Protein Science (1993), *2*, 1798-1810. Cambridge University Press. Printed in the USA.

REVIEW

Knowledge-based model building of proteins: Concepts and examples

JÜRGEN BAJORATH,[1] RONALD STENKAMP,[2] AND ALEJANDRO ARUFFO[1,2]

[1] Bristol-Myers Squibb, Pharmaceutical Research Institute, Seattle, Washington 98121
[2] Department of Biological Structure, University of Washington, Seattle, Washington 98195

(RECEIVED July 20, 1993; ACCEPTED August 23, 1993)

Abstract

We describe how to build protein models from structural templates. Methods to identify structural similarities between proteins in cases of significant, moderate to low, or virtually absent sequence similarity are discussed. The detection and evaluation of structural relationships is emphasized as a central aspect of protein modeling, distinct from the more technical aspects of model building. Computational techniques to generate and complement comparative protein models are also reviewed. Two examples, P-selectin and gp39, are presented to illustrate the derivation of protein model structures and their use in experimental studies.

Keywords: protein modeling; protein structure; sequence similarity; sequence-structure compatibility; structural similarity

The gap between the number of available amino acid sequences and three-dimensional structures of proteins elucidated by crystallography or NMR techniques is expanding rapidly. The rate of sequence determination is at least 50-fold higher than the rate of structure determination (Bowie et al., 1991). It is therefore not surprising to note an increasing interest in predictive methods to derive three-dimensional protein models (Thornton et al., 1991; Fetrow & Bryant, 1993; Rost et al., 1993). We will review current protein modeling techniques, with a focus on knowledge-based methods (Blundell et al., 1987; Greer, 1991). Central to the review will be the question of how meaningful template structures for protein modeling can be identified and used for model building. Furthermore, we will report on two recent examples of knowledge-based model building carried out in our laboratory, P-selectin and gp39, the human ligand for CD40, and will discuss the role of these models for the rationalization and design of experiments.

Homology versus similarity

Knowledge-based model building is often called "modeling by homology." Such modeling techniques start from the premise that known three-dimensional structures can be used to model unknown structures of proteins that exhibit distinct structural similarity to the known structural templates. Homology in its narrow biological meaning, however, defines the relation of proteins that have a common evolutionary origin (Lewin, 1987; Reeck et al., 1987). Homologous sequences and structures, related by divergent evolution, may be more or less similar. In turn, structures related by convergent evolution, by definition nonhomologous, may display high similarity. In this review, we focus on the identification and utilization of structural similarity and not on the question of whether proteins are homologous or not. We therefore use the the terms "comparative model building" rather than "modeling by homology," and "structural similarity" instead of "structural homology." For the same reasons, the terms "sequence identity" and "sequence similarity" are used, the latter taking into account conservative residue replacements.

Comparison of structural models

Methods to compare three-dimensional structures of proteins (Bryant, 1989; Taylor & Orengo, 1989, Šali & Blundell, 1990; Vriend & Sander, 1991; Orengo et al., 1992; Zhu et al., 1992) are essential for assessing the degree of structural similarity. One conventional, simple, and widely

Reprint requests to: Jürgen Bajorath, Bristol-Myers Squibb, 3005 First Avenue, Seattle, Washington 98121.

used quantitative measure of the structural similarity of two macromolecules is the root mean square (rms) distance between equivalent atoms after rigid-body superposition of the molecules. Values of the rms distance ranging from fractions of Ångstroms up to 2 or 3 Å have been taken as evidence for structural similarity. However, it is often difficult to assess such results due to uncertainties about how the numbers are calculated in any particular case. For example, different investigators use different sets of atoms for the superposition (α carbons, main-chain atoms, all atoms, etc.) and calculate the rms distance over different sets of atoms (which may be included in or excluded from the particular superposition set). In addition, the identification and treatment of gaps and loops can affect the rms distance between equivalent atoms significantly.

As an example of the effects of these considerations, we have superposed two well-refined, high-resolution crystallographic rubredoxin models using different sets of equivalent atoms and calculated the rms distance between the models. The proteins used were rubredoxin from *Clostridium pasteurianum* (Watenpaugh et al., 1979), containing 54 amino acids, and rubredoxin from *Desulfovibrio desulfuricans* (Stenkamp et al., 1990), containing 45 residues. When only the α-carbon atoms are used in the superposition and rms distance calculations, and the extra seven-amino acid loop in the larger molecule is omitted, the rms distance between the models is 0.66 Å. When the main-chain atoms are used in the calculation, the rms deviation only increases slightly to 0.68 Å, and when the carbonyl oxygen atoms are included, the rms distance becomes 0.71 Å. The rms value is highly dependent on correct identification of the residues involved in the missing loop in the smaller protein. If two residues in the loop (one at each end) are included in the superposition, the rms distance rises to 0.96 Å, and if four residues are included, the rms distance becomes 1.72 Å. This illustrates that small mistakes in the identification and superposition of equivalent residues in loops or gaps can have significant effects on the calculated rms distance that do not correlate with structural similarity.

From sequence to structure

The ab initio prediction of the three-dimensional structure of a protein from its amino acid sequence is hindered by the fact that the protein-folding problem remains to be understood. The basic question underlying the protein-folding problem and ab initio structure predictions is the same: "How is the three-dimensional structure of a protein determined by its amino acid sequence?" Even a short polypeptide or small protein can, in principle, adopt so many conformations that their systematic generation and evaluation on a computer is an impossible task. Therefore, methods that attempt to predict the structure of a protein de novo, i.e., without the use of a structural template, start from secondary structure predictions, alignments of a number of similar sequences followed by secondary structure assignments (Garnier et al., 1978; Cohen et al., 1986; Bazan, 1990; Benner & Gerloff, 1991), or analysis of hydrophobic patterns in sequences (Henrissat et al., 1990).

Promising predictions of secondary-structure elements have been made recently (Benner & Gerloff, 1991) for the catalytic subunit of the cAMP-dependent protein kinase (Knighton et al., 1991), the Src homology 3 (SH3) domains (Benner et al., 1993; Musacchio et al., 1993), and cytokine receptors (Bazan, 1990, 1992). Some of these examples suggest that the inclusion of evolutionary criteria (i.e., homologous sequences) into the analysis of multiple sequence alignments is likely to improve the accuracy of secondary structure predictions. Despite these recent results, the average accuracy of secondary structure predictions per residue remains between 62% and 70% for all current methods (Rost et al., 1993). Once secondary-structure elements are assigned in approximate fashion, the difficult step of correct spatial assembly of the secondary-structure elements is required (Cohen et al., 1979; Benner, 1992) in order to build a three-dimensional model.

Different from ab initio methods, knowledge-based approaches to protein-structure prediction attempt to relate protein sequences to known three-dimensional protein structures. The basic idea is to identify at least one parent structure that displays essentially the same fold as the protein to be modeled. How such structural relationships are established and how a model can be built and refined will be discussed in some detail below. It is important to note that comparative modeling methods are by definition unable to predict a structure with a fold not yet described experimentally.

Protein folds

Approximately 50% of newly solved experimental structures appear to be related to known folding motifs (Blundell & Doolittle, 1992), consistent with the idea that only a limited number of building blocks of protein structures, i.e., defined spatial combinations of secondary-structure elements, have been generated during evolution. The maximum number of protein families has been estimated to be approximately 1,000 (Chothia, 1992), and the number of families with distinct topology to be 500–700 (Blundell & Johnson, 1993). Thornton and coworkers, using a combination of sequence (Needleman & Wunsch, 1970) and structure (Taylor & Orengo, 1989; Orengo et al., 1992) comparison methods, have reported that the number of proteins with "nonanalogous" folds deposited in the Brookhaven Protein Data Bank (April 1992; Bernstein et al., 1977) is 112 (Thornton, 1992; Orengo et al., 1993). If more rigorous structure comparison criteria are used, 150 "nonhomologous" folds are detected (Orengo et al.,

1993). Clearly, the assessment of protein-fold similarity is dependent on the criteria and "cutoffs" being used. Studies like these illustrate the relation between the more general similarity of folding motifs and specific differences when comparing single structures and their topologies. Topology differences in otherwise similar folding motifs, often found in similar structures related by convergent evolution, are critical for the assignment of protein sequences to "library folds" (Chothia & Finkelstein, 1990).

