

Knowledge-Based Protein Modeling

Mark S. Johnson, Narayanaswamy Srinivasan, Ramanathan Sowdhamini, and Tom L. Blundell

Imperial Cancer Research Fund Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, London

Referee: Dr. Michael Gribskov, 10100 Hopkins Drive, San Diego Super Computer, La Jolla, CA 92093

ABSTRACT: Knowledge, both from the three-dimensional structures of homologous proteins and from the general analysis of protein structure, is of value in modeling a protein of known sequence but unknown structure. While many models are still constructed at least in part by manual methods on graphics devices, automated procedures have come into greater use. These procedures include those that assemble fragments of structure from other known structures and those that derive coordinates for the model from the satisfaction of restraints placed on atomic positions.

KEY WORDS: comparative modeling, proteins, three-dimensional structures, amino acid sequences, structure analysis, data banks, searching, alignments.

I. INTRODUCTION

Knowledge of protein three-dimensional structures is a basic prerequisite for understanding function. It may give clues, not apparent from the sequence, about distant relatives that share a catalytic mechanism or recognize the same ligand. It may thus provide a basis for further studies of substrate or ligand interactions that are essential for a full and detailed understanding of protein function.

Much progress has been made in the definition of protein three-dimensional structures by X-ray analysis and nuclear magnetic resonance. The latest issue of the bulletin of the Brookhaven Protein Data Bank (Brookhaven National Laboratories)^{1,2} records the deposition of more than 1000 sets of atomic coordinates for protein structures. However, many of these are site-directed mutants or inhibitor complexes of the same proteins, and many are related members of families of proteins — globins, serine proteinases, immunoglobulins, etc. — with similar sequences and closely related three-dimensional structures. Even quite different proteins, at least in terms of sequence, can have very similar folds; for example, the sulfate and phosphate binding pro-

teins, the transferrins, and the porphobilinogen deaminases have similar bilobal anion binding structures but no significant sequence identities.^{3,4} Protein topologies such as the $\alpha\beta$ -nucleotide binding motif, the $\alpha\beta$ -barrel (TIM barrel), the β -jelly roll, the four- α -helix bundle, and the β -immunoglobulin domain (Ig fold) have been found in a wide range of protein structures.

There have been many thoughtful discussions of the number of different folds adopted by globular proteins.^{5,6} New experimental determinations of proteins indicate that about 50% may be known, but this is likely to be an overestimate as protein crystallographers and NMR spectroscopists tend to select similar proteins that are amenable to their techniques. We estimate the number to be between 500 and 700.⁷ This estimate implies that, with the increasing number of new structures defined each year, we should move toward an experimental definition of one example of each common fold. If methods to identify the folds from their sequences can be developed and if comparative modeling can be extended to more distantly related protein topologies, then we should be able to provide at least rough indications for most sequences as they become available.^{7,8}

In this review we discuss modeling procedures that involve knowledge of proteins with a common fold. Such procedures can be envisaged as two steps.⁹ The first step is to solve the inverse folding problem: to define all those sequences that can adopt a particular tertiary fold (Figure 1). Operationally, this is more usefully posed as defining whether a new sequence belongs to any of the "known" folds. It involves projecting restraints from a three-dimensional structure onto a one-

dimensional sequence. The second step is to use the sequence, together with the knowledge that the protein belongs to a family of known fold, to construct a model (Figure 1). This form of protein modeling or prediction involves placing restraints from a known fold on the three-dimensional structure postulated for a new protein sequence. The two steps require similar knowledge of the structures of protein families that can be expressed as rules; these relate both local and global three-

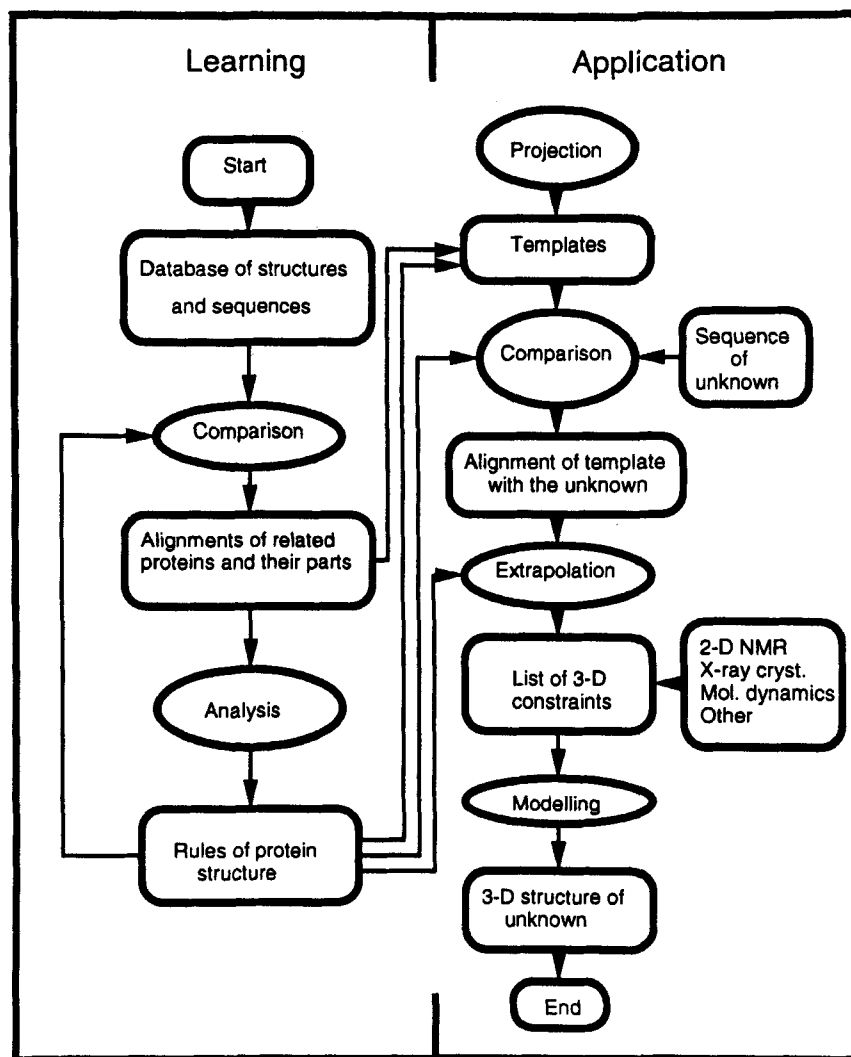


FIGURE 1. A scheme for the knowledge-based modeling of proteins. This approach involves the derivation of rules from the comparisons of sequences and three-dimensional structures and their use in the generation of a template and the construction of a three-dimensional model. Operations involved in the comparison, analysis, projection, and extrapolation with modeling are described in detail within the text. (From Šali, A. et al., *Trends Biochem. Sci.*, 15, 235, 1990. With permission.)

dimensional structure to patterns in the sequence of amino acids in the polypeptide chain. The method is comparative but exploits the broader knowledge base of nonhomologous protein structures.

A. Early Modeling Studies

Early modeling studies frequently relied on the construction of wire or plastic models and only later incorporated interactive computer graphics. Our main emphasis within this review is on rule-based construction of protein models. Nevertheless, it is instructive to make a historical analysis of some of the early modeling studies based on knowledge of homologous or other proteins with a common fold. For many of the following examples, the crystal structures are now available and it is therefore possible to access how well the models predicted the three-dimensional structure.

The first models produced from homologous proteins were constructed by taking the existing coordinates of a single known structure and then altering those side chains that were not identical in the protein to be modeled. This approach to protein modeling is still employed today with considerable success, especially when the proteins are similar. When the sequences are more dissimilar (i.e., >30% sequence identity), models constructed on this basis can have significant problems. Most obviously, the backbone of the model closely resembles that of crystal structure employed in the modeling, deletions and especially insertions are not handled well and there are often clashes between side-chains. When relationships are very distant, there is not much hope of producing a reliable model.

Some of the first models made using information of homologous proteins with known structure are listed in Table 1. Browne et al.¹⁰ published the first model using an X-ray-derived structure with a similar sequence; they modeled bovine α -lactalbumin on the three-dimensional structure of hen egg-white lysozyme. The sequences contain identical patterns of disulfide bonds and share 39% sequence identity. No insertions of "new" polypeptide were needed because

the polypeptide chain is shortened in α -lactalbumin and only deletions needed to be modeled. Warne et al.¹¹ used their procedures designed for the refinement of X-ray coordinates to produce a model for α -lactalbumin using the structure of lysozyme. This model was then compared with that generated by Browne et al.,¹⁰ revealing large differences in some portions of the two models, especially from residue 100 onward. When the structure of α -lactalbumin was solved, Acharya et al.¹² reported that the two independent models were generally correct except for the carboxyl-terminal portion of the models where they differed from each other and from the α -lactalbumin crystal structure.

Hartley¹³ and colleagues considered the similarities among sequences of the serine proteinase family and rationalized the observed features in terms of the one known structure: bovine α -chymotrypsin. They modeled both trypsin and elastase to fit the electron density map for chymotrypsin, while leaving both the main chain and identical side chains largely untouched. McLachlan and Shotton¹⁴ modeled the α -lytic proteinase of *Myxobacter* 495 based on the structures of both chymotrypsin and elastase. This was a more difficult challenge; the sequence identity between the two proteinases was only 18%. Furthermore, an alignment between the two sequences is fragmented by a large number of gaps, including five that are between 6 and 19 residues in length (Figure 2).

When Brayer et al.¹⁵ solved the structure of the α -lytic proteinase, they made comparisons¹⁶ with the model. Although they found that portions of both domains were constructed correctly, misalignment of the sequence with those of the known 3-D structures led to incorrect regions. For residues 214–220 of the binding cleft, there was a 4-residue offset that disrupted the active site itself and serine-214, an invariant residue in serine proteinases, was replaced by an asparagine (Figure 2). As a result of their comparison, Delbaere et al.¹⁶ suggested that protein structure predictions should be used to establish homology but not structural features, especially of catalytic or substrate binding sites. In retrospect, this can be seen to be a consequence not only of the difficulty in aligning the sequences (Figure 2) but also to

TABLE 1
A Survey of Models Obtained from Comparative Modeling

Protein modeled	Proteins used for modeling	Ref.
Bovine α -lactalbumin	Hen egg-white lysozyme	10
Trypsin	α -Chymotrypsin	13
Elastase	α -Chymotrypsin	13
α -Lytic protease	Elastase	14
<i>Streptomyces griseus</i>	Bovine trypsin and	312
trypsin	other pancreatic serine proteases	25
Porcine relaxin	Insulin	21
Shark relaxin	Insulin	21
		313
Relaxin	2-Zinc insulin	22
Insulin-like growth factor	Insulin	19
		314
Rat relaxin	2Zn Insulin	315
Haptoglobin (heavy chain)	Chymotrypsin	24
	Trypsin	
	Elastase	
β -Crystallin	γ I-Crystallin	30
Casirugua insulin	Porcine insulin	18
Prothrombin factor Xa	γ -Chymotrypsin	26
	<i>Streptomyces griseus</i> protease B	
Blood coagulation factor XA, 1XA and thrombin	Serine proteases	316, 317
Urokinase	Chymotrypsin	25, 170
Rat relaxin	4Zn Insulin	318
Murine EGF	Wheat germ agglutinin and snake venom neurotoxins	319
Renin	Endothiapepsin	36
	Rhizopuspepsin	37
	Penicillopepsin	
	Endothiapepsin	
Insulin-like growth factor	Insulin	320
Urokinase	Chymotrypsin	321
Tissue-type plasminogen activator	Chymotrypsin	321
Mouse, rat testis LDH isoenzyme	Mouse Apo-LDH	322
HLA membrane proximal α 2 and β 2 domains	Immunoglobulin constant domain	323
Human renin	Rhizopuspepsin	40
Human renin	Rhizopuspepsin	41
Renin	Penicillopepsin	324
Hystricomorph insulins and insulin-like growth factor	Porcine, human hagfish insulin	325
G loop antibodies	v FAB	326
Rhodopsin	Bacteriorhodopsin	285
Human renin	Pepsin, Penicillopepsin	42
	Rhizopuspepsin	43
	Endothiapepsin	
Prothoraciotropic hormone	Insulin	327

TABLE 1 (continued)
A Survey of Models Obtained from Comparative Modeling

Protein modeled	Proteins used for modeling	Ref.
(PTTH)		
HyHEL-10 lysozyme-binding antibody	McPC603 variable region of immunoglobulin	328
Modified protein C inhibitor	Modified α 1-antitrypsin	329
HIV-1 proteinase	Endothiapepsin	330
		171
Zinc-binding domain from transcription factor III	Other metalloprotein structures	331
Human renin	Pepsin, Penicillopepsin Rhizopuspepsin Endothiapepsin, Chymosin	35
HIV-binding domains of CD4 antigen	Immunoglobulins	332
HIV protease	Rous sarcoma virus	333
Calcium vector protein	Calmodulin	334
	Troponin C	
Chymopapain M	Papain	335
	Actinidin	
Chymopapain B	Papain	336
	Actinidin	
Stem bromelain	Papain	337
	Actinidin	
Class I-E subtilases like furin	<i>Subtilisin BPN'</i>	338
Cytochrome P450 ₁₇ α	Cytochrome P450 CAM	339
RNase Pch1	RNase T1	340
Photosystem II	Photosystem from purple bacteria	341
Human apolipoprotein D	Insecticyanin	342
RNA-binding domain	Acylophosphatase from horse muscle	343
Human defensin HNP-3	Rabbit neutrophil HNP-5	168
Human plasma kallikrein	Bovine α -chymotrypsin, tonin rat mast cell proteinase, porcine elastase bovine trypsin, porcine kallikrein	168
NH2-domain of intercellular adhesion molecule 1	Four IgG constant domains	344
Neutral protease from <i>B. subtilis</i>	Thermolysin	345
G-domain of chloroplast elongation factor Tu	<i>E. coli</i> EF-Tu	346
Core proteins D1 and D2 of the photosynthetic center of pea	Photosynthetic center of <i>R. viridis</i> and <i>R. sphaeroides</i>	347
Papaya proteinase Ω	Papain	349
	Actinidin	
Formaldehyde dehydrogenase	Alcohol dehydrogenase	350
Monocyte chemoattractant	Interleukin-8	351

TABLE 1 (continued)
A Survey of Models Obtained from Comparative Modeling

Protein modeled	Proteins used for modeling	Ref.
Activating protein MCAF/MCP1	Interleukin-8	351
Mammalian aspartate transcarbamylase		352
Lactococcal proteinase (LLSK11)	<i>Subtilisin BPN'</i> <i>Subtilisin Carlsberg</i> <i>Subtilisin thermotase</i>	353
Mouse mAb 425 variable region	HYHEL-5 antibody variable region	354
DNA binding domain of Myb oncoprotein	NMR structure of <i>Antennapedia homeodomain</i>	355
Amylin (human)	Glyceraldehyde phosphate dehydrogenase (145–180)	356
Amylin (Rat)	Insulin	356
α -CGRP (human)	Amylin	356
Cytochrome-P450 (human)	Cytochrome P450 P450 CAM (<i>Pseudomonas putida</i>)	357
Lignin peroxidase LIII from <i>Phlebia radiata</i>	Cytochrome c peroxidase	120, 121 358
Ferredoxin from <i>Methanococcus thermolithotrophicus</i>	Ferredoxin from <i>Peptococcus aerogenes</i>	359
Ribonuclease H domains in reverse transcriptases from retroviruses	RNAse H from <i>E. coli</i>	360
C ₁ subunit of β -crustacyanin	A2 subunit of retinol binding protein	361
Peptide binding domain of hsp70	Peptide binding domain of class I HLA	362
Anti-carbohydrate antibody (YsT9.1)	Fv regions of McPC603 J539 and human REI	363
ATP-binding domain of periplasmic permease	Adenylate kinase, p21 ras and EfTu	364
C5a receptor	Bacteriorhodopsin	284
Guanine nucleotide-binding regulatory protein-coupled receptors	Bacteriorhodopsin	291
Onconase (P30 protein)	Bovine ribonuclease A	365
Thyroxine-binding globulin (TBG)	α -Antitrypsin	366 367
Carcinoembryonic Antigen (CEA)	NMR structure of rat CD2 First domain of human CD4 and REI	368 369
<i>E. coli</i> tyrosine aminotransferase	<i>E. coli</i> aspartate aminotransferase	370
Mitochondrial inorganic pyrophosphatase from <i>S. cerevisiae</i>	Cytoplasmic enzyme from <i>S. cerevisiae</i>	371

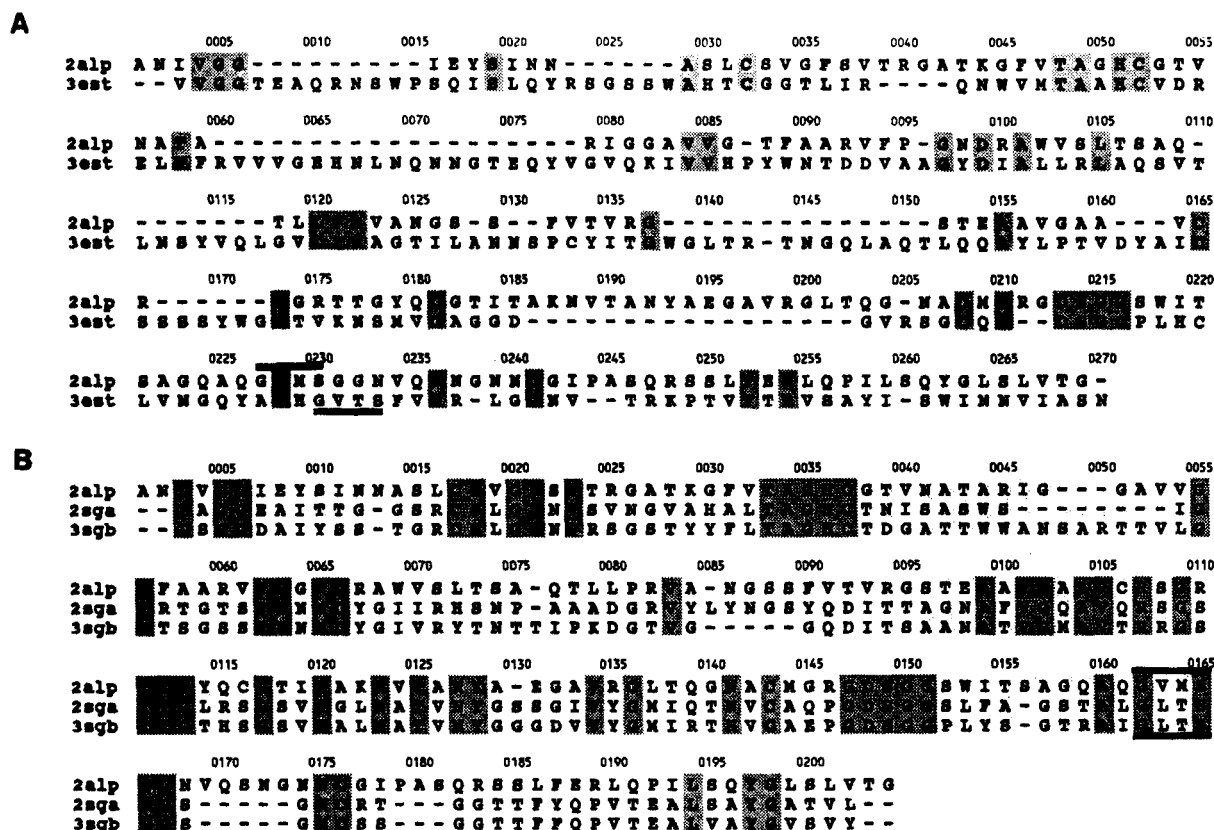
TABLE 1 (continued)
A Survey of Models Obtained from Comparative Modeling

Protein modeled	Proteins used for modeling	Ref.
Cytoplasmic inorganic pyrophosphatase from <i>S. pombe</i>	Cytoplasmic enzyme from <i>S. cerevisiae</i>	371
cGMP-binding domain of cyclic GMP-gated ion channel from photoreceptors	cAMP binding domain of catabolite gene activator protein (CAP)	372
Core proteins D1 and D2 of the photosynthetic center of pea	Photosynthetic center of <i>R. viridis</i> and <i>R. sphaeroides</i>	289
Human β 2-adernoreceptor	Bacteriorhodopsin	290
Cathepsin D	Several aspartic proteinases	303
Mast cell chymases	Serine proteases	373
Carbohydrate recogn. domain of human E-selectin	Rat mannose-binding protein	374
EGF and NGF binding protein	Porcine kallikrein, rat tonin	375
Hyperthermophilic rubredoxin	Three other rubredoxins	376

the difficulty in modeling the variable, mainly loop regions (Figure 3).

The structure of the first polypeptide hormone, porcine pancreatic insulin,¹⁷ led to modeling of other insulins, insulin-like growth factors, relaxins, and the prothoracicotropic hormone of *Bombyx mori*. The insulins from casiragua and coypu¹⁸ and the insulin-like growth factors¹⁹ were straightforward modeling targets as the disulfide pattern, structurally important glycines, and the hydrophobic core were identical. A more recent study of the structure of human insulin-like growth factor confirmed the general tertiary structure but found that the connecting peptide was disordered in solution.²⁰ Bedarkar et al.²¹ and Isaacs et al.²² modeled porcine relaxin, which is more distantly related but has a two-chain structure similar to insulin, with a conserved pattern of three disulfide bridges and two invariant glycines. Although the sequence identity did not extend to the core, this was found to be conserved as hydrophobic, giving support to the proposal that relaxin adopted an insulin-like fold. The insulin-like fold has been confirmed by X-ray analysis of human relaxin crystals.²³

Greer²⁴ was the first to realize the power of modeling variable regions by abstracting appropriate conformations from a family of homologous proteins of known structure. He used the family of serine proteinases to illustrate this approach. On alignment of the structures of trypsin, elastase, and chymotrypsin, many C α -carbons were found within 1.0 Å (1.0 Å = 10⁻¹⁰ cm) of each other; all of the remaining positions corresponded to solvent-exposed loop regions where all of the insertions/deletions were located. When the hemoglobin heavy chain, the sequence to be modeled, was aligned with these structures, the insertions and deletions relative to the known structures also corresponded to the loops. Greer's strategy was to build the main chain of both the spatially conserved and variable loop regions from fragments of each of the three known structures. The construction of the loop regions, however, was not as straightforward as that for the structurally conserved regions. Although deletions of one or two residues were easily accommodated and similar-length loops were extrapolated from one of the homologous structures, one long loop was only partially modeled. Side chains were modeled according to the conformation found at the



A and B

FIGURE 2. Alignments among the bacterial and mammalian serine proteinases. In (A) the α -lytic proteinase (Brookhaven code: 1.2 2alp) and elastase (3est) are aligned according to McLachlan and Shotton.¹⁴ The alignments¹⁵⁴ of the bacterial and mammalian serine proteinases obtained by comparing¹⁵³ their 3-D structures are shown in B and C. The region about the glycine and serine shown to be conserved in both the (B) bacterial and (C) mammalian serine proteinases and critical to their catalytic activity (thick under- and overlines) is misaligned in A; (2sga, proteinase A, *S. griseus*; 3sgb, proteinase B, *S. griseus*; 1thr, thrombin, *Bos taurus*; 1ton, tonin, *Rattus rattus*; 2pka, kallikrein A, *Sus scrofa*; 1trm, trypsin, *Rattus rattus*; tptn, trypsin, *Bos taurus*; 2gch, γ -chymotrypsin, *Bos taurus*; 1hne, neutrophil elastase, *Homo sapiens*; 3rp2, mast-cell proteinase II, *Rattus rattus*; 1sgt, trypsin, *S. griseus*). Numbering is by position in the alignment.

same positions for those identical side chains. Greer^{25,26} subsequently applied his approach to the modeling of a number of different serine proteinases. Read et al.²⁷ compared models constructed for *Streptomyces griseus* trypsin with the crystal structure. They concluded that the regions involved in substrate binding were most poorly determined by the protein modeling procedures.

The γ II-crystallin of the vertebrate eye lens is comprised of four Greek key motifs organized as two globular domains. Each domain consists of

two motifs related by a local twofold axis.^{28,29} Only two residues are invariant in the large family of γ - and β -crystallins: a glycine that is required to allow a β -hairpin to fold onto an antiparallel β -sheet and a serine that is hydrogen bonded to the folded hairpin. The invariant residues, together with a pattern of residues conserved as hydrophobic, permitted this pattern of four Greek key motifs to be recognized in all γ - and β -crystallins^{29,30} and in an unrelated protein from the coat of the spores of the bacterium *Myxococcus xanthus*.³¹ For the γ -crystallins, the models have

C

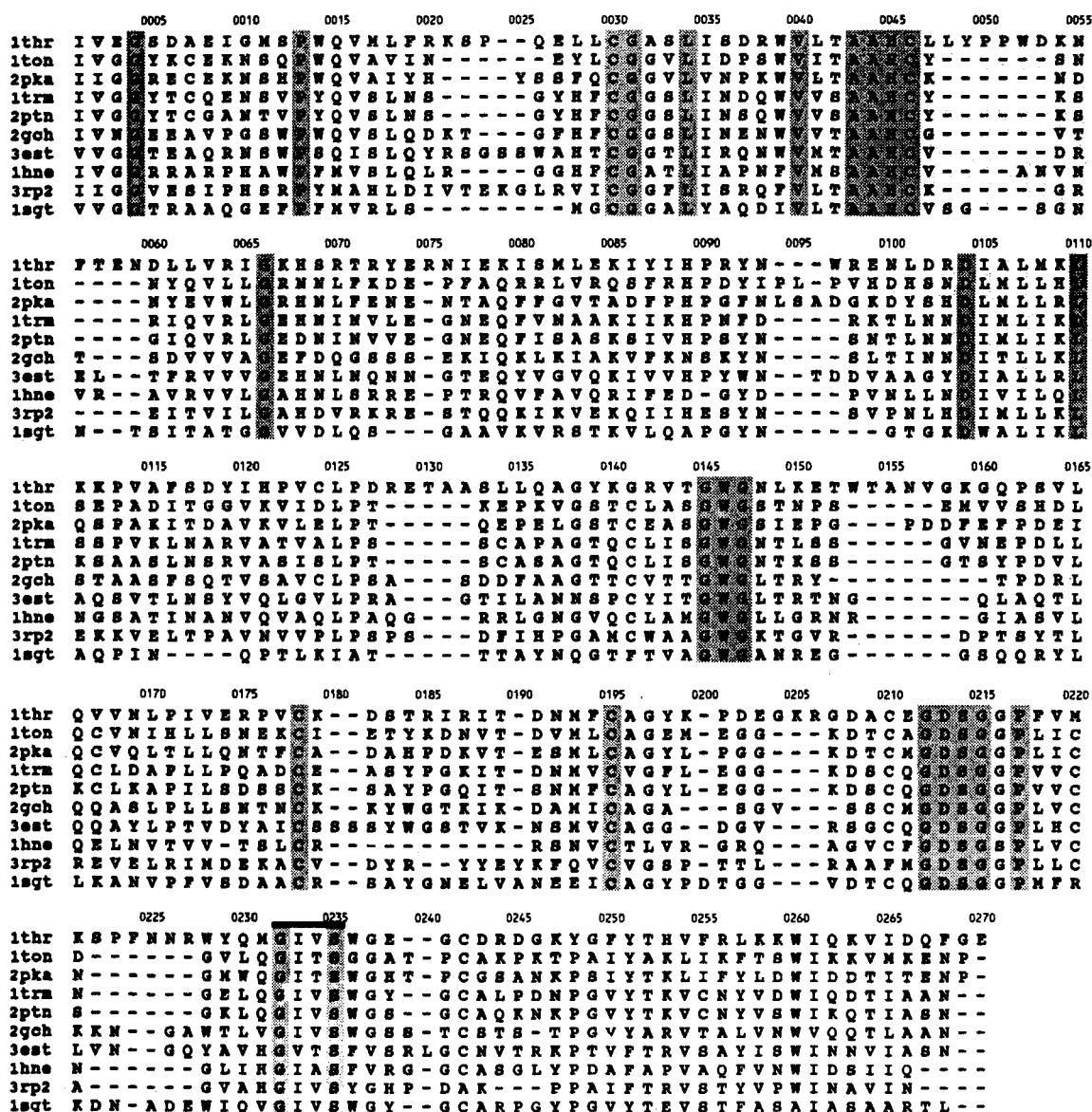


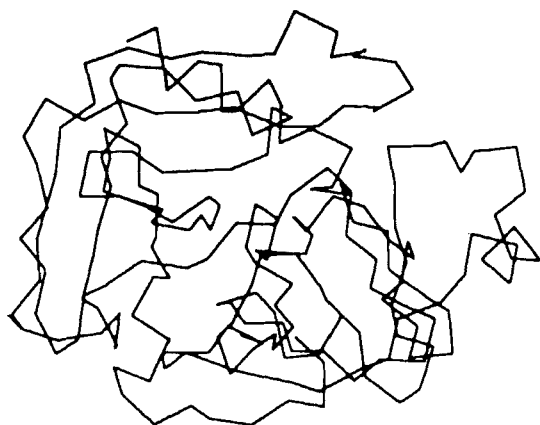
FIGURE 2C

proven to be generally correct.^{32,33} However, in the dimeric β Bp-crystallin, the domains of the same subunit are separated by a linker polypeptide and domains from two *different* subunits are associated in the same way as the single chain of the monomeric γ -crystallins (Figure 4).^{30,34}

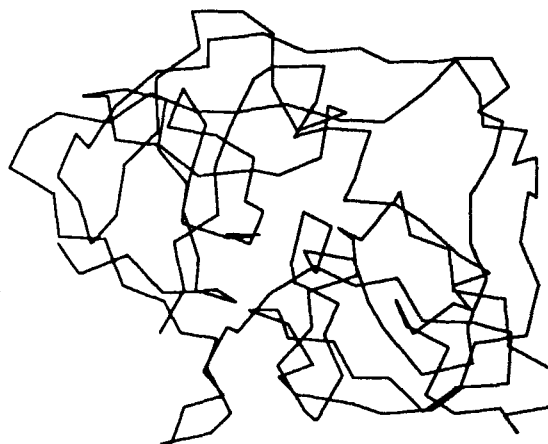
Because of its involvement in the release of the hormone angiotensin from angiotensinogen, and the role of angiotensin in blood pressure regulation, renin and renin-inhibitor complexes were obvious targets for modeling (for a review, see

Reference 35). Models for renin were first constructed on the basis of the murine sequence³⁶ and later on the human sequence³⁷ when this became available. These were constructed using the three-dimensional structure of the distantly related fungal proteinase, endothiapepsin, as no refined, high-resolution structure of a mammalian enzyme was available. Carlson et al.^{38–40} and Akahane et al.⁴¹ built models of human renin using the structures of other fungal aspartate proteinases such as rhizopuspepsin and penicillopepsin. Plattner et al.⁴²

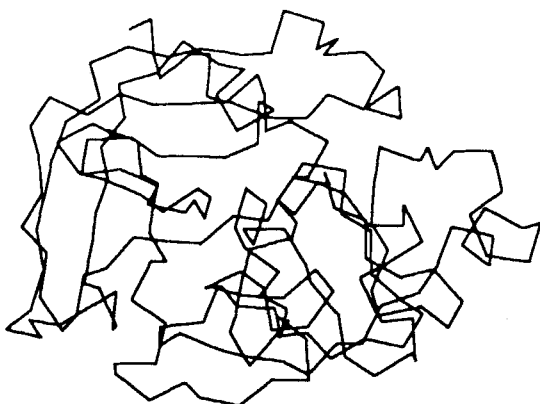
(a) Chymotrypsin (4cha)



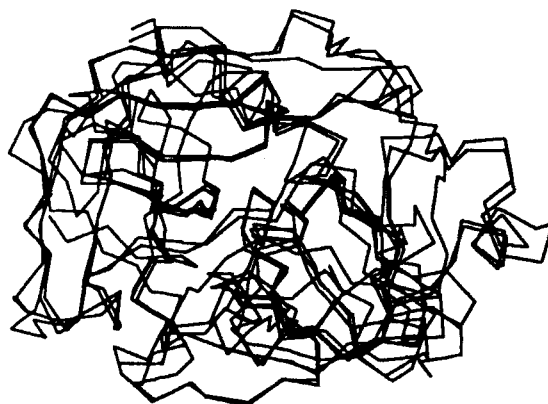
(c) α -lytic proteinase (2alp)



(b) Elastase (3est)



(d) Fitted structures (4cha, 3est, 2alp)



(e) Topologically equivalent regions

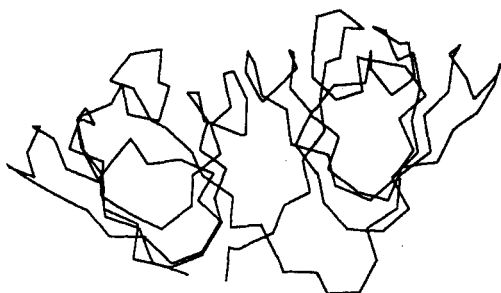


FIGURE 3. Similarity and reduction in the conserved core among the bacterial and mammalian serine proteinases: (a) chymotrypsin (Brookhaven code: 4cha, C^α -trace); (b) elastase (3est); (c) α -lytic proteinase (2alp); (d) the three structures were superposed using MNYFIT;¹⁴² (e) the conserved framework — C^α s within 2.5 Å of each other.

and Sham et al.⁴³ built their models on the basis of rhizopuspepsin, penicillopepsin, and endothia-pepsin, as well as the partially refined structures

of pepsin. Later structures were constructed using the structure of pepsin and chymosin, as they became available in the refined forms.^{35,44} The

(a) γ -crystallin (1gcr)



(b) β B-crystallin (1bb2)



FIGURE 4. Crystal structures of eye-lens crystallins: (a) γ l-crystallin (1gcr)^{28,29} and (b) β Bp-crystallin (1bb2).³⁴ Models of the β -crystallins constructed on the basis of the four Greek key motifs of the γ -crystallins had the same monomer structure. The actual β -crystallin structure (b) has contributions of two Greek key motifs, each from two separate but identical chains.

models suffered from the shortcomings arising from the differences in the “framework” — the arrangement of helices and strands — between the mammalian and fungal aspartic proteinases, as well as the rather different variable regions that are found in renins. Nevertheless, the catalytic and active site cleft in general was well modeled and the models have been widely and usefully exploited in the design of antihypertensives (Figure 5).³⁵

Since the mid-1980s, a large number of models of other proteins has been constructed and reported in the literature. Many of these are listed in Table 1.

