Comparative Protein Modelling by Satisfaction of Spatial Restraints

Andrej Šali[†] and Tom L. Blundell

ICRF Unit of Structural Molecular Biology
Department of Crystallography
Birkbeck College, London WC1E 7HX, England

(Received 30 March 1993; accepted 13 July 1993)

We describe a comparative protein modelling method designed to find the most probable structure for a sequence given its alignment with related structures. The three-dimensional (3D) model is obtained by optimally satisfying spatial restraints derived from the alignment and expressed as probability density functions (pdfs) for the features restrained. For example, the probabilities for main-chain conformations of a modelled residue may be restrained by its residue type, main-chain conformation of an equivalent residue in a related protein, and the local similarity between the two sequences. Several such pdfs are obtained from the correlations between structural features in 17 families of homologous proteins which have been aligned on the basis of their 3D structures. The pdfs restrain C^{α}–C^{α} distances, main-chain N–O distances, main-chain and side-chain dihedral angles. A smoothing procedure is used in the derivation of these relationships to minimize the problem of a sparse database. The 3D 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 non-hydrogen atoms. The method is automated and is illustrated by the modelling of trypsin from two other serine proteinases.

Keywords: comparative protein modelling; restraints; optimization; protein database; serine proteinases

1. Introduction

Approaches to determination and prediction of protein three-dimensional (3D[‡]) structure can be classified on the basis of the predominant information that is used to calculate the model. The experimental methods include X-ray crystallography (Blundell & Johnson, 1976) and multi-dimensional nuclear magnetic resonance (NMR) techniques (Bax, 1989). The theoretical approaches (Fasman, 1989) can be divided into physical and empirical methods. The physical prediction methods are based on interactions between atoms and include molecular dynamics and energy minimization (Brooks III *et al.*, 1988), whereas the empirical methods depend on the protein structures that have been already determined by experiment. They include combinatorial (Cohen & Kuntz, 1989) and comparative modelling (Blundell *et al.*, 1987; Šali *et al.*, 1990; Swindells & Thornton, 1991).

Comparative modelling uses experimentally determined protein structures to predict conformation of other proteins with similar amino acid sequences. This is possible because a small change in the sequence usually results in a small change in the 3D structure (Hubbard & Blundell, 1987; Lesk & Chothia, 1986). The accuracy of protein models obtained by comparative modelling compares favourably with that of models calculated by other theoretical methods. The comparative method produces models with an r.m.s. error as low as 1 Å for sequences that