Comparative model building

The first comparative model building study was carried out by Phillips and colleagues, who derived a model for bovine α-lactalbumin based on the crystal structure of hen egg-white lysozyme (Browne et al., 1969). Comparative modeling as a general method to build three-dimensional models for members of a family of proteins (for which crystal structures of several other members are known) was introduced by Greer (1981, 1990). The procedure begins with careful superposition of the known crystal structures to identify their structurally conserved regions (usually well-defined secondary structure elements) and to distinguish these from the structurally variable regions (mostly loops on the protein surface). A template of combined structurally conserved regions, defining the core of the protein, is created. The sequence of the family member whose structure is to be modeled is included in the structure-based sequence alignment of the particular protein family. This provides the basis for construction of the model. In a meaningful alignment, amino acid insertions and deletions are observed in loop regions (or at most at the termini of secondary-structure elements) but not within α-helices or β-strands. Model structures derived using this methodology include renin (Blundell et al., 1983) and the C5a inflammatory protein (Greer, 1985). Model building of the HIV protease based on the structure of Rous sarcoma virus protease (Weber et al., 1989) and on the more distantly related eukaryotic aspartyl protease family (Pearl & Taylor, 1987) represent attractive examples of comparative modeling because subsequent determination of the crystal structure of the HIV protease (Wlodawer et al., 1989) has allowed a detailed assessment of the generated models (Weber, 1990).

How to build a model

Given the identity of at least one parent structure, how can a model of the related protein be constructed? Again, the first and crucial step is to identify the protein core or structurally conserved regions. If several template structures are available, an average or consensus backbone template may be generated (Sutcliffe et al., 1987). In the next step, nonconserved loop regions need to be modeled. Using the method of Jones and Thirup (1986), candidate loops with the same length as the loop to be modeled and with a reasonable spatial fit onto the adjacent core regions of the template structure can be extracted from the Brookhaven Protein Data Bank. Candidate loops that do not exhibit steric overlaps with the template are examined for sequence similarities with the loop region in the protein to be modeled. Because this procedure will not always lead to convincing solutions for loop regions, the knowledge-based model may have to be complemented with loops generated by conformational search techniques (Moult & James, 1986; Bruccoleri & Karplus, 1987; Shenkin et al., 1987). Loops generated by conformational search are usually filtered using criteria such as low force-field energy, low exposed hydrophobic surface area, low solvent-accessible surface, or solvation free energy (Bruccoleri et al., 1988; Mas et al., 1992).

After the overall backbone model structure is complete, side-chain conformations must be modeled. Conservative side-chain replacements can be carried out using the most similar side-chain conformation found in the rotamer libraries (Ponder & Richards, 1987; Schrauber et al., 1993). Combinatorial methods are available (Novotny et al., 1988; Holm & Sander, 1992; Mas et al., 1992) to generate low (or lowest) energy combinations of side-chain conformers, either using rotamer conformations or, alternatively, systematic conformational search. Such techniques are applicable to model nonconservative side-chain replacements. Close examination of the spatial position of nonconservative residue changes is generally useful for modeling conformations of clusters of spatially close or directly interacting residues. It also helps to understand which part(s) of the model may diverge significantly from the template structures. Distance geometry methods, starting from interatomic distance constraints derived from a selected template structure, have also been used to generate the backbone and some (core) side-chain conformations of model structures (Srinivasan et al., 1993).

Model refinement is the last step in the general model-building procedure. Refinement strategies aim to improve intramolecular contacts, relieve steric strain, and optimize the stereochemistry of the generated model. Therefore, the model is usually subjected to energy minimization calculations. Model refinement is often more critical than it may appear and is far from being a routine step. For example, the degree of such energy minimizations can be critical, and the structural deviations observed in the final model will depend greatly on the specifics of the minimization protocol employed. Conventional force field-based calculations with, for example, their approximate treatment of electrostatic interactions, may induce artificial structural effects and in turn may lead to significant structural deviations relative to the template structure. If the template is a high-quality crystal structure, large deviations should be avoided, yet good stereochemistry and some conformational relaxation of the model should be achieved. This can usually be obtained by applying har-

monic constraints to the backbone atoms of the model during the minimization protocol.

The confidence level of protein models

The procedure outlined above describes the more technical aspects of model building. It can be regarded as a general route for construction of a protein model if it is possible to identify at least one (more or less closely related) parent structure. Nonconserved loop regions and nonconservative side-chain substitutions can be expected to be the least reliable parts of the model structure, provided the modeling study was based on the meaningful selection of at least one template structure. The overall quality of the model depends greatly on the quality of available template structure(s), i.e., resolution, degree of refinement, and potential disorder of parts of the crystallographic (or NMR) model (Bränden & Jones, 1990). These criteria need to be considered when selecting template structures for model-building exercises, particularly because the accuracy of the model is generally lower than the accuracy of the parent structure(s). The accuracy of the model determines the extent to which it can be utilized. Approximate or "low-resolution" models may be sufficient to select potential target residues for site-specific mutagenesis studies or to evaluate the spatial arrangements of the N- and C-termini in the modeled protein. In contrast, the use of model structures to analyze protein–ligand interactions, such as in the design of renin inhibitors (Hutchins & Greer, 1991), requires an accuracy as high as possible, i.e., approaching that of structural models determined experimentally.

Similarity of sequences and structures

It is important to stress that the more technical aspects of protein model building, i.e., the structural manipulations that can be carried out on a computer graphics display and the computational and refinement procedures, are distinct from the identification of structural templates. Protein families such as serine proteases or antibodies, which have often been targeted using standard structure-based modeling techniques (as described above), usually have a common feature: they display significant sequence similarity, often 50–80% or more. In other words, sequence similarity has, in such cases, served as a direct measure for structural similarity. This reflects a generally true assumption: high sequence identity corresponds to distinct structural similarity. For example, the rms deviation for core regions of proteins that display 50% sequence identity can be expected to be approximately 1 Å (Chothia & Lesk, 1986). In cases of high sequence similarity between template and target structure, the assumptions underlying the comparative model-building approach are generally valid.

Three-dimensional structure is, however, significantly more conserved than sequence (Chothia & Lesk, 1986). Sequence similarity scores between proteins in the "twilight zone" (Doolittle, 1985), for example, of ~20% or less, can frequently be observed in sequence searches. What do such sequence similarities mean? Analysis of fragment pairs of protein structures in the Brookhaven Protein Data Bank suggests that sequence identities of ~25% correspond to structural similarity of fragments consisting of 80 residues or more (Sander & Schneider, 1991). For protein structures, such sequence similarities may indicate more distant structural relationships where, for example, secondary structure elements have different length and are somewhat shifted relative to one another. Details in the structures of such proteins may differ considerably. However, core structures of proteins can be very similar despite low or even insignificant sequence similarities. This is known for some well-established folding motifs such as the eight-stranded α/β-barrel, or TIM-barrel (Farber & Petsko, 1990), or the immunoglobulin superfamily (Williams, 1987), including recently "structurally confirmed" members such as the prokaryotic chaperone PapD (Holmgren & Bränden, 1989), CD4 (Ryu et al., 1990; Wang et al., 1990), CD2 (Driscoll et al., 1991; E.Y. Jones et al., 1992), and CD8 (Leahy et al., 1992). In such cases, significant topology differences may be present, despite the similarity of the core structure, and may prohibit the assignment of templates for detailed model building.