II. PRELIMINARY CONCERNS

A. The Data

Deposition of the coordinates of protein structures, derived from X-ray diffraction, NMR spec-

troscopy, and neutron diffraction with the Protein Data Bank^{1,2} is increasingly a requirement for publication by journals. The Protein Data Bank (Table 2) can either be obtained by contacting Brookhaven National Laboratories (Protein Data Bank, Chemistry Department, Bldg. 555, Upton, N.Y. 11973) or directly using computer networks: using the anonymous FTP (file transfer protocol) server allows access not only to the current entries in the database but also to those that are in preparation for future release.

Protein modelers, like anyone using the protein data bank, need to be aware of the limitations of the data. Most errors in positions will occur for those atoms that lie at the protein surface. Surface side chains and loops are most disordered as their interactions are largely with the solvent and there are relatively few constraining contacts involved in packing within the crystal (where there are they may be distorted, for example, compare the X-ray structure of temdamistat⁴⁵ with its NMR

* The reliability factor or R-factor represents the differences between the observed F_o and calculated F_c structure factor amplitudes: $R = \sum_{hkl} K |F_o| - |F_c| / \sum_{hkl} K |F_o|$, where K is a scaling factor.

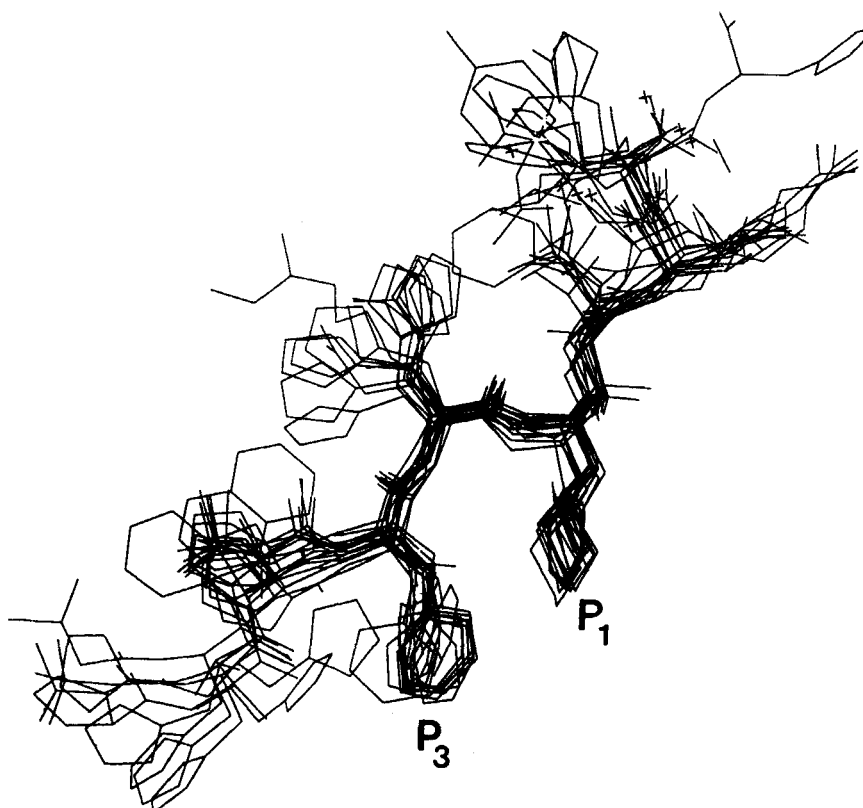


FIGURE 5. Superposition of 20 inhibitors complexed with the aspartic proteinase (endothiapepsin). P1 and P3 refer to the specificity subsites. (From Dhanaraj, V. et al., *Innovations on Proteases and Their Inhibitors*, Avilés, F. X., Ed., Walter de Gruyter, Berlin, 1993. With permission.)

structure⁴⁶). On the other hand, buried residues are constrained by the surrounding main chain and side chains, and their average positions are well defined and conserved. Even so, side chains and even the main chain “backbone” occupy no single fixed position, but an ensemble of possibilities generally exists.

Two numerical values give some indication of the reliability of a particular structure: the resolution expressed in Å and the R-factor. The resolution describes the minimum interatomic spacing for which X-ray data contribute toward the structure analysis. At a resolution of less than 5 Å for an all α -helical protein and less than 3.5 to 4.0 Å for a β -protein it is impossible to identify the elements of secondary structure and only at a resolution of 3.0 Å can the protein backbone be traced with any confidence. With very high resolution, of the order of atomic

bond distances (1.0 to 1.5 Å), individual atoms begin to appear. The R-factor,* which describes the difference between the observed and calculated diffraction amplitudes, gives some indication of the error in the reconstructed image, although its absolute value depends on the method of refinement and the ratio of observations to parameters refined. A “better” structure is normally taken as that with a higher resolution (>2.0 Å) and a lower R-factor ($<20\%$). Islam et al.⁴⁷ have considered to what extent van der Waals radii occupy the same regions of space and how these bad contacts relate to resolution and R-factor. They found that some high-resolution structures, refined to a low R-factor, still had fairly large errors.

But these are gross descriptors of the dataset used in reconstructing the electron density and do not pinpoint where the data, and subsequently the

TABLE 2
Structure, Sequence, and Alignment Databases

Database	Contact details
Structure databases	
PBD (Protein Data Bank)	Chemistry Department Building 555, Brookhaven National Laboratory Upton, New York, NY 11973, USA Anonymous ftp at pdb.pdb.bnl.gov
CSD (Cambridge Crystal Structure Data Centre) (Small molecule structures)	University Chemical Laboratories Lensfield Road Cambridge CB2 1EW
Sequence databases	
PIR (Protein Identification Resource)	National Biomedical Research Foundation Georgetown University Medical Center 3900, Reservoir Road NW Washington D.C., 20007 USA e-mail- pirmail@gunbrf.bitnet
GENBANK (Genetic sequence data bank)	T-10, MS K710 Los Alamos National Laboratory, Los Alamos, NM 87545, USA e-mail- genbank@genbank.bio.net
EMBL database (includes SWISSPROT and PROSITE)	European Molecular Biology Laboratory Meyerhof Straße 1, D-6900, Heidelberg, Germany Network server by e-mail- net-serv@embl.heidelberg.de
SBASE (domain library)	International centre for Genetic Engineering and Biotechnology Area Science Park 34012 Trieste Italy Anonymous ftp file server- ftp.icgeb.trieste.it
Derived alignment databases	
NRL-3D (Database of 3-D structures and related sequences)	US Naval Research Laboratory Washington, DC 20375 USA
HSSP (Homology-derived secondary structure of proteins)	EMBL Meyerhof Straße 1, D-6900 Heidelberg Germany Network server by e-mail- net-serv@embl.heidelberg.de
3D-ALI (Structural superpositions and multiple sequence alignments)	EMBL Meyerhof Straße 1, D-6900 Heidelberg Germany e-mail- argos@embl-heidelberg.de
FSSP (Families of structurally similar proteins)	EMBL Meyerhof Straße 1, D-6900 Heidelberg Germany Network server by e-mail- net-serv@embl.heidelberg.de

TABLE 2 (continued)
Structure, Sequence, and Alignment Databases

Database	Contact details
Homologous proteins aligned on the basis of 3-D structural features	J. P. Overington Pfizer Central Research Sandwich Kent CT13 9NJ U.K. e-mail-Overingtonj@pfizer.com

structure, are uncertain locally. The X-ray structural data may reflect this in high values of the thermal parameters (B-values), which indicate the extent of both static and thermal disorder. Phillips et al.,⁴⁸ Schreuder et al.,⁴⁹ Stout,⁵⁰ and Adman⁵¹ discuss the features of correct and incorrectly modeled X-ray-derived structures for turkey egg-white lysozyme, RuBisCO, and ferredoxins.

Fitting a polypeptide chain into an electron density map is greatly assisted by the knowledge of the protein's sequence. In our experience, if no sequence is available, then the success in correctly identifying amino acid types from the density is approximately 60%. It is impossible to

differentiate between atoms of similar atomic numbers, notably carbon, nitrogen, and oxygen. This can pose difficulties in distinguishing between some amino acid types: Asp and Asn, Glu and Gln, and even Leu and Asn, although the environmental context (buried or exposed to solvent; hydrogen bonding potential) can sometimes help one to make an educated guess. Lysine can also be confused for a shorter side chain because it is often highly mobile and therefore has "invisible" terminal atoms. Some portions of polypeptide chain, particularly solvent-exposed loops, will be so disordered that neither the side chains nor the main chain can be built. Adman⁵¹ discusses a

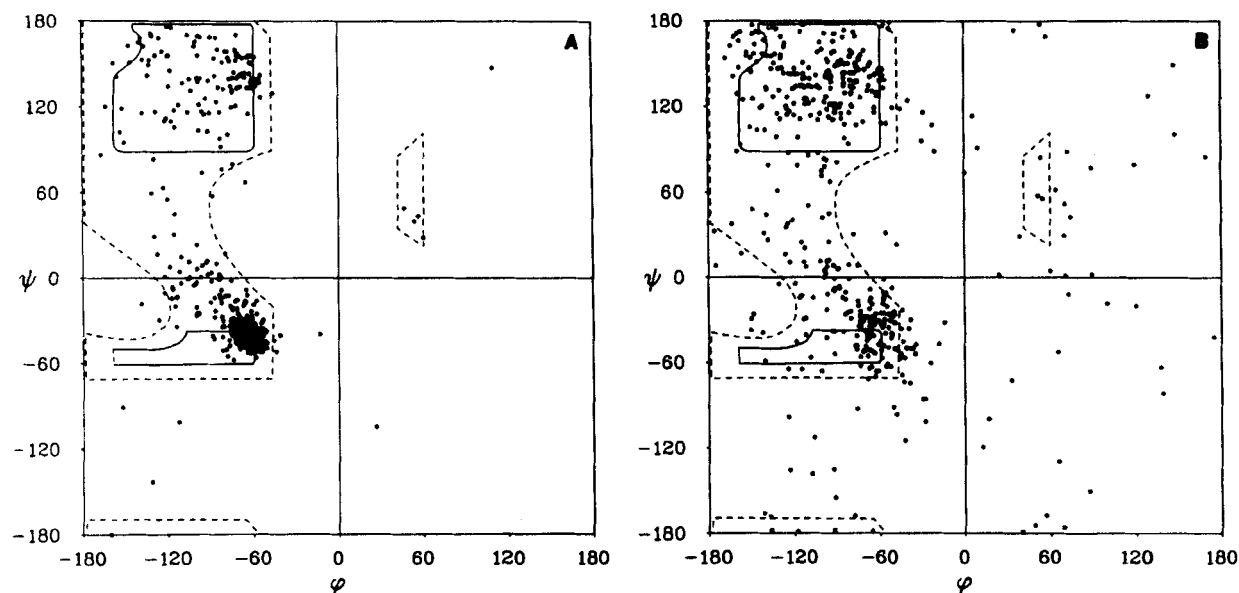


FIGURE 6. Main-chain torsion angles (ϕ, ψ) of non-glycyl, non-prolyl residues in known protein structures determined at various crystallographic resolution: (a) 1.6 Å or better resolution (Brookhaven codes: 5cpv, 1ccr, 351c, 1ecd, and 1cse); (b) 2.3–2.8 Å resolution (1tpk, 3fx, 1hdd, 155c, and 1mib). Note that the lower resolution structures have more points in the disallowed regions of the Ramachandran map.

number of these uncertainties in the context of the structure solution and subsequent refinements of *P. aerogenes* ferredoxin.

Thornton and co-workers^{52,53} have carried out an analysis of highly refined, high-resolution structures in order to assess the expected precision of X-ray structures; it has been shown that the coordinate data deposited can have many errors.^{47,52-53} Bond lengths and bond angles may have values distant from those observed in other proteins through careless refinement of the structure. Main-chain ϕ , ψ , and ω torsion angles may lie outside the acceptable bounds described by the Ramachandran map constructed for the database as a whole. In Figure 6, we show the main chain ϕ and ψ torsion angles for proteins with resolutions between 2.3 and 2.8 Å and those with resolutions higher than 1.6 Å. Side-chain χ_1 and χ_2 torsion angles may not correspond to peaks of the distribution of angles observed for a particular side chain. The analysis of the entire data bank⁵⁴ has revealed other areas of concern, including missing atoms, mislabeled atoms, incorrect

chirality, and unusual disulfide χ_3 torsion angles.

The selection of structures in the data bank is biased toward those proteins that have been easily obtainable, available in large amounts, stable, and, of course, those that can be crystallized. Recent advances in molecular biology have led to the expression and solution of 3-D structures of less-abundant proteins from low-copy messenger RNA. For the most part, these are globular proteins; there are few examples of proteins that are integral to the lipid bilayer as these proteins are particularly difficult to crystallize. However, it is encouraging to the protein modeler that about one quarter of the sequences determined from genome studies can be found to have at least 25% sequence identity with protein folds that have already been deposited in the data bank.^{55,56} Thus, many of the 50,000 protein sequences available are probably amenable to comparative modeling (Figure 7).

Sequence data (Table 2) are available either as nucleic acid sequences (e.g., EMBL and GenBank) or amino acid sequences (e.g., PIR,

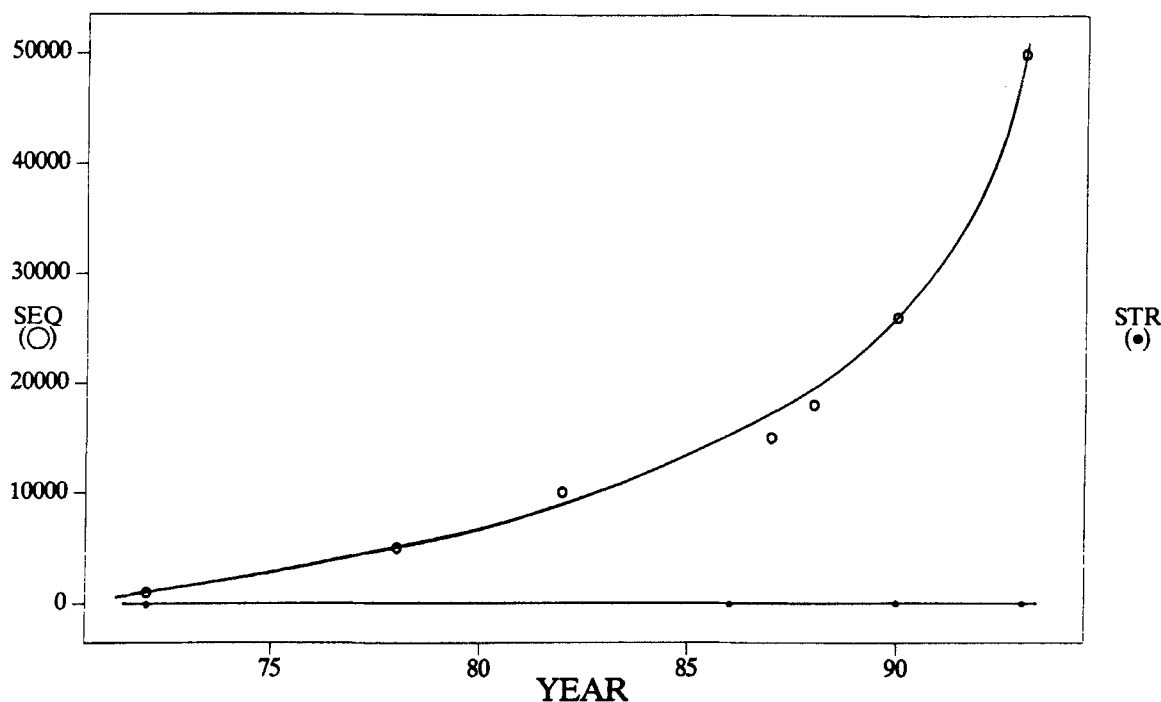


FIGURE 7. The known protein sequences are now over 50,000 in number, while the number of crystal structure is over 1000. Unique folds account for less than 200 three-dimensional structures. Even so, a large proportion of the known sequences are available to comparative modeling approaches.

SWISS-PROT, NEWAT, OWL).⁵⁷⁻⁶⁰ Because most official data banks now incorporate electronic data entry, there is little backlog between the appearance of sequences in the literature and their availability in data collections (which are many times accessible through anonymous FTP servers over electronic networks).

B. Locating and Aligning Homologues

We wish to extrapolate knowledge about one or more related protein structures to a homologous or analogous sequence of unknown structure. When a new fold is described or a new sequence appears in the literature, it is necessary to identify related sequences and three-dimensional structures using computerized searches. In this section we consider alignment and data bank search techniques for both sequences and three-dimensional structures.

1. Protein Data Bank Searches

A variety of sequence searching tools have been available for some time.^{57-59,61-65} Often, the most efficient approach is a simple search of the sequence data bank entries by title, key words, and, for enzymes, the EC (Enzyme Commission) numbers. A search for patterns of residues or motifs that distinguish a family (residues critical to binding, catalysis or a structurally important feature) can be used to identify homologues; the PROSITE data bank⁶⁶ contains many protein patterns that are characteristic of particular families. More generalized data bank searching, where full-length sequences are compared, involves methods that either compare overlapping segments or rely on dynamic programming algorithms.

The sequence comparison approach of Needleman and Wunsch⁶⁷ and derivatives thereof⁶⁸⁻⁷² can be used repetitively to compare a sequence against an entire data bank of sequences. A score is attributed to each of the tens-of-thousands of alignments and the scores are ranked. There is difficulty in recognizing distantly related matches with low similarity from the background noise of the unrelated comparisons. Only homologues with

percentage sequence identities greater than about 20 to 25% are certain to be among the top matches.

With dynamic programming approaches, the global optimal alignment is sought. This, however, may not be so useful when comparing sequences that vary greatly in size or when locating independent folding domains, which are often embedded within a much larger sequence (e.g., epidermal growth factor, fibronectin types I and II, proproteinase kringle, complement C9, etc.). Dynamic programming approaches also fail to recognize the presence of duplicated regions within the compared sequences. An alternative dynamic programming approach, due to Smith and Waterman,⁷² aligns the best single matching region that may or may not encompass a majority of the residues from the sequences and thus allows smaller domains to be located. This algorithm has been incorporated into many of the available search procedures.

The comparison of segments can circumvent some of these disadvantages.^{57,59,73} The differences in length between compared sequences is no longer a problem; internal duplications are also easily found and embedded domains can be located.^{57,74} The comparison of fixed-length segments of sequences is much less efficient computationally than dynamic programming methods. While gaps are not a problem, one is left with a series of scores for segments that must be interpreted, but an alignment is generally not produced. No one search procedure is perfect and most will miss related proteins when the sequence similarity is very low. One means to increase the sensitivity of a search is to consider additional information. This might include knowledge of the sequence variability found within a family of proteins on the basis of the multiple alignment of their sequences, knowledge of residues essential to the protein and conserved throughout the family, and structural information if this is available for a sequence or related protein (see later).

Searching with a single sequence is unlikely to be the best approach when the sequence similarity is low. A number of alternative strategies have been developed to improve the chances of finding homologues. Templates, consensus sequences, profile analysis, or family composite scores derived from multiple alignments can all

be used to enhance the efficiency of searching procedures.⁷⁵⁻⁷⁹ For instance, a data bank search with the sequence of the erythrocrucorin from *Chironomus thummi thummi* was able to locate 239 out of the 624 globins in a sequence data bank prior to the first non-globin. When the search was made with the alignment obtained from 15 globin sequences, 602 of 624 were found, and the remaining ones included the 15 half-size globin fragments within the data bank (Figure 8).⁸⁰

The method of scoring comparisons can also be very important to both searches and alignments (see later). Searching for regions with high levels of matched identical residues works well for very similar proteins but rapidly loses utility as the sequence identity dips below about 35%. Other methods of scoring consider matches for conservative changes and nonconservative changes alike.⁸¹⁻⁸⁵ Some have considered the substitutions that actually occur in homologous families of proteins, the most well-known being the matrix of Dayhoff and co-workers.⁸⁶ Recently, a number of groups have compared data banks of sequences⁸⁷⁻⁸⁹ and structures^{79,90} in order to try to improve on the Dayhoff approach. Indeed, comparisons of the methods^{79,88} indicate that the 1978 version of the Dayhoff matrix performs considerably less well than more recent adaptations (see Reference 79 for a comparison of 14 scoring procedures). Table 3 presents a log-odds matrix suitable for scoring sequences and based on the analysis of 65 families of aligned 3-D structures.⁷⁹

2. Sequence Alignments

Sequence-based alignments are useful in determining invariant features (key residues) and those regions where variability is high (positions of gaps and loops) in a protein family. If the sequences are not very similar to each other, then family alignments can help to reduce gross errors in alignment: residues essential to a protein's function are rigorously conserved throughout its members and will be conserved spatially within the three-dimensional structures. With sufficient sequences and their alignment, the error in the model of the α -lytic proteinase about the invariant Ser-214 would have been detected (see above and Figure 2).

We have already discussed a number of the alignment techniques in the context of data bank comparisons. Again these procedures involve segment-based comparisons or Needleman and Wunsch-type dynamic programming techniques, although the focus has shifted to the multiple alignment of sequences that leads to more accurate alignments.^{79,91-93} Techniques have included methods that seek the global optimum by consideration of higher dimensional forms of the Needleman and Wunsch approach,^{71,91,92,94,95} segment-based approaches,⁹² and iterative binary comparisons,^{96,97} some of which have exploited preliminary phylogenetic relationships to enhance tree construction.^{80,96,98,99} Alternatively, alignments can be made with those procedures that employ consensus templates, profiles, and residue patterns.^{75-78,100,101}

In comparison with the results of multiple alignments, pairwise comparisons have considerably more error in the alignments when viewed in light of aligned 3-D structures (Figure 9).^{79,102} Obviously, regions with many matched identical residues will be more certain than those where gaps are seen to occur or where the sequence similarity is low. A number of groups have detailed approaches that can pinpoint strongly aligned regions and areas where alternative local alignments are nearly as probable as the optimal one.¹⁰³⁻¹⁰⁶ Indeed, the approach of Saqi and Sternberg¹⁰⁵ has been combined with side-chain volume requirements in order to assist in modeling and determination of alignments compatible with the known structure.¹⁰⁷

3. Recognizing Common Folds from Sequences

The association of a new sequence with a characterized protein is a major difficulty in using the common fold as a framework for protein modeling. This has been addressed successfully on many occasions by using structural information to identify key features in protein architecture and associating these with invariant or conserved sequences.

One of the first attempts to use structural information systematically was by Taylor,⁷⁵ who

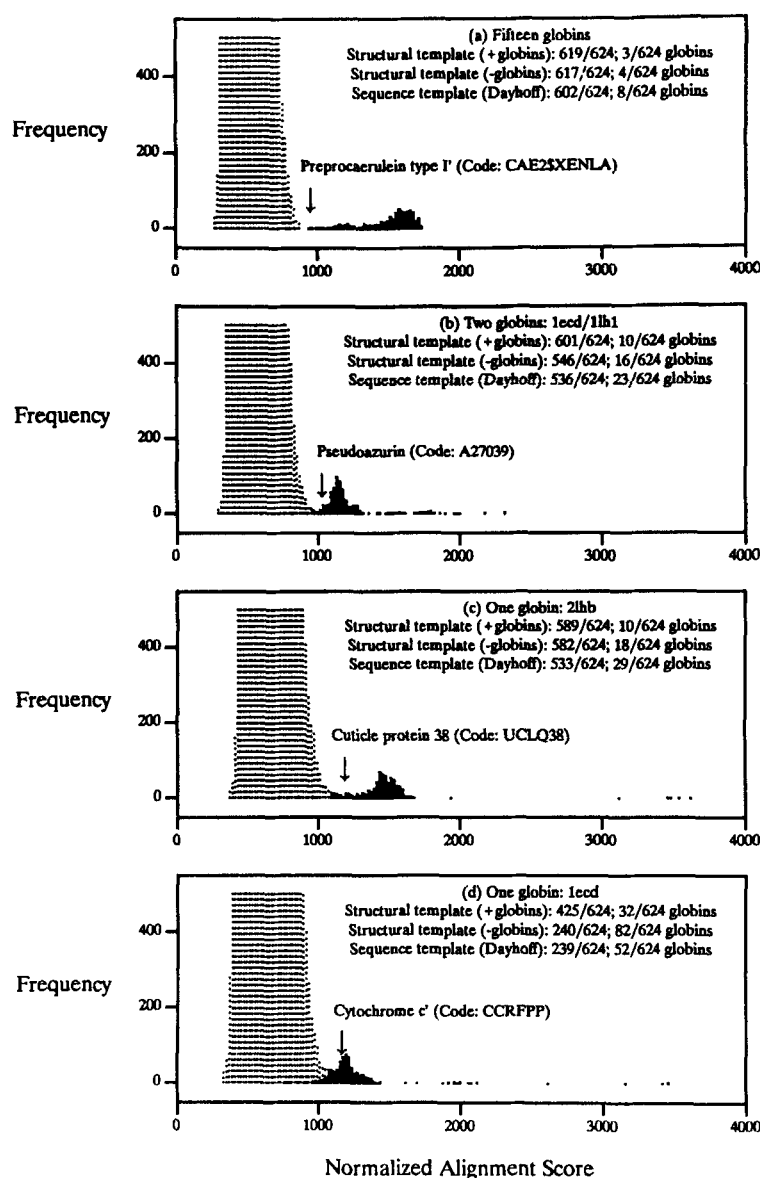


FIGURE 8. Search of globin structural-templates and sequence-templates against an amino acid sequence data bank of 21,049 entries greater than 55 residues. The frequency of the alignment scores, normalized by the length of the alignment, are plotted as dotted lines; the frequency of normalized scores for comparisons involving the 624 globins in the database are superposed (solid lines). Plots are provided for the structural templates only, but the number of globins located prior to the first non-globin are shown for both the structural, with and without globin contributions toward the substitution frequency data bank, and sequence-based searches of the same proteins. The second set of values reflect the number of globins not located within the first 624 highest ranked scores. (a)–(d) were obtained using structural templates (residue solvent accessibility, secondary structure considerations, and hydrogen bond capability); a constant gap penalty of 35 was used. Sequence templates were computed using the Dayhoff 250 PAM amino acid scoring matrix⁸⁶ scaled between 0 and 100; a penalty of 30 was assessed for gaps. Alignments were made using a dynamic programming approach that is part of the program PSLAVE.⁸⁰ In (a) structural and sequence templates were searched based on the structural and sequence alignment of fifteen globins: (Brookhaven codes, 2hbb, the α -chain of human hemoglobin; 2hbb, the β -chain of human hemoglobin; 2mhb, the α -chain of equine hemoglobin; 2mhb, the β -chain of equine hemoglobin; 1mba, the myoglobin of *Aplysia liminacia*; 1mbs, myoglobin of *Phoca vitulina*; 2mm1, myoglobin of *Homo sapiens*; 4mbn, sperm whale myoglobin; 21hb, the hemoglobin of the sea lamprey; 1pbx, hemoglobin of *Pagothenia bernacchii*; 1pmb, porcine myoglobin; 1ecd, erythrocrurin of *Chironomus thummi thummi*; 1lh1, leghemoglobin of *Lupinus luteum*; 1sdh, hemoglobin of *Scapharca inaequivalvis*; 1lth, hemoglobin of *Urechis caupo*. (b): an alignment of 1ecd, erythrocrurin of *Chironomus thummi thummi* and 1lh1, leghemoglobin of *L. luteum*. (c): 2lhb, the hemoglobin of the sea lamprey. (d): 1ecd, erythrocrurin of *Chironomus thummi thummi*. (From Johnson, M. S. et al., *J. Mol. Biol.*, 231, 735, 1993. With permission.)

developed a method of generating templates for each part of the framework of a protein generated from superposition on the basis of known three-dimensional structures of proteins in a family. Ponder and Richards¹⁰⁸ used a library of side-chain rotamers and sought to find combinations of sequence and side-chain conformation that would allow retention of a known three-dimensional structure. This provides a powerful approach to identifying closely related sequences but it is computationally very expensive. It is limited by the fact that distantly related proteins evolve through relative translations and reorientations of the elements of secondary structures in order to allow changes in side-chain shapes and volumes.

Sippl¹⁰⁹ and Jones et al.¹¹⁰ have sought to overcome this problem by using knowledge-based potentials.¹⁰⁹ They thread (match sequence positions with those in the structure; effectively an alignment or set of alignments, although gaps are not considered in some applications and the sequence may be incremented along that of the structure one residue at a time or as overlapping segments) a sequence through a known structure and ask for each alignment: can the sequence of interest adopt that three-dimensional fold? Similar potentials have been derived by Maiorov and Crippen¹¹¹ and Godzik et al.¹¹² that recognize the correct folding of globular proteins. The latter approach evaluates the match in terms of rough 3-D models.

Sippl¹⁰⁹ has derived pseudopotentials for interresidue interactions that are then applied to the detection of protein folds.^{113,114} The interresidue distances ($C^\alpha-C^\alpha$ or $C^\beta-C^\beta$) for 400 possible pairs of amino acids in a large number of known structures were compiled and classified as short, medium, and long-range interactions depending on the number of intervening residues along the sequence. The observed frequencies of occurrence were then converted to potentials of mean force by using the well-known Boltzmann device. The problem of sparse data was minimized with a smoothing function.¹⁰⁹

Sippl and Weitckus¹¹⁴ use these potentials to detect native folds of amino acid sequences with unknown 3-D structure. The sequence in question is modeled by aligning it with all the known folds in the protein data bank. For every model, the

pseudo-energy (sum of individual residue pair energies) is computed and the one with lowest energy is suggested to be the native fold. This method has been tested by modeling sequences of the α and β chains of human hemoglobin and the sea lamprey hemoglobin, proteins whose crystal structures are all known. The lowest energies were obtained when sequences from these three proteins were matched with their corresponding crystal structures, followed by matches to other members of the globin family. This method was then applied to globin sequences of unknown structure where the sequence similarity varied between 17 and 49% with the known 3-D structures. With one exception in ten, globin sequences were associated with a globin fold as the lowest energy match. The RMS error in the models built using this approach can be as large as 5.5 Å. This is largely a consequence of the threading procedure where gaps are not allowed in the alignment. This can cause severe problems in terms of modeling, while using this procedure to detect folds would be much less affected.

Maiorov and Crippen¹¹¹ devised a similar potential from interresidue interactions observed in known structures (in addition to distances between the side chains, they also considered main-chain to main-chain and main-chain to side-chain interactions), as well as for a large number of native-like alternate "folds", which they generated by threading smaller sequences onto native structures. They devised a function that will differentiate between the true X-ray crystal structures and the protein-like "folds" derived from comparing unrelated proteins. Their function is adjusted in such a way that the crystal structures will have a lower potential energy. In their test with a large number of native structures and homologues that were not used in the construction of the contact potential, the compact structures are identified to have the native fold. A major emphasis of this work is to use knowledge about incorrect "folds" in addition to knowledge of the native fold in formulating the potential energy function.

