative could help bring about this vision soon.

Acknowledgments

We are very grateful to the many people with whom we have discussed these ideas and in particular to David Eisenberg, Sung-Hou Kim, Norman Doggett, Michael Altherr, Goutam Gupta, Min Park, Hans Frauenfelder, and Morton Bradbury.

References

- Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known 3-dimensional structure. *Science* 253:164–170.
- Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Venter JC. 1996. Complete genome sequence of the methanogenic archaeon, *methanococcus-jannaschii*. *Science* 273:1058–1073.
- Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823–826.
- Collins FS, Guyer MS, Chakravarti A. 1997. Variations on a theme: Cataloging human dna-sequence variation. *Science* 278:1580–1581.
- Demayer M, Desmet J, Lasters I. 1997. All-in-one: A highly detailed rotamer library improves both accuracy speed in the modeling of side-chains by dead-end elimination. *Folding & Design* 2:53–66.
- Dev SB. 1997. The changing face of biology. *Interdisciplinary Science Reviews* 22:29–36.
- Firestone SM, Nixon AE, Benkovic SJ. 1996. Threading your way to protein function. *Chemistry & Biology* 3:779–783.
- Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb J-F, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Watthey L, McDonald L, Artiach P, Bowman C, Garland S, Fujii C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC. 1997. Genomic sequence of a lyme-disease spirochete, *borrelia-burgdorferi*. *Nature* 390:580–586.
- Frishman D, Mewes HW. 1997. Protein structural classes in 5 complete genomes. *Nature Struct Biol* 4:626–268.
- Gilbert W. 1991. Towards a paradigm shift in biology. *Nature* 349:99.
- Hendrickson WA, Ogata C. 1997. Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol* 276:494–523.
- Hieter P, Boguski M. 1997. Functional genomics: It's all how you read it. *Science* 278:601–602.
- Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–634.
- Pinto AL, Hellinga HW, Caradonna JP. 1997. Construction of a catalytically active iron superoxide-dismutase rational protein design. *Proc Natl Acad Sci USA* 94:5562–5567.
- Rapley R. 1995. The biotechnology and applications of antibody engineering. *Molecular Biotechnology* 3:139–154.
- Rost B, Sander C. 1995. Progress of 1d protein-structure prediction at last. *Proteins Struct Funct Genet* 23:295–300.
- Rowen L, Mahairas G, Hood L. 1997. Sequencing the human genome. *Science* 278:605–607.
- Smith DR, Doucettstamm LA, Deloughery C, Lee HM, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lum W, Pothier B, Qiu DY, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwan N, Caruso A, Bush D, Safer H, Patwell D, Prabhakar S, McDougall S, Shimer G, Goyal A, Petrokovski S, Church GM, Daniels CJ, Mao JI, Rice P, Nolling J, Reeve JN. 1997. Complete genome sequence of *methanobacterium-thermoautotrophicum* delta-h: Functional-analysis and comparative genomics. *J Bacteriol* 179:7135–7155.
- Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou LX, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzgerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC. 1997. The complete genome sequence of the gastric pathogen *helicobacter-pylori*. *Nature* 388:539–547.
- Velinde CLMJ, Hol WGF. 1994. Structure-based drug design: Progress, results and challenges. *Structure* 2:577–587.
- Vogt G, Woell S, Argos P. 1997. Protein thermal-stability, hydrogen-bonds, and ion-pairs. *J Mol Biol* 269:631–643.