More and more examples of structural similarity with moderate to low or virtually nonexistent sequence similarity are being elucidated. Mandelate racemase and muconate lactonizing enzyme display 26% sequence identity, yet their overall structures are strikingly similar (Neidhart et al., 1990). The structures of transforming growth factor-β2 and nerve growth factor exhibit a similar core topology in the absence of sequence similarity (Swindells, 1992). The B-subunit of heat-labile enterotoxin, verotoxin-1, the anticodon-binding domain of Asp-tRNA synthetase, and staphylococcus nuclease show very similar structures but insignificant sequence similarity (Murzin, 1993). A similar observation has been made for the L-arabinose, D-glucose, and D-ribose binding proteins (Vyas et al., 1991). The sequence identity between the heat shock protein fragment HSC70 and actin is less than 15%, but their folds are strikingly similar (Flaherty et al., 1991) and are also similar to the structure of hexokinase, a more distantly related structure (Bränden, 1990). Thus, conventional sequence comparisons are no longer a reliable tool to estimate the degree of structural relationships if the sequence similarity is moderate or low.

Sequence–structure alignments

For protein modeling, therefore, a key question arises. How can structural similarities (that may allow the iden-

tification of a template structure) be detected or confirmed in cases of low to insignificant sequence identities? Characteristic sequence motifs (Bairoch, 1991), which often indicate more functional relationships, may be used to assign sequences to protein families. If at least moderate sequence similarities to a protein with known three-dimensional structure are detectable, perhaps 20–25%, an alignment of the sequence relative to the known crystal structure(s) can be attempted (Cygler et al., 1993; Story et al., 1993). This is analogous to the procedure used to identify structurally conserved regions in a family of proteins with members of known and unknown three-dimensional structure. It may then be possible to evaluate the significance of the conservation or nonconservation of certain residues to the integrity of the particular core structure. The main task is to determine whether residues are conserved in the hydrophobic core region(s) of the protein; whether conserved disulfide bridges, metal coordination sites, or active-site residues can be found; or whether residues are conserved that adopt unusual torsional space, such as glycines at positions where structural or packing constraints would not permit the accommodation of other residues. Sequence–structure alignments of this kind are likely to aid in the identification of residues that are important or characteristic for a particular fold ("folding determinants") and estimation of the degree of structural similarity between proteins with known and unknown structure. Sequence–structure alignments remain critically dependent on a detailed analysis of the crystallographic or NMR structure used and may require significantly more time than the model-building protocol itself.

Sequence–structure alignments are in general more meaningful if multiple sequences (for many members of a protein family) and somewhat distantly related protein structures can be included in the comparison. This emphasizes the importance of methods for comparison of distantly related structures (Bryant, 1989; Taylor & Orengo, 1989; Šali & Blundell, 1990; Vriend & Sander, 1991; Hobohm et al., 1992; Orengo et al., 1992; Zhu et al., 1992) and of methods to create libraries of sequence–structure comparisons and alignments (Sander & Schneider, 1991; Levitt, 1992; Pascarella & Argos, 1992). Alignment of multiple sequences relative to crystal structures is a very informative way to utilize sequence–structure alignments because sequence variability can be assessed relative to three-dimensional constraints. Multiple sequence–structure alignments allow evaluation of the range of residue substitutions permitted at a given position in a structure. It may be recognized that certain positions in a structure require the presence of small hydrophobic, large and bulky, or charged residues. Such classifications take tolerated sequence diversity into account and are much more reliable than sequence alignments for the assignment of sequences to three-dimensional folds. The attraction of this approach was demonstrated by Blum et al. (1993), who found that 60% of the residues in two variant surface glycoproteins of *Trypanosoma brucei* are structurally equivalent despite only 16% sequence identity. This structural comparison was applied in order to modify and refine multiple sequence alignments for a class of variant surface glycoproteins. The generated multiple sequence–structure alignment enabled the authors to predict some detailed structural features of variant surface glycoproteins with unknown structure (Blum et al., 1993).

Absence of sequence similarity

The importance of recognizing structural similarities in virtual absence of any significant sequence similarity is illustrated by a recent analysis of the 182 deduced amino acid sequences of the entire yeast chromosome III (Bork et al., 1992). Only ~13% of these sequences were found to belong to sequence families for which three-dimensional information is available. It is, however, not possible to conclude from this analysis that only 13% of these sequences represent structures that are related to known three-dimensional folds. Conventional alignment of sequences relative to crystal structures are usually based on a "first hint" of some (even low) sequence similarity or, alternatively, require initial recognition of structural similarity in different experimental structures. The identification of structural similarity in the absence of obvious sequence relations and in the absence of direct structural comparisons requires a different approach.

The inverse folding approach

Recently, progress has been made in the development of methods that start from what is called the "inverse folding problem" (Blundell & Johnson, 1993; Bowie & Eisenberg, 1993; Fetrow & Bryant, 1993; Wodak & Rooman, 1993). In contrast to the protein folding problem, the initial question of the inverse folding approach is "Because we are presently unable to understand how a sequence determines a protein fold, can we determine which amino acid sequences are compatible with a given three-dimensional structure?" In contrast to the protein folding approach, the starting point here is structure, not amino acid sequence. For comparative protein modeling, the question asked would be "Given the sequence of a protein with unknown structure, is this sequence compatible with a known three-dimensional fold?"

Investigation of the inverse folding problem started from the idea that one could design a sequence that would be consistent with a given protein structure based on the interactions that are found within this structure (Pabo, 1983). Novotny and colleagues (1984, 1988) generated and analyzed protein models that were misfolded deliberately. They fitted the sequence of hemerythrin, a four-helix bundle structure, onto the β-sheet fold of an immunoglobulin variable domain and found that force field-based

calculations were unable to discriminate between models folded correctly and those folded incorrectly (Novotny et al., 1984). But criteria such as solvent-exposed hydrophobic surface area and solvation effects were able to do so (Novotny et al., 1988), consistent with the results obtained by other groups (Eisenberg & McLachlan, 1986; Chiche et al., 1990; Sander & Holm, 1992). More direct sequence–structure compatibility studies were carried out by Ponder and Richards (1987), who analyzed which combinations of amino acids were able to reproduce the interior packing of some protein structures. The inverse folding approach has, in fact, led to the development of methods that allow the detection of structural similarities in the absence of sequence similarity. All methods developed to date determine the compatibility of sequences with protein folds and vice versa. The specific approach to the problem differs, however, and may be divided into two groups.

One approach, introduced by Eisenberg and coworkers (Bowie et al., 1991) and Blundell and coworkers (Overington et al., 1992), translates a given protein structure from three-dimensional into one-dimensional space, not as a sequence of residues but rather as a sequence of specific residue environments at each position of the structure. In order to define categories of residue environments, criteria such as solvent-accessible surface, buried polar surface area, and secondary structure are used. Each of the 20 amino acids has a certain probability, based on the chemical nature of its side chain, to be found in one of the defined environmental classes, presently 18 in the implementation of Eisenberg and coworkers. An overall compatibility score for a certain amino acid sequence with the "3D-profile" (Bowie et al., 1991) of a given three-dimensional structure is calculated essentially by adding the probability of occurrence for each residue in the specific residue environment defined at a position in the three-dimensional structure. In principle, it is expected that the higher the respective score, the more distinct the similarity of the structures. Eisenberg and coworkers were able to detect the structural similarity between HSC70 and actin and to discriminate between the correct and intentionally misfolded protein models of Novotny and colleagues (Lüthy et al., 1992). Profile methods are dependent on finding the best alignment of a sequence of residues with a sequence of environmental classes (i.e., a 3D-profile) by the use of dynamic programming techniques. This involves the assignment of insertions and deletions, critical parameters for inverse folding methods. Profile methods also depend on the "coarseness of the grid" of the defined environmental classes and on the validity of the criteria used to define them. Sequences with high similarity to the sequence of the structure for which a 3D-profile is generated are expected to score high (Bryant & Lawrence, 1993). This reflects the fact that high sequence identity is correlated with distinct structural similarity.