Skolnick and his colleagues¹¹² propose an approach for recognizing sequences compatible with a 3-D fold and identifying the fold of a sequence. Potentials are again extrapolated from residue-residue distances from 3-D struc-

TABLE 3
Structure-Based Amino Acid Scoring Table (Upper Triangle); All Positive Table (Lower Triangle); All Values Have Been Multiplied by 10⁷⁹

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6.0	-3.4	-1.6	-0.7	-3.2	-0.5	-3.1	-2.2	-0.9	-3.3	-1.5	-1.4	-1.0	-0.6	-1.6	0.0	-0.8	-0.5	-5.8	-4.0
C	15.8	16.1	-9.7	-6.9	-4.4	-8.2	-8.2	-7.7	-8.7	-8.7	-4.4	-7.6	-8.9	-6.9	-5.6	-7.7	-6.0	-4.8	-9.1	-7.7
D	6.4	25.9	8.5	2.4	-7.0	-2.1	-2.3	-4.8	-1.5	-8.0	-5.9	2.6	-1.0	-1.1	-3.4	-0.2	-1.8	-5.2	-6.0	-3.8
E	8.2	0.1	18.3	8.6	-6.4	-2.5	-1.7	0.5	-5.6	1.8	-0.6	-3.8	-5.0	-6.4	-6.0	-4.8	-0.5	-4.2	-7.6	-3.7
F	9.1	2.9	12.2	18.4	10.4	-8.6	-1.7	0.5	-3.5	-7.2	-5.2	-1.4	-2.5	-2.8	-2.8	-1.3	-3.8	-5.6	-6.3	-5.4
G	6.6	5.4	2.8	3.4	20.2	8.0	-3.2	-5.5	-3.5	-4.2	-2.3	1.7	-4.3	1.4	0.1	-2.6	-3.0	-3.9	-4.0	-0.4
H	9.3	1.6	7.7	7.3	1.2	17.8	12.7	-5.1	-4.7	2.6	2.6	-4.7	-5.7	-7.0	-5.4	-4.7	-3.2	3.9	-3.3	-2.5
I	6.7	1.6	9.1	7.5	8.1	6.6	22.5	8.1	7.6	-3.4	-1.9	0.1	-0.6	1.1	3.2	-1.5	-0.2	-3.7	-5.4	-3.7
K	7.6	2.1	5.0	5.0	10.3	4.3	4.7	17.9	17.4	7.3	4.4	-4.8	-2.8	-4.4	-3.7	-5.2	-4.6	1.8	-1.0	-2.4
L	8.9	1.1	8.3	10.9	4.2	6.3	9.9	5.1	6.4	17.1	11.2	-3.7	-9.8	-0.6	-4.2	-4.8	-3.2	0.7	-0.9	-1.3
M	6.5	1.1	1.8	4.2	11.6	2.6	5.6	12.4	7.9	14.1	20.9	8.0	-2.4	-0.8	-1.5	1.0	0.1	-5.7	-6.1	-1.3
N	8.3	5.4	3.9	7.0	9.2	4.6	7.5	12.4	9.9	5.0	6.1	17.8	10.3	-3.6	-3.6	-1.0	-2.0	-5.2	-7.4	-7.0
P	8.4	2.2	12.4	9.1	6.0	8.4	11.5	5.1	9.2	7.0	0.0	7.4	20.1	9.0	2.1	-1.2	-0.4	-4.9	-3.8	-2.1
Q	8.8	0.9	8.8	8.3	4.8	7.3	5.5	4.1	9.2	5.4	9.2	9.0	6.2	18.8	10.0	-0.6	-1.4	-4.3	-6.2	-3.4
R	9.2	2.9	8.7	12.2	3.4	7.0	11.2	2.8	10.9	6.1	5.6	8.3	6.2	11.9	19.8	5.8	2.0	-4.3	-6.2	-2.7
S	8.2	4.2	6.4	9.6	3.8	7.0	9.9	4.4	13.0	5.4	5.0	10.8	8.8	8.6	9.2	15.6	6.8	-1.9	-9.3	-2.7
T	9.8	2.1	9.6	7.6	5.0	8.5	7.2	5.1	8.3	4.6	6.6	9.7	7.8	9.4	8.4	11.8	16.6	7.0	-4.9	-1.8
V	9.0	3.8	8.0	9.3	4.8	6.0	6.8	6.6	9.6	5.2	6.6	10.5	4.1	6.2	4.9	5.5	7.9	16.8	15.2	2.3
W	9.3	5.0	4.6	5.6	8.5	4.2	5.9	13.7	6.1	11.6	10.5	4.1	4.6	6.2	6.0	3.6	0.5	4.9	25.0	10.5
Y	4.0	0.7	3.8	2.2	13.2	3.5	5.8	6.5	4.4	8.8	8.9	3.7	2.4	1.6	7.7	6.4	7.1	8.0	12.1	20.3
Y	5.8	2.1	6.0	6.1	13.2	4.4	9.4	7.3	6.1	7.4	8.5	8.5	2.8	4.7	7.7	6.4	7.1	8.0	12.1	20.3

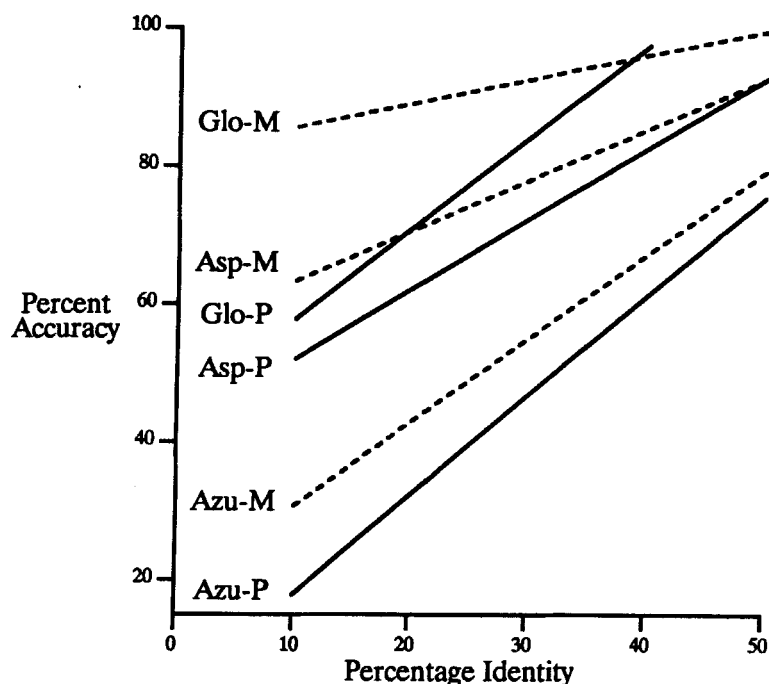


FIGURE 9. Accuracy of globin, aspartate proteinase, and azurin alignments: percentage of positions correctly aligned (in comparison with their aligned 3-D structures¹⁵⁴) and plotted against the percentage sequence identity. Sequences were aligned as pairs (Glo-P, globins; Asp-P, aspartyl proteinases; Azu-P, azurins) and as multiple alignments using MALIGN.⁸⁰ The results are averaged over 14 scoring procedures.⁷⁹ (From Johnson, M. S. and Overington, J. P., *J. Mol. Biol.*, 232, 1993. With permission.)

tures; in this case 59 nonhomologous structures. The total energy of a protein is calculated as the sum of contributions arising from (1) exposed polar groups and buried hydrophobic residues, (2) pairwise interaction between all side chains that are in contact, and (3) aggregates of triplets of interacting side chains. "Topology fingerprints" were then computed for 125 protein structures; sequences matched with a structure can then be evaluated in terms of its energy using the three contributions to the energy described above. The alignment procedure incorporates gap penalty terms to account for insertions and deletions. In the search procedure, a sequence can be tested against all 125 fingerprints and the best scoring structure selected. A fingerprint can also be searched against a data bank of sequences.

Their procedure has been applied to identify some examples of distant relationships such as

plastocyanin and azurin, globin and phycocyanin, and proteins adopting the TIM fold.¹¹² A search in the sequence database for a plastocyanin fingerprint, for example, identified all of the plastocyanins in the database followed by the amicyanins, H.8 outer membrane precursor proteins, azurins, and pseudoazurins, all having a common fold. However, different folds, such as the Ig κ -chain, cytochrome *c* oxidase, dihydrofolate reductase, and "protease B", were also identified possibly because of uncertain alignments or because of short, very good alignments.

An alignment between the azurin sequence and the fingerprint of plastocyanin led to a model constructed using their lattice refinement method.¹¹⁵ However, the RMSD (root mean squared deviation) between the model and crystal structure is of the order of 7 Å. Their search procedure performed with a similar

success rate for the globin/phycoerythrin family and for TIM folds, where some nonnative folds were identified among the correct ones.

In order to escape from the limitations of the three-dimensional structure of any one member of the family, it may be necessary to “project” the restraints of the three-dimensional fold onto the one dimension of the sequence⁹ (Figure 1) and to work by comparing sequence templates or profiles. This can be approached by determining the propensity of an amino acid to occur in each class of local structural environment defined by solvent accessibility and secondary structure as shown by Eisenberg and his colleagues.¹¹⁶ Bowie et al.¹¹⁶ reclassified the 20 amino acid types in terms of 18 local structural environments based on the analysis of sequence alignments and the extrapolation of environments from representative 3-D structures. Alternatively, it can be achieved by calculating substitution tables as a function of local environment.^{117,118}

The method of Johnson and co-workers^{80,119} uses expanded amino acid substitution tables that take into account the local environment in tertiary structures (Figure 10). Analysis of families of structurally aligned 3-D structures led to grouping of amino acids into 766 different classes. These approaches can be used not only to compare a structure or structural alignment against a data bank of protein sequences but also to search a sequence or alignment of sequences against a data bank of all available structures. This approach has been exploited in the alignment and modeling of lignin peroxidase on the basis of the structure of cytochrome *c* peroxidase.^{120,121} It can also be used to establish relationships between sequence and structure when the similarity is very low (Figure 8). For example, whereas a sequence search of a bacterial serine proteinase will not locate any of the mammalian proteinases (not shown), a search of the structural templates with a bacterial sequence (or a mammalian sequence), as shown in Figure 11, will find all of the mammalian proteinases (or all of the bacterial sequences) within the top scoring matches,⁸⁰ something not achieved by some threading approaches.

The methods of Bowie et al.¹¹⁶ and Johnson et al.⁸⁰ are both able to detect distantly related sequences that adopt a particular protein fold,

even when the sequence identity is between 10 and 20%. But none of the methods is successful at identifying all members of a family. A major problem is in introducing useful gap penalties. Johnson et al.⁸⁰ have done this by introducing structure-dependent gap penalties as used in the comparisons of protein three-dimensional structures. However, when there are long insertions and deletions within equivalent domains or where one structure contains extra domains on either end, then the comparisons are problematic. For instance, the difficulty in identifying mammalian serine proteinases on the basis of the bacterial enzymes and vice versa is, in part, a consequence of the long insertions in the sequences of one group relative to the other (Figures 2, 3).

4. Alignment and Searching of Three-Dimensional Structures

Fast, quantitative measurement of structural similarity and the comparison of protein structures are central to comparative modeling. The first systematic methods to quantitatively identify structural similarity were used to compare homologous proteins^{122–126} by superposition of main-chain C^α -atoms. The “distance” or “diagonal” plot, where the distances between a pair of C^α -atoms are plotted as a matrix, provides a rapid way for visual recognition of domains as well as to identify structural similarity between two proteins.^{127–130}

Rossmann and colleagues (Reference 131 for review) extended methods of rigid-body least-squares superposition to the search for similar substructures between more distantly related proteins. A primary problem with these comparative techniques is the identification and assignment of the “initial equivalences”. This they achieved by visual inspection of the two protein molecules.^{132–135} An update of the “initial equivalences” is made by attaching a probability to each inter-protein C^α -atom pair to be structurally equivalenced, based on the distance between the two C^α -atoms and the local main-chain orientations. The initial structural superposition is then refined through either a repetitive update of equivalences until there are no further increases in their number, or with a

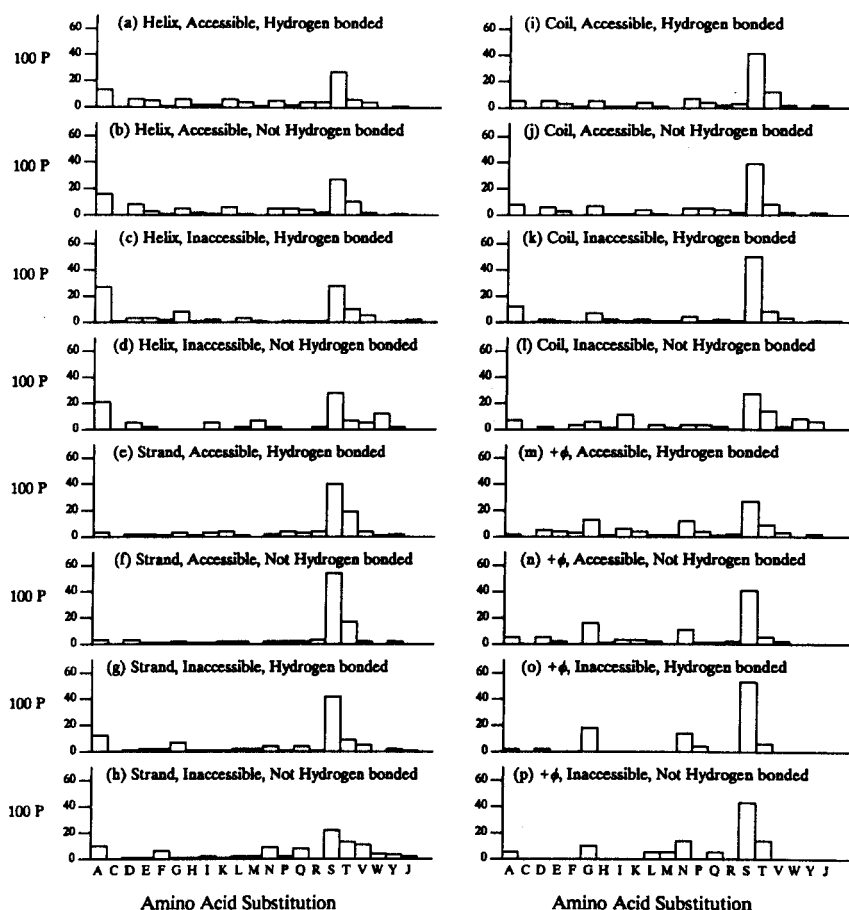


FIGURE 10. Sixteen environment-dependent distributions for one amino acid: serine. The distributions represent the probability of replacement to each of the 21 amino acids (Cys was segregated into cysteine, "J", and half-cystine, "C") given a serine in a particular local environment within a 3-D structure. Scores were obtained from the analysis of 72 aligned families of 3-D structures.^{80,117,119,154} The local protein environments included the secondary structure (main-chain torsions angles), the residue solvent accessibility (<7% or ≥7% relative accessibility) and hydrogen bonding between the side chain and another side chain, main-chain amide or main-chain carbonyl. (From Johnson, M. S. et al., *J. Mol. Biol.*, 231, 735, 1993. With permission.)

systematic variation of the three Eulerian angles that describe the relative orientation of the two structures. In the method of Remington and Matthews,^{136,137} a moving window is used to compare all possible segments between two proteins with a least-squares procedure.

The success of these methods has been documented clearly by the comparison of distantly related structures and the identification of common NAD-binding motifs in unrelated proteins. However, while the results of the former analysis

are severely dependent on the choice of the initial equivalences, the latter method can suffer where there are insertions and deletions between the pair of proteins under comparison.¹³⁸ Additionally, both methods are slow, computationally expensive, and do not converge reliably.¹³⁹

The method of Murthy¹⁴⁰ makes use of regular elements of secondary structure and considers their axes as vectors. The structural comparison of two proteins is then performed much like the method of Rao and Rossmann.¹³² The "goodness"

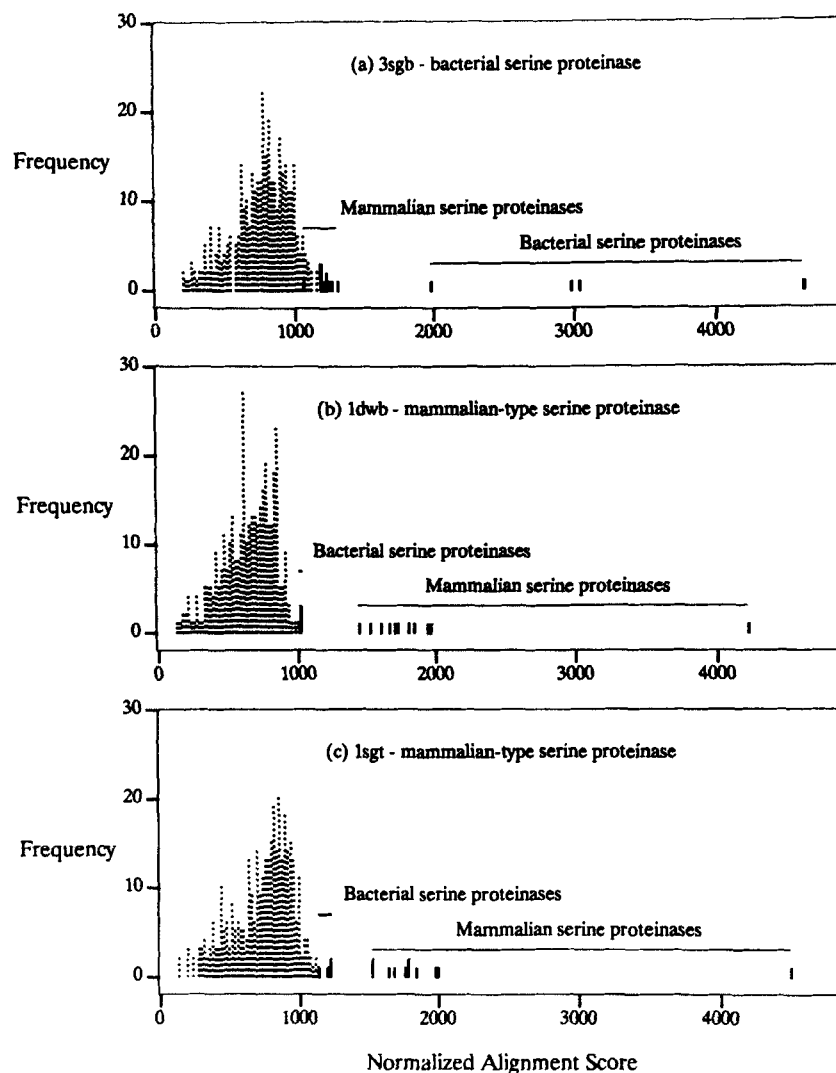


FIGURE 11. Search of bacterial and mammalian serine proteinases against the structural template data bank:⁸⁰ (a) the sequence corresponding to the structure of the bacterial-type proteinase (proteinase B of *Streptomyces griseus*; Brookhaven code: 3sgb; in (b) the sequence of a mammalian-type proteinase (1dwb, bovine thrombin), and in (c) the mammalian-type trypsin of *S. griseus* (1sgt). The templates constructed for the structures in the Brookhaven protein data bank do not include contributions from the serine proteinase structures. (From Johnson, M. S. et al., *J. Mol. Biol.*, 231, 735, 1993. With permission.)

of every superposition is evaluated by aligning the two proteins by means of dynamic programming⁶⁷ and the differences in intra-segment-segment distances. This method has the major advantage in that it is very fast and provides preferred packing arrangements of secondary structures. Sippl¹⁴¹ has used the distance diagonal plot as a tool to compare two protein structures for similar-

ity by comparing segments. The RMSD is computed between all pairs of segments on the basis of the C^α - C^α distances over the segments. Similarities are indicated by high-scoring strips parallel to the main diagonal.

In order to perform comparative modeling based on several known structures, Sutcliffe et al.¹⁴² devised a method for the simultaneous

rigid-body superposition of two or more structures. This program MNYFIT, which makes use of McLachlan's method,¹²⁶ superposes C^α -atoms and uses an iterative procedure to provide an average structure (a "framework") for modeling purposes (Figures 3 and 12). This procedure was subsequently modified to trace-out matched equivalent positions and an alignment of all residues by applying dynamic programming to matrices of C^α - C^α distances.^{143,144} Recently, May and Johnson (unpublished results) have developed an

approach that employs a genetic algorithm combined with dynamic programming in the search for the appropriate rotation matrix necessary to superpose one structure upon another. The major advantage with this procedure is that initial equivalences are not specified and comparisons can be made among analogous proteins in the search for similar substructures (Figure 13).

Barton and Sternberg¹⁴⁵ have used a combination of dynamic programming and intermolecular distance matrices for the comparison of two

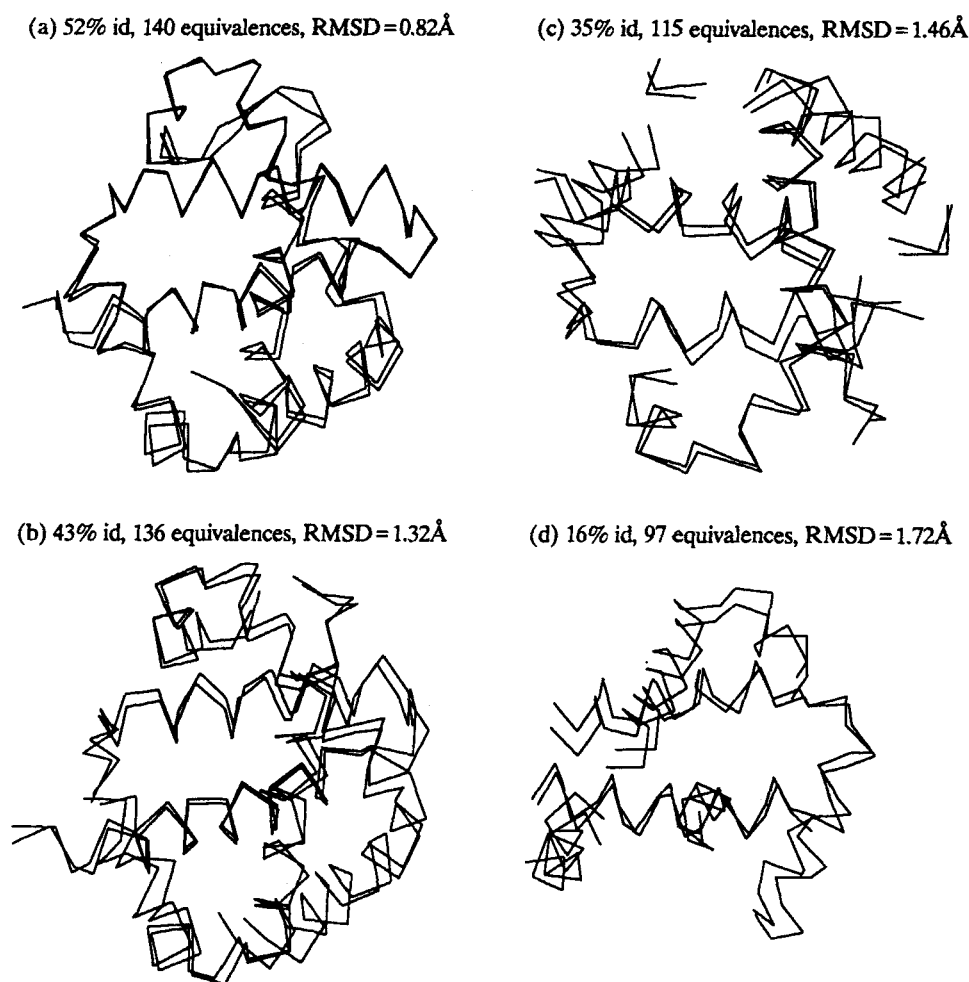
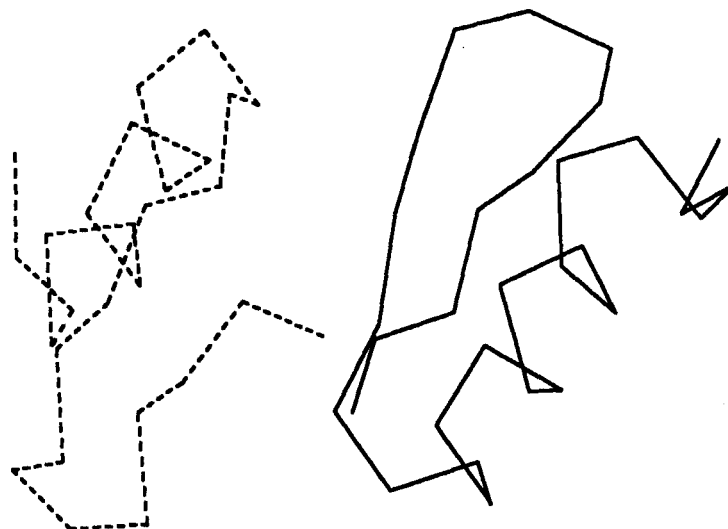


FIGURE 12. Superposition of pairs of structures and reduction in the "framework" contribution as a function of decreasing percentage sequence identity. Traces of the backbones are shown for C^α -positions within 2.5 Å after optimal superposition: (a) the hemoglobin α -chain of *Pagothenia bernacchii* (Brookhaven code: 1pbx) and the α -chain of equine hemoglobin (2mhb); (b) *P. bernacchii* globin α -chain (1pbx) and the β -chain of human hemoglobin (2hhb); (c) the human hemoglobin β -chain (2hhb) and the sea lamprey globin (2lhb); (d) the erythrocrucorin of *Chironomus thummi thummi* (1ecd) and the leghemoglobin of *Lupinus luteum* (1lh1). Structures were superposed with MNYFIT.¹⁴²

(a) Zinc-fingers before comparison



(b) Zinc-fingers after comparison

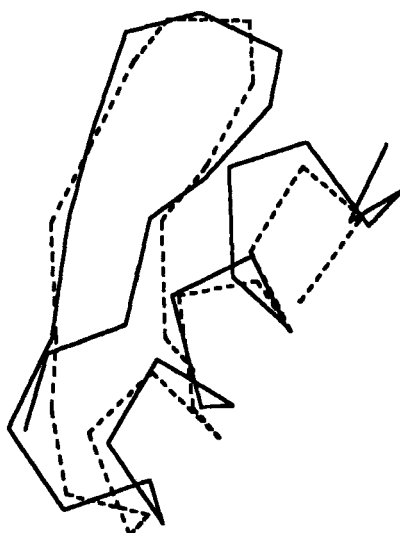


FIGURE 13. Superposition of zinc-finger domains (Brookhaven codes: 5znf and 1bbo) using an approach that combines the use of genetic algorithms and dynamic programming. In (a) the initial orientation of the structures is shown followed by (b) their optimal superposition (May and Johnson, unpublished results).

equivalent loop regions in proteins. This procedure, LOPAL, was used successfully for the structural comparison of loops from the immunoglobulin family.

Another approach based on the distance plot is the method of Richards and Kundrot,¹⁴⁶ who have derived patterns for helices, extended strands, and their combinations. A search for similar pat-

terns can then be made against the structural data bank.

The procedure WHATIF by Vriend and Sander¹⁴⁷ identifies similar structural fragments between two proteins, followed by a 3-D superposition to check for false positives. The identification of similarity is based on comparing patterns of intrafragment distances and their

representation on a diagonal plot. This program is not sensitive to insertions and deletions between fragments or the topology of the fragments.

Taylor and Orengo^{148,149} have used dynamic programming to perform structure comparisons at the level of interatomic distances. They consider main-chain dihedral angles, solvent accessibility, hydrogen bonding, and characteristics of the residues. The method has been tested on several protein families. Faster (computationally) versions of this procedure compare only those residues that share similar features¹⁴⁹ or elements of secondary structure.¹⁵⁰

Karpen et al.¹⁵¹ focus on substructural comparison between two proteins based on RMSD in the main chain dihedral angles and is otherwise similar to the method of Remington and Matthews.^{136,137}

Zuker and Somarjai¹⁵² superpose several segments of one protein on equivalent segments of another. This method seems to be critically dependent on the features used for optimal superposition, which may include the consideration of any gaps between segments, the number of residues within segments, and the RMSD between equivalent segments.

The program COMPARER developed by Šali and Blundell¹⁵³ uses conserved features at various levels of structural organization and can compare two or more structures. These features include the residue similarity, secondary structure, solvent accessibility, chain direction, and hydrogen-bonding patterns. Thus, every residue pair is compared with respect to one protein feature at a time. The optimal alignment of the structures is performed with a combination of dynamic programming (individual features are weighted and combined together) and simulated annealing (used to equivalence relationships among patterns of hydrogen bonding). Once aligned, conservation of key features among the proteins can be easily identified with the program JOY¹¹⁷ (Figure 14). Indeed, nearly 90 family alignments have been constructed for 347 protein structures;¹⁵⁴ for example, one of these families, the azurins and plastocyanins, are shown in Figure 14. COMPARER also includes a gap-penalty function to help improve alignments when gaps are introduced between helices and strands.¹⁵⁵ Indeed, by considering many features

of protein structure, COMPARER can align the more distantly related proteins. Figure 15 shows a comparison of the alignments for some aspartic proteinase domains on the basis of sequence, rigid-body structural superposition (MNYFIT¹⁴²) and the COMPARER approach.

The structure comparison method of Rose and Eisenmenger¹⁵⁶ also employs dynamic programming. The initial alignment is automatic and based on geometric criteria such as the curvature and the torsion of cubic splines through C α -atoms. The procedure has been applied to serine proteases, globins, lysozymes, and small copper-binding proteins. However, in the case of NAD-binding proteins, the performance of the method is less satisfactory, as the initial alignment proves to be difficult.

Subbarao and Haneef¹⁵⁷ have applied techniques from graph theory to the automatic identification of topological equivalences between pairs of molecules. The equivalences are defined by maximizing the number of residues with similar environments. This procedure has been applied to approximately 600 protein structures to obtain 107 "unique" folds.

Graph theory has also been exploited successfully to identify secondary structural motifs in proteins. A protein structure is interpreted as a labeled graph, where the secondary structural elements and their relationships with respect to length and angle correspond to the node and edges of the graph, respectively. Koch et al.¹⁵⁸ describe a " β -graph", the graph notation for β -sheet structural regions in proteins and their procedure to extract β -sheets in the form of a β -graph for various proteins in the Brookhaven Protein data bank.

The program POSSUM, which is an extension of Ulmann's subgraph isomorphism algorithm,¹⁵⁹ is used to search graph representations for known patterns.¹⁶⁰ This procedure was tested for its ability to identify β -strand substructures described by Richardson.¹⁶¹ While the procedure was able to pick up as many as 37 β -strand rich patterns, the method failed to identify another 18 patterns. This was attributed to the Kabsch and Sander¹⁶² method of identifying secondary structures, which is the major input to the program. Some strands are not identified, particularly where they are distorted.

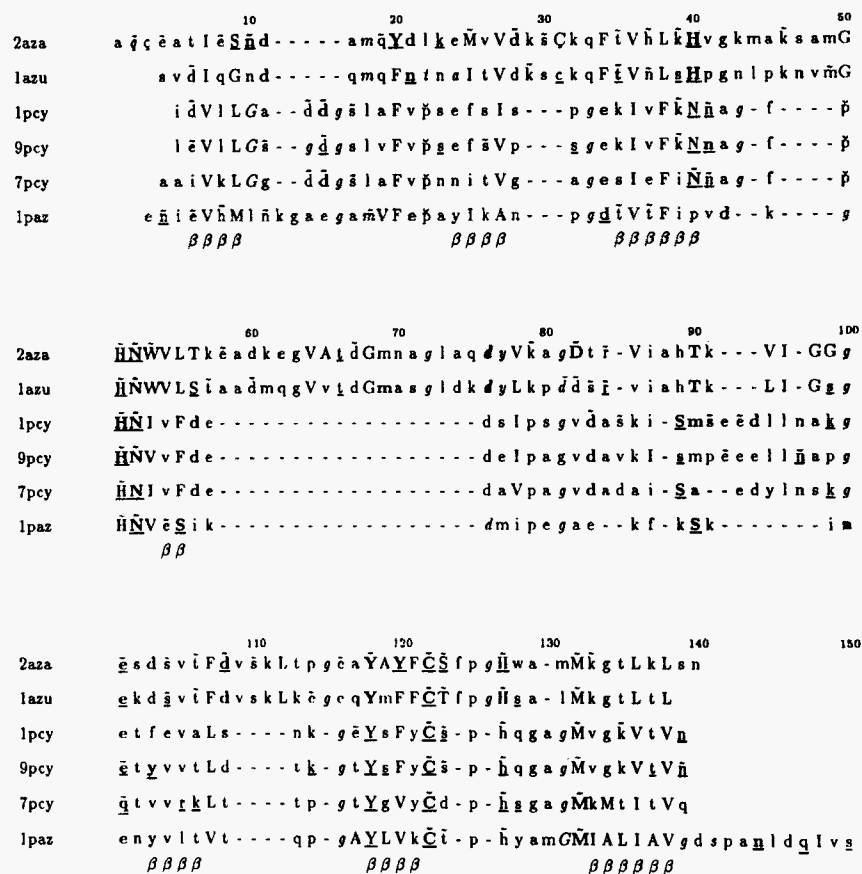


FIGURE 14. Structural alignment of one family from the structural alignment data bank.¹⁵⁴ The azurin/plastocyanin family were aligned with the program COMPARE¹⁵³ and encoded by JOY^{117,119} to display local structural features (Brookhaven codes: 2aza, azurin of *Alcaligenes denitrificans*; 1azu, azurin of *Pseudomonas aeruginosa*; 1pcy, plastocyanin of *Populus nigra Italica*; 9pcy, plastocyanin of *Phaseolus vulgaris*; 7pcy, plastocyanin of *Enteromorpha prolifera*; 1paz, pseudoazurin of *Alcaligenes faecalis*). The standard one-letter amino acid code is used ("C" = half-cystine; "-" = a deletion) but with the following additions made to exhibit the similarities and differences in structural environments within the proteins:^{117,119} UPPER CASE, solvent inaccessible (i.e., L: leucine); lower case, solvent accessible (i.e., d: aspartic acid); ϕ and *italic*, positive phi torsion angle (i.e., g: glycine); *cis*-peptide, breve (i.e., \breve{p} : proline); hydrogen bond to another side chain, tilde (i.e., \tilde{N} : asparagine); hydrogen bond to main-chain amide, **bold** (i.e., **e**: glutamic acid); hydrogen bond to main-chain carbonyl, underline (i.e., d: aspartic acid); disulfide bond, cedilla (i.e., ç: half-cystine); β : beta-strand main-chain. Numbering is by position in the alignment.

A query graph constructed for the chemotaxis protein Che Y consists of a five-stranded β -sheet, three helices on one side of the β -sheet and two on the other.¹⁶³ Interestingly, it pointed to the structural similarity of Che Y to the elongation factor Tu (EF-Tu) of *Escherichia coli* (Figure 16). An extension of this method to consider the

maximum common subgraph¹⁶⁴ was used to show that carboxypeptidase A shares structural similarity with leucine aminopeptidase.