For comparative model-building attempts, profile methods are attractive because they permit the calculation of 3D-profiles for each unique structure deposited in the Brookhaven Protein Data Bank and the examination of novel sequences relative to the calculated 3D-profiles. The indication of certain structural similarity is, however, not always sufficient to build a reasonably accurate model for a novel protein because similar folds may still have some significant structural or topology differences, as mentioned above, such as observed in the second domains of CD4 and PapD compared to an immunoglobulin constant domain. This represents a general problem that is difficult to circumvent in building models of more distantly related proteins. The 3D-profile method offers, at the least, a means of assessing the accuracy of a protein model, be it derived from experiment or theoretical modeling (Lüthy et al., 1992). By profile comparison of a model to its own sequence, a global incompatibility of sequence and structure (a mistraced crystallographic model or a knowledge-based model based on an incorrect initial hypothesis) can be identified, and more localized errors, such as β-strands that are "out of register" or buried single charged residues, may be detected in otherwise valid structural models. For comparative modeling, this suggests a method to assess model structures "in retrospect": once a model is generated based on the fold of a given structure, a profile comparison can be carried out with the (experimental) template structure scored against its own sequence and the model structure (the same fold) against its sequence. Comparison of these profiles can frequently identify errors in a model structure that may, for example, correspond to some topology differences in the template and the target structure.

The second class of methods based on the inverse folding approach (Finkelstein & Reva, 1991; Godzik & Skolnick, 1992; D.T. Jones et al., 1992; Maiorov & Crippen, 1992; Sippl & Weitckus, 1992; Bryant & Lawrence, 1993) essentially starts from the premise that a three-dimensional structure determines which residues in the sequence interact with one another and which do not. At each position in a three-dimensional structure, a residue interacts specifically with spatially related residue positions. Using distance criteria, pairs of interacting residues can be determined and classified for each position in a structure. If a sequence is fitted, or "threaded," onto a fold, the residues that occupy certain spatial positions may display more or less favorable interaction patterns, depending on their chemical nature and on the nature of their interaction partners. The pairwise interactions are defined mainly by backbone distances and are expressed using pairwise interaction potentials, approximate conformational energy terms such as the potentials introduced by Sippl (1990). Pseudoconformational energies are summed over all residue positions and result in a more favorable (more negative) or less favorable (more positive) energy, a measure for sequence–structure compatibility. All in-

teraction potentials are derived principally from the analysis of residue interaction patterns found in protein structures. The specific details and the number of parameters used by the different groups for the formulation of residue interactions vary considerably. The results are promising. For example, Thornton and coworkers (D.T. Jones et al., 1992) were able to detect the structural relation between actin and HSC70 and reported the detection of similarity between actin and the structurally more distantly related hexokinase (which was not found in the profile searches of Eisenberg and coworkers). Recently, the 3D-profile method was extended to include neighbor residue preferences (Wilmanns & Eisenberg, 1993), somewhat analogous to the pairwise interaction energies used in residue contact methods. This combination has allowed the identification of sequences that display the α/β-barrel fold (Wilmanns & Eisenberg, 1993).

Examples of comparative model building: P-selectin

Having reviewed the concepts and recent developments in the field of structure-based model building, we will now discuss two examples of protein model structures that were built in our laboratory. Both examples are proteins with low sequence identity to known structures.

As the first example, we discuss P-selectin, a member of the selectin family of cell adhesion molecules (Springer, 1990; Lasky, 1992). The selectins are type I membrane glycoproteins with similar molecular organization and function. The extracellular domains of the selectins are composed of an amino terminal C-type lectin-like domain followed by an EGF-like domain and a variable number of repeats homologous to complement regulatory proteins. These molecules mediate their adhesion function by binding interactions between the lectin domain of the selectin and carbohydrate ligands on the opposing cell. P-selectin is expressed in the α-granules of platelets and the Weibel-Palade bodies of the vascular endothelial cells. Following platelet or endothelial cell activation, P-selectin is rapidly redistributed from intracellular stores to the cell surface, where it mediates leukocyte-platelet or leukocyte-endothelial cell adhesion. Recently, a series of elegant in vitro studies by Lawrence and Springer (1991) have provided convincing evidence that P-selectin is in part responsible for mediating the rolling of leukocytes on activated vascular endothelium, a prerequisite for leukocyte extravasation at sites of inflammation. These experiments have been corroborated by the observation that anti-P-selectin monoclonal antibodies can block acute inflammatory responses in vivo. The exact composition of the physiological ligand of P-selectin remains to be elucidated. However, binding studies in vitro have shown that P-selectin can bind to Le^x, sialyl-Le^x, sulfatides, sulfoglucuronyl glycosphingolipids, and an ~250-kDa glycoprotein expressed by leukocytes.

Model building of selectin-ligand binding domains was long hindered by the fact that no meaningful structural template could be identified. The situation changed when the crystal structure of the lectin domain of a rat mannose-binding protein was solved (Weis et al., 1991). The structure of the C-type lectin domain of the mannose-binding protein revealed a novel protein fold with a significant content of nonclassical secondary structure. The sequence identity between the rat mannose-binding protein and the selectin lectin domains is ~25%. Weis et al. (1991) presented a sequence-structure alignment that suggested that the disulfide bonds, many of the residues in the two hydrophobic core region of the lectin domain, and the residues of one of the two calcium-binding sites are conserved in the selectin family. This demonstrated the potential significance of sequence–structure alignments and provided an ideal basis to generate a corresponding alignment of P-selectin relative to the lectin domain of the rat mannose-binding protein. Starting from this alignment (Fig. 1A), model building of P-selectin was conducted (Hollenbaugh et al., 1993).

The model structure of the P-selectin ligand-binding domain (Fig. 1B), constructed from the coordinates of the mannose-binding protein at 2.5 Å resolution, was complemented by conformational search calculations for loop regions whose conformations could not be modeled from known crystallographic structures (Hollenbaugh et al., 1993). A 3D-profile analysis was then used to assess the compatibility of the P-selectin sequence with its model structure relative to the mannose-binding protein sequence and structure. This analysis is shown in Figure 1C. Comparison of the profiles of the model and crystal structure suggested the global compatibility of the P-selectin sequence with the fold of the C-type lectin domain of the mannose-binding protein. The Z-scores were comparable for both structural models. This was consistent with the conclusions reached by the sequence–structure alignment. Potential local inconsistencies in the P-selectin model could not be detected. Local errors may, however, still be present in the P-selectin model. For example, residues in a surface loop may still remain in a favorable environmental class (and, therefore, score high) if the conformation of the loop were modeled incorrectly. Such local errors are unlikely to be detected by comparative profile analysis. The novel fold of this protein family, with its unusually high content of nonclassical secondary structure, illustrates how crucial the identification of a closely related structural template has been for model building of P-selectin. Using the mannose-binding protein as structural template, models of E-selectin (Erbe et al., 1992; Mills, 1993) and P-selectin (Erbe et al., 1993) were also derived by similar approaches.

The P-selectin model was used in order to identify residues in P-selectin crucial for the binding to its cellular ligand. Different from other carbohydrate-binding proteins, the crystal structure of the mannose-binding pro-

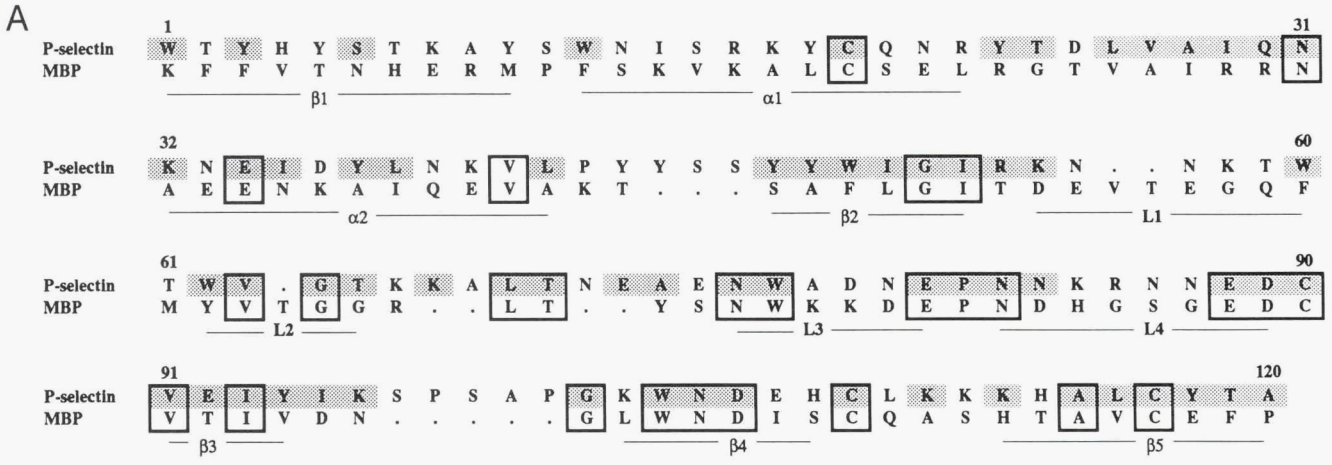

Fig. 1. A: Structure-based sequence alignment of C-type lectin domains of P-selectin and the mannose-binding protein (MBP). Secondary-structure elements in MBP are labeled. Residues conserved in MBP and P-selectin are boxed and residues conserved in all selectins are shaded. This sequence–structure alignment provided the basis for comparative model building of the P-selectin ligand-binding domain. **B:** Model structure of the ligand-binding domain of P-selectin. The model, represented as a solid ribbon, was built based on identified structural similarity to the crystal structure of the C-type lectin domain of the rat mannose-binding domain, which revealed a previously unknown protein fold. The view is along α-helix 2, located below the loop region colored in yellow. A conserved calcium position is shown in red. Analysis of the model suggested a shallow depression proximal to the conserved calcium as a potential ligand-binding site. This region is flanked on the left by a loop, colored in yellow, with a five-residue insertion relative to the mannose-binding protein. Residues in the proposed ligand-binding region of P-selectin were subjected to site-specific mutagenesis analysis, confirming the hypothesis regarding the location of residues in P-selectin critical for binding to its cellular ligand. The residues are shown in a color-coded fashion: magenta, crucial for binding; lavender, significant contribution to binding; blue, minor contribution to binding. **C:** Assessment of the P-selectin model by comparison of 3D-profiles of the MBP crystal structure (relative to the MBP sequence) and of the P-selectin model structure (relative to its sequence). The profiles were calculated using a 21-residue window for score averaging. The calculated *Z*-score for the MBP sequence and crystal structure is 30.8, and the *Z*-score for the P-selectin sequence and model is 34.9. No negative profile values were observed that would indicate local inconsistencies in the structural models. The analysis suggests an equivalent compatibility of sequence and structure for the model and the X-ray structure.

tein did not display a distinct crevice or cavity on its surface (Weis et al., 1991), an attractive region for binding of a carbohydrate. One of the major structural differences between the selectin-lectin domains and the mannose-binding protein is a five-residue insertion in a surface loop after residue 96 in P- and E-selectin (highlighted in Fig. 1B). The extension of this loop results in the formation of a shallow groove on the P-selectin surface that is apparent when the solvent-accessible surface of the model is inspected. This region, proximal to the conserved calcium-binding site, was selected as an attractive region for ligand binding to P-selectin, and residues in this region were subjected to site-specific mutagenesis (Hollenbaugh et al., 1993).

The mutagenesis analysis identified several residues crucial for the binding of P-selectin to its cellular ligand and generated a picture of the P-selectin ligand-binding site shown in Figure 1B. Independent studies identified the same region in P-selectin (Erbe et al., 1993) and the corresponding region in E-selectin (Erbe et al., 1992) to be crucial for E- and P-selectin binding to immobilized glycolipids. A subsequent crystallographic analysis on the mannose-binding protein in complex with high mannose (Weis et al., 1992) has shown that a mannose residue coordinates directly to the conserved calcium and also has suggested how a fucose residue, a part of sialylated Lewis X (sLe^x), may bind to this calcium in the selectins (Weis et al., 1992). These results, combined with the identification of the P-selectin ligand-binding site region and with the identification of residues crucial for ligand binding, provide the basis for studying selectin-ligand interactions in more detail.

Model building of gp39, the ligand of CD40

The second example is gp39, the ligand for CD40 (Hollenbaugh et al., 1992). The ligand gp39 is a type II membrane protein expressed by activated T cells (Armitage et al., 1992; Hollenbaugh et al., 1992; Spriggs et al., 1992), whose extracellular domain displays some sequence similarity to tumor necrosis factor (TNF). In vitro studies have shown that B cells expressing CD40 can be driven to proliferate and differentiate into antibody-secreting cells when incubated with fibroblasts expressing membrane-bound gp39 or a soluble recombinant form of gp39 and cytokines, such as IL-4 and IL-10. This suggests that gp39 provides important T cell-dependent signals for B cells. Further evidence for the critical role of gp39-CD40 interactions in T cell-dependent B cell responses has come from the identification of the molecular defect responsible for the human antibody production deficiency known as hyper-IgM syndrome (HIM) (Aruffo et al., 1993; Hill & Chapel, 1993; Marx, 1993). The B cells from HIM patients are unable to isotype switch and to mount an effective humoral immune response against foreign antigens. We and other groups have shown that T cells from HIM patients either do not express gp39, or express a defective form of gp39, following activation. Lack of CD40 engagement during T cell-B cell interactions in these patients results in inadequate B cell stimulation and defective antibody production.

The extracellular domain of gp39 contains approximately 150 amino acids. Sequence searches revealed similarity to TNFα and TNFβ, whose crystal structures have been solved (Eck & Sprang, 1989; Eck et al., 1992) and shown to display the same fold: a homotrimeric all-β structure consisting of β-sandwich monomers with "jellyroll" topology (Eck & Sprang, 1989). The trimer is formed along a threefold symmetry axis through the base of the "bell-shaped" molecule. The sequences of gp39 and TNFα are ~20% identical. The availability of the crystal structure of TNFα in the Brookhaven Protein Data Bank enabled us to attempt a detailed sequence-structure alignment of gp39 relative to TNFα (Aruffo et al., 1993). The analysis revealed the conservation or conservative replacement of the majority of putative "folding determinants" of the TNFα structure (Eck & Sprang, 1989), i.e., residues involved in β-sheet packing interactions and/or dimer or trimer subunit interactions. Insertions and deletions were accommodated in loop regions, and the alignment was supported further by the fact that two cysteine residues in gp39 were in appropriate spatial positions to form a disulfide bond. The disulfide bonds of TNFα are not conserved in gp39 or TNFβ.

Overall, the analysis suggested that the degree of structural similarity between TNFα and gp39 was more distinct than was suggested by its sequence identity. This provided the basis for model building of gp39, beginning with the crystallographic coordinates of TNFα at 2.6 Å resolution. Four loop regions of the model, with relative insertions and deletions, were reconstructed using systematic conformational search, and the trimeric model of the gp39 was refined using a constrained energy minimization protocol. The model (Fig. 2A) was then assessed using 3D-profiles relative to the TNFα crystal structure and its sequence (Fig. 2B), analogous to the analysis shown for P-selectin. The 3D-profile analysis, carried out for the monomeric as well as the trimeric forms, confirmed the compatibility of the gp39 sequence with its assumed fold. As was found for P-selectin and the mannose-binding protein, the Z-scores for TNFα and gp39 were comparable.