Grindley et al.¹⁶⁵ have applied the method of Mitchell et al.¹⁶⁰ to the identification of common structural patterns from distantly related proteins such as the azurins and immunoglobins and to the

(a) F-STR

```

*****
4APE-N ---STGSATTTIDSLDDAYITPVQ-IGT-----PAQTINLDFDTGSSDLNVFSSETTASEVDGQTIYTPSK
2APP-N --AASGVATNTPTA-NDEEYITPVT-IG-----GTTINLNFDTGSADLNVFSTELPASQQSGHVSYPNSA
2APR-N --AGVGTVPMTDYG-NDIEYQGVT-IG-----PGKKFNLDFDTGSSDLNIASTLCT-NCSSGQTKYDPNQ

4APE-C YTGSIITYTAVSTKQ---GFWEWTSTGYAVGSGTFFKSTSIDGIADTGTTLLYLPA TVVSA-----YNAQ
2APP-C YTGSLTYTGVDNSQ---GFWSFNVDSTAGSQ-SGDG-FSGIADTGTTLLLDSDSVVSQ-----YYSQ
2APR-C FKGSLTTPIDNSR---GWWGITVDRATVGTSTVAS-SFDGILDGTGTTLLILPNNIAAS-----VARA

*****
4APE-N STTAKLLSGATWSISYGDGSSSSGD---VYTDVTVSGGLTVTGQ-----AVESA KKV5
2APP-N --TGKELSGYTWSISYGDGSSASGN---VPTDSVTVGCVTAHQ-----AVQAAQQIS
2APR-N SSTYQAD-GRTWSISYGDGSSASGI---LAKDNVNLGGLLIKQ-----TIELAKREA

4APE-C VSGAKSSSSV-----GGYVFPESA-TLPSFTFVGVSARIVIPGDYIDFGPISTGSSSCFPGGIQSSA---
2APP-C VSGAQQDSNA-----GGYVFDCT-NLPDFSVSISGYTATVPGSLINVGPSGD-GSTCLGQISNS---
2APR-C Y-GASDNGD-----GTYTISCDSAFKPLVFSINGASFQVSPDLVFEEF---QGQCIAGFGYG---

*****
4APE-N SSFTEDSTIDGLLGLAFSTLNTVSPQQTFFDNAKAS--LDSPVFTADLGY---HAPGTYNFGFIDTTA
2APP-N AQFQQTNDNDGLLGLAFSSINTVQPSQTTFFDTVKSS--LAQPLFAVALKH---QPGVYDFGFISSK
2APR-N ASFASG-PNDGLLGLFDTITTVRG--VKTPMDNLISQGLISRPIFGVYLGAKNGGGGEYIFGGYDSTK

4APE-C -----GIGINIFGD-----VALKAA-----FVVFNGA-----TPTLGFASK---
2APP-C -----GIGSIFGD-----IFLKSQ-----YVVFDSG-----G-PQLGFAPA---
2APR-C -----NWGFIIIGD-----TFLKNN-----YVVFNGQ-----V-PEVQIAPVA---E

```

(b) SEQ

```

4APE-N -STGSATTTIDSLD-----DAYITPVQIGT-P-AQTINLDFDTGSSDL-----WVFSSETTAS
2APP-N AASGVATNTPTAN-D-----EYITPVTIG-----GTTINLNFDTGSADL-----WVFSSTELPAS
2APR-N AGVGTVPMTDYG-N-D-----IEYQGVTIGT-P-GKKFNLDFDTGSSDL-----WI-ASTLCTN

4APE-C -YTGSIITYTAVSTKQGFWEWTSTGY--AVGSGTFFK-STSIDGIADTGTTLLYLPA TVVSAVNAQVSGAKSS
2APP-C -YTGSLTYTGVDNSQGFWSFNVDSTAGSQSG-----DGPSGIADTGTTLLLDSDSVVSQYYSQVSGAQDQ
2APR-C -FKGSLTTPIDNSRGWN-----GITVDRATVGTSTVASSFDGILDGTGTTLLILPNNIAASV-ARAYGASDN

4APE-N EVDGQTIYT-PSKSTTAKLLSGATWSISYG-----DGSS---SSGDVYTD--TVS VGGGLTVTGQAVESAKK
2APP-N QQSGHVSYN-P-SATGKELSGYTWSISYG-----DGSS---ASGNVFTD--SVTVGGVTAHQAVQAAQ
2APR-N CGSGQTKYD-PNQSSTYQA DGRTWISYSG-----DGSS---ASGILAKD--NVNLGGLLIKQGTIELAKR

4APE-C SSVGG--YVFPD-SAT-LP-----SPTFG-----VGSARIVIPGD-YIDFGPISTGSSSCFPGGIQSSAGI
2APP-C SNAGG--YVFD-S-T-N-LPDFSIS-GYTATVPGSL--INVGPSGD-----G-STCLGQISNSGI
2APR-C GD-GT--YTI---SCDTSAFKPLVFSI-----NGASFQVSPDLVFEEF---G-QCIAG-----F-GV

4APE-N VSSFTEDSTIDGLLGLAFSTLNTVSPQQTFFDNAKASLDSPVFTADL---GYHAPGTYNFGFIDTTA
2APP-N ISAQFQQTNDNDGLLGLAFSSINTVQPSQTTFFDTVKSSLAQPLFAVAL---KHQQPGVYDFGFISSK
2APR-N EAA SFASGPN-DGLLGLFDTITTVRGVKTMDNLISQGLISRPIFGVYLGAKNGGGGEYIFGGYDSTK

4APE-C GINIFG-----DVALKAAF-----VVFNGA-----TTP---TL-----G---FASK--
2APP-C GFSIFG-----DIFLKSQY-----VVFDS-----DGP---QL-----G---FAPQA-
2APR-C GHWGFIIIG-DTFLKNNY-----VVFNQ-----GVP-----EVQIAPVAE

```

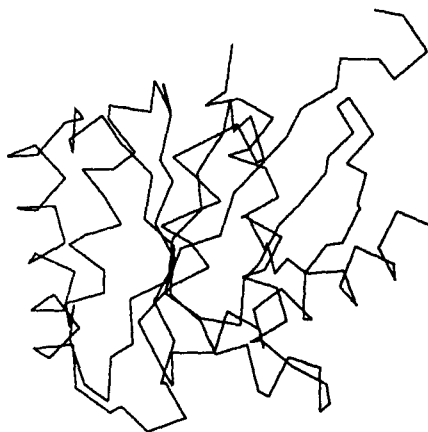
FIGURE 15. Alignments of the aspartic proteinase amino- and carboxyl-terminal domains from multifunctional¹⁵³ (F-STR) and multiple-sequence (SEQ) comparisons. Asterisks indicate those positions among the structures that were found to be equivalent under rigid-body superposition with the program MNYFIT.¹⁴² Note the lack of correspondence between the structure-based and sequence-based alignments. Brookhaven codes: 4APE, endo-thiapepsin; 2APP, penicillopepsin; 2APR, rhizopuspepsin; amino- and carboxyl-terminal domains are labeled with an "N" or "C", respectively. (From Johnson, M. S., Šali, A., and Blundell, T. L., *Methods Enzymol.*, 83, 670, 1990. With permission.)

search for folds common to ubiquitin, thioredoxin, the heat-shock cognate protein, and chymotrypsin. The procedure works well where there are insertions and deletions, where the sequence similarity is low, and where there is circular permutation of sequence as in the lectins. Furthermore, it does not require structural equivalences to be defined prior to the comparison.

C. Protein Relationships

Once aligned, sequences and protein structures can be clustered. A matrix of distances computed from alignment scores for all pairs of proteins can be used to construct trees that describe the relationships among them. While there is extensive methodology for tree construction from

(a) Elongation factor TU (1etu)



(b) Bacterial chemotaxis regulator CHE-Y (2chy)



FIGURE 16. C α -Traces of (a) elongation factor TU (Brookhaven code: 1etu) and (b) bacterial chemotaxis regulator Che Y (2chy) in similar orientations. Both these structures are doubly wound parallel β -sheets surrounded by α -helices.

protein sequences,^{166,167} the clustering of protein 3-D structures has received less attention. Rossmann and co-workers^{132,133} first constructed dendrograms based on structural features alone to describe distant phylogenetic relationships among the nucleotide binding proteins. Johnson et al.^{143,144} have shown that the structural distance metric, defined on the basis of topological equivalences and RMSDs between superposed family members, can give useful dendrograms that correlate well with those derived from sequence (Figure 17). This approach has been extended to include additional sequence and structural features.^{144,153} Because these features can include relationships such as hydrogen bonding, which are known to be conserved in evolution, structures, that bear little similarity in sequence can be compared and classified at statistically significant levels (Figure 18). The correlation of sequence-based and structure-based phyletic trees for the same family of proteins has been used in the selection of structures in which to model an unknown.¹⁶⁸

III. AUTOMATED AND RULE-BASED APPROACHES

In Section I, we have described how modeling using knowledge of proteins with a common fold was applied to give accurate models

as early as 1969 (Reference 10; for additional reviews see References 169, 170). However, for many years it was carried out manually¹¹ or using interactive computer graphics.^{22,24,26,28,29} To minimize the subjective manual decisions in different stages, automatic and rule-based procedures have been developed that exploit knowledge of protein structures in a systematic way (some packages for the display, manipulation, and modeling of proteins are listed in Table 4). Such modeling techniques generally fall into two classes: (a) the assembly of rigid fragments and (b) the use of distance geometry to construct models that are in best agreement with distance constraints.

A. Modeling — Fragment Based

Many approaches depend on the assembly of rigid fragments from existing known structure.^{8,24–26,171,172} They borrow local main-chain and side-chain structures from equivalent fragments in known structures and extrapolate these features to the sequence of the unknown.

1. Databases of Fragment Conformations

Jones and Thirup¹⁷³ showed that a protein structure can be built from a combination of seg-

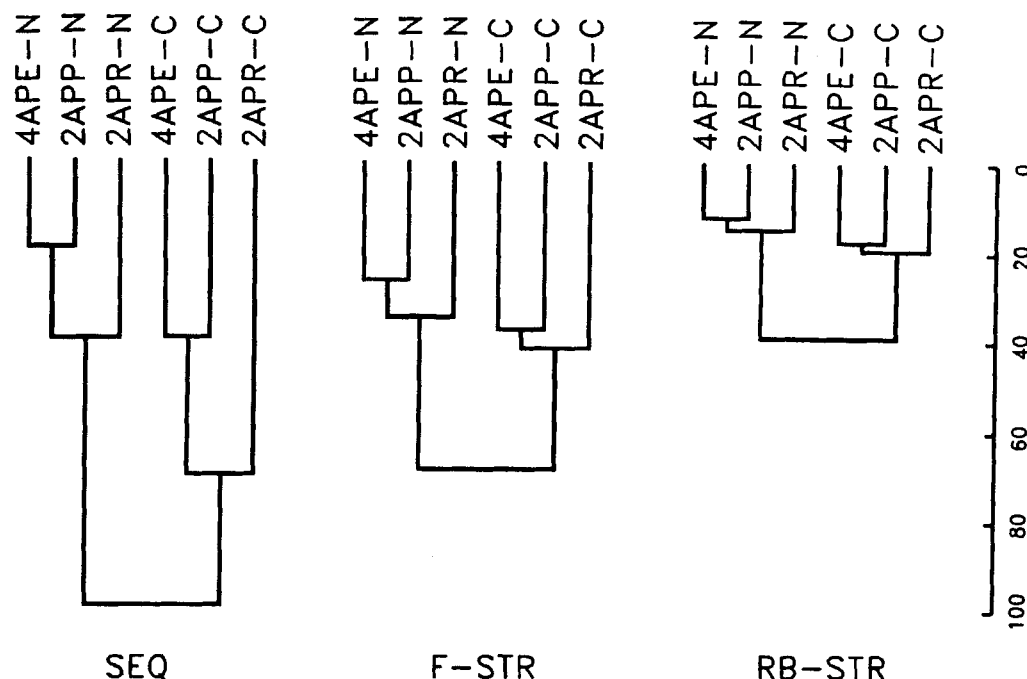


FIGURE 17. Dendrograms derived for the amino- and carboxyl-terminal domains of the aspartic proteinase from multiple-sequence alignment (SEQ), multifeature structural alignment (F-STR) and pairwise rigid-body structural alignment (RB-STR). The proteinases are 4APE, endothiapepsin; 2APP, penicillopepsin; 2APR, rhizopuspepsin. The amino- and carboxyl-terminal domains are indicated with an "N" and "C", respectively. (From Johnson, M. S., Šali, A., and Blundell, T. L., *Methods Enzymol.*, 83, 670, 1990. With permission.)

ments from other proteins. Unger et al.¹⁷⁴ clustered six residue stretches in known protein structures into 100 building blocks, each representing a family of similar structures. This database of hexapeptide structures can then be used to replace 76% of all hexapeptides in known structures with an error of less than 1 Å, and can be joined together to cover 99% of the residues. The model for a protein is constructed by melding together these overlapping "building blocks".

The similar approaches of Claessens et al.¹⁷⁵ and Levitt¹⁷⁶ focus on building local pieces of main chain from C^α -positions. For example, Claessens et al.¹⁷⁵ use various local substructures ("spare parts"), identified from a dataset of 66 high-resolution X-ray structures, to reproduce main-chain conformations successfully (RMSD <1 Å). They also show that this approach can be used to model insertions and deletions. For short fragments, simple geometry-based criteria are

sufficient to reproduce native conformations, but, for long segments, especially those with glycyl residues, one must also use the sequence similarity as a constraint in searching for a suitable fragment.

Several other approaches are available to build a backbone based on the C^α -trace.¹⁷⁷⁻¹⁸¹ Most methods generate backbone coordinates by searching for short segments in known structures with the best overlap of C^α -coordinates. Holm and Sander^{182,183} make explicit use of this approach in comparative modeling of the main chain and side chains when the C^α -trace of the unknown is derived on the basis of known homologues.

2. Selection of Fragments to Construct a Framework

The framework of a family of related structures is defined as the structurally conserved re-

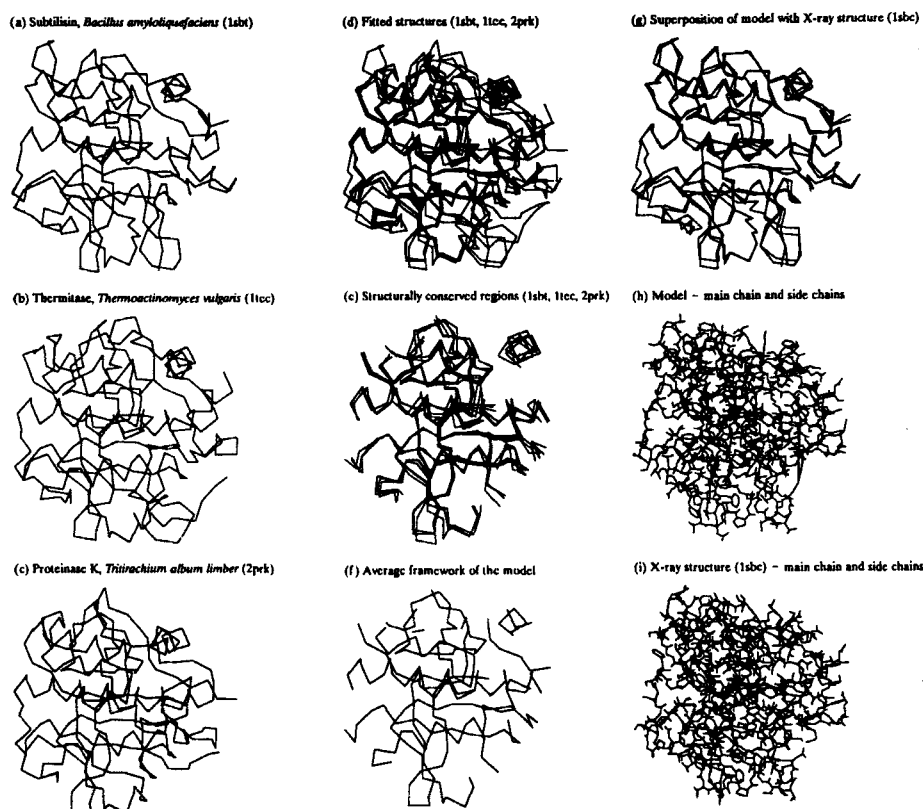


FIGURE 18. Comparative model building of subtilisin Carlsberg¹⁸⁶ using the known structures of (a) subtilisin BPN' (Brookhaven code: 1sbt) (b) thermitase (1tec) and (c) proteinase K (2prk). The multiple superposition of known structures¹⁴² is shown in (d) and the structurally conserved regions (SCRs) in (e). The framework that represents the mean or weighted mean of the SCRs is shown in (f). The superposition of C^α -positions in the modeled and X-ray structure (1sbc) is shown in (g); (h) and (i) correspond to the model and X-ray structures, respectively, with all nonhydrogen atoms displayed. (See Reference 378 for an excellent study of subtilisin modeling.)

gions (SCRs), which are often helices and extended strands. In the comparative modeling approach, COMPOSER,^{142,184,185} the framework is identified by rigid body superposition of homologues of known structure. The topologically equivalent C^α -atoms are defined by a distance cut-off (2.5 Å) and the mean of these positions constitutes the framework. For every SCR in the unknown, the corresponding SCR (in a known structure) with the greatest sequence similarity is superposed on the framework (Figure 18).

The contributions of each of the homologues should be weighted, preferably by the square of sequence identity with the unknown. This results in more accurate models especially where the structures of the homologues include both closely re-

lated and distantly related proteins.¹⁸⁶ Figure 18 illustrates the superposition of subtilisin BPN', proteinase K, and thermitase, which have been used in modeling subtilisin Carlsberg. Models derived from the weighted and unweighted selection of frameworks show that the weighted selection leads to more accurate models.¹⁸⁶

3. Classification of Irregular Regions

Comparisons of structures from proteins^{154,187} reveal that whereas secondary structural elements tend to be well conserved, loop regions accommodate most replacements, deletions, and insertions.¹⁸⁷⁻¹⁹⁰ Hence, the loop regions have proven

TABLE 4
Some Useful Software for the Analysis, Display, and Modeling of Biomolecules

Contact address of company/person	Software	Remarks
BIOSYM Technologies 9685 Scranton Road San Diego CA 92121-3752 Phone 619/458-9990	Insight II Discover DMol DelPhi Homology Apex-3D Ludi Turbomole	Modeling and graphics molecular mechanics and dynamics package quantum mechanics package package for analysis of electrostatic properties <i>Comparative modeling software</i> Activity prediction system <i>De Novo</i> Ligand design Quantum chemistry program
Molecular Simulations, Inc. 16 New England Executive Park Burlington MA 01830 Phone (617) 229-9800	QUANTA, CHARMM and XPLOR NMR Workbench	Sequence searches and molecular modeling and minimization NMR structure determination
Oxford Molecular, Ltd The Magdalen Centre Oxford Science Park Sanford-on-Thames Oxford OX4 4GA Phone (0865) 784600	STAR UHBD Polaris AbM Iditis PROCHECK	Structure refinement of macromolecules Solvation free energy Electrostatic properties Modeling antibodies Protein structure-derived database and search tools Stereochemical quality of protein structures
AUTODESK 2320 Marinship Way Sausalito CA 94965 Phone (415) 332-2344	HyperChem	Software for design, visualization of molecular structures
Tripos Associates 1699 S. Hanley Road, Suite 303, St. Louis MO 63144-2913 Phone (800) 323-2960	COMPOSER (SYBYL) Leapfrog Receptor Unity	Comparative modeling package docking and drug design Chemical information discovery
A. Jones Department of Molecular Biology Uppsala University Biomedical Center, Box 590, S-751 Uppsala, Sweden	FRODO, O	Interactive graphics analysis and modeling of proteins
Langridge, R. Univ. California San Francisco School of Pharmacology Dept. Pharmaceutical Chemistry Computer Graphics Laboratory San Francisco CA 94143	MIDAS	Graphics display of macromolecules
Vriend, G. European Molecular Biology Laboratory, Heidelberg Germany	WHAT IF	Graphical interface for structural alignments of proteins
Evans, S.V. (Present) National Research Council of Canada Institute for Biological Sciences M54, Ottawa, Canada, K1A0R6 e-mail - elmo@nrcbsa.bio.nrc.ca	SETOR	Hardware-lighted 3-D display

particularly difficult to model accurately. However, the accurate modeling of loops may be crucial, as they are often involved in recognition processes, for example, the hypervariable regions of immunoglobulins.

a. Elements That Bring About Sharp Chain Reversal — the β -Turns

Venkatachalam¹⁹¹ showed that certain combinations of backbone dihedral angles of a tetrapeptide give rise to a sharp reversal of the polypeptide chain. Such tetrapeptides, the β -turns, were identified by the presence of a 4 \rightarrow 1 hydrogen bond $[(i + 3) \text{ N-H} \cdots \text{O} = \text{C} (i)]$. Model calculations have revealed six types of β -turns (designated type I, II, III and I', II', and III'), which were shown to occur widely in proteins.^{192–194} Figure 19 shows ideal values of ϕ and ψ at the two central residues for the six types of β -turns. While

the type I β -turn is most frequently observed, the type II' turn is the rarest.^{195,196}

Lewis et al.¹⁹⁴ relaxed the definition for β -turns by introducing more geometric freedom in the assignment of hydrogen bonds as well as introducing distance criteria for β -turn identification. Types I and III and types I' and III' form very similar pairs and are often not distinguished in analyses. A reclassification of the β -turns into I, I', II, II', VIa, VIb, and IV by Richardson¹⁶¹ is widely followed. Wilmot and Thornton¹⁹⁵ used a dataset of 59 high-resolution protein structures and a C_i^α to C_{i+3}^α distance criterion; this study led to the identification of the type VIII turn, which does not possess a stabilizing H-bond but is quite abundant in proteins.

Amino acid preferences for the two central residues were used for protein secondary structure prediction¹⁹⁷ (for classic reviews see References 198–201). Gly, Pro, Asn, and polar residues have a high propensity to lie in β -turns.¹⁹⁵

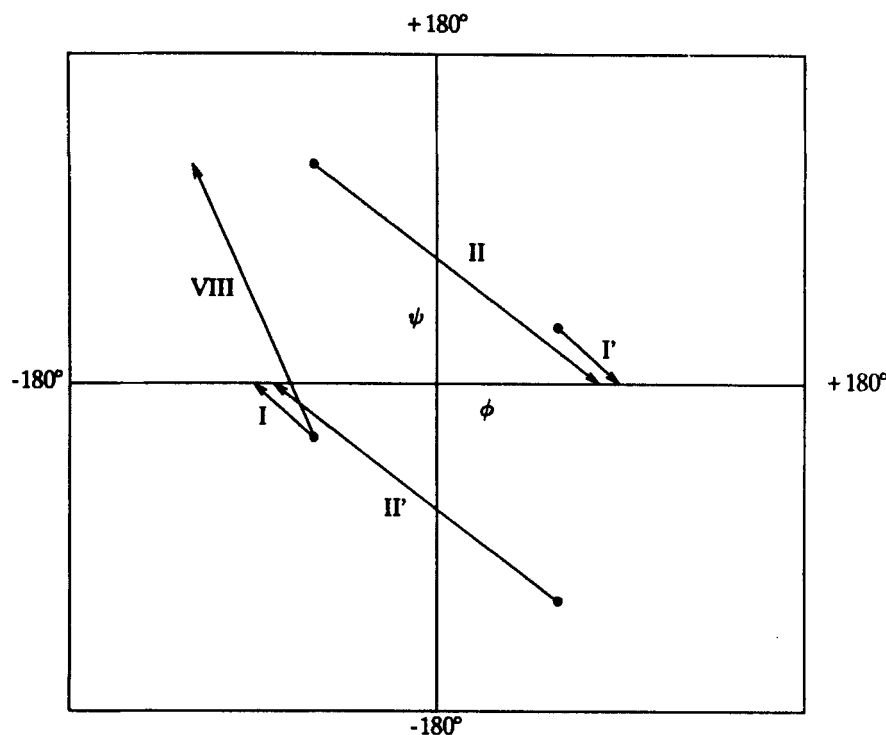


FIGURE 19. The (ϕ, ψ) values at the middle two positions of some of the common β -turn types in protein structures. For every turn type “•” indicates the (ϕ, ψ) values at the first of the two positions of the turn and the tip of the arrow represents the (ϕ, ψ) values at the following position. The β -turn types are indicated at each arrow.

Residue preferences were observed for different turn types in positions flanking turn forming residues. For example, Gly with positive ϕ torsion angle is often found at position $i+3$ of type I β -turns.

b. Short Loops Connecting Secondary Structural Motifs — the Example of $\beta\beta$ Motifs

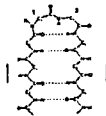
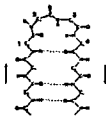
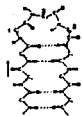
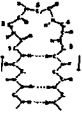
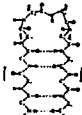
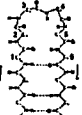
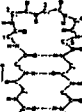
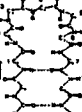
An analysis of the loops in α - α , α - β , β - α , and β - β motifs indicate that distinct patterns exist in these loop regions.^{202–205} Loops in β -hairpins have been classified according to the number of residues and the hydrogen bonding patterns.²⁰⁶ When equivalent hairpins in homologous proteins are compared, it is often found that the number of residues and hydrogen bonding are conserved,

despite changes in amino acid sequence. Glycines are often conserved in short loops and can be used to distinguish particular conformations.²⁰⁷ However, interconversion of β -turn types I and II is quite common within β -hairpins belonging to the same protein family. Insertions of a residue in a β -hairpin can disrupt the β -ladder hydrogen-bonding pattern by the presence of a β -bulge. Although the sequence patterns are not able to recognize particular turns and hairpins in a general sequence search, they can be used to select between conformations if the adjacent secondary structures are defined from a comparison with homologues. These studies provide rules that are useful in comparative modeling (Figure 20).

Loop conformations are strongly dependent on the relative orientation of flanking elements of secondary structure.²⁰⁵ Using a dataset of 65 largely

SYSTEMATIC MODELLING OF β -HAIRPINS

← REPLACEMENT →

SET	DOUBLE H-BOND				SINGLE H-BOND		ALTERNATIVE
A	2:2 	★ Type I' Gly - Asn - Gly - Asp oL yL	★ Type II' - Gly - Ser - Thr c yR	Type I - X - X - oR oR	2:4 	unusual— Various	6:6 6:8 10:10 10:12
B + 1	3:3 	Rare - Various			3:5 	★ Type I [1-4] • GI β -bulge - X - X - X - Gly - X - s oR yR yL s	Various 7:7 7:9 11:11 11:13
C + 2	4:4 	★ Type I [1-4] - X - X - X - Gly oR oR yR oL	Various		4:6 	Many different conformations	8:8 8:10 12:12 12:14
D + 3	5:5 	Many different conformations			5:7 	Many different conformations	9:9 9:11 13:13 13:15

DELETION
 ↑
 ↓
 INSERTION

FIGURE 20. A schematic classification of β -hairpin structures. Structures in horizontal lines have identical numbers of residues but different conformations and sequences; these are useful in modeling replacements. Structures in columns represent β -hairpin conformations with differing numbers of residues and these are used in modeling insertions and deletions. (From Sibanda, B. L., Blundell, T. L., and Thornton, J. M., *J. Mol. Biol.*, 206, 759, 1989. With permission.)

nonhomologous proteins, it has been shown that similar L-shaped $\beta\beta$ -motifs are present in several unrelated proteins. For short L-loop structures with ≤ 5 residues, type II and type VIII β -turns are frequently found.²⁰⁸

c. Features of Other Loops

Several groups^{161,188,204–206,208–213} suggest that characteristic conformations are observed for each class of loop. A systematic analysis of backbone conformational characteristics in loops from 15 protein crystal structures has shown that irregular substructures are combinations of small standard structures, with sequence patterns that are characteristic of the conformation and the relative disposition of the strands and helices. For example, a three-residue linker connecting two contiguous helices whose conformation is $\alpha_L\beta\beta$ results in an arrangement of two helices at right angles to each other.²¹⁴

Loops between 6 and 16 residues long and a short distance between the segment termini (<10 Å) have been analyzed by Lezczynski and Rose.¹⁸⁸ “ Ω ” loops are compact substructures, which are often present at the molecular surface and probably act as independent folding units. Ω -Shaped loops from various proteins were examined for regularities in residue composition, size and shape, compactness, accessibility, and their role in protein taxonomy.

Ring et al.²¹³ have adopted a different approach to the classification of loops in proteins in order to identify common patterns. A morphological definition of loops based on virtual torsion angles joining four consecutive C^α -atoms is utilized. From a dataset of 67 high-resolution, largely nonhomologous protein structures, 432 loops (4 to 20 residues in length) were identified. Of these, 205 loops were classified as linear, 133 as non-linear but planar, 86 as non-linear and non-planar, and 8 were designated *compound* loops. Loops are characterized by virtual torsion angles computed over tetrapeptide segments. Every tetrapeptide is designated a one-letter alphabet that represents the structural descriptors. Thus, a loop of i residues is converted into a one-dimensional string of alphabets $i-3$ long. This kind of

classification has led to the establishment of consensus structures of functionally important loops as the calcium-binding EF-hands and the phosphate-binding p-loops. However, the most significant point that has emerged from this study is that most loops appear to be linear or flat.

By means of a comparative study of immunoglobulin structures and sequences, Chothia and Lesk²¹⁵ show that only a small fraction of possible main-chain conformations is adopted by the hypervariable loops. Based on the known structures, some of the hypervariable loops were modeled.²¹⁶ Using the criterion that certain key residues are important in determining the conformation, they predict that about 90% of the hypervariable regions in V_κ domains and about 70% of the H1 and H2 regions in V_H domains have a structure very similar to the variable heavy domains for which crystal structures are available.

d. Uncommon Structures

Classically, all segments in proteins except α -helices and β -strands were considered to be irregular. However, other regular repeating structures do occur in proteins. Examples include 3_{10} helices (for a recent paper see Reference 217), the π -helix (a short stretch in catalase²¹⁸), left-handed α -helices (found in thermolysin²¹⁹), and the ϵ -helix (found in α -chymotrypsin²²⁰).

The recently determined crystal structure of pectate lyase C by Yoder et al.²²¹ reveals a new motif, three parallel β -strands coiled into a larger helix, the parallel β -helix. The β -helix is characterized by a pitch of 0.22 Å with 22 residues per turn. Cohen²²² comments that the spiral is right-handed and runs through such that the i th carbonyl oxygen is hydrogen-bonded to the $i + 22$ nd amide hydrogen, while the interior is packed by the side chains of asparagines, serines, and several aliphatic hydrophobic residues.

Blundell et al.²²³ identified a nine-residue polyproline helix with mean (ϕ, ψ) values around ($-70, 138$), in the crystal structure of avian pancreatic polypeptide. During the study of protein conformations based on virtual parameters centered around the C^α -atoms,^{224,225} it was noted that

a single helix of the collagen type exists in the bacteriochlorophyll protein at residues 277–284; other examples included the short segments in BPTI and in cytochrome c_{551} . An examination of the crystal structures of 40 globular proteins led to the realization that these collagen-like helical segments are common in proteins.²²⁶ Adzhubei and Sternberg²²⁷ show that although prolines are frequently found in these helices, polar residues like serine and threonine are also common and most polyproline helices are solvent exposed. These studies emphasize that the polyproline helix should be considered as a separate class of secondary structure in globular proteins.

4. Searching for Linkers

Most approaches to modeling variable regions involve a search for fragments of suitable length and end-to-end distances with a check that the modeled loop does not clash with the rest of the protein (Figure 21A).^{171,173,176,185} The identified segment is usually fitted to the anchor regions (the ends of the intervening regions in the model that are mainly the helices and strands) and subtle “tweaking” of single bonds are sometimes permitted to achieve the best superposition (F. Eisenmenger, unpublished results). The selection of the correct conformation can be improved by considering the RMS difference in the overlap (anchor) regions and the sequence similarity between the identified segment and the one to be modeled. Candidate loops can also be ranked by using structural templates.¹⁸⁵ The templates reflect the amino acid substitutions that are compatible with the local structural environment for each amino acid defined in terms of main chain conformation, solvent accessibility, hydrogen bonding, disulfide bonding, and *cis*-peptide conformations.^{117,119}

A retrospective analysis of COMPOSER-built protein models whose crystal structures are now available confirms that errors are greatest in the loops.¹⁸⁶ The main reasons for this are that they are often variable in length, sequence, and conformation, even among the proteins in a given family, and they undergo large movements about the mean position in the family as they are situated at

the molecular surface and not in a tightly packed hydrophobic region. The temperature factors are usually high and the electron density poorly defined.

If a loop is modeled on the equivalent segment in a homologous protein, it provides a more accurate structure than if it is modeled on an unrelated protein. If no segment of the same length is found in homologous proteins, a loop of slightly different length but similar sequence may be useful. Sibanda and Thornton¹⁸⁹ show that in homologous proteins, β -hairpin conformations are often conserved for most of the loop, even if there are insertions and deletions (Figure 20). This observation is exploited in the collar extension approach,¹⁸⁶ in which use is made of a loop from a homologue that differs by one or two residues at the most and has greater than 40% sequence identity with the loop to be modeled. The framework is extended to include similar regions and the rest of the loop is modeled on a fragment obtained from a search of all known protein structures. This approach minimizes the error in the loop region. Figure 21B shows an example from a model of subtilisin Carlsberg, where a loop is constructed using the extended-collar approach with the equivalent region in thermitase.

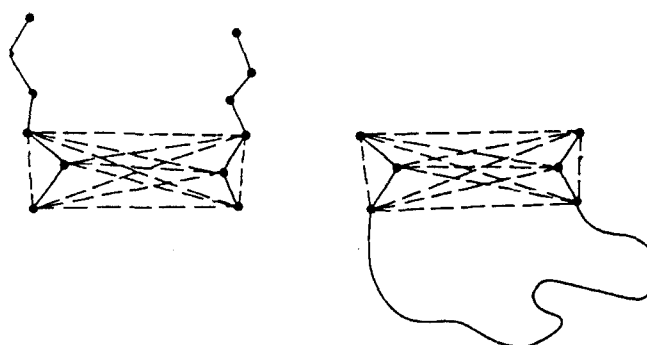
5. Side Chains — Features and Modeling

The proper conformations of side chains are important in the packing of amino acids. In modeling side chains in homologous proteins, preferred side-chain torsion angles, close-packing, covalent cross-links, hydrogen bonds, ion-pairs, and other electrostatic interactions must be considered.

a. Side Chain Torsions

The side chain torsion angles, $N-C^\alpha-C^\beta$ - $C^\gamma(\chi_1)$, $C^\alpha-C^\beta-C^\gamma-C^\delta(\chi_2)$ etc. of amino acid residues in proteins prefer values that correspond to the three staggered orientations about a single bond, viz., $+60^\circ$ (gauche minus or *g*[−]), -60° (gauche plus or *g*⁺) and 180° (trans or *t*). This has

A



B

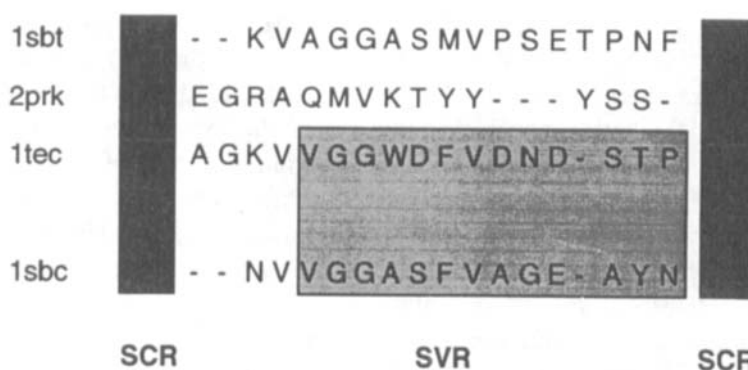
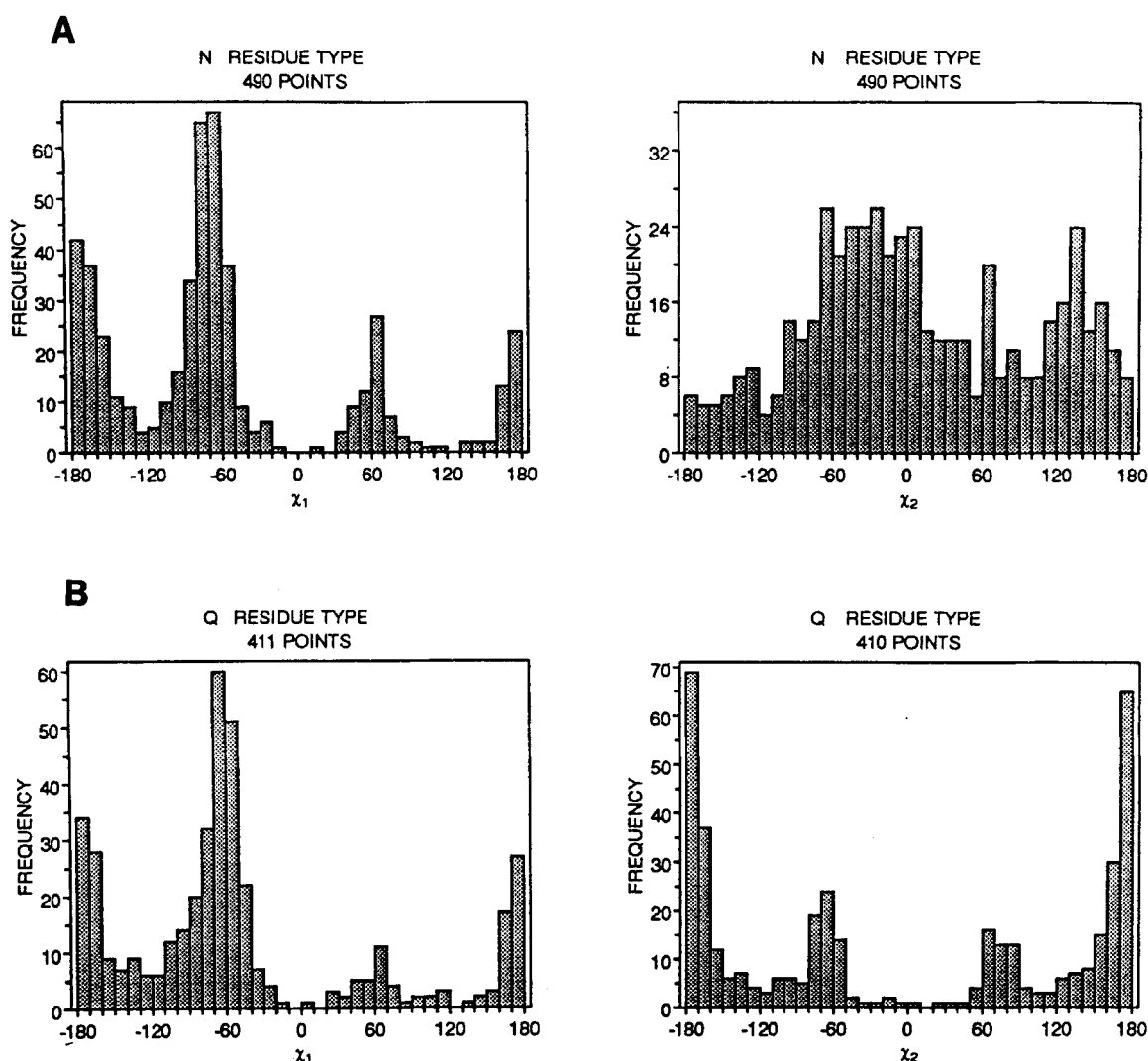


FIGURE 21. Modeling of structurally variable regions (SVRs). The two figures at the top demonstrate the ring-closure procedure. All possible distances between two sets of C^α -atoms (three in each) in the anchor region of the structurally conserved regions (SCRs) flanking the loop are used to search for possible loops from known 3-D structures (top left). Segments satisfying these distance restraints and with an appropriate number of residues are screened to eliminate loops that clash with the rest of the model. Finally, candidate loops are ranked based on local sequence similarity with the unknown or by template matching¹⁸⁵ and according to the RMSD at the fitted anchor regions. The selected loop is melded in the anchor region (top right) and, if necessary, subtle tweaking of backbone torsion angles are made (F. Eisenmenger, unpublished results) in order to arrive at the best fit. Loops are most accurately modeled if they are selected from the corresponding segments from homologous known structures. This is fairly straightforward if the segment has the same number of residues as the unknown. Where the number of residues are slightly different, the collar extension procedure can be used (bottom). Here, subtilisin Carlsberg, which is modeled on other members of the subtilisin family, but the loop has no *exact* equivalent in the known structures. One of the SCRs (darkly shaded) of thermitase (1tec) is extended (lightly shaded) in order to model most of the residues in the loop of subtilisin Carlsberg. Note the sequence similarity in the extended regions. The final two residues of the loop are modeled based on a search for fragments with compatible end-to-end and anchor distances from a database of known 3-D structures. This procedure is likely to result in more accurate loop conformation than one where the entire loop is modeled on an unrelated protein.