Recently, we showed that defective forms of gp39, which lead to severe immunodeficiencies in patients, were caused by naturally occurring site-specific mutations in gp39 (Aruffo et al., 1993). Mutations in several patients were analyzed. We used our gp39 model structure to examine the spatial positions of these mutations. Interestingly, these mutations map to spatially distant surface regions in gp39 and are, somewhat different from mutant forms isolated by us and other groups, not likely to prevent correct gross folding or trimer formation of the gp39

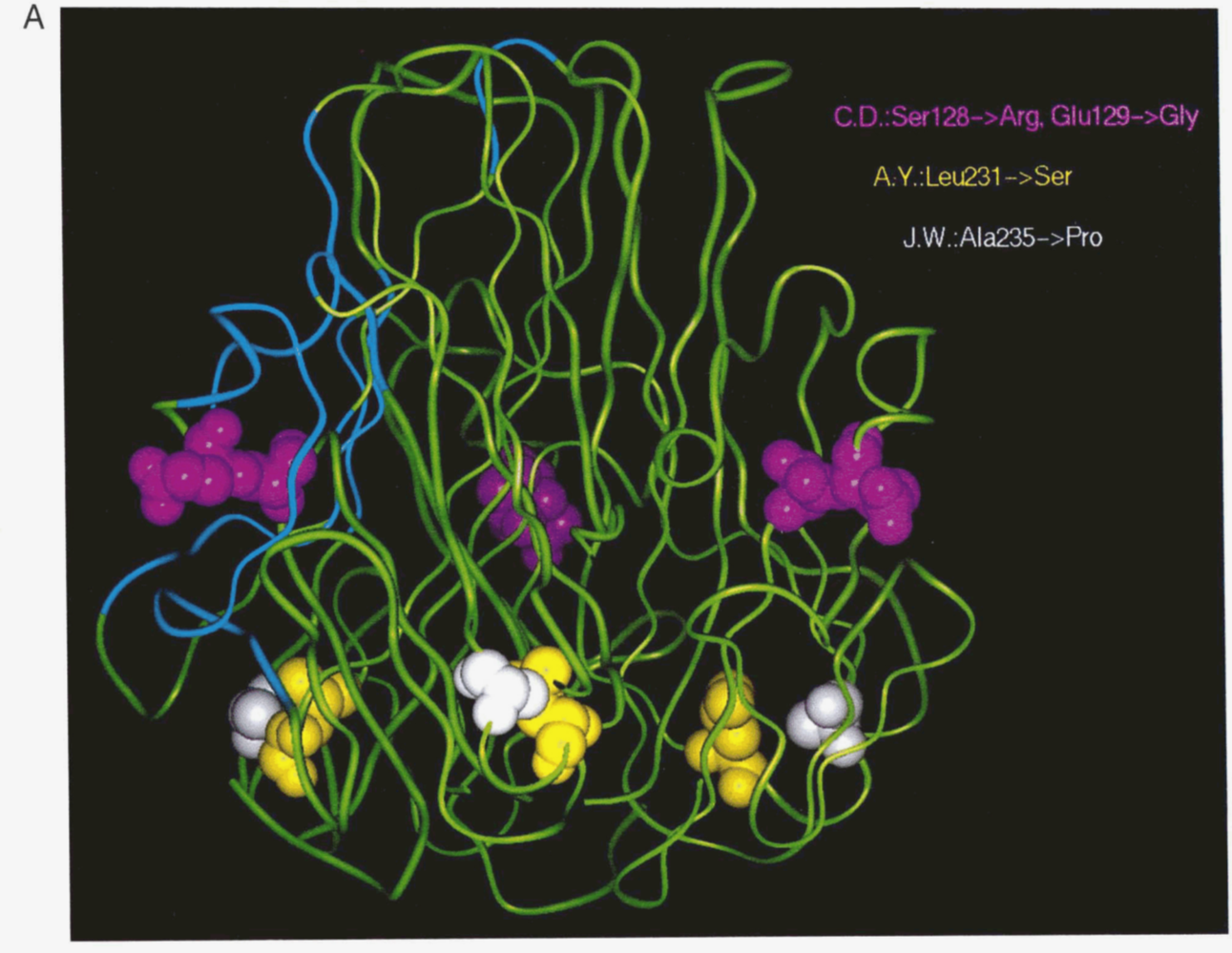

Fig. 2. A: Model structure of the extracellular domain of gp39. The structure of the homotrimer is shown in green. The *y*-axis is approximately parallel to the threefold molecular symmetry axis. The locations of site-specific natural gp39 mutants in three patients (C.D., A.Y., and J.W.) are shown. These mutations impair the function of gp39 in these patients. One of the three symmetry-related regions in gp39 that correspond to the receptor-binding sites in tumor necrosis factor beta (TNFβ) is colored in blue. Locations of the three mutations shown here suggests that these residue changes may interfere with the binding of gp39 to its receptor CD40. **B:** Comparison of calculated 3D-profiles, using a 21-residue window, for the trimeric tumor necrosis factor alpha (TNFα) crystal and the trimeric gp39 model structure relative to their sequences. The calculated *Z*-scores for the TNFα crystal structure and sequence and for the gp39 model and sequence were 30.8 and 32.8, respectively. The analysis supports the proposed structural similarity of gp39 and TNFα.

molecule. The location of these mutants in the gp39 trimer is shown in Figure 2A.

How do these mutants impair the function of gp39? A hypothesis was developed by further evaluation of the relationship between gp39 and TNFα. Site-specific mutagenesis analysis on TNFα has led to the identification of a surface region in TNFα important for receptor binding (Zhang et al., 1992). A coherent receptor-binding region in TNFα is formed by residues of two monomers. It follows that, in the trimer, three equivalent spatially distant and symmetry-related receptor-binding regions should exist. Consistent with this assumption, it was reported that

TNFβ-receptor crystals were obtained with one TNFβ-trimer in complex with three symmetry-related soluble receptor fragments (D'Arcy et al., 1993). The crystal structure of the complex (Banner et al., 1993) confirmed the binding of three receptor fragments to the proposed binding sites, each formed by two monomers. Figure 2A shows the region in gp39 corresponding to the identified receptor-binding region of TNFα and the location of the gp39 mutants relative to this region. As can be seen, the location of the mutants essentially maps to the analogous region in gp39. This suggests the testable hypothesis that the receptor-binding regions in TNFα and gp39 correspond spatially and, furthermore, suggests a possible explanation for the role of the isolated gp39 mutants that lead to nonfunctional gp39 molecules. These residue changes may in fact interfere, directly or indirectly, with the binding of gp39 to its receptor.

Conclusions

We have reviewed how protein models can be derived from structural templates and have emphasized the significance of concepts and methods that allow the detection of structural similarity in cases of moderate to insignificant sequence identity. Assessment of the degree of structural similarity between template and unknown structures is important because models based on more distant structural relationships are limited in their use for detailed experimental design. Evaluation of structural similarity is distinct from the more technical aspects of model building. The inverse folding approach offers additional ways to assess the confidence level of model structures.

We also have discussed two examples of protein model building in cases of low sequence similarity and have shown how these models were used for the rationalization and the design of experiments. The P-selectin model structure has allowed us to identify residues that are crucial for the binding of P-selectin to its cellular ligand. The gp39 model suggests that some naturally occurring nonfunctional forms of gp39 have specific mutations that may interfere with the binding of gp39 to its receptor CD40.

References

Armitage, R.J., Fanslow, W.C., Strockbine, L., Sato, T.A., Clifford, K.N., Macduff, B.M., Anderson, D.M., Gimpel, S.D., Davis-Smith, T., Maliszewski, C.R., Clark, E.A., Smith, C.A., Grabstein, K.H., Cosman, D., & Spriggs, M.K. (1992). Molecular and biological characterization of a murine ligand for CD40. *Nature 357*, 80–82.

Aruffo, A., Farrington, M., Hollenbaugh, D., Xu, L., Milatovich, A., Nonoyama, S., Bajorath, J., Grosmaire, L.S., Stenkamp, R., Neubauer, M., Roberts, R.L., Noelle, R.J., Ledbetter, J.A., Francke, U., & Ochs, H.D. (1993). The CD40 ligand, gp39, is defective in activated T cells from patients with X-linked hyper-IgM syndrome. *Cell 72*, 1–20.

Bairoch, A. (1991). PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res. 16*, 2241–2245.

Banner, D.W., D'Arcy, A., Janes, W., Gentz, R., Schoenfeld, H.-J., Broger, C., Loetscher, H., & Lesslauer, W. (1993). Crystal structure of the soluble human 55 kd TNF receptor-human TNFβ complex: Implications for TNF receptor activation. *Cell 73*, 431–445.

Bazan, J.F. (1990). Structural design and molecular evolution of a cytokine receptor family. *Proc. Natl. Acad. Sci. USA 87*, 6934–6938.

Bazan, J.F. (1992). Unraveling the structure of IL-2. *Science 257*, 410–412.

Benner, S.A. (1992). Predicting de novo the folded structure of proteins. *Curr. Opin. Struct. Biol. 2*, 402–412.

Benner, S.A., Cohen, M.A., & Gerloff, D. (1993). Predicted secondary structure for the Src homology 3 domain. *J. Mol. Biol. 229*, 295–305.

Benner, S.A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: The catalytic domain of protein kinases. *Adv. Enzyme Regul. 31*, 121–181.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Blum, M.L., Down, J.A., Gurnett, A.M., Carrington, M., Turner, M.J., & Wiley, D.C. (1993). A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*. *Nature 362*, 603–609.

Blundell, T.L. & Doolittle, R.F. (1992). Sequences and topology—An inverse approach to the old folding problem. *Curr. Opin. Struct. Biol. 2*, 381–383.

Blundell, T.L. & Johnson, M.S. (1993). Catching a common fold. *Protein Sci. 2*, 877–883.

Blundell, T.L., Sibanda, B.L., & Pearl, L. (1983). Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature 304*, 273–275.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J., & Thornton, J.M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature 326*, 347–352.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., & Sonnhammer, E. (1992). What's in a genome? *Nature 358*, 287.

Bowie, J.U. & Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol. 3*, 437–444.

Bowie, J.U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science 235*, 164–170.

Bränden, C.-I. (1990). Founding fathers and families. *Nature 346*, 607–608.

Bränden, C.-I. & Jones, T.A. (1990). Between subjectivity and objectivity. *Nature 343*, 687–698.

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C., & Hill, R.L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg white lysozyme. *J. Mol. Biol. 42*, 65–86.

Bruccoleri, R.E., Haber, E., & Novotny, J. (1988). Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature 335*, 564–568.

Bruccoleri, R.E. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers 26*, 137–168.

Bryant, S.H. (1989). PKB: A program system and data base for analysis of protein structure. *Proteins Struct. Funct. Genet. 5*, 233–247.

Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins Struct. Funct. Genet. 16*, 92–112.

Chiche, L., Gregoret, L.M., Cohen, F.E., & Kollman, P.A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci. USA 87*, 3240–3243.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature 357*, 543–544.

Chothia, C. & Finkelstein, A.V. (1990). The classification and origin of protein folding patterns. *Annu. Rev. Biochem. 53*, 1007–1039.

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823–826.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., & Fletterick, R.J. (1986). Turn prediction using a pattern matching approach. *Biochemistry 25*, 266–275.

Cohen, F.E., Richmond, T.J., & Richards, F.M. (1979). Protein folding: Evaluation of simple rules for the assembly of helices into ter-

tiary structure with myoglobin as an example. *J. Mol. Biol. 132*, 275–288.
Cygler, M., Schrag, J.D., Sussman, J.L., Harel, M., Silman, I., Gentry, A.K., & Doctor, B.P. (1993). Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins. *Protein Sci. 2*, 366–382.
D'Arcy, A., Banner, D.W., Janes, W., Winkler, F.K., Loetscher, H., Schönfeld, H.-J., Zulauf, M., Gentz, R., & Lesslauer, W. (1993). Crystallization and preliminary crystallographic analysis of a TNF-β-55 kDa TNF receptor complex. *J. Mol. Biol. 229*, 555–557.
Doolittle, R.F. (1985). The genealogy of some recent evolved vertebrate proteins. *Trends Biochem. Sci. 10*, 233–237.
Driscoll, P.C., Cyster, J.G., Campbell, I.D., & Williams, A.F. (1991). Structure of rat T lymphocyte CD2 antigen. *Nature 353*, 762–765.
Eck, M.J. & Sprang, S.R. (1989). The structure of tumor necrosis factor alpha at 2.6 Å resolution. *J. Biol. Chem. 264*, 17595–17604.
Eck, M.J., Ultsch, M., Rinderknecht, E., de Vos, A.M., & Sprang, S.R. (1992). The structure of human lymphotoxin (tumor necrosis factor-β) at 1.9 Å resolution. *J. Biol. Chem. 267*, 2119–2122.
Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature 319*, 199–203.
Erbe, D.V., Watson, S.R., Presta, L.G., Wolitzky, B.A., Foxall, C., Brandley, B.K., & Lasky, L.A. (1993). P- and E-selectin use common sites for carbohydrate recognition and cell adhesion. *J. Cell Biol. 120*, 1227–1235.
Erbe, D.V., Wolitzky, B.A., Presta, L.G., Norton, C.R., Ramos, R.J., Burns, D.K., Rumberger, J.M., Rao, B.N.N., Foxall, C., Brandley, B.K., & Lasky, L.A. (1992). Identification of an E-selectin region critical for carbohydrate recognition and cell adhesion. *J. Cell Biol. 119*, 215–227.
Farber, G.K. & Petsko, G.A. (1990). The evolution of alpha/beta barrel enzymes. *Trends Biochem. Sci. 15*, 228–234.
Fetrow, J.S. & Bryant, S.H. (1993). New programs for protein structure prediction. *Biotechnology 11*, 479–484.
Finkelstein, A.V. & Reva, B.A. (1991). A search for the most stable fold of proteins. *Nature 351*, 497–499.
Flaherty, K.M., McKay, D.B., Kabsch, W., & Holmes, K.C. (1991). 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.
Garnier, J., Osguthorpe, D.J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol. 120*, 97–120.
Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA 89*, 12098–12102.
Greer, J. (1981). Comparative model building of mammalian serine proteases. *J. Mol. Biol. 153*, 1027–1042.
Greer, J. (1985). Model structure for the inflammatory protein C5a. *Science 228*, 1055–1060.
Greer, J. (1990). Comparative modeling methods: Applications to the family of the mammalian serine proteases. *Proteins Struct. Funct. Genet. 7*, 317–334.
Greer, J. (1991). Comparative modeling of homologous proteins. *Methods Enzymol. 202*, 239–252.
Henrissat, B., Saloheimo, M., Lavaitte, S., & Knowles, J.K.C. (1990). Structural homology among the peroxidase enzyme family revealed by hydrophobic cluster analysis. *Proteins Struct. Funct. Genet. 8*, 251–257.
Hill, A. & Chapel, H. (1993). The fruits of cooperation. *Nature 361*, 494.
Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci. 1*, 409–417.
Hollenbaugh, D., Bajorath, J., Stenkamp, R., & Aruffo, A. (1993). Interaction of P-selectin (CD62) and its cellular ligand: Analysis of critical residues. *Biochemistry 32*, 2960–2966.
Hollenbaugh, D., Grosmaire, L.S., Kullas, C.D., Chalupny, N.J., Braesch-Andersen, S., Noelle, R.J., Stamenkovic, I., Ledbetter, J.A., & Aruffo, A. (1992). The human T cell antigen gp39, a member of the TNF gene family, is a ligand for the CD40 receptor: Expression of a soluble form of gp39 with B cell co-stimulatory activity. *EMBO J. 11*, 4313–4321.
Holm, L. & Sander, C. (1992). Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins Struct. Funct. Genet. 14*, 213–223.
Holmgren, A. & Bränden, C.-I. (1989). Crystal structure of chaperone protein PapD reveals an immunoglobulin fold. *Nature 342*, 248–251.
Hutchins, C. & Greer, J. (1991). Comparative modeling of proteins in the design of novel renin inhibitors. *Crit. Rev. Biochem. Mol. Biol. 26*, 77–127.
Jones, D.T., Taylor, W.R., & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature 358*, 86–89.
Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K., & Stuart, D.I. (1992). Crystal structure of a soluble form of the cell adhesion molecule CD2. *Nature 360*, 232–239.
Jones, T.A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J. 5*, 819–822.
Knighton, D.R., Zheng, J., Ten Eyck, F.F., Ashford, V.A., Xuong, N.H., Taylor, S.S., & Sowadski, J.M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science 253*, 404–414.
Lasky, L.A. (1992). Selectins: Interpreters of cell-specific carbohydrate information during inflammation. *Science 258*, 964–969.
Lawrence, M.B. & Springer, T.A. (1991). Leukocytes roll on a selectin at physiologic flow rates: Distinction from and prerequisite for adhesion through integrins. *Cell 65*, 859–873.
Leahy, D.J., Axel, R., & Hendrickson, W.A. (1992). Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 Å resolution. *Cell 68*, 1145–1162.
Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol. 226*, 507–533.
Lewin, R. (1987). When does homology mean something else? *Science 237*, 1570.
Lüthy, R., Bowie, J.U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature 356*, 83–85.
Maiorov, V.N. & Crippen, G.M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol. 227*, 876–888.
Marx, J. (1993). Cell communication failure leads to immune disorder. *Science 259*, 896–897.
Mas, M.T., Smith, K.C., Yarmush, D.L., Aisaka, K., & Fine, R.M. (1992). Modeling the anti-CEA antibody combining site by homology and conformational search. *Proteins Struct. Funct. Genet. 14*, 483–498.
Mills, A. (1993). Modelling of the carbohydrate recognition domain of human E-selectin. *FEBS Lett. 319*, 5–11.
Moult, J. & James, M.N.G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins Struct. Funct. Genet. 1*, 146–163.
Murzin, A.G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J. 12*, 861–867.
Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., & Saraste, M. (1992). Crystal structure of a SRC-homology 3 (SH3) domain. *Nature 359*, 851–855.
Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol. 48*, 443–453.
Neidhart, D.J., Kenyon, G.L., Gerlt, J.A., & Petsko, G.A. (1990). Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature 347*, 692–694.
Novotny, J., Bruccoleri, R.E., & Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure prediction. *J. Mol. Biol. 177*, 787–818.
Novotny, J., Rashin, A.A., & Bruccoleri, R.E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct. Funct. Genet. 4*, 19–30.
Orengo, C.A., Brown, N.P., & Taylor, W.R. (1992). Fast structure alignment for protein databank searching. *Proteins Struct. Funct. Genet. 14*, 139–167.
Orengo, C.A., Flores, T.P., Taylor, W.R., & Thornton, J.M. (1993). Identification and classification of protein fold families. *Protein Eng. 6*, 485–500.
Overington, J., Donnelly, D., Johnson, M.S., Šali, A., & Blundell, T.L. (1992). Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci. 1*, 216–226.
Pabo, C. (1983). Designing proteins and peptides. *Nature 301*, 200.
Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng.*, 121–137.