A and B

FIGURE 22. Distributions of χ_1 and χ_2 side-chain torsion angles for three residues: (a) asparagine, (b) glutamine, and (c) arginine. These distributions were obtained from 61 high-resolution crystal structures.³⁷⁹

been confirmed by analyses using unrelated protein structures.^{228–230} Figure 22 shows the side chain torsion angle distribution for several amino acids obtained from examination of 61 protein structures. Of the three values, torsion angles corresponding to *t* and *g* are preferred for χ_1 , while *t* is preferred for χ_2 . When the C^γ -atom is trigonal as in Asp, Phe, Tyr, Trp, Asn, or His, preferred values of χ_2 are either $+90^\circ$ or -90° (Reference 228).

The preferred side chain torsion angle distribution depends on the secondary structure of the

polypeptide.²³¹ An analysis of 61 proteins with resolution of 2 Å or better shows that the preferred conformation of χ_1 for residues in an α -helix is *t*, although polar residues with short side chains have a preferred χ_1 value of *g*+.²³¹ Rotamer libraries¹⁰⁸ have been constructed to indicate the preferred side chain torsions for all the amino acids,²³⁰ while Tuffery et al.²³² derive an expanded set of rotamers using dynamic cluster analysis. More recently, a rotamer library derived from 132 known structures has been used to correlate (ϕ, ψ) values with side-chain torsion angle probabili-

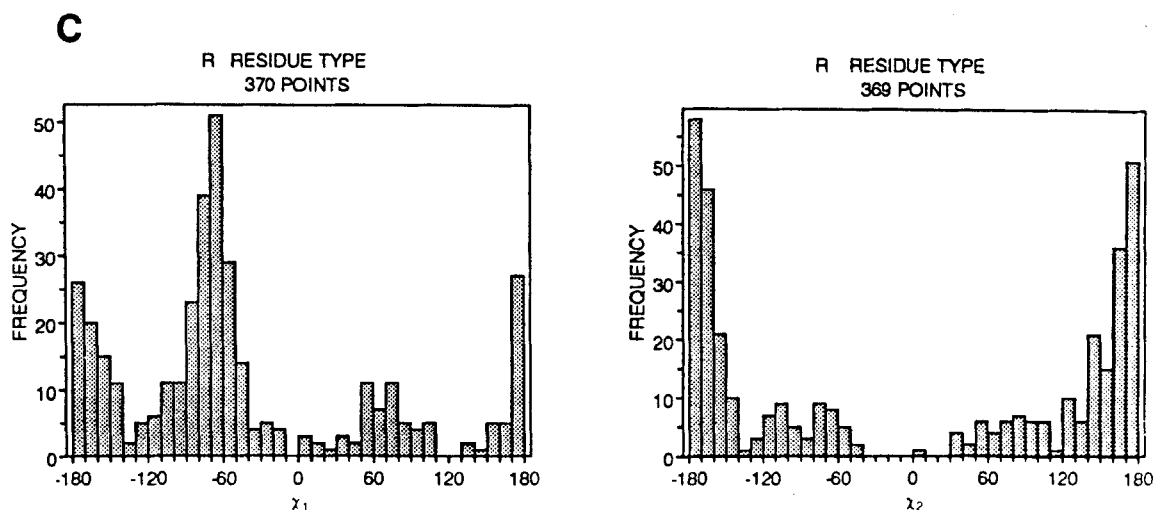


FIGURE 22C

ties.²³³ Well-defined rules for side-chain building have been derived¹⁸⁴ that depend on the main-chain torsion angles and the side-chain orientation in the known homologues.

It has been shown that for a given protein crystal structure, the “pooled standard deviation of χ_1 ”, which is a measure of the deviation of the observed χ_1 from ideality, decreases dramatically as the resolution of the crystal structure improves.⁵³ The calculation of this parameter has been incorporated into the program PROCHECK.⁵⁴ Schrauber et al.²³⁴ reinvestigated a library of side-chain rotamers by analyzing side-chain conformations in light of the accuracy of the crystal structures. They find that 70 to 95% of various amino acids have side-chain conformations within $\pm 20^\circ$ of preferred values and the deviation decreases in more accurate structures. Eisenmenger et al.²³⁵ suggest discrete or limited searches of side-chain conformational space, combined with energy minimization.

b. Volume Effects and Packing

In the interior of a protein, the side-chain atoms of various amino acids are packed efficiently with a small void volume (a measure of the unoccupied space in the interior of proteins). Various analyses have been made to derive rules for efficient packing, and this knowledge can be

extended to protein structure prediction and modeling. Several rules for side-chain–side-chain recognition have been derived.^{236–242} For example, aromatic side chains prefer to pack in a geometry where electronegative atoms are close to the edge of the rings.²³⁸

An algorithm to identify dense clusters of side-chain atoms in proteins²⁴³ finds that clusters with 3 or 4 residues are localized on the surface rather than in the interior, and more than half of the clusters involve residue pairs with oppositely charged atoms within 4.5 Å of each other. This method, which has been applied to a dataset of 157 proteins falling into the three groups (all- α , α/β and all- β proteins, shows that most dense clusters are located at the end of helices and strands in all- α and α/β proteins); the all- β proteins were found to have dense clusters along the middle portion of extended strands.

A side-chain rotamer library has been used to explore optimal packing of amino acid side-chain clusters in globular proteins.^{108,244} This can be used to identify all possible sequences compatible with a tertiary template. Side-chain conformations can be predicted using simulated annealing to optimize packing;²⁴⁵ this approach to energy minimization has been used to position core residues in variants of λ -repressor and to provide an energy-based view of the stability and activity of mutants.²⁴⁶

c. Electrostatic Interactions

While hydrophobic forces play a dominant role in protein folding, electrostatic interactions have been recognized to be quite important in the packing of side-chain atoms.^{236,241,247,248} Baker and Hubbard²⁴⁹ and Stickle et al.²⁵⁰ have derived statistical information on preferred patterns of hydrogen bonding useful in modeling. They show that hydrogen bonds are often formed between near-neighbors along the amino acid sequence, except for salt bridges and side-chain–main-chain hydrogen bonds that are clustered in helix-capping residues. Barlow and Thornton²⁴⁷ and Rashin and Honig²⁵¹ show that interatomic distances between C^α -atoms of residues involved in salt bridges are usually less than 4 Å apart.

A survey of solvent interactions in 16 high-resolution, nonhomologous proteins by Thanki et al.²⁵² showed that solvent side-chain interactions were not random. They find that the proportion of residues whose side-chain atoms are interacting with water is inversely proportional to the hydrophobicity of the amino acid.

d. Covalent Cross-Links

Most disulfide bonds in proteins are inaccessible to the solvent. Cystines have specific stereochemical preferences at the bridge torsions and small inter-Cys C^α - C^α and C^β - C^β distances, which are generally smaller than 6 Å and 4.5 Å, respectively.²⁵³ The torsion angle about the S-S bond (χ_{ss}) has a strong preference for -90° and $+90^\circ$ and χ_1 and χ_2 prefer to be in one of the three staggered conformations. The bridge conformation corresponds to χ_{ss} of -90° and the χ_1 and χ_2 values of -60° (g+) is often preferred and is known as “a left-handed spiral”;^{161,254} seven other distinct less populated families also exist.²⁵⁵

e. Modeling Side Chains

An exhaustive search of all combinations of side-chain rotamers is computationally expensive and methods have been proposed to limit

the conformational search.^{108,245} Holm and Sander¹⁸³ use simulated annealing with a Monte Carlo algorithm to rapidly sample the rotamer library and optimize packing. They predict the correct χ_1 values in 81% of the cases when compared with high-resolution crystal structures. Success falls with the decreasing accuracy of backbone coordinates, as reflected by the resolution of the crystal structure. However, if the backbone is modeled on an accurate homologous structure, 70% of the χ_1 values are predicted correctly if the unknown has high sequence similarity with the template. Desmet et al.²⁵⁶ propose a “dead-end elimination theorem” for identifying and eliminating incompatible rotamers with respect to the given backbone and the surrounding side chains. This approach can dramatically reduce the number of conformations to be searched.

Abagyan and Argos²⁵⁷ have performed conformational searches of side-chain torsions in enkephalins using the Monte Carlo procedure combined with energy minimization. They find that the Monte Carlo selection of random dihedral angles accompanied by energy minimization at a constant temperature performs better than simulated annealing. They also find that searches around preferred torsion angles did not improve the search.

Schiffer et al.²⁵⁸ use energetics to identify the correct side-chain conformation in comparative modeling. The initial orientations of side chains are obtained from a homologue. The solvation energy term was found to be essential in predicting the structure of surface residues, whereas a simpler molecular mechanics calculation is sufficient to correctly model those in the core. Eleven of the solvent inaccessible residues in trypsin were targeted as a test of their procedure. The average volume overlap error for the modeled residues, in comparison with the crystal structures, is 14% ($\pm 8\%$) and this is not significantly different from the average volume overlap error ($13\% \pm 10\%$) for the entire protein. Tests with seven solvent-accessible residues indicate that low energy conformations may not always correspond to those seen in the crystal structure in cases where the residues are associated with ordered water molecules.

Wilson et al.²⁵⁹ describe a similar approach using “energetics” as a tool to position side chains in comparative modeling. They use a library of rotamers and allow an average of 5 to 6 different conformations per residue. In their modeling of the α -lytic protease, correct prediction could be obtained for 90% of the side chains. By testing with several pairs of homologous proteins, they show that accuracy in side-chain orientation is not extremely sensitive to the correct backbone conformation. The RMSD of predicted positions of side-chain atoms rises from 1.31 Å in a test case with the correct backbone to 2.68 Å in a test case with <35% sequence similarity.

Dunbrack and Karplus²³³ utilize a minimization routine that orients a side chain by considering the correlation with backbone torsion angles and packing with other side chains. Their procedure predicted 81% of the side-chain torsion angles correctly in crambin and 61% in lysozyme. But they do point out that results for a single protein do not test a prediction scheme!

In contrast to the Monte Carlo search techniques that attempt to perform simultaneous side-chain prediction, the method of Eisenmenger et al.²³⁵ considers the side-chain conformation based solely on the interactions of side-chain atoms with those of the backbone. The RMSD for side-chain atoms from six protein models lies in the range of 1.0 Å to 2.1 Å when compared with their crystal structures.

In COMPOSER, the side chains are modeled depending on the orientation of side chains in the equivalent positions in the known homologues or based on a large number of rules derived for their preferred conformations in various secondary structures.¹⁸⁴ A side chain in the model is oriented in a similar way to that in an equivalent position of the known structural homologue, so long as conserved features are observed (Figure 23). In cases where the side-chain orientation is not conserved in a topologically equivalent position in the known homologues, then the knowledge derived from a general analysis on side chain conformation in helices and sheets is used.^{184,230} It is emphasized that the building of side chains within COMPOSER is to provide their initial orientations. Other techniques, including energy minimization and localized molecular dynamics, can then be applied to the model.

B. Modeling — Restraint Based

Relatively less attention has been given to modeling procedures using distance restraints (Figure 24).⁹ Havel and Snow²⁶⁰ perform multiple sequence alignment of homologues of the unknown structure and derive distance and chirality restraints based on alignments with the homologue(s) of known structure, referred to here as the template. This is analogous to the derivation of interproton distances from nuclear Overhauser effects (NOEs) in multidimensional NMR experiments. Distance geometry procedures (DISGEO) are used to derive an ensemble of structures compatible with input restraints. Their model of the silver pheasant ovomucoid third domain has a RMSD of 0.78 Å (for C^α -atoms) with the template structure of the Japanese quail ovomucoid third domain (JQOM3) from which it was built. The RMSD between the model and its corresponding experimental structure (SPOM3) was 2.36 Å over the C^α -atoms and 2.60 Å for the backbone atoms. This difference is comparable to that between the experimental structures of SPOM3 and JQOM3, which is 2.28 Å for the C^α -atoms. Thus, the method has produced a model that is very close to the template structure rather than the actual structure of SPOM3. Most errors originated in the amino-terminal segment; if this segment is ignored, the RMSD is 1.1 Å, which is smaller than the structural difference between X-ray structures of SPOM3 and JQOM3.²⁶⁰

Srinivasan et al.²⁶¹ also derive distance restraints from homologous structures but use restraints from steric contacts and the globular dimensions of the template structure from which the model will be built. Optimal packing of side chains in the core is tested. For a model of the α -chain of horse hemoglobin, based on the α -chain of human hemoglobin, which has 84% sequence identity, the RMSD between all atoms in crystal and modeled structures was 1.0 Å.

Taylor,²⁶² who has developed a method for modeling α -helical proteins, first predicts the positions of helices based on multiple-sequence alignments. All packing arrangements of α -helices²⁶³ are generated and are assessed for both general and specific structural rules, involving distance and motif restraints. Application of this

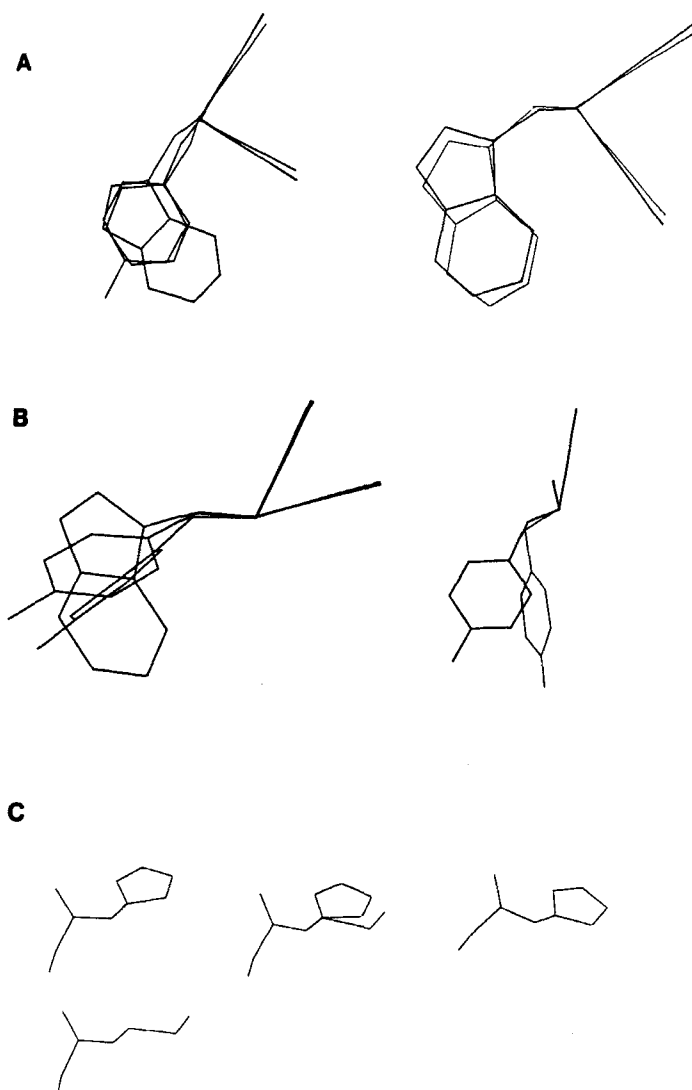


FIGURE 23. Side-chain modeling in COMPOSER. (A) Topologically equivalent side chains, Trp-113 of subtilisin BPN' (1sbt), Phe-113 of proteinase K (2prk), and Tyr-121 of thermitase (1tec) are shown superposed (left) and the superposition of modeled (thin lines) and X-ray (thick lines) conformations of Trp-112 in the equivalent position of subtilisin Carlsberg (1sbc) is shown (right); B similar to A for Trp-104 of 1sbt, Tyr-104 of 2prk, and Trp-112 of 1tec (left) and Tyr-103 of 1sbc, modeled and crystal orientation (right). Note that the accuracy in the side-chain modeling is better where the side-chain orientations are conserved in the structures used to build the model. (C) Illustration of how a side chain is substituted. In this example, His-82 of myoglobin is being built. The two plots on the left show His in the most probable conformation and the Met from which it is being built. The middle plot shows C^β , C^γ , and $C^{\delta 2}$ of His are least squares fitted to the C^β , C^γ , and $S^{\delta 2}$ of Met with weights of 1.0, 1.0, and 0.01, respectively. This ensures that the plane of His and the C^β and C^γ positions of His are defined. The plot on the right shows that the S^δ and C^ϵ of Met are discarded to leave the modeled His residue. (C, from Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L., *Protein Eng.*, 1, 385, 1987. With permission.)

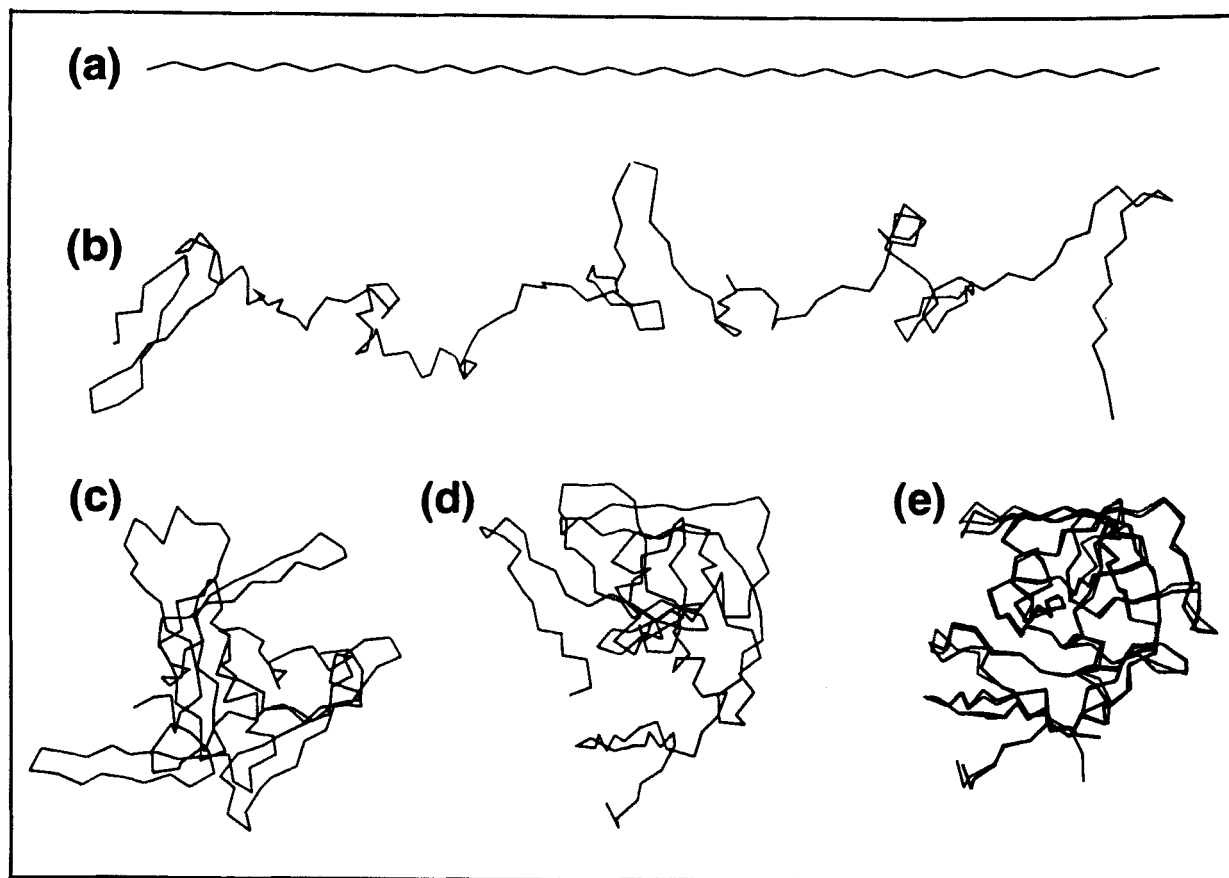


FIGURE 24. Stepwise generation of a model of endothiapepsin using information from several homologous aspartate proteinases and from protein structures in general and expressed as distance constraints on atomic positions. (a) The initial extended main chain ($C\alpha$ -trace), (b) local constraints are considered first, c and d are intermediate structures that arise as global constraints are applied and e is the final model compared with its crystal structure. (From Šali, A. et al., *Trends Biochem. Sci.*, 15, 235, 1990. With permission.)

method to myoglobin and parvalbumin ranked the native folds within the top 4% of the structures generated. Incorporation of protein-specific restrictions like haem-binding in myoglobin and the E-F hand arrangement in parvalbumin selected the native fold as one of two possibilities. An extension of this procedure to non- α -helical structures has been made in modeling the nucleotide binding domain of the cytochrome b_{245} β -chain (W. R. Taylor, D. T. Jones, and A. W. Segal, personal communication). Taylor²⁶⁴ has also developed a method for constructing a 3-D model starting from an idealized fold. This method estimates and scales the pairwise residue distances based on multiple

sequence alignments and conserved hydrophobicity. The scaled distances are found to be compatible with the native forms and are used to pack the hydrophobic core and build a rough fold based on abstract representations of protein architecture.

Šali and co-workers^{9,265} describe a comparative modeling procedure that arrives at a 3-D model by optimally satisfying restraints extrapolated from homologous 3-D structures to the sequence to be modeled. These restraints are expressed as probability density functions (pdfs) for the features to be restrained (Figure 24). For example, the probabilities for main-chain conformation of an equivalent residue in a related

protein are expressed as a function of the local similarity between the two sequences. Several such pdfs are obtained from the correlations between structural features in 17 families of homologous proteins that have been aligned on the basis of their 3-D structures. The pdfs restrain C^α - C^α distances, main-chain N-O distances, and main-chain and side-chain dihedral angles. A smoothing procedure (adapted from Sippl¹⁰⁹) is used in the derivation of these relationships to minimize the problem of a sparse database. The 3-D model of a protein is obtained by optimization of the molecular pdf such that the model violates the input restraints as little as possible. The molecular pdf is derived as a combination of pdfs restraining individual spatial features of the whole molecule. The optimization procedure is a variable target function method that applies the conjugate gradients algorithm to positions of all nonhydrogen atoms. The method is automated and is illustrated by (1) the model of a domain of endothiapepsin⁹ constructed on the basis of other aspartate proteinase domains (Figure 24), and (2) the modeling of trypsin on elastase and tonin that have about 40% sequence identity with trypsin. In the first case, the RMSD with the crystal structure was 0.76 Å, although only C^α - C^α restraints were used. In the second example, the best of 11 very similar models has an RMSD of 0.7 Å for 195 topologically equivalent C^α -atoms, with the crystal structure.

Other methods based on the satisfaction of distance constraints resemble the methods described above, although they have not been applied in the context of comparative modeling. For example, Saitoh et al.²⁶⁶ predict two-dimensional distance diagonal plots based on the Ooi number predicted from the amino acid sequence.²⁶⁷ The Ooi number at any given residue is the number of C^α -atoms in a sphere of 14 Å radius and whose center is at the C^α -atom of interest.²⁶⁸ Once the 2-D distance plot is constructed, the three-dimensional structure is modeled on the constraints incorporated from the two-dimensional distance plot. Applicability of this procedure could be improved by incorporation of knowledge from homologues of known tertiary structure.

Sowdhamini et al.²⁶⁹ have built stereochemically acceptable 3-D models for small

disulfide-rich systems like conotoxin. A large number of random conformations is generated that are then used for screening. This method exploits a database of conformations from disulfide-containing segments together with segments in proteins of known 3-D structure that can accommodate a disulfide bridge.²⁵³ The multiple conformations obtained, for example, in the enterotoxins and conotoxins, undoubtedly arise through paucity of restraints, but they may also reflect the conformational heterogeneity that characterizes these small proteins in solution.

Fujiyoshi-Yoneda et al.²⁷⁰ apply molecular dynamics simulations on distance constraints derived from the homologue of known 3-D structure. Secondary and tertiary folds start from a fully extended polypeptide chain. This method is applied to model *C. atrox* venom PLA₂ on bovine pancreatic PLA₂ and to model trypsin on the basis of elastase. Bohr et al.²⁷¹ derive distance constraints for the main-chain backbone by training a neural network on the basis of a functionally similar protein of known 3-D structure. The binary distances between the C^α -atoms (a value of 0 when the distance is less than the threshold, 1 otherwise) are predicted by the trained neural network and these are utilized to generate the protein backbone using minimization methods; the predicted and X-ray structures of BPTI have a RMSD of 1.2 Å for C^α -atoms. The model of trypsin has a large error of 3 Å, which the authors attribute to errors in the loops. Friedrichs et al.²⁷² propose a method for tertiary structure modeling based on the construction of an associative memory Hamiltonian, which incorporates the knowledge base of amino acid sequences and C^α -traces of proteins of known structure. The Hamiltonian is then used by a molecular dynamics routine with simulated annealing to recall the structure of a protein from the memory database. The application to the comparative modeling of cytochromes results in errors larger than expected compared with the crystal structure, for example, for methods that use the assembly of rigid fragments.

The method by Brocklehurst and Perham²⁷³ also derives restraints from known homologous structures. The structures are aligned on the basis of structural features like main-chain conformation, hydrogen bonds, and secondary structural

features. They primarily use restraints that are directly related to the stabilization of the folded state, namely, main-chain–main-chain hydrogen bonds and interactions between non-polar groups. In view of minimal set of restraints, this method can result in fairly accurate models particularly if the template structures cluster around the target structure. Three-dimensional models of the biotinylated domain from the pyruvate carboxylase of yeast and the lipoylated H-protein from the glycine cleavage system of pea leaf were constructed using this method, on the basis of the known structures of two lipoylated domains of 2-oxo acid dehydrogenase multienzyme complexes. The RMSD between the known structures is 2.5 Å. The RMSD between the one of the unknowns (H-protein) and the two knowns are 1.9 and 2.9 Å, and those between the biotinylated-domain model and the knowns are 1.7 and 1.4 Å.

C. Modeling Integral Membrane Proteins

Comparative modeling strategies have also been applied to the modeling of transmembrane proteins. Structure prediction in this class of proteins is hampered by the limited number of known structures. However, several attempts have been made to model integral membrane proteins by exploiting the major constraint that these proteins are two-dimensional and mostly helical.

The problem has been further simplified by first identifying the sequence corresponding to the helices. For example, the transmembrane helices have been identified using six different secondary structure prediction methods.^{274,275} The validity of this approach was first tested on bacteriorhodopsin followed by the carboxyl-terminal domain of rhodopsin. Two transmembrane helices were identified in the rhodopsin domain, and this method was found to be particularly sensitive in identifying breaks and distortions in helices. Bovine rhodopsin was also modeled after combining results from secondary structure prediction together with data on chemical and enzymatic modifications.²⁷⁶

Because of the interactions of residues from integral membrane proteins with the bilayer, there exists a pattern arising from interaction of vari-

able hydrophobic residues with the bilayer. Hence, helices may be identified either by the periodicity in occurrence of hydrophobic residues (as in References 277–280) or by the periodicity of conserved and variable residues (as in References 281–283). Of the two, the pattern of buried and exposed residues has been found to be weak when compared with globular proteins and hence less reliable in the exact identification of helices.²⁸² In the structure prediction of the C5a receptor,²⁸⁴ a member of the rhodopsin superfamily, positioning of the helices was determined by considering various factors, including Cys pairing, interdigitation of helices, and aromatic cluster formation.

In order to identify residues in the lipid face and to establish the topology of helices, chemical probes were employed in the modeling of ovine rhodopsin.²⁸⁵ In addition, site-directed mutagenesis techniques have been used to identify ligand-binding sites.²⁸⁶ Deletion mutants of the hamster β_2 -adrenergic receptor were able to show that the hydrophobic core was found to be involved in ligand-binding, while most of the hydrophilic residues were not directly involved. With the availability of the bacteriorhodopsin structure derived from cryoelectron microscopic studies,²⁸⁷ transmembrane helix regions as well as potential lipid-facing residues were identified based on the assumption that the core of the transmembrane helix-bundle is roughly the same in all of these proteins.

A suite of programs, PERSCAN, has been developed in order to identify potential helical regions, their buried and exposed faces, and also to predict the polar interface at the borders of the lipid bilayer (Figure 25).^{283,288} PERSCAN can base its analysis on the differences in amino acid substitutions seen for lipid- and water-soluble proteins. Amino acid substitutions were tabulated (a 20-by-20 matrix containing 3853 amino acid replacements) based on the known structures of the two homologous photo-reaction centers and alignments with several sequences.²⁸⁸ A difference probability matrix was then constructed by comparing this matrix with one obtained from water-soluble proteins. This matrix indicates that the lipid accessible residues are less conserved than the inaccessible residues of the water-soluble in-

A

```

br  NIETLLFMVLDVSAKVGFLILLR
sr  AGVALTYVFLDVLAQVPYVYFFYA
hr  GVTSWAYSVLDFVAKYVFVAFILLR
      -----

```

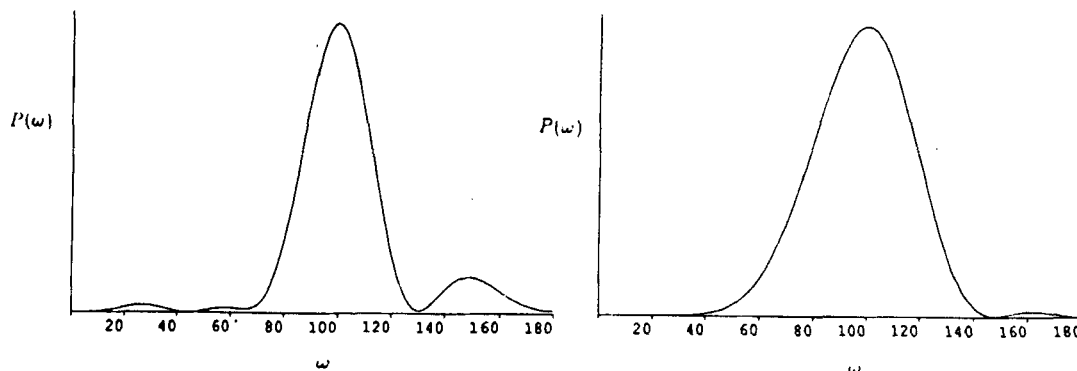
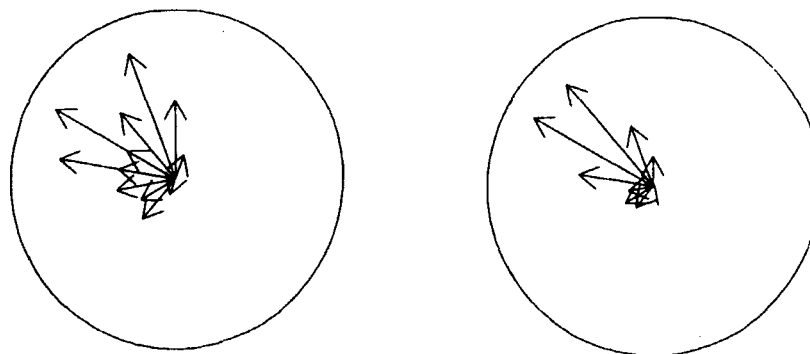
B**C****A, B and C**

FIGURE 25. Fourier transform analysis of bacteriorhodopsin. A: alignment of helix 7 from bacteriorhodopsin (br) with the equivalent regions in halorhodopsin (hr) and sensory rhodopsin (sr). The dashed line indicates the optimal window (range = 7–12 residues) that produced the optimal value for the prediction. B: optimal power spectrum showing a peak (periodicity in amino acid patterns) about 110° (a helix). C: helical wheel plot of vectors that indicate the inside face of the proposed helix. D: helical wheel showing amino acids and the vector indicating the helix internal face. E: vertical representation of the helical wheel with the amino-terminal at the top; horizontal lines to the right indicate buried residues and lines to the left exposed ones. Residues within the optimal window are indicated with solid lines. (From Donnelly, D. et al., *Protein Sci.*, 2, 55, 1993. With permission.)