Pearl, L.H. & Taylor, W.R. (1987). A structural model for the retroviral proteases. *Nature 329,* 351–354.

Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences of allowed sequences for different structural classes. *J. Mol. Biol. 193,* 775–791.

Reeck, G.R., de Häen, C., Teller, D.C., Doolittle, D.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., & Zuckerkandl, E. (1987). "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell 50,* 667.

Rost, B., Schneider, R., & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem. Sci. 18,* 120–123.

Ryu, S.-E., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.-h., Axel, R., Sweet, R.W., & Hendrickson, W.A. (1990). Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature 348,* 419–425.

Šali, A. & Blundell, T.L. (1990). The definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol. 212,* 403–428.

Sander, C. & Holm, L. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol. 225,* 93–105.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet. 9,* 56–68.

Schrauber, H., Eisenhaber, F., & Argos, P. (1993). Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol. 230,* 592–612.

Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H., & Levinthal, C. (1987). Predicting antibody hypervariable loop conformation I. Ensembles of random conformations for ringlike structures. *Biopolymers 26,* 2053–2085.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol. 216,* 859–883.

Sippl, M.J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins Struct. Funct. Genet. 13,* 258–271.

Spriggs, M.K., Armitage, R.J., Strockbine, L., Clifford, K.N., Macduff, B.M., Sato, T.A., Maliszewski, C.R., & Fanslow, W.C. (1992). Recombinant human CD40 ligand stimulates B cell proliferation and immunoglobulin E secretion. *J. Exp. Med. 176,* 1543–1550.

Springer, T.A. (1990). Adhesion receptors of the immune system. *Nature 346,* 425–434.

Srinivasan, S., March, C.J., & Sudarsanam, S. (1993). An automated method for modeling proteins on known templates using distance geometry. *Protein Sci. 2,* 277–289.

Stenkamp, R.E., Sieker, L.C., & Jensen, L.H. (1990). The structure of rubredoxin from *Desulfovibrio desulfuricans* strain 27774 at 1.5 Å resolution. *Proteins Struct. Funct. Genet. 8,* 352–364.

Story, R.M., Bishop, D.K., Kleckner, N., & Steitz, T.A. (1993). Structural relationship of bacterial RecA proteins to recombination proteins from bacteriophage T4 and yeast. *Science 259,* 1892–1896.

Sutcliffe, M.J., Haneef, I., Carney, D., & Blundell, T.L. (1987). Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng. 1,* 377–384.

Swindells, M.B. (1992). Structural similarity between transforming growth factor-β2 and nerve growth factor. *Science 258,* 1160–1161.

Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol. 208,* 1–22.

Thornton, J.M. (1992). Lessons from analyzing protein structures. *Curr. Opin. Struct. Biol. 2,* 888–894.

Thornton, J.M., Flores, T.P., Jones, D.T., & Swindells, M.B. (1991). Prediction of progress at last. *Nature 354,* 105–106.

Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins Struct. Funct. Genet. 11,* 52–58.

Vyas, N.K., Yyas, M.N., & Quiocho, F.A. (1991). Comparison of the periplasmatic receptors for L-arabinose, D-glucose/D-galactose, and D-ribose. *J. Biol. Chem. 266,* 5226–5237.

Wang, J., Yan, Y., Garrett, T.P.J., Liu, J., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L., & Harrison, S.C. (1990). Atomic structure of human CD4 containing two immunoglobulin-like domains. *Nature 348,* 411–418.

Watenpaugh, K.D., Sieker, L.C., & Jensen, L.H. (1979). The structure of rubredoxin at 1.2 Å resolution. *J. Mol. Biol. 131,* 509–522.

Weber, I.T. (1990). Evaluation of homology modeling of HIV protease. *Proteins Struct. Funct. Genet. 7,* 172–184.

Weber, I.T., Miller, M., Jaskolski, M., Leis, J., Skalka, A.M., & Wlodawer, A. (1989). Molecular modeling of the HIV-1 protease and its substrate binding site. *Science 243,* 928–931.

Weis, W.I., Drickamer, K., & Hendrickson, W.A. (1992). Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature 360,* 127–134.

Weis, W.I., Kahn, R., Fourme, R., Drickamer, K., & Hendrickson, W.A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science 254,* 1608–1615.

Williams, A.F. (1987). A year in the life of the immunoglobulin superfamily. *Immunol. Today 8,* 298–303.

Wilmanns, M. & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. *Proc. Natl. Acad. Sci. USA 90,* 1379–1383.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J., & Kent, S. (1989). Conserved folding in retroviral proteases. Crystal structure of a synthetic HIV-1 protease. *Science 245,* 616–621.

Wodak, S.J. & Rooman, M.J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol. 3,* 247–259.

Zhang, X.-M., Weber, I., & Chen, M.-J. (1992). Site-directed mutational analysis of human tumor necrosis factor alpha receptor-binding site and structure-functional relationship. *J. Biol. Chem. 267,* 24069–24075.

Zhu, Z.-Y., Šali, A., & Blundell, T.L. (1992). A variable gap penalty function and feature weights for protein 3-D structure comparison. *Protein Eng. 5,* 43–51.