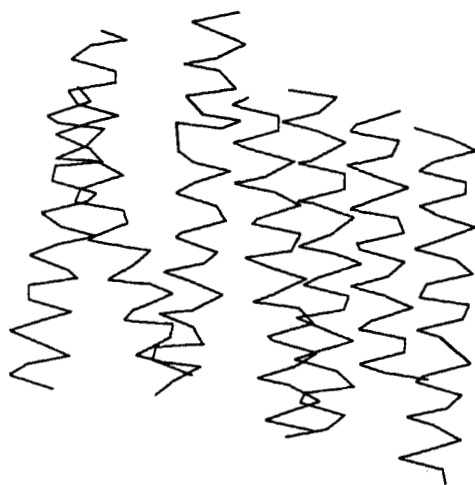
teriors. PERSCAN has been tested by the 3-D modeling of bacteriorhodopsin and subsequent comparison with the structure (Figure 25). PERSCAN has also been applied to the modeling of the β 2-adrenergic receptor (Figure 26), which belongs to a family of membrane spanning proteins that includes the α - and β -adrenergic receptors.

A three-dimensional model of the photosystem II reaction center of pea has been made using

COMPOSER.²⁸⁹ The two crystal structures of the photosynthetic reaction centers from bacterial sources (*R. viridis* and *R. sphaeroides*) have been used as structural templates for this study. Alignment of the pea sequences with the bacterial ones were made using the environment-dependent substitution tables,⁸⁰ which improves the quality of the alignment. The core proteins D1 and D2, including co-factors, have been modeled.

Model of human β_2 -adrenergic receptor

(a) Side view



(b) Top view

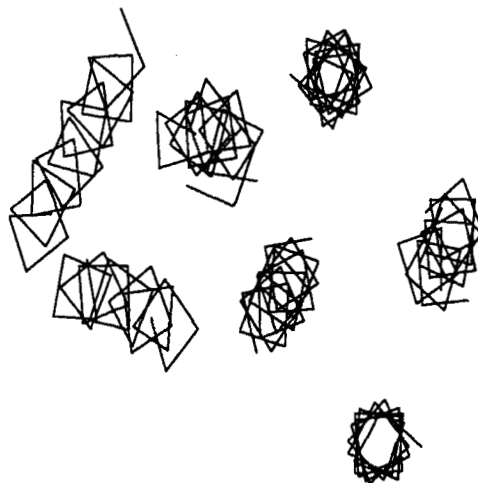


FIGURE 26. Model of the β_2 -adrenergic receptor modeled on the basis of the prediction of transmembrane helices using PERSCAN²⁸⁸ and the projection map of rhodopsin.³⁸⁰ The lipid bilayer is perpendicular to the view in a; b, top view. (From D. Donnelly, J. B. C. Findlay, A. M. MacLeod, and T. L. Blundell, unpublished results.)

check stereochemical parameters like bond lengths, bond angles, and hydrogen-bond geometry.

Assessment of the compatibility of a sequence to a given fold is a challenging task,²⁹³ particularly if the model is built by a procedure that does not make use of homologous structures. Several attempts have been made to identify structural descriptors that discriminate between incorrectly and correctly folded structures. Novotny et al.²⁹⁴ constructed incorrect models deliberately; they built the sequence of hemerythrin, which is predominantly α -helical, on to a domain of immunoglobulin, which is predominantly β -sheet and vice versa (as shown using an alternative procedure in Figure 27). After energy minimization, they compared these incorrect models with native protein folds and found that the side chains could be readily accommodated into the incorrect folds. However, they found that with the incorrect folds more non-polar groups were exposed to the surroundings and more polar side chains were buried within the core than expected from experimentally defined protein structures.²⁹⁵ If these solvent

effects are not taken into consideration, the empirical energy difference between the folded and misfolded forms is not significant.

Bryant and Amzel²⁹⁶ made an analysis of non-bonded contacts in known protein structures for nearest neighbor preferences. They found that hydrophobic residues make twice as many contacts with themselves as one would expect on a random basis. This was confirmed by comparison of this feature among native protein folds and incorrect computer-generated folds. Novotny et al.²⁹⁷ reinvestigated their original "incorrect" models based on hemerythrin and immunoglobulin, using a conformation sampling program (CONGEN) to model side chains. They found that several features could help distinguish between correct and incorrect folds: (1) non-polar side-chains exposed to the solvent, (2) buried ionizable groups, and (3) empirical free-energy functions that incorporate solvent effects.

Baumann et al.²⁹⁸ analyzed 128 3-D structures and compared the distributions of non-polar and polar residues found in their analyses with

those in some “designed” proteins. They identified three of the putative folds to be incorrectly designed on the basis of the unusual distribution of non-polar and polar side chains in the buried core and the solvent accessible surface.

Hendlich et al.¹¹³ have used potentials of mean force derived for interactions between C^β -atoms in 3-D structures¹⁰⁹ to calculate the conformational energy of different folds of a given amino acid sequence. Their procedure can identify native protein folds among a large number of incorrect models, although the discrimination may not be satisfactory for proteins with large prosthetic groups or iron-sulfur clusters. A similar approach by Sippl and Weitckus¹¹⁴ detects native-like folds. They have tested their method by identifying globin folds to be compatible with globin sequences. Chiche et al.²⁹⁹ have found a linear relationship between the solvation free energy of folding and protein size for known crystal structures. The misfolded structures were found to show a higher solvation free energy than predicted. This approach could discriminate between correctly folded and misfolded structures, but in some cases the misfolded structures have values close to the predicted value and hence need very careful analysis.

Lüthy et al.³⁰⁰ extend the work of Bowie et al.¹¹⁶ for generating 3-D profiles and developed a method to assess 3-D models. The compatibility of a sequence to a 3-D fold is tested by calculating scores for the association of a residue with its structural environment (solvent accessibility and local secondary structure). The incorrect models can be identified by examining the profile scores calculated over a fixed-length window moving along the polypeptide chain. Others have calculated propensity tables from a database of aligned homologous proteins (in more than 70 families^{117,119,154}) and used a greater range of local environments (Figure 27); this was used to assess a model of the dust mite allergen Der pI (N. A. Kalsheker, N. Srinivasan, C.J. Thorpe, J. P. Overington, and C. M. Topham, unpublished results).

If many sequences that are homologous to the modeled protein are available, substitution tables can be used to assess the model. Overington et al.^{117,119} align the sequence of the model with all

similar proteins in a sequence data bank. The residue substitution pattern derived from this alignment is viewed in light of the structural environment at every residue position in the model, and this substitution pattern is compared with the one derived from the alignment of known homologous tertiary structures. The errors in local regions in a correctly folded structure are more difficult to recognize.³⁰¹

Bond lengths and bond angles need to be precise in order to make reliable use of a protein model in studies of molecular recognition. Morris et al.⁵³ have analyzed the (ϕ, ψ) values, peptide planarity, bond lengths and bond angles, hydrogen-bond geometry, and side-chain conformations of known protein structures as a function of the atomic resolution. From these data they arrived at expected values of these parameters depending on the resolution (for X-ray structures) of the 3-D structure. For example, with a model whose accuracy can be compared with a typical 2 Å resolution crystal structure, more than 80% of the (ϕ, ψ) values are expected to lie within the allowed regions in the Ramachandran map. The program PROCHECK⁵⁴ gives graphic and quantitative information about various structural features and assessment of the model.

While the compatibility of a sequence to the modeled fold may be judged highly by each of the model evaluation procedures, local inaccuracies are difficult to identify. A positive health-check by one method does not imply that the model is fully correct, although by using many methods working on various different principles it may be possible to uncover a large number of the inaccuracies in a model. It is, however, ironic that the final evaluation of any model can only be made when a structure analysis has itself been completed.

V. ACCURACY OF MODELS DERIVED FROM COMPARATIVE MODELING

Retrospective analyses of a number of three-dimensional models built using COMPOSER (Figure 28) have been reported (References 44, 186, 301; Guruprasad and Blundell, unpublished results). In Table 5 are shown comparisons between modeled structures and their corresponding crys-

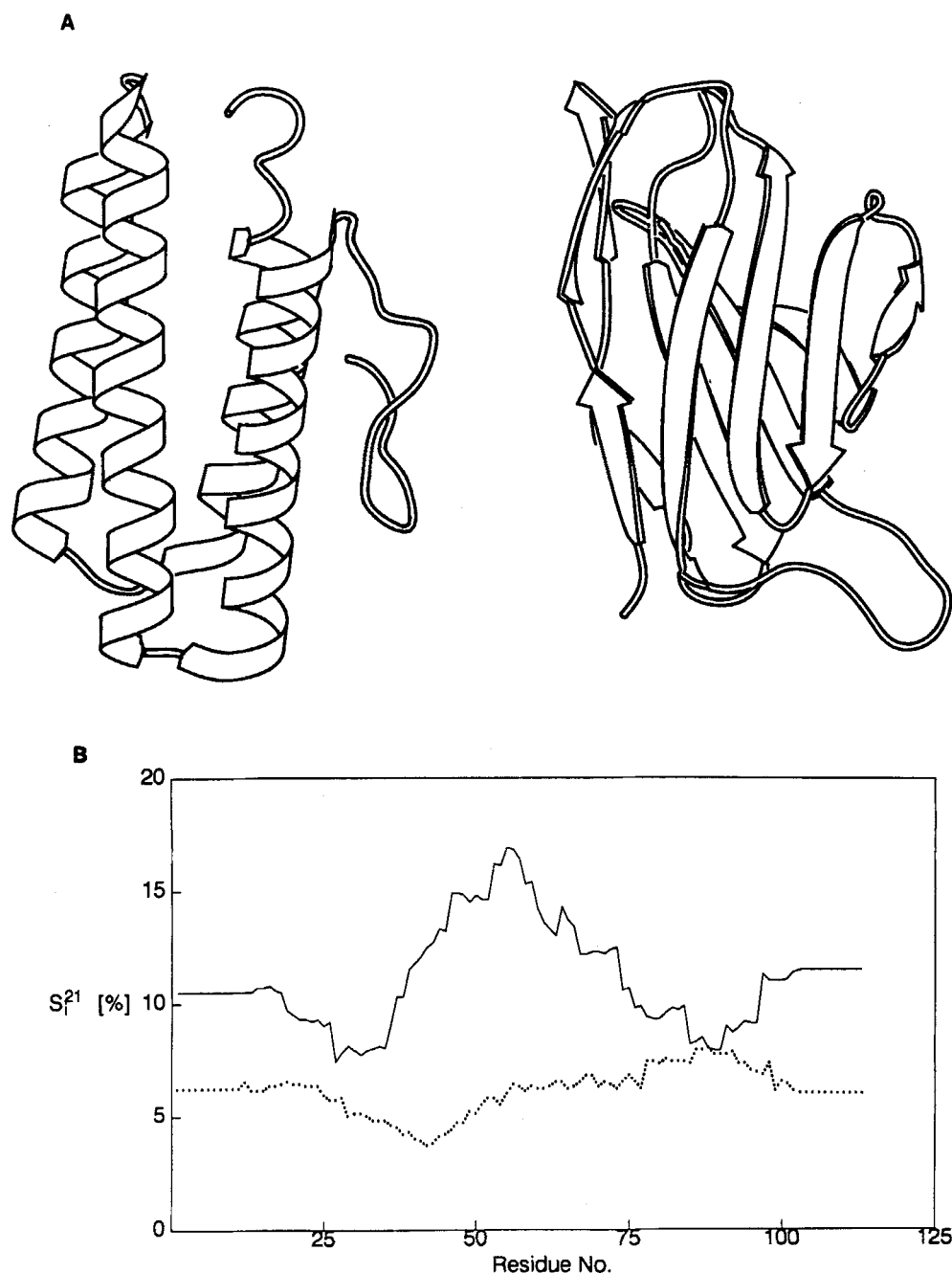


FIGURE 27. Identification of an incorrect fold. The sequence of an immunoglobulin variable domain, which is predominantly β -sheet (top right), is threaded onto the fold of hemerythrin, which is a four-helical bundle (top left). The 3-D profile scores for the correct (solid line) and incorrect (dotted line) folds calculated using a 21-residue moving window are shown at the bottom.

tal structures (both for the framework regions and over all C^{α} -atoms),¹⁸⁶ as well as the expected error in the crystal structures based on their resolution.³⁰² As indicated by the smaller errors seen

between the model and its X-ray structure over the framework region, the models resemble the X-ray structure of the modeled protein more than the X-ray structure of the closest homologue.

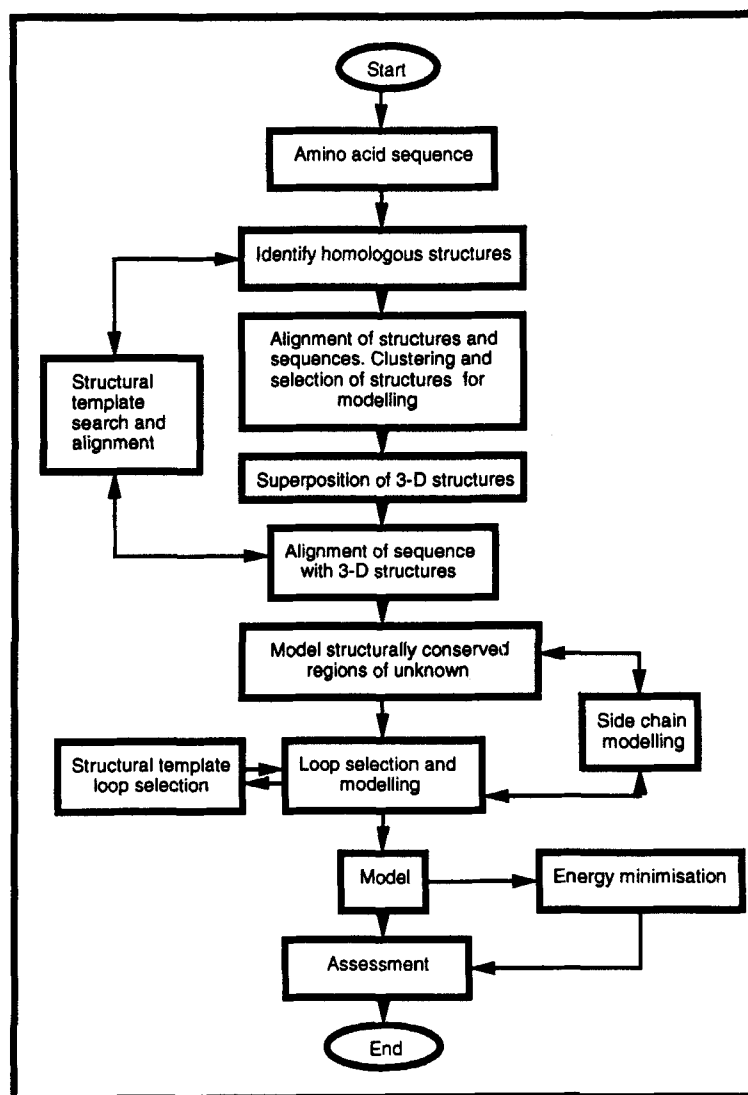


FIGURE 28. A flow-chart for the comparative modeling procedure, COMPOSER. Various steps in COMPOSER and the tools used are indicated.

The RMSD (all C^{α} -atoms) for superpositions between the modeled and X-ray structures lies in the range of 0.7 to 1.7 Å for most of the models. The errors are generally greatest in the loop regions, particularly if they are built on an unrelated protein, but these errors can be minimized when it is possible to extend the structurally conserved regions of one of the known homologues on the basis of a loop of similar sequence but different length (Figure 21). This has resulted in more accurate models for the variable regions in myoglobin,¹⁸⁶ subtilisin Carlsberg¹⁸⁶ and in cathepsin D.³⁰³

The error in the core region can be reduced by generating a framework in which contributions from the known homologues are weighted on the basis of their sequence similarity with the unknown.¹⁸⁶ This is particularly important if the percentage sequence identities of the unknown with the known structures vary greatly. For example, the azurin from *P. aeruginosa* was modeled on the azurin from *A. denitrificans*, pseudo-azurin from *A. faecalis*, and the plastocyanins from *Populus Algra* and *E. prolifara*. The percentage sequence identities of the modeled protein

TABLE 5
Known Structures Built Using COMPOSER, Their Accuracy and Comparison of Their X-ray Structure with Homologues

PDB code of the protein modeled	Closest homologue (A)		Homologue with best resolution		Homologue with least resolution		RMSD (Å) between X-ray str. of unknown and A (equivalences)	RMSD (Å) between the model and X-ray str. (equivalences)	
	Code	% seq.sim.	Code	Exp. error (Å)	Code	Exp. error (Å)		Framework	Whole chain
5MBN	2HHB	25	1ECA	0.18	2LHB	0.40	1.37 (138)	0.83 (107)	0.99 (146)
1MPP	2APR	30	2APR	0.33	1YPA	0.96	1.85 (282)	1.71 (282)	3.47 (354)
1DTX	1AAP	36	5PT1	0.03	1AAP	0.21	0.91 (56)	0.72 (54)	0.85 (56)
2PTN	4CHA	44	3EST	0.27	3RP2	0.36	1.52 (219)	0.84 (177)	1.15 (223)
1BBC	1PPA	47	1BP2	0.29	1PP2	0.59	1.30 (118)	1.10 (94)	1.56 (124)
1LZT	1LZ1	60	1LZ1	0.21	1ALC	0.29	0.76 (129)	0.65 (115)	0.73 (129)
1AZU	2AZA	62	1PAZ	0.23	2AZA	0.55	0.88 (123)	0.68 (50)	0.90 (126)
1SBC	2SBT	70	2PRK	0.21	1SBT	0.59	0.73 (272)	0.67 (213)	1.07 (274)

From Srinivasan, N. and Blundell, T. L., *Protein Eng.*, 6, 501, 1993. With permission.

with these homologues is 62, 12, 20, and 30%, respectively. A model constructed from equal contributions from these homologues differs from the true structure over the framework region by an RMSD of 1.43 Å. The error is reduced to 0.68 Å if the homologues are weighted using the square of their sequence identity to the unknown.

Frazao et al.⁴⁴ have analyzed a model of human renin constructed using the known structures of three fungal proteinase (rhizopuspepsin, endothiapepsin, and penicillopepsin) and two mammalian enzymes (porcine pepsin and calf chymosin) before the crystal structures of human and mouse renins³⁰⁴ were determined. A comparison of the X-ray and modeled human renin structures shows that the 280 C α -atoms forming the framework have a RMSD of 0.84 Å, which is smaller than the RMSD between the X-ray structure of human renin and the structures used in the model building with COMPOSER. Some of the errors, for example, from the flap that covers the active site, arise from the fact that the model was constructed on the basis of the unliganded forms of pepsin and chymosin, while the crystal structure of renin has a bound inhibitor. Errors in the relative orientation of both domains in the renin model are difficult to avoid in view of the rigid body shifts known to occur with the aspartic proteinases.^{305–308} Although the catalytic residues are well modeled, there are some problems in modeling the specificity pockets as a result of alterations to well-defined loops in the other proteinases. Other features of renin, such as the

conformation of a large loop containing several prolines including two in the *cis*-conformation, were not modeled correctly. However, an increase in the number of different structures for a family increases the accuracy of the models. For example, consider the modeling of cathepsin D³⁰³ in which a similar prolyl-rich loop was modeled on that of renin. Such problems, due to unique features in the unknown (often at loops), were also noted by Read et al.,²⁷ Remington et al.,³⁰⁹ Frommel et al.,³¹⁰ and Weber³¹¹ in the context of the comparison of the modeled and experimental structures of trypsin from *Streptomyces griseus*, rat mast cell protease, thermolysin, and HIV protease, respectively. In the case of the HIV protease, the substrate binding site is modeled accurately and could be used to design inhibitors.³¹¹

Eighteen models constructed over the years with the procedure COMPOSER (Figure 28) are shown in Figure 29.

VI. SUMMARY COMMENTS

Although knowledge-based modeling procedures were first developed by protein crystallographers with an interest in homologues of the proteins solved in their laboratories, the wealth of sequences available, together with the interest in protein structure as a basis for design, has created interest in the technique in many laboratories. The wide availability of the molecular graphics workstations, with user-friendly interfaces, has

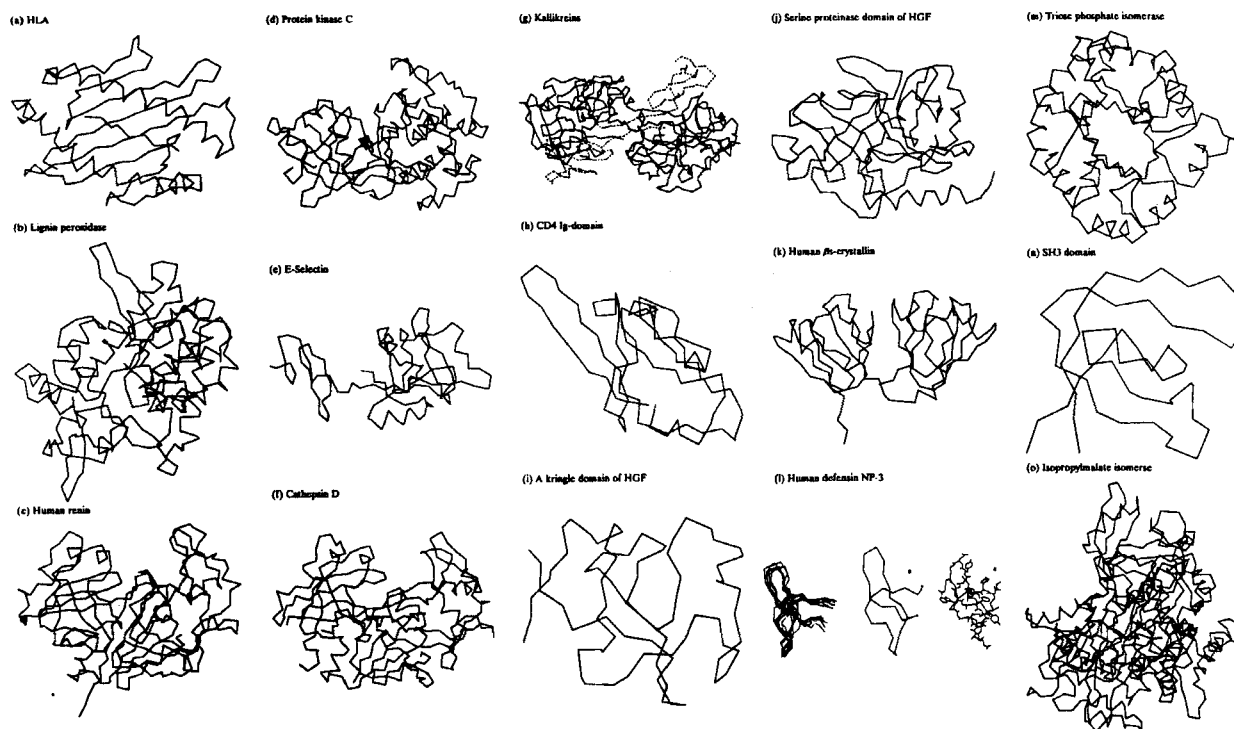


FIGURE 29. A survey of models built with COMPOSER. (a) Model of HLA-B53 built using the known structures of HLA-A2, HLA-Aw68, and HLA-B27 (C. J. Thorpe, D. S. Moss, and P. J. Travers, unpublished results). The bound molecule is a malarial peptide and shown in the plane of the β -sheet; (b) lignin peroxidase LIII constructed on the basis of its similarity with cytochrome c peroxidase;^{120,121} (c) human renin model⁴⁴ constructed using the known structures of rhizopuspepsin, endothiapepsin, penicillopepsin, porcine pepsin, and calf chymosin; (d) protein kinase C built using the crystal structure of cyclic-AMP dependent protein kinase (N. Srinivasan, P. Parker, and B. Bax, unpublished results); (e) the carbohydrate recognition domain of human E-selectin modeled on rat mannose binding protein;³⁷⁴ (f) the model of cathepsin D³⁰³ generated using the known structures of pepsin and chymosin and remodeled using the known structure of renin; (g) hypothetical structure of the high-molecular-weight complex of epidermal growth factor (EGF, dashed) with its binding protein (EGF-BP, solid).³⁷⁵ EGF-BP, a glandular kallikrein, was modeled on porcine pancreatic kallikrein and rat tonin. The two EGF chains in the complex are NMR structures.^{381,382} (h) Ig-domain of CDC4 (D. Carney, unpublished results); (i) model of one of the four kringle-domains in human hepatocyte growth factor (HGF) built using the three known kringle structures (L. E. Donate, N. Srinivasan, R. Sowdhamini, T. L. Blundell, and E. Gherardi, unpublished results); (j) serine proteinase domain of human HGF constructed from the structures of elastase, trypsin, and rat mast cell proteinase (L. E. Donate, N. Srinivasan, R. Sowdhamini, T. L. Blundell, and E. Gherardi, unpublished results); (k) the model of β s-crystallin (now designated γ s-crystallin) generated using the coordinates of γ ll-crystallin and β b2-crystallin (N. Srinivasan, S. Zarina, and C. Slingsby, private communication); (l) human defensin:¹⁶⁸ left, representative structures generated for rabbit defensin NP-5 from published 2D-NMR distance constraints³⁸³ with the program DISGEO,³⁸⁴ middle, main-chain C^α trace for the human model derived from the NMR structures using COMPOSER; right, all-atom model for human defensin; (m) Triose phosphate isomerase (N. Srinivasan, P. Balaram, and H. Balaram, unpublished results); (n) SH3-domain, (R. Guillory and B. Bax, unpublished results); (o) Isopropylmalate isomerase (A. May, M. S. Johnson, and R. Viner, unpublished results).

further contributed to the popularity of the approach.

In this review we have described progress from a partly subjective, mainly interactive exer-

cise carried out by experts to a more automated and rule-based approach, which exploits our extensive knowledge of protein structure and interactions. We have emphasized the importance of

correct sequence alignment of the protein to be modeled and its homologues. Incorrect alignment is the source of the largest errors. We have also shown how knowledge of the structures of proteins in general can be used to assist in the construction of useful models.

Such models have a value that is operationally defined. In some cases a model that defines the approximate disposition of groups relative to a ligand-binding site may be of value in suggesting a mechanism or an explanation for specificity. Approximate models may also be useful in the early stages of novel ligand design. However, for many purposes, especially in drug design, a more precise model will be required in order to assess the likely effects of small changes in the chemistry of the ligand or the sequence of a protein. Precise models can now be constructed in favorable circumstances, most importantly the existence of homologues of known structure. Most proteins, of interest in the design of drugs, pesticides and vaccines, will be present in very low copy numbers in the organism. This means that quantities of the protein will be not directly available for experimental studies and that the sequence will be defined from cloning and sequencing the cDNA. This will provide a route for expression of the protein for crystallization or solution NMR studies. However, there will always be a period in which a sequence, but not a three-dimensional structure, is available for a protein of interest. This may be several weeks, months, or even years for a large, flexible, or unstable protein. It is during this period that a model will be of greatest value.

Protein models have played an important role in many design processes. In the design of inhibitors, for example, of HIV proteinase for AIDS antivirals or human renin for antihypertensives, models were exploited for several years before X-ray crystal structures of target enzyme complexes were available. For HIV proteinase the relationship with aspartic proteinases of known structure was distant, and the models were a very rough guide. For renin the models were quite precise and allowed useful elaboration of lead compounds. This included suggestions of cyclization to decrease flexibility, of decreases in size to

improve oral availability, of removal of peptide bonds to decrease proteolysis, of addition of groups to improve lipid solubility, and of modification of groups to improve the specificity of ligand binding.

Models are also useful in protein engineering. They suggest sites where mutations might be introduced effectively. They have provided a guide to the construction of useful chimeric molecules where parts of one protein are grafted onto another. This has been particularly effective in the production of humanized antibodies or chimeric growth factors, for example, hybrid neurotrophic factors comprised of fragments of nerve growth factor, brain-derived neurotrophic factor (BDNF), and neurotrophic factor 3 (NT3).

The increasing interest on the part of medicinal chemists, molecular biologists, pharmacologists, plant biotechnologists, and many others who have little direct knowledge of protein structure is underlining the importance of procedures that use, in an automatic way, rules about and knowledge of protein structure. Such approaches are certain to play an important role, alongside experimental techniques, in protein modeling and design.

ACKNOWLEDGMENTS

We would like to thank our colleagues who have kindly provided coordinates of their models: K. Guruprasad (cathepsin D), Christopher Thorpe (HLA), Alan Mills (E-selectin), Ben Bax (protein kinase C and kallikrein), Devon Carney (CDC4), Ansku Hoffrén (lignin peroxidase), V. Dhanaraj (human renin), Dan Donnelly (β_2 -adrenergic receptor), Alex May (isopropylmalate isomerase), S. Zarina and Christine Slingsby (γ s-crystallin), P. Balaram and H. Balaram (triose phosphate isomerase), Luis E. Donate (domains of HGF), and Richard Guillory (SH3-domain), and figures: Andrej Šali, Lynn Sibanda, Dan Donnelly, V. Dhanaraj, and Alex May. We thank John Overington for the use of his display program (JDRAW) used to display C^α -traces in many of the figures. Support for this work has been provided by the Imperial Cancer Research Fund and Tripos Associates. We would also like to extend our thanks to those experimental scientists who

are responsible for the freely accessible structures and sequences that are literally a gold mine for the scientific community.

REFERENCES

- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M., Protein data bank: a computer based archival file for macromolecular structures, *J. Mol. Biol.*, 112, 535, 1977.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., and Weng, J., Protein data bank, in *Crystallographic Databases — Information Content, Software Systems, Scientific Applications*, Allen, F. H., Bergerhoff, G., and Sievers, R., Eds., Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987, 107.
- Louie, B. V., Brownlie, P. D., Lambert, R., Cooper, J. B., Blundell, T. L., Wood, S. P., Warren, M. J., Woodcock, S. C., and Jordan, P. M., Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site, *Nature*, 359, 33, 1992.
- Louie, G. V., Porphobilinogen deaminase and its structural similarity to the bidomain binding proteins, *Curr. Opin. Struct. Biol.*, 3, 401, 1993.
- Dayhoff, M. O., Barker, W. C., and Hunt, L. T., Establishing homologies in protein sequences, *Methods Enzymol.*, 91, 524, 1983.
- Chothia, C., One thousand families for the molecular biologist, *Nature*, 357, 543, 1992.
- Blundell, T. L. and Johnson, M. S., Catching the common fold, *Protein Science*, 2, 877, 1993.
- Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M., Knowledge-based prediction of protein structure and the design of novel molecules, *Nature*, 326, 347, 1987.
- Šali, A., Overington, J. P., Johnson, M. S., and Blundell, T. L., From comparisons of protein sequences and structures to protein modeling and design, *Trends Biochem. Sci.*, 15, 235, 1990.
- Browne, W. J., North, A. C. T., Phillips, D. C., Drew, K., Vanaman, T. C., and Hill, R. L., A possible three-dimensional structure of bovine alpha-lactalbumin based on lysozyme, *J. Mol. Biol.*, 42, 65, 1969.
- Warne, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W., and Sheraga, H. A., Computation of structures of homologous proteins. α -Lactalbumin from lysozyme, *Biochemistry*, 13, 768, 1974.
- Acharya, K. R., Stuart, D. I., Walker, N. P. C., Lewis, M., and Phillips, D. C., Refined structure of baboon α -lactalbumin at 1.7 Å resolution. Comparison with C-type lysozyme, *J. Mol. Biol.*, 208, 99, 1989.
- Hartley, B. S., Homologies in serine proteinases, *Phil. Trans. R. Soc. Lond.*, B257, 77, 1970.
- McLachlan, A. D. and Shotton, D. M., Structural similarities between A-lytic protease of *Myxobacter* 495 and elastase, *Nature New Biol. (London)*, 229, 202, 1971.
- Brayer, G. D., Delbaere, L. J. T., and James, M. N. G., Molecular structure of the α -lytic protease from *Myxobacter* 495 at 2.8 angstroms resolution, *J. Mol. Biol.*, 131, 743, 1979.
- Delbaere, L. T. J., Brayer, G. D., and James, M. N. G., Comparison of the predicted model of α -lytic protease with the X-ray structure, *Nature*, 279, 165, 1979.
- Blundell, T. L., Dodson, G. G., Hodgkin, D. C., and Mercola, D. A., Insulin: the structure in the crystal and its reflection in chemistry and biology, *Adv. Prot. Chem.*, 26, 279, 1972.
- Blundell, T. L. and Horuk, R., Monomeric insulin from the casiragua: molecular model building using computer graphics, *Hoppe-Seyler's Z. Physiol. Chem.*, 362, 727, 1981.
- Blundell, T. L., Bedarkar, S., Rinderknecht, E., and Humbel, R. E., Insulin-like growth factor: a model for tertiary structure accounting for immunoreactivity and receptor binding, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 180, 1978.
- Cooke, R. M., Harvey, T. S., and Campbell, I. D., Solution structure of IGF-like growth factor I. A NMR and restrained MD study, *Biochemistry*, 30, 5484, 1991.
- Bedarkar, B., Turnell, W. G., Schwabe, C., and Blundell, T. L., Relaxin has conformational homology with insulin, *Nature*, 270, 449, 1977.
- Isaacs, N., James, R., Niall, H., Bryant-Greenwood, G., Dodson, G., Evans, A., and North, A. C. T., Relaxin and its structural relationship to insulin, *Nature*, 271, 278, 1978.
- Eigenbrot, C., Randal, M., Quan, C., Burnier, J., O'Connell, L., Rinderknecht, E., and Kossiakoff, A. A., X-ray structure of human relaxin at 1.5 Å. Comparison to insulin and implications for receptor binding determinants, *J. Mol. Biol.*, 221, 15, 1991.
- Greer, J., Model for the haptoglobin heavy chain based upon structural homology, *Proc. Natl. Acad. Sci. U.S.A.*, 77, 3393, 1980.
- Greer, J., Comparative model building of mammalian serine proteinases, *J. Mol. Biol.*, 153, 1027, 1981.
- Greer, J., Model of a specific interaction. Salt bridges between prothrombin and its activating enzyme blood clotting factor X_a, *J. Mol. Biol.*, 153, 1043, 1981.
- Read, R. J., Brayer, G. D., Jurášek, L., and James, M. N. G., Critical evaluation of comparative model building of *Streptomyces griseus* trypsin, *Biochemistry*, 23, 6570, 1984.
- Blundell, T. L., Lindley, P., Miller, L., Moss, D., Slingsby, C., Tickle, I., Turnell, B., and Wistow, G., The molecular structure and stability of the eye

- lens: X-ray analysis of γ -crystallin II, *Nature*, 289, 771, 1981.
29. Wistow, G., Turnell, B., Summers, L., Slingsby, C., Moss, D., Miller, L., Lindley, P., and Blundell, T., X-ray analysis of the eye lens protein γ -II crystallin at 1.9 Å resolution, *J. Mol. Biol.*, 170, 175, 1983.
 30. Wistow, G., Slingsby, C., Blundell, T., Driessen, H., Dejong, W., and Bloemendal, H., Eye-lens proteins — 3-dimensional structure of β -crystallin predicted from monomeric γ -crystallin, *FEBS Letts.*, 133, 9, 1981.
 31. Wistow, G., Summers, L., and Blundell, T., *Myxococcus xanthus* spore coat protein-S may have a similar structure to vertebrate lens β - γ -crystallins, *Nature*, 315, 771, 1985.
 32. Summers, L., Wistow, G., Narebor, M., Moss, D. S., Lindley, P., Slingsby, C., Blundell, T., Bartunik, H., and Bartels, K., X-ray studies of the lens specific proteins: the crystallins, *Pept. Prot. Rev.*, 3, 147, 1984.
 33. White, H. E., Driessen, H. P. C., Slingsby, C., Moss, D. S., and Lindley, P. F., Packing interactions in the eye lens: structural analysis internal symmetry and lattice interactions of bovine γ -IVa-crystallin, *J. Mol. Biol.*, 207, 217, 1989.
 34. Bax, B., Lapatto, R., Nalini, V., Driessen, H., Lindley, P. F., Mahadevan, D., Blundell, T. L., and Slingsby, C., X-ray analysis of β -B2-crystallin and evolution of oligomeric lens proteins, *Nature*, 347, 776, 1990.
 35. Hutchins, C. and Greer, J., Comparative modeling of proteins in the design of novel renin inhibitors, *Crit. Rev. Biochem. Mol. Biol.*, 26, 77, 1991.
 36. Blundell, T. L., Sibanda, B. L., and Pearl, L., The three-dimensional structure, specificity and catalytic mechanism of renin, *Nature*, 304, 273, 1983.
 37. Sibanda, B. L., Blundell, T., Hobart, P. M., Fogliand, M., Bindra, J. S., Dominy, B. W., and Chirgwin, J. M., Computer-graphics modeling of human renin: specificity, catalytic activity and intron-exon junctions, *FEBS Letts.*, 174, 102, 1984.
 38. Carlson, W., Hanschumacher, M., Karplus, M., and Haber, E., Studies of the three-dimensional structure of human renin and its inhibitors, *J. Hypertension*, 2, 281, 1984.
 39. Carlson, W., Haber, E., Feldmann, R., and Karplus, M., A model for the three-dimensional structure of renin, in *Proceedings of the Eight American Peptide Symposium*, Hruby, V. J. and Rich, D. J., Eds., Pierce Chemical Co., Rockford, IL, 1984, 821.
 40. Carlson, W., Karplus, M., and Haber, E., Construction of a model for the 3-dimensional structure of human renal renin, *Hypertension*, 7, 13, 1985.
 41. Akahane, K., Nakagawa, S., and Umeyama, H., Three-dimensional structure of human renin, *Hypertension*, 7, 3, 1985.
 42. Plattner, J. J., Greer, J., Fung, A. K., Stein, H., Kleinert, H. D., Sham, H. L., Smital, J. R., and Perun, T. J., Peptide analogues of angiotensinogen. Effect of peptide chain length on renin inhibition, *Biochem. Biophys. Res. Commun.*, 139, 982, 1986.
 43. Sham, H. L., Bolis, G., Stein, H. H., Fesik, S. W., Marcotte, P. A., Plattner, J. J., Rempel, C. A., and Greer, J., Renin inhibitors. Design and synthesis of a new class of conformationally restricted analogues of angiotensinogen, *J. Med. Chem.*, 31, 284, 1988.
 44. Frazao, C., Topham, C., Dhanaraj, V., and Blundell, T. L., Comparative modeling of human renin. A retrospective evaluation of the model with respect to the X-ray crystal structure, 1993 (in press).
 45. Billeter, M., Kline, A. D., Braun, W., Huber, R., and Wüthrich, K., Comparison of the high-resolution structures of α -amylase inhibitor tendamistat determined by nuclear magnetic resonance in solution and by X-ray diffraction in single crystals, *J. Mol. Biol.*, 206, 677, 1989.
 46. Kline, A. D., Braun, W., and Wüthrich, K., Determination of the complete three-dimensional structure of the α -amylase inhibitor tendamistat in aqueous solution by nuclear magnetic resonance and distance geometry, *J. Mol. Biol.*, 204, 675, 1988.
 47. Islam, S. A. and Sternberg, M. J. E., Bad contacts in protein structures, in *Accuracy and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, Henrick, K., Moss, D. S., and Tickle, I. J., Eds., 1990, 53.
 48. Phillips, S. E. V., Somers, W. S., Bhat, T. N., and Parsons, M. R., Structure determination of turkey egg lysozyme, in *Accuracy and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, Henrick, K., Moss, D. S., and Tickle, I. J., Eds., 1990, 63.
 49. Schreuder, H. A., Curmi, P. M. G., Cascio, D., and Eisenberg, D., The RuBisCO saga, in *Accuracy and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, Henrick, K., Moss, D. S., and Tickle, I. J., Eds., 1990, 73.
 50. Stout, C. D., Hallmarks of a wrong structure, in *Accuracy and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, Henrick, K., Moss, D. S., and Tickle, I. J., Eds., 1990, 91.
 51. Adman, E. T., A tale of four iron-sulfur proteins: sequence errors and other matters, in *Accuracy and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, Henrick, K., Moss, D. S., and Tickle, I. J., Eds., 1990, 97.
 52. Thornton, J. M., McArthur, M. W., Smith, D. K., Gardner, S. P., Hutchinson, E. G., Morris, A. L., and Sibanda, B. L., Analysis of errors found in protein structure coordinates in the Brookhaven data bank, in *Accuracy and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, Henrick, K., Moss, D. S., and Tickle, I. J., Eds., 1990, 39.

53. Morris, A. L., MacArthur, M. W., Hutchinson, E. G., and Thornton, J. M., Stereochemical quality of protein structure coordinates, *Proteins*, 12, 345, 1992.
54. Laskowski, P. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M., PROCHECK — a program to check the stereochemical quality of protein structures, *J. Appl. Cryst.*, 26, 283, 1993.
55. Sander, C. and Schneider, R., Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins*, 9, 56, 1991.
56. Pascarella, S. and Argos, P., A data bank merging related protein structures and sequences, *Protein Eng.*, 5, 121, 1992.
57. Doolittle, R. F., *Of URFs and ORFs. A Primer on How to Analyze Derived Amino Acid Sequences*, University Science Books, Mill Valley, CA, 1986.
58. Lesk, A. M., Ed., *Computational Molecular Biology*, Oxford University Press, Oxford, 1988.
59. Doolittle, R. F., Ed., *Molecular evolution: computer analysis of protein and nucleic acid sequences*, *Methods Enzymol.*, 183, 1990.
60. Bleasby, A. J. and Wooton, J. C., Construction of validated, non-redundant composite protein sequence databases, *Protein Eng.*, 3, 153, 1990.
61. Lipman, D. J. and Pearson, W. R., Rapid and sensitive protein similarity searches, *Science*, 227, 1435, 1985.
62. Altschul, S. F. and Lipman, D. J., Protein database searches for multiple alignments, *Proc. Natl. Acad. Sci. U.S.A.*, 87, 5509, 1990.
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., Basic local alignment search tool, *J. Mol. Biol.*, 215, 403, 1990.
64. Etzold, T. and Argos, P., SRS — an indexing and retrieval tool for flat file data libraries, *Comp. Appl. Biosci.*, 9, 49, 1993.
65. Brutlag, D. L., Dautricourt, J. P., Maulik, S., and Relp, J., Improved sensitivity of biological sequence database searches, *Comp. Appl. Biosci.*, 6, 237, 1990.
66. Bairoch, A., PROSITE — a dictionary of sites and patterns in proteins, *Nucl. Acids Res.*, 19, 2241, 1991.
67. Needleman, S. B. and Wunsch, C., A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, 48, 444, 1970.
68. Sellers, P. H., On the theory and computation of evolutionary distances, *SIAM J. Appl. Math.*, 26, 787, 1974.
69. Gotoh, O., An improved method for matching biological sequences, *J. Mol. Biol.*, 162, 705, 1982.
70. Smith, T. F., Waterman, M. S., and Fitch, W. F., Comparative biosequence metrics, *J. Mol. Evol.*, 18, 38, 1981.
71. Fredman, M. L., Computing evolutionary similarity measures with length independent gap penalties, *Bull. Math. Biol.*, 46, 553, 1984.
72. Smith, T. F. and Waterman, M. S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147, 195, 1981.
73. Karlin, S. and Altschul, S. F., Applications and statistics for multiple high scoring segments in molecular sequences, *Proc. Natl. Acad. Sci. U.S.A.*, 90, 5873, 1993.
74. Fitch, W. M., An improved method of testing for evolutionary homology, *J. Mol. Biol.*, 16, 9, 1966.
75. Taylor, W. R., Identification of protein sequence homology by consensus sequence alignment, *J. Mol. Biol.*, 188, 233, 1986.
76. Patthy, L., Detecting homology of distantly related proteins with consensus sequences, *J. Mol. Biol.*, 198, 567, 1987.
77. Bork, P., Recognition of functional regions in primary structures using a set of property patterns, *FEBS Letts.*, 257, 191, 1989.
78. Barton, G. J. and Sternberg, M. J. E., A sensitive method to detect weak structural similarities, *J. Mol. Biol.*, 212, 389, 1990.
79. Johnson, M. S. and Overington, J. P., A structural basis for sequence comparisons: an evaluation of scoring methodologies, *J. Mol. Biol.*, 233, 716, 1993.
80. Johnson, M. S., Overington, J. P., and Blundell, T. L., Alignment and searching for common protein folds using a data bank of structural templates, *J. Mol. Biol.*, 231, 735, 1993.
81. McLachlan, A. D., Tests for comparing related amino acid sequences. Cytochrome c and cytochrome c551, *J. Mol. Biol.*, 6, 409, 1971.
82. Feng, D.-F., Johnson, M. S., and Doolittle, R. F., Aligning amino acid sequences: comparison of commonly used methods, *J. Mol. Evol.*, 21, 112, 1985.
83. Grantham, R., Amino acid difference formula to help explain protein evolution, *Science*, 185, 862, 1974.
84. Rao, J. K. M., New scoring matrix for amino acid exchanges based on residue characteristic physical parameters, *Int. J. Pept. Protein Res.*, 29, 276, 1987.
85. Miyata, T., Miyazawa, S., and Yasunaga, T., Two types of amino acid substitutions in protein evolution, *J. Mol. Evol.*, 12, 219, 1979.
86. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C., *Atlas of Protein Sequence and Structure*, Vol. 5, Dayhoff, M. O., Ed., National Biomedical Research Foundation, Washington, D.C., 1978.
87. Gonnet, G. H., Cohen, M. A., and Benner, S. A., Exhaustive matching of the entire protein database, *Science*, 256, 1443, 1992.
88. Henikoff, S. and Henikoff, J. G., Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. U.S.A.*, 89, 10915, 1992.
89. Jones, D. T., Taylor, W. R., and Thornton, J. M., The rapid generation of mutation matrices, *CABIOS*, 8, 275, 1992.
90. Risler, J. L., Delormo, M. O., Delacroix, H., and Henaut, A., Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix, *J. Mol. Biol.*, 204, 1019, 1988.

91. Jue, R. A., Woodbury, N. W., and Doolittle, R. F., Sequence homologies among *E. coli* ribosomal proteins: evidence for evolutionary related groupings and internal duplications, *J. Mol. Evol.*, 15, 129, 1980.
92. Johnson, M. S. and Doolittle, R. F., A method for the simultaneous alignment of three or more amino acid sequences, *J. Mol. Evol.*, 23, 267, 1986.
93. Barton, G. J. and Sternberg, M. J. E., Evaluation and improvement in the automatic alignment of protein sequences, *Protein Eng.*, 1, 89, 1987.
94. Murata, M., Richardson, J. S., and Sussman, J. L., Simultaneous comparison of three protein sequences, *Proc. Natl. Acad. Sci. U.S.A.*, 82, 3073, 1985.
95. Lipman, D. J., Altschul, S. F., and Kececioglu, J. D., A tool for multiple sequence alignment, *Proc. Natl. Acad. Sci. U.S.A.*, 86, 4412, 1989.
96. Feng, D.-F. and Doolittle, R. F., Progressive sequence alignment as a prerequisite to correct phylogenetic trees, *J. Mol. Evol.*, 25, 351, 1987.
97. Barton, G. J. and Sternberg, M. J. E., A strategy for the rapid multiple alignment of proteins, *J. Mol. Biol.*, 198, 329, 1987.
98. Higgins, D. G. and Sharp, P. M., CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, *Gene*, 73, 237, 1988.
99. Higgins, D. G., Bleasby, A. J., and Fuchs, R., CLUSTAL-V — improved software for multiple sequence alignment, *Comp. Appl. Biosci.*, 8, 189, 1992.
100. Gribskov, M., McLachlan, A. D., and Eisenberg, D., Profile analysis detection of distantly related proteins, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 4355, 1987.
101. Smith, R. F. and Smith, T. F., Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative modelling, *Protein Eng.*, 5, 35, 1992.
102. Barton, G. J. and Sternberg, M. J. E., Evaluations and improvements in the automatic alignment of protein sequences, *Protein Eng.*, 1, 89, 1987.
103. Gotoh, O., Consistency of optimal sequence alignments, *Bull. Math. Biol.*, 52, 509, 1990.
104. Vingron, M. and Argos, P., Determination of reliable regions in protein sequence alignments, *Protein Eng.*, 3, 565, 1990.
105. Saqi, M. A. and Sternberg, M. J. E., A simple method to generate non-trivial alternate alignments of protein sequences, *J. Mol. Biol.*, 219, 727, 1991.
106. Zuker, M., Suboptimal sequence alignment in molecular biology — alignment with error analysis, *J. Mol. Biol.*, 221, 403, 1991.
107. Saqi, M. A. S., Bates, P. A., and Sternberg, M. J. E., Toward an automatic method of predicting protein structure by homology: an evaluation of suboptimal sequence alignments, *Protein Eng.*, 5, 305, 1992.
108. Ponder, J. W. and Richards, F. M., Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes, *J. Mol. Biol.*, 193, 775, 1987.
109. Sippl, M. J., Calculation of conformational ensembles from potentials of mean force, *J. Mol. Biol.*, 213, 859, 1990.
110. Jones, D. T., Taylor, W. R., and Thornton, J. M., A new approach to protein fold recognition, *Nature*, 358, 86, 1992.
111. Maiorov, N. V. and Crippen, G. M., Contact potential that recognizes the correct folding of globular proteins, *J. Mol. Biol.*, 227, 876, 1992.
112. Godzik, A., Kolinski, A., and Skolnick, J., Topology fingerprint approach to the inverse folding problem, *J. Mol. Biol.*, 227, 227, 1992.
113. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Caseri, G., and Sippl, M. J., Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force, *J. Mol. Biol.*, 216, 167, 1990.
114. Sippl, M. J. and Weitckus, S., Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations, *Proteins*, 13, 258, 1992.
115. Skolnick, J. and Kolinski, A., Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. 6-member, Greek key β -barrel proteins, *J. Mol. Biol.*, 212, 787, 1990.
116. Bowie, J. U., Lüthy, R., and Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structure, *Science*, 253, 164, 1991.
117. Overington, J. P., Johnson, M. S., Šali, A., and Blundell, T. L., Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction, *Proc. R. Soc. Lond.*, B241, 146, 1990.
118. Lüthy, R., McLachlan, A. D., and Eisenberg, D., Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities, *Proteins*, 10, 229, 1991.
119. Overington, J. P., Donnelly, D., Šali, A., Johnson, M. S., and Blundell, T. L., Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds, *Protein Sci.*, 1, 216, 1992.
120. Hoffrén, A. M., Saloheimo, M., Thomas, P., Overington, J., Johnson, M. S., and Blundell, T. L., Modeling the lignin peroxidase LIII of *Phlebia radiata* using a knowledge-based approach, *J. Chim. Phys.*, 88, 2659, 1991.
121. Hoffrén, A. M., Saloheimo, M., Thomas, P., Overington, J. P., Johnson, M. S., Knowles, J. K. C., and Blundell, T. L., Modeling of the lignin peroxidase LIII of *Phlebia radiata*: use of a sequence template generated from a 3-D structure, *Protein Eng.*, 6, 177, 1993.
122. Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., and Xuong, Ng. H., Chymotrypsinogen: 2.5 Å

- crystal structure, comparison with α -chymotrypsin, and implications for zymogen activation, *Biochemistry*, 9, 1997, 1970.
123. Huber, R., Epp, O., Steigemann, W., and Formanek, H., The atomic structure of erythrocrucorin in the light of the chemical sequence and its comparison with myoglobin, *Eur. J. Biochem.*, 19, 42, 1971.
124. McLachlan, A. D., A mathematical procedure for superimposing atomic coordinates of proteins, *Acta Cryst.*, A28, 656, 1972.
125. McLachlan, A. D., Gene duplication in the structural evolution of chymotrypsin, *J. Mol. Biol.*, 128, 49, 1979.
126. McLachlan, A. D., Rapid comparison of protein structures, *Acta Cryst.*, A38, 781, 1982.
127. Go, M., Correlation of DNA exonic regions with protein structural units in hemoglobin, *Nature*, 291, 90, 1981.
128. Go, M., Modular structural units, exons, and function in chicken lysozyme, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 1964, 1983.
129. Nishikawa, K. and Ooi, T., Comparison of homologous tertiary structures of proteins, *J. Theor. Biol.*, 43, 351, 1974.
130. Rossmann, M. G. and Liljas, A. P., Recognition of structural domains in globular proteins, *J. Mol. Biol.*, 85, 177, 1974.
131. Matthews, B. W. and Rossmann, M. G., Comparison of protein structures, *Methods Enzymol.*, 115, 397, 1985.
132. Rao, S. T. and Rossmann, M. G., Comparison of super-secondary structures in proteins, *J. Mol. Biol.*, 76, 241, 1973.
133. Eventoff, W. and Rossmann, M. G., The evolution of dehydrogenases and kinases, *Crit. Rev. Biochem.*, 3, 111, 1975.
134. Rossmann, M. G. and Argos, P., Exploring structural homology of proteins, *J. Mol. Biol.*, 105, 75, 1976.
135. Rossmann, M. G. and Argos, P., The taxonomy of protein structure, *J. Mol. Biol.*, 109, 99, 1977.
136. Remington, S. J. and Matthews, B. W., A general method to assess similarity of protein structures, with application to T4 bacteriophage lysozyme, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 2180, 1978.
137. Remington, S. J. and Matthews, B. W., A systematic approach to the comparison of protein structures, *J. Mol. Biol.*, 140, 77, 1980.
138. Taylor, W. R. and Orengo, C. A., Protein structure alignment, *J. Mol. Biol.*, 208, 1, 1989.
139. Weaver, L. H., Gruetter, M. G., Remington, S. J., Gray, T. M., Isaacs, N. W., and Matthews, B. W., Comparison of goose-type, chicken-type and phage-type lysozymes illustrates the changes that occur in both amino acid sequence and three-dimensional structures during evolution, *J. Mol. Evol.*, 21, 97, 1985.
140. Murthy, M. R. N., A fast method of comparing protein structure, *FEBS Letts.*, 168, 97, 1984.
141. Sippl, M. J., On the problem of comparing protein structures: development and applications of a new method for the assessment of structural similarity, *J. Mol. Biol.*, 156, 359, 1982.
142. Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L., Knowledge-based modeling of homologous proteins. I. Three-dimensional frameworks derived from simultaneous superposition of multiple structures, *Protein Eng.*, 1, 377, 1987.
143. Johnson, M. S., Sutcliffe, M. J., and Blundell, T. L., Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins, *J. Mol. Evol.*, 30, 43, 1990.
144. Johnson, M. S., Šali, A., and Blundell, T. L., Phylogenetic relationships from three-dimensional protein structures, *Methods Enzymol.*, 83, 670, 1990.
145. Barton, G. J. and Sternberg, M. J. E., LOPAL and SCAMP: techniques for the comparison and display of protein structures, *J. Mol. Graph.*, 6, 190, 1988.
146. Richards, F. M. and Kundrot, C. E., Identification of structural motifs from protein coordinate data: secondary structure and first level super-secondary structure, *Proteins*, 3, 71, 1988.
147. Vriend, G. and Sander, C., Detection of common three-dimensional substructures in proteins, *Proteins*, 11, 52, 1991.
148. Taylor, W. R. and Orengo, C. A., A holistic approach to protein structure alignment, *Protein Eng.*, 2, 505, 1989.
149. Orengo, C. A. and Taylor, W. R., A rapid method of protein structure alignment, *J. Theor. Biol.*, 147, 517, 1990.
150. Orengo, C. A., Brown, N. P., and Taylor, W. R., Fast structure alignment for protein databank searching, *Proteins*, 14, 139, 1992.
151. Karpen, M. E., de Haseth, P. L., and Neet, K. E., Comparing short protein substructures by a method based on backbone torsion angles, *Proteins*, 6, 155, 1989.
152. Zuker, M. and Somorjai, R. L., The alignment of protein structures in three dimensions, *Bull. Math. Biol.*, 51, 55, 1989.
153. Šali, A. and Blundell, T. L., The definition of topological equivalence in homologous and analogous structures: a procedure involving a comparison of local properties and relationships, *J. Mol. Biol.*, 212, 403, 1990.
154. Overington, J. P., Zhu, Z.-Y., Šali, A., Johnson, M. S., Sowdhamini, R., Louie, G. V., and Blundell, T. L., Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins, *Biochem. Soc. Trans.*, 21, 597, 1993.
155. Zhu, Z.-Y., Šali, A. and Blundell, T. L., A variable gap penalty function and feature weights for protein 3-D structure comparison, *Protein Eng.*, 5, 43, 1992.
156. Rose, J. and Eisenmenger, F., A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm, *J. Mol. Evol.*, 32, 340, 1991.

157. Subbarao, N. and Haneef, I., Defining topological equivalences in macromolecules, *Protein Eng.*, 4, 877, 1991.
158. Koch, I., Kaden, F., and Selbig, J., Analysis of protein sheet topologies by graph theoretical methods, *Proteins*, 12, 314, 1992.
159. Ullman, J. R., An algorithm for subgraph isomorphism, *J. Assoc. Comput. Mach.*, 23, 31, 1976.
160. Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willett, P., Use of techniques derived from graph theory to compare secondary structure motifs in proteins, *J. Mol. Biol.*, 212, 151, 1989.
161. Richardson, J. S., The anatomy and taxonomy of protein structure, *Adv. Prot. Chem.*, 34, 167, 1981.
162. Kabsch, W. and Sander, C., Dictionary of protein secondary structure. Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22, 2577, 1983.
163. Artymiuk, P. J., Rice, D. W., Mitchell, E. M., and Willett, P., Structural resemblance between the families of bacterial signal-transduction proteins and of G-proteins revealed by graph theoretical techniques, *Protein Eng.*, 4, 39, 1990.
164. Artymiuk, P. J., Grindley, H. M., Park, J. E., Rice, D. W., and Willett, P., 3-Dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase-A revealed by graph theoretical techniques, *FEBS Letts.*, 303, 48, 1992.
165. Grindley, H. M., Artymiuk, P. J., Rice, D. W., and Willett, P., Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm, *J. Mol. Biol.*, 229, 707, 1993.
166. Felsenstein, J., Phylogenies from molecular sequences: inference and reliability, *Annu. Rev. Genet.*, 22, 521, 1988.
167. Shearer, A. C. and Johnson, M. S., Confidence limits on the branching order of phylogenetic trees, *Protein Sci.*, 2, 1686, 1993.
168. Johnson, M. S., Overington, J., and Šali, A., Knowledge-based protein modeling: human plasma kallikrein and human neutrophil defensin, in *Current Research in Protein Chemistry: Techniques, Structure and Function*, Villafranca, J. J., Ed., Academic Press, San Diego, 1990, 567.
169. Swindells, M. B. and Thornton, J. M., Structure and prediction and modeling, *Curr. Opin. Biotechnol.*, 2, 512, 1991.
170. Greer, J., Comparative modeling of homologous proteins, *Methods Enzymol.*, 202, 239, 1991.
171. Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M., Knowledge-based protein modeling and design, *Eur. J. Biochem.*, 172, 513, 1988.
172. Greer, J., Comparative modeling methods — applications to the family of the mammalian serine proteinases, *Proteins*, 7, 317, 1990.
173. Jones, T. H. and Thirup, S., Using known substructures in protein model building and crystallography, *EMBO J.*, 5, 819, 1986.
174. Unger, R., Harel, D., Wherland, S., and Sussman, J. L., A 3-D building blocks approach to analyzing and predicting structure of proteins, *Proteins*, 5, 335, 1989.
175. Claessens, M., Cutsen, E. V., Lasters, I., and Wodak, S., Modeling the polypeptide backbone with "spare parts" from known protein structures, *Protein Eng.*, 4, 335, 1989.
176. Levitt, M., Accurate modelling of protein conformation by automatic segment matching, *J. Mol. Biol.*, 226, 507, 1992.
177. Bassolino-Klimas, D. and Bruccoleri, R. E., Application of a directed conformational search for generating 3-D coordinates for protein structures from α -carbon coordinates, *Proteins*, 14, 465, 1992.
178. Correa, P. E., The building of protein structures from α -carbon coordinates, *Proteins*, 7, 366, 1990.
179. Luo, Y., Jiang, X., Lai, L., Qu, C., Xu, X., and Tang, Y., Building protein backbones from C^α coordinates, *Protein Eng.*, 5, 147, 1992.
180. Reid, L. S. and Thornton, J. M., Rebuilding flavodoxin from C^α coordinates: a test study, *Proteins*, 5, 170, 1989.
181. Rey, A. and Skolnick, J., Efficient algorithm for the reconstruction of a protein backbone from the α -carbon coordinates, *J. Comp. Chem.*, 13, 443, 1992.
182. Holm, L. and Sander, C., Database algorithm for generating protein backbone and sidechain coordinates from C^α trace: application to model building and detection of coordinate errors, *J. Mol. Biol.*, 218, 183, 1991.
183. Holm, L. and Sander, C., Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology, *Proteins*, 14, 213, 1992.
184. Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L., Knowledge-based modeling of homologous proteins. II. Rules for the conformations of substituted sidechains, *Protein Eng.*, 1, 385, 1987.
185. Topham, C. M., McLeod, A., Eisenmenger, F., Overington, J. P., Johnson, M. S., and Blundell, T. L., Fragment ranking in modeling of protein structures: conformationally constrained environmental amino acid substitution tables, *J. Mol. Biol.*, 229, 194, 1993.
186. Srinivasan, N. and Blundell, T. L., An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure, *Protein Eng.*, 6, 501, 1993.
187. Bajaj, M. and Blundell, T. L., Evolution and the tertiary structure of proteins, *Ann. Rev. Biophys. Bioeng.*, 13, 453, 1984.
188. Leszczynski, J. F. and Rose, G. D., Loops in globular proteins — a novel category of secondary structure, *Science*, 234, 849, 1986.

189. **Sibanda, B. L. and Thornton, J. M.**, Conformation of beta-hairpins in protein structures: classification and diversity in homologous structures, *Methods Enzymol.*, 202, 59, 1991.
190. **Pascarella, S. and Argos, P.**, Analysis of insertions/deletions in protein structures, *J. Mol. Biol.*, 224, 461, 1992.
191. **Venkatachalam, C. M.**, Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, *Biopolymers*, 6, 1425, 1968.
192. **Kuntz, I. D.**, Protein folding, *J. Am. Chem. Soc.*, 94, 4009, 1972.
193. **Crawford, J. L., Lipscomb, W. N., and Schellman, C. G.**, The reverse turn as a polypeptide conformation in globular proteins, *Proc. Natl. Acad. Sci. U.S.A.*, 70, 538, 1973.
194. **Lewis, P. N., Momany, F. A., and Scheraga, H. A.**, Chain reversals in proteins, *Biochem. Biophys. Acta*, 393, 211, 1973.
195. **Wilmot, C. M. and Thornton, J. M.**, Analysis and prediction of the different types of β -turn in proteins, *J. Mol. Biol.*, 203, 221, 1988.
196. **Wilmot, C. M. and Thornton, J. M.**, β -Turns and their distortions: a proposed new nomenclature, *Protein Eng.*, 3, 479, 1990.
197. **Chou, P. Y. and Fasman, G. D.**, Prediction of protein conformation, *Biochemistry*, 13, 222, 1974.
198. **Chou, P. Y. and Fasman, G. D.**, β -Turns in proteins, *J. Mol. Biol.*, 115, 135, 1977.
199. **Smith, P. A. and Pease, L. G.**, Reverse turns in peptides and proteins, *Crit. Rev. Biochem.*, 8, 315, 1980.
200. **Rose, G. D., Gierasch, L. M., and Smith, J. A.**, Turns in peptides and proteins, *Adv. Prot. Chem.*, 37, 1, 1985.
201. **Milner-White, E. J. and Poet, R.**, Loops, bulges, turns and hairpins in proteins, *Trends Biochem. Sci.*, 12, 189, 1987.
202. **Sibanda, B. L. and Thornton, J. M.**, β -Hairpin families in globular proteins, *Nature*, 316, 170, 1985.
203. **Edwards, M. S., Sternberg, M. J. E., and Thornton, J. M.**, Structural and sequence patterns in the loops of β - α - β units, *Protein Eng.*, 1, 173, 1987.
204. **Thornton, J. M., Sibanda, B. L., Edwards, M. S., and Barlow, D. J.**, Analysis, design and modification of loop regions in proteins, *BioEssays*, 8, 63, 1988.
205. **Srinivasan, N., Sowdhamini, R., Ramakrishnan, C., and Balaram, P.**, Analysis of short loops connecting secondary structural elements in proteins, in *Molecular Conformation and Biological Interactions*, Balaram, P. and Ramaseshan, S., Eds., Indian Academy of Sciences, Bangalore, 1991.
206. **Sibanda, B. L., Blundell, T. L., and Thornton, J. M.**, Conformation of β -hairpins in protein structures: a systematic classification with applications to modelling by homology, electron density fitting and protein engineering, *J. Mol. Biol.*, 206, 759, 1989.
207. **Sibanda, B. L. and Thornton, J. M.**, Accommodating sequence changes in β -hairpins in proteins, *J. Mol. Biol.*, 229, 428, 1993.
208. **Sowdhamini, R., Srinivasan, N., Ramakrishnan, C., and Balaram, P.**, Orthogonal β - β motifs in proteins, *J. Mol. Biol.*, 223, 845, 1992.
209. **Efimov, A. V.**, Standard conformations of polypeptide chains in irregular regions of proteins, *Mol. Biol. (USSR)*, 20, 208, 1986.
210. **Efimov, A. V.**, Standard structures in protein molecules. α - β hairpins, *Mol. Biol. (USSR)*, 20, 258, 1986.
211. **Efimov, A. V.**, Standard structures in protein molecules. β - α hairpins, *Mol. Biol. (USSR)*, 20, 267, 1986.
212. **Efimov, A. V.**, Patterns of loop regions of proteins, *Curr. Opin. Struct. Biol.*, 3, 379, 1993.
213. **Ring, C. S., Kneller, D. G., Langridge, R., and Cohen, F. E.**, Taxonomy and conformational analysis of loops in proteins, *J. Mol. Biol.*, 224, 685, 1992.
214. **Efimov, A. V.**, Structure of $\alpha\alpha$ hairpins with short connections, *Protein Eng.*, 4, 245, 1991.
215. **Chothia, C. and Lesk, A. M.**, Canonical structures for the hypervariable regions of immunoglobulins, *J. Mol. Biol.*, 196, 901, 1987.
216. **Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Maiuzza, R. A., Phillips, S. E. V., and Poljak, R. A.**, The predicted structure of immunoglobulin — D1.3 and its comparison with the crystal structure, *Science*, 233, 755, 1986.
217. **Karpen, M. E., Dehaseth, P. L., and Neet, K. E.**, Differences in the amino acid substitutions of 3(10) helices and alpha helices, *Protein Sci.*, 1, 1333, 1992.
218. **Vainshtein, B. K., Melikadamyam, W. R., Barynin, V. V., Vagin, A. A., Grevenko, A. I., Borisov, V. V., Bartels, K. S., Fita, I., and Rossmann, M. G.**, 3-Dimensional structure of catalase from *Penicillium-vitale* at 2.0 Å resolution, *J. Mol. Biol.*, 188, 49, 1986.
219. **Matthews, B. W., Weaver, L. H., and Kester, W. R.**, The conformation of thermolysin, *J. Biol. Chem.*, 249, 8030, 1974.
220. **Srinivasan, R., Balasubramanian, R., and Rajan, S. S.**, Extended helical conformation newly observed in protein folding, *Science*, 194, 720, 1976.
221. **Yoder, M. D., Keen, N. T., and Juranak, F.**, New domain motif: the structure of pectate lyase C, a secreted plant virulence factor, *Science*, 260, 1503, 1993.
222. **Cohen, F. E.**, The parallel β -helix of pectate lyase C: something to sneeze at, *Science*, 260, 1444, 1993.
223. **Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P., and Wu, C. W.**, X-ray analysis (1.4 Å resolution) of avian pancreatic polypeptide — small globular protein hormone, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 4175, 1981.
224. **Ramakrishnan, C. and Soman, K. V.**, Identification of secondary structures in globular proteins — a new algorithm, *Int. J. Pept. Prot. Res.*, 20, 218, 1982.

225. Soman, K. V. and Ramakrishnan, C., Occurrence of single helix of the collagen type in globular proteins, *J. Mol. Biol.*, 170, 1045, 1983.
226. Ananthanarayanan, V. S., Soman, K. V., and Ramakrishnan, C., A novel supersecondary structure in globular proteins comprising the collagen-like helix and β -turn, *J. Mol. Biol.*, 198, 705, 1987.
227. Adzhubei, A. A. and Sternberg, M. J. E., Left-handed polyproline II helices commonly occur in globular proteins, *J. Mol. Biol.*, 229, 472, 1993.
228. Janin, J., Wodak, S., Levitt, M., and Maigret, B., Conformations of amino acid side chains in proteins, *J. Mol. Biol.*, 125, 357, 1978.
229. Bhat, T. N., Sasisekharan, V., and Vijayan, M., An analysis of side chain conformation in proteins, *Int. J. Peptide Prot. Res.*, 13, 170, 1979.
230. Summers, N. L., Carlson, W. D., and Karplus, M., Analysis of side chain orientations in homologous proteins, *J. Mol. Biol.*, 196, 175, 1987.
231. McGregor, M. J., Islam, S. A., and Sternberg, M. J. E., Analysis of the relationship between side chain conformation and secondary structure in globular proteins, *J. Mol. Biol.*, 198, 295, 1987.
232. Tuffrey, P., Etchebest, C., Hazout, S., and Lavery, R., A new approach to the rapid determination of protein side chain conformations, *J. Biomolec. Str. Dyn.*, 8, 1267, 1991.
233. Dunbrack, R. L. and Karplus, M., Backbone dependent rotamer library for proteins — application to side chain prediction, *J. Mol. Biol.*, 230, 543, 1993.
234. Schrauber, H., Eisenhaber, F., and Argos, P., Rotomers — to be or not to be — an analysis of amino acid side chain conformations in globular proteins, *J. Mol. Biol.*, 230, 592, 1993.
235. Eisenmenger, F., Argos, P., and Abagyan, R., A method to configure protein side chains from the mainchain trace in homology modeling, *J. Mol. Biol.*, 231, 849, 1993.
236. Warme, P. K. and Morgan, R. S., A survey of amino acid side-chain interactions in 21 proteins, *J. Mol. Biol.*, 118, 289, 1978.
237. Burley, S. K. and Petsko, G. A., Aromatic — aromatic interaction — a mechanism of protein structure stabilization, *Science*, 229, 23, 1985.
238. Singh, J. and Thornton, J. M., The interaction between phenylalanine rings in proteins, *FEBS Letts.*, 191, 1, 1985.
239. Singh, J. and Thornton, J. M., SIRIUS — an automated method for the analysis of the preferred packing arrangements between protein groups, *J. Mol. Biol.*, 211, 595, 1990.
240. Blundell, T., Singh, J., Thornton, J., Burley, S. K., and Petsko, G. A., Aromatic interactions, *Science*, 234, 1005, 1986.
241. Singh, J., Thornton, J. M., Snarey, M., and Campbell, S. F., The geometries of interacting arginine-carboxyls in proteins, *FEBS Letts.*, 224, 161, 1987.
242. Hunter, C. A., Singh, J., and Thornton, J. M., π - π Interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins, *J. Mol. Biol.*, 218, 837, 1991.
243. Heringa, J. and Argos, P., Side chain clusters in protein structures and their role in protein folding, *J. Mol. Biol.*, 220, 151, 1991.
244. Richards, F. M., Areas, volumes, packing, and protein structures, *Ann. Rev. Biochem. Bioeng.*, 6, 151, 1977.
245. Lee, C. and Subbiah, S., Prediction of protein side chain conformation by packing optimization, *J. Mol. Biol.*, 217, 373, 1991.
246. Lee, C. and Levitt, M., Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core, *Nature*, 352, 448, 1991.
247. Barlow, D. J. and Thornton, J. M., Ion pairs in proteins, *J. Mol. Biol.*, 168, 867, 1983.
248. Burley, S. K. and Petsko, G. A., Weakly polar interactions in proteins, *Adv. Prot. Chem.*, 39, 125, 1988.
249. Baker, E. N. and Hubbard, R. E., Hydrogen bonding in globular proteins, *Prog. Biophys. Mol. Biol.*, 44, 97, 1984.
250. Stickle, D. F., Presta, L. G., Dill, K. A., and Rose, G. D., Hydrogen bonding in globular proteins, *J. Mol. Biol.*, 226, 1143, 1992.
251. Rashin, A. A. and Honig, B., On the environment of ionizable groups in globular proteins, *J. Mol. Biol.*, 173, 515, 1984.
252. Thanki, N., Thornton, J. M., and Goodfellow, J. M., Distributions of water around amino acid residues in proteins, *J. Mol. Biol.*, 202, 637, 1988.
253. Sowdhamini, R., Srinivasan, N., Shoichet, B., Santi, D. V., Ramakrishnan, C., and Balaram, P., Stereochemical modeling of disulfide bridges — criteria for introduction into proteins by site-directed mutagenesis, *Protein Eng.*, 3, 95, 1989.
254. Thornton, J. M., Disulphide bridges in globular proteins, *J. Mol. Biol.*, 151, 261, 1981.
255. Srinivasan, N., Sowdhamini, R., Ramakrishnan, C., and Balaram, P., Conformations of disulfide bridges in proteins, *Int. J. Pept. Prot. Res.*, 36, 147, 1990.
256. Desmet, J., Demaeyer, M., Hazes, B., and Lasters, I., The dead-end elimination theorem and its use in protein side chain positioning, *Nature*, 356, 539, 1992.
257. Abagyan, R. and Argos, P., Optimal protocol and trajectory visualization for conformational searches of peptides and proteins, *J. Mol. Biol.*, 225, 519, 1992.
258. Schiffer, C. A., Caldwell, J. W., Kollman, P. A., and Stroud, R. M., Prediction of homologous protein structures based on conformational searches and energetics, *Proteins*, 8, 30, 1990.
259. Wilson, C., Gregoret, L. M., and Agard, D. A., Modeling side chain conformation for homologous proteins using an energy-based rotamer search, *J. Mol. Biol.*, 229, 996, 1993.

260. **Havel, T. F. and Snow, M. E.**, A new method for building protein conformations from sequence alignments with homologues of known structure, *J. Mol. Biol.*, 217, 1, 1991.
261. **Srinivasan, S., March, C. J., and Sudarsanam, S.**, An automated method for modeling proteins on known templates using distance geometry, *Protein Sci.*, 2, 277, 1993.
262. **Taylor, W. R.**, Toward protein tertiary fold prediction using distance and motif constraints, *Protein Eng.*, 4, 853, 1991.
263. **Finkelstein, A. V. and Ptitsyn, O. B.**, Why do globular proteins fit the limited set of patterns?, *Prog. Biophys. Mol. Biol.*, 50, 171, 1987.
264. **Taylor, W. R.**, Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data, *Protein Eng.*, 1993 (in press).
265. **Sali, A. and Blundell, T. L.**, Comparative modeling by satisfaction of spatial constraints, *J. Mol. Biol.*, 1993 (in press).
266. **Saitoh, S., Nakai, T., and Nishikawa, K.**, A geometrical constraint approach for reproducing the native backbone conformation of a protein, *Proteins*, 15, 191, 1993.
267. **Nishikawa, K. and Ooi, T.**, Radial locations of amino acid residues in a globular protein — correlation with the sequence, *J. Biochem.*, 100, 1043, 1986.
268. **Nishikawa, K. and Ooi, T.**, Prediction of the surface-interior diagram of globular proteins by an empirical method, *Int. J. Pept. Protein Res.*, 16, 19, 1980.
269. **Sowdhamini, R., Ramakrishnan, C., and Balaram, P.**, Modeling multiple disulfide loop containing polypeptides by random conformation generation. The test cases of α -conotoxin GI and endothelin I, *Protein Eng.*, 6, 873, 1993.
270. **Fujiyoshi-Yoneda, T., Yoneda, S., Kitamura, K., Amisaki, T., Ikeda, K., Inoue, M., and Ishida, T.**, Adaptability of restrained molecular dynamics for tertiary structure prediction — application to *Crotalus-atrox* venom phospholipase — A2, *Protein Eng.*, 4, 443, 1991.
271. **Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B., and Petersen, S. B.**, A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks, *FEBS Letts.*, 261, 43, 1990.
272. **Friedrichs, M. S., Goldstein, R. A., and Wolynes, B. G.**, Generalized protein tertiary structure recognition using associative memory Hamiltonians, *J. Mol. Biol.*, 222, 1013, 1991.
273. **Brocklehurst, S. M. and Perham, R. N.**, Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipolated H-protein from the peas leaf glycine cleavage system: a new automated method for the prediction of protein tertiary structure, *Protein Sci.*, 2, 626, 1993.
274. **Eliopoulos, E., Geddes, A. J., Brett, M., Pappin, D. J. C., and Findlay, J. B. C.**, A structural model for the chromophore binding domain of ovine rhodopsin, *Int. J. Biol. Macromol.*, 4, 263, 1982.
275. **Pappin, D. J. C., Eliopoulos, E., Brett, M., and Findlay, J. B. C.**, A structure model for ovine rhodopsin, *Int. J. Biol. Macromol.*, 6, 73, 1984.
276. **Hargrave, P. A., McDowell, J. H., Feldman, R. J., Atkinson, P. H., Rao, J. K. M., and Argos, P.**, Rhodopsins protein and carbohydrate structure — selected aspects, *Vision Res.*, 24, 1487, 1984.
277. **Eisenberg, D., Weiss, R. M., and Terwilliger, T. C.**, The hydrophobic moment detects periodicity in protein hydrophobicity, *Proc. Natl. Acad. Sci. U.S.A.*, 81, 140, 1984.
278. **Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisa, C.**, Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, *J. Mol. Biol.*, 195, 659, 1987.
279. **Jennings, M. J.**, Topography of membrane proteins, *Annu. Rev. Biochem.*, 58, 999, 1989.
280. **Bowie, J. U., Clarke, N. D., Pabo, C. O., and Sauer, R. T.**, Identification of protein folds — matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures, *Proteins*, 7, 257, 1990.
281. **Komiya, H., Yeates, T. O., Rees, D. C., Allen, J. P., and Feher, G.**, Structure of the reaction center from *R. sphaeroides* R-26 and 2.4.1: symmetry relations and sequence comparisons between different species, *Proc. Natl. Acad. Sci. U.S.A.*, 85, 9012, 1988.
282. **Rees, D. C., DeAntonio, L., and Eisenberg, D.**, Hydrophobic organization of membrane proteins, *Science*, 245, 510, 1989.
283. **Donnelly, D., Johnson, M. S., Blundell, T. L., and Saunders, J.**, An analysis of the periodicity of conserved residues in sequence alignments of G-protein coupled receptors. Implications for the three-dimensional structure, *FEBS Letts.*, 251, 109, 1989.
284. **Grotzinger, J., Engels, M., Jacoby, E., Wollmer, A., and Strassburger, W.**, A model for the C5a receptor and for its interaction with the ligand, *Protein Eng.*, 4, 767, 1991.
285. **Findlay, J. B. C. and Pappin, D. J. C.**, The opsin family of proteins, *Biochem. J.*, 238, 625, 1986.
286. **Dixon, R. A. F., Sigal, I. S., Rands, E., Register, R. B., Candelore, M. R., Blake, A. D., and Strander, C. D.**, Ligand-binding to the β -adrenergic receptor involves its rhodopsin-like core, *Nature*, 326, 73, 1987.
287. **Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., and Downing, K. H.**, Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy, *J. Mol. Biol.*, 213, 899, 1990.
288. **Donnelly, D., Overington, J. P., Ruffle, S. V., Nugent, J. H. A., and Blundell, T. L.**, Modeling α -helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues, *Protein Sci.*, 2, 55, 1993.

289. Ruffle, S. V., Donnelly, D., Blundell, T. L., and Nugent, J. H. A., A 3-dimensional model of the photosystem-II reaction center of *pisum-sativum*, *Photosynthesis Res.*, 34, 287, 1992.
290. Cronet, P., Sander, C., and Vriend, G., Modeling of transmembrane seven helix bundles, *Protein Eng.*, 6, 59, 1993.
291. Pardo, L., Ballesteros, J. A., Osman, R., and Weinstein, H., On the use of the transmembrane domain of bacteriorhodopsin as a template for modeling the three-dimensional structure of guanine nucleotide-binding regulatory protein-coupled receptors, *Proc. Natl. Acad. Sci. U.S.A.*, 89, 4009, 1992.
292. Vogel, H. and Jahnig, F., Models for the structure of outer-membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction methods, *J. Mol. Biol.*, 190, 191, 1986.
293. Thornton, J. M., Protein structure — the shape of things to come, *Nature*, 335, 10, 1988.
294. Novotny, J., Bruccoleri, R. E., and Karplus, M., An analysis of incorrectly folded models, *J. Mol. Biol.*, 177, 787, 1984.
295. Chothia, C., The nature of the accessible and buried surface in proteins, *J. Mol. Biol.*, 105, 1, 1976.
296. Bryant, S. H. and Amzel, L. M., Correctly folded proteins make twice as many hydrophobic contacts, *Int. J. Pept. Prot. Res.*, 29, 46, 1987.
297. Novotny, J., Rashin, J. J., and Bruccoleri, R. E., Criteria that discriminate between native proteins and incorrectly folded model, *Proteins*, 4, 19, 1988.
298. Baumann, G., Frommel, C., and Sander, C., Polarity as a criterion in protein design, *Protein Eng.*, 2, 329, 1989.
299. Chiche, L., Gregoret, L. M., Cohen, F. E., and Kollman, P. A., Protein model structure evaluation using the solvation free energy of folding, *Proc. Natl. Acad. Sci. U.S.A.*, 87, 3240, 1990.
300. Lüthy, R., Bowie, J. U., and Eisenberg, D., Assessment of protein models with three-dimensional profiles, *Nature*, 356, 83, 1992.
301. Topham, C. M., Thomas, P., Overington, J. P., Johnson, M. S., Eisenmenger, F., and Blundell, T. L., An assessment of COMPOSER: a rule-based approach to modeling protein structure, *Biochem. Soc. Symp.*, 57, 1, 1991.
302. Hubbard, T. J. P. and Blundell, T. L., Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modeling, *Protein Eng.*, 1, 159, 1987.
303. Scarborough, P. E., Guruprasad, K., Topham, C., Richo, G. R., Conner, G. E., Blundell, T. L., and Dunn, B. M., Exploration of subsite binding specificity of human cathepsin D through kinetics and rule-based molecular modeling, *Protein Sci.*, 2, 264, 1993.
304. Dhanaraj, V., Dealwis, C. G., Frazao, C., Badasso, M., Sibanda, B. L., Tickle, I. J., Cooper, J. B., Driessen, H. P. C., Newman, M., Aguilar, C., Wood, S. P., Blundell, T. L., Hobart, P. M., Geoghegan, K. F., Ammirati, M. J., Danley, D. E., O'Connor, B. A., and Hoover, D. J., X-ray analyses of peptide inhibitor complexes define the structural basis of specificity for human and mouse renins, *Nature*, 357, 466, 1992.
305. Šali, A., Veerapandian, B., Cooper, J. B., Foundling, S. I., Hoover, D. J., and Blundell, T. L., High-resolution X-ray diffraction study of the complex between endothiapepsin and an oligopeptide inhibitor: an analysis of the inhibitor binding and description of the rigid body shift in the enzyme, *EMBO J.*, 8, 2179, 1989.
306. Šali, A., Veerapandian, B., Cooper, J. B., Moss, D. S., Hofmann, T., and Blundell, T. L., Domain flexibility in aspartic proteinases, *Proteins*, 12, 158, 1992.
307. Abad-Zapatero, C., Rydel, T. J., and Erickson, J., Revised 2.3 Å structure of porcine pepsin: evidence for a flexible subdomain, *Proteins*, 8, 62, 1990.
308. Sielecki, A. R., Federov, A. A., Boodhoo, A., Andreeva, N. A., and James, M. N. G., Molecular and crystal structures of monoclinic porcine pepsin refined at 1.8 Å resolution, *J. Mol. Biol.*, 214, 143, 1990.
309. Remington, S. J., Woodbury, R. G., Reynolds, R. A., Matthews, B. W., and Neurath, H., The structure of rat mast cell protease II at 1.9 Å resolution, *Biochemistry*, B27, 8097, 1988.
310. Frommel, C. and Sander, C., Thermitase. A thermostable subtilisin: comparison of predicted and experimental structures and the molecular cause of thermostability, *Proteins*, 5, 22, 1989.
311. Weber, I. T., Evaluation of homology modeling of HIV protease, *Proteins*, 7R, 172, 1990.
312. Jurasek, L., Olafson, R. W., Johnson, P., and Smillie, L. B., *Miami Winter Symp.*, 11, 93, 1976.
313. Bedarkar, S., Blundell, T. L., Gowan, L. K., McDonald, K., and Schwabe, C., On the three-dimensional structure of relaxin, *Annal. N.Y. Acad. Sci.*, 380, 22, 1982.
314. Blundell, T. L., Conformation and molecular biology of polypeptide hormones. I. Insulin, insulin-like growth factor and relaxin, *Trends Biochem. Sci.*, 4, 51, 1979.
315. Dodson, E. J., Dodson, C. J., Hodgkin, D. C., and Reynolds, C. D., Structural relationships in the two-zinc insulin hexamer, *Can. Biochem.*, 57, 469, 1979.
316. Furie, B., Bing, D. H., Feldmann, R. J., Robison, D. J., Burnier, J. P., and Furie, B. C., Computer-generated models of blood-coagulation factor-XA, factor-IXA and thrombin based on structural homology with other serine proteases, *J. Biol. Chem.*, 257, 3875, 1982.
317. Furie, B., Bing, D. H., Furie, B. C., Robison, D. J., Burnier, J. P., and Feldman, R. J., 3-Dimensional computer graphics models of bovine factor XA, factor IXA and thrombin, *Thrombosis Hemostasis*, 46, 14, 1981.
318. Dodson, G. G., Eliopoulos, E. E., Isaacs, N. W., McCall, M. J., Niall, H. D., and North, A. C. T., Rat relaxin — insulin-like fold predicts a likely receptor-binding region, *Int. J. Biol. Macromol.*, 4, 399, 1982.

319. Komoriya, A., Meyers, C., Acton, N., Lehrman, S. R., Sporn, M., and Feldmann, R., Computer modeled tertiary structure of murine epidermal growth factor and properties of some synthetic analogs and fragments, *Fed. Proc.*, 41, 1188, 1982.
320. Blundell, T. L., Bedarkar, S., and Humbel, R. E., Tertiary structures, receptor binding and antigenicity of insulin-like growth factors, *Fed. Proc.*, 42, 2592, 1983.
321. Straßburger, W., Wiollmer, A., Pitts, J. E., Glover, I. D., Tickle, I. J., Blundell, T. L., Steffens, G. J., Gunzler, W. A., Otting, F., and Flohe, L., Adaptation of plasminogen-activator sequences to known protease structures, *FEBS Letts.*, 157, 219, 1983.
322. Li, S. S. L., Fitch, W. M., Pan, Y. C. E., and Sharief, F. S., Evolutionary relationships of vertebrate lactate dehydrogenase isozyme-A4 (muscle), isozyme-B4 (heart) and isozyme-C4 (testis), *J. Biol. Chem.*, 258, 7029, 1983.
323. Travers, P., Blundell, T. L., Sternberg, M. J. E., and Bodmer, W. F., Structural and evolutionary analysis of HLA-D-region products, *Nature*, 310, 235, 1984.
324. Raddatz, E., Schittenhelm, C., and Barnickel, G., Computer-graphics methods in pharmaceutical research: visualization of renin-inhibitor complexes, *Kontakte*, (Darmstadt), 3, 1985.
325. Dalgard, E., Bajaj, M., Honegger, A. M., Pitts, J., Wood, S., and Blundell, T. L., The conformation of insulin-like growth factors: relationships with insulins, *J. Cell. Sci. Suppl.*, 3, 53, 1985.
326. Rees, A. R. and de la Paz, P., Investigating antibody specificity using computer graphics and protein engineering, *Trends Biochem. Sci.*, 11, 144, 1986.
327. Jhoti, H., McLeod, A. N., Blundell, T. L., Ishizaki, H., Nagasawa, H., and Suzuki, A., Prothoracicotrophic hormone has an insulin-like tertiary structure, *FEBS Letts.*, 219, 419, 1987.
328. Smith-Gill, S. J., Mainhart, C., Lavoie, T. B., Feldmann, R. J., Drohan, W., and Brooks, B. R., A 3-dimensional model of an anti-lysozyme antibody, *J. Mol. Biol.*, 194, 713, 1987.
329. Toma, K., Yamamoto, S., Deyashiki, Y., and Suzuki, K., 3-Dimensional structure of protein-C inhibitor predicted from structure of α -1-antitrypsin with computer graphics, *Protein Eng.*, 1, 471, 1987.
330. Pearl, L. H. and Taylor, W. R., A structural model for the retroviral proteases, *Nature*, 329, 351, 1987.
331. Berg, J. M., Proposed structure for the zinc-binding domains from transcription factor-III α and related proteins, *Proc. Natl. Acad. Sci. U.S.A.*, 85, 99, 1988.
332. Bates, P. A., McGregor, M. J., Islam, S. A., Sattentau, Q. J., and Sternberg, M. J. E., A predicted 3-dimensional structure for the human immunodeficiency virus binding domains of CD4 antigen, *Protein Eng.*, 3, 13, 1989.
333. Weber, I. T., Miller, M., Jaskolski, M., Leis, J., Skalka, A. M., and Wlodawer, A., Molecular modeling of the HIV-1 protease and its substrate binding site, *Science*, 243, 928, 1989.
334. Cox, J. A., Alard, P., and Schaad, P., Comparative modeling of amphioxus calcium vector protein with calmodulin and troponin C, *Protein Eng.*, 4, 23, 1990.
335. Topham, C. M., Overington, J., Thomas, M., Kowlessur, D., Thomas, E. W., and Brocklehurst, K., Three-dimensional structure and thiol reactivity characteristics of chymopapain M (papaya proteinase IV), *Biochem. Soc. Trans.*, 18, 934, 1990.
336. Topham, C. M., Overington, J., O'Driscoll, M., Salih, E., Thomas, M., Thomas, E. W., and Brocklehurst, K., Three-dimensional structure of a B-type chymopapain, *Biochem. Soc. Trans.*, 18, 933, 1990.
337. Topham, C. M., Overington, J., Kowlessur, D., Thomas, M., Thomas, E. W., and Brocklehurst, K., Investigation of mechanistic consequences of natural structural variation within the cysteine proteinases by knowledge-based modeling and kinetic methods, *Biochem. Soc. Trans.*, 18, 579, 1990.
338. Van de Ven, W. J. M., Voorberg, J., Fontijn, J., Pannekoek, H., Van denouwe, A. M. W., Van duijnhoven, Roebroek, A. J. M., and Seizen, R. J., Furin is a subtilisin-like proprotein processing enzyme in higher eukaryotes, *Mol. Biol. Rep.*, 14, 265, 1990.
339. Laughton, C. A., Neidle, S., Zvelebil, M. J. J. M., and Sternberg, M. J. E., A molecular model for the enzyme cytochrome-P45017- α , a major target for the chemotherapy of prostatic cancer, *Biochem. Biophys. Res. Commun.*, 171, 1160, 1990.
340. Floegel, R., Zielenkiewicz, P., and Saenger, W., Tertiary structure of RNase Pchl predicted from the model structure of RNase MS and the crystal structure of RNase T1 — comparison among the model structures — testing the limits of modeling by homology, *Eur. Biophys. J.*, 18, 225, 1990.
341. Svensson, B., Vass, I., Cedergren, E., and Styring, S., Structure of donor side components in photosystem II predicted by computer modeling, *EMBO J.*, 9, 2051, 1990.
342. Peitsch, M. C. and Boguski, M. S., The first lipocalin with enzymatic activity, *Trends Biochem. Sci.*, 16, 363, 1991.
343. Ghetti, A., Bolognesi, M., Cobianchi, F., and Morandi, C., Modeling by homology of RNA-binding domain, *Mol. Biol. Rep.*, 14, 87, 1990.
344. Giranda, V. L., Chapman, M. S., and Rossmann, M. G., Modeling of the human intercellular adhesion molecule-1, the human rhinovirus major group receptor, *Proteins*, 7, 227, 1990.
345. Signor, G., Vita, C., Fontana, A., Frigerio, F., Bolognesi, M., Toma, S., Gianna, R., Degregoriis, E., and Grandi, G., Structural features of neutral protease from *Bacillus subtilis* deduced from model building and limited proteolysis experiments, *Eur. J. Biochem.*, 189, 221, 1990.
346. Lapadat, M. A., Deerfield, D. W., Pedersen, L. G., and Spemulli, L. L., Generation of potential structures for the G-domain of chloroplast EF-Tu using comparative modeling, *Proteins*, 8, 237, 1990.

347. Bowyer, J., Hilton, M., Whitelegge, J., Jewess, P., Camilleri, P., Crofts, A., and Robinson, H., Molecular modelling studies on the binding of phenylurea inhibitors to the D1 protein of photosystem II, *Z. Naturforsch.*, 45c, 379,
348. Tietjen, K. G., Kluth, J. F., Andree, R., Haug, M., Lindig, M., Muller, K. H., Wroblowsky, H. J., and Trebst, A., The herbicide binding niche of Photosystem II — a model, *Pestic. Sci.*, 31, 65, 1991.
349. Topham, C. M., Salih, E., Frazao, C., Kowlessur, D., Overington, J. P., Thomas, M., Brocklehurst, S. M., Patel, M., Thomas, E. W., and Brocklehurst, K., Structure-function relationships in the cysteine proteinases actinidin, papain and papaya proteinase omega deduced by knowledge-based modeling and active-center characteristics determined by 2-hydronic state reactivity probe kinetics and kinetics of catalysis, *Biochem. J.*, 280, 79, 1991.
350. Lapatto, R., Model for the structure of formaldehyde dehydrogenase based on alcohol dehydrogenase, *Int. J. Biol. Macromol.*, 13, 73, 1991.
351. Gronenborn, A. M. and Clore, G. M., Modeling the three-dimensional structure of the monocyte chemoattractant and activating protein MCAF/MCP-1 on the basis of the solution structure of interleukin-8, *Protein Eng.*, 4, 263, 1991.
352. Scully, J. L. and Evans, D. R., Comparative modeling of mammalian aspartate-transcarbomylase, *Proteins*, 9, 191, 1991.
353. Vos, P., Boerrigter, I. J., Buist, G., Haandrikman, A. J., Nijhuis, M., Dereuver, M. B., Siezen, R. J., Veneman, G., De Vos, W. M., and Kok, J., Engineering of the *Lactococcus lactis* serine proteinase by construction of hybrid enzymes, *Protein Eng.*, 4, 479, 1991.
354. Kettleborough, C. A., Saldanha, J., Heath, V. J., Morrison, C. J., and Bendig, M. M., Humanization of a mouse monoclonal antibody by CDR grafting — the importance of framework residues on loop conformation, *Protein Eng.*, 4, 773, 1991.
355. Frampton, J., Gibson, T. J., Ness, S. A., Doderlein, G., and Graf, T., Proposed structure for the DNA-binding domain of the *myb* oncoprotein based on model building and mutational analysis, *Protein Eng.*, 4, 891, 1991.
356. Saldanha, J. and Mahadevan, D., Molecular model building of amylin and α -calcitonin gene related polypeptide hormones using a combination of knowledge sources, *Protein Eng.*, 4, 539, 1991.
357. Zvelebil, M. J. J. M., Wolf, C. R., and Sternberg, M. J. E., A predicted 3-dimensional structure of human cytochrome P450 — implications for substrate specificity, *Protein Eng.*, 4, 271, 1991.
358. Du, P., Collins, J. R., and Loew, G. H., Homology modeling of a heme protein, lignin peroxidase, from the crystal structure of cytochrome-C peroxidase, *Protein Eng.*, 5, 679, 1992.
359. Bruschi, M., Bonicel, J., Hatchikan, E. C., Fardeau, M. L., Belaich, J. P., and Frey, M., Amino acid sequence and molecular modeling of a thermostable 2(4Fe-4S) ferredoxin from the archaeobacterium *Methanococcus thermolithotrophicus*, *Biochim. Biophys. Acta*, 1076, 79, 1991.
360. Nakamura, H., Katayanagi, K., Morikawa, K., and Ikehara, M., Structural models of ribonuclease H domains in reverse transcriptases, *Nucl. Acids Res.*, 19, 1817, 1991.
361. Keen, J. N., Caceres, I., Eliopoulos, E. E., Zagalsky, P. F., and Findlay, J. B. C., Complete sequence and model for the C1 subunit of the carotenoprotein, crustacyanin and model for the dimer β -crustacyanin formed from the C1 and A2 subunits with astaxanthin, *Eur. J. Biochem.*, 202, 31, 1991.
362. Rippmann, F., Taylor, W. R., Rothbard, J. B., and Green N. M., A hypothetical model for the peptide binding domain of hsp70 based on the peptide binding domain of HLA, *EMBO J.*, 10, 1053, 1991.
363. Oomen, R. P., Young, N. M., and Bundle, D. R., Molecular modeling of antibody antigen complexes between the *Brucella abortus* O-chain polysaccharide and a specific monoclonal antibody, *Protein Eng.*, 4, 427, 1991.
364. Mimura, C. S., Holbrook, S. R., and Ames, G. F., Structural model of the nucleotide-binding conserved component of periplasmic permeases, *Proc. Natl. Acad. Sci. U.S.A.*, 88, 84, 1991.
365. Mosimann, S. C., Johns, K. L., Ardelt, W., Mikulski, S. M., Shogen, K., and James, M. N. G., Comparative molecular modeling and crystallization of p-30 protein — a novel antitumor protein of ranapiens oocytes and early embryos, *Proteins*, 14, 392, 1992.
366. Terry, C. J. and Blake, C. C. F., Comparison of the modeled thyroxine binding-site in TBG with the experimentally determined site in transthyretin, *Protein Eng.*, 5, 505, 1992.
367. Jarvis, J. A., Munro, S. L. A., and Craik, D. J., Homology model of thyroxine binding globulin and elucidation of the thyroid-hormone binding site, *Protein Eng.*, 5, 61, 1992.
368. Bates, P. A., Luo, J. C., and Sternberg, M. J. E., A predicted 3-dimensional structure for the carcino-embryonic antigen (CEA), *FEBS Letts.*, 301, 207, 1992.
369. Mas, M. T., Smith, K. C., Yarmush, D. L., Aisaka, K., and Fine, R. M., Modeling the anti-CEA antibody combining site by homology and conformational search, *Proteins*, 14, 483, 1992.
370. Jager, J., Solmajer, T., and Jansonius, J. N., Computational approach toward the 3-dimensional structure of *Escherichia coli* tyrosine aminotransferase, *FEBS Letts.*, 306, 234, 1992.
371. Vihinen, M., Lundin, M., and Baltscheffsky, H., Computer modeling of 2 inorganic pyrophosphatases, *Biochem. Biophys. Res. Commun.*, 186, 122, 1992.
372. Kumar, V. D. and Weber, I. T., Molecular model of the cyclic GMP-binding domain of the cyclic GMP-gated ion channel, *Biochemistry*, 31, 4643, 1992.

373. Šali, A., Matsumoto, R., McNeil, H. P., Karplus, M., and Stevens, R. L., Three-dimensional models of four mouse mast cell chymases, *J. Biol. Chem.*, 268, 9023, 1993.
374. Mills, A., Modeling the carbohydrate recognition domain of human E-selectin, *FEBS Letts.*, 319, 5, 1993.
375. Bax, B., Blaber, M., Ferguson, G., Sternberg, M. J. E., and Walls, P. H., Prediction of the three-dimensional structures of the nerve growth factor and epidermal growth factor binding proteins (kallikreins) and an hypothetical structure of the high-molecular-weight complex of epidermal growth factor with its binding protein, *Protein Sci.*, 2, 1229, 1993.
376. Wampler, J. E., Bradley, E. A., Stewart, D. E., and Adams, M. W. W., Modeling the structure of *Pyrococcus furiosus* rubredoxin by homology to other X-ray structures, *Protein Sci.*, 2, 640, 1993.
377. Dhanaraj, V., Dealwis, C., Bailey, D., Cooper, J. B., and Blundell, T. L., The three-dimensional structures of inhibitor complexes of monomeric aspartic proteinases, in *Innovations on Proteases and Their Inhibitors*, Avilés, F. X., Ed., Walter de Gruyter, Berlin, 1993, in press.
378. Siezen, R. J., de Vos, W. M., Leunissen, J. A. M., and Dijkstra, B. W., Homology modeling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases, *Protein Eng.*, 4, 719, 1991.
379. Šali, A., Ph.D. thesis, University of London, 1990, 140.
380. Schertler, G. F. X., Villa, C., and Henderson, R., Projection structure of rhodopsin, *Nature*, 362, 770, 1993.
381. Montelione, G. T., Wütrich, K., Nice, E. C., Burgess, A. W., and Sheraga, H. A., Solution structure of epidermal growth factor: determination of the polypeptide backbone chain-fold by nuclear magnetic resonance and distance geometry, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 5226, 1987.
382. Montelione, G. T., Wütrich, K., Burgess, A. W., Nice, E. C., Wagner, G., Gibson, K. D., and Sheraga, H. A., Solution structure of murine epidermal growth factor determined by NMR spectroscopy and refined by energy minimization with restraints, *Biochemistry*, 31, 236, 1992.
383. Pardi, A., Hare, D. R., Selsted, M. E., Morrison, R. D., Bassolino, D. A., and Bach, A. C., Solution study of the rabbit neutrophil defensin NP-5, *J. Mol. Biol.*, 201, 625, 1988.
384. Havel, T. F., Kuntz, I. D., and Crippen, G. M., The theory and practice of distance geometry, *Bull. Math. Biol.*, 45, 665, 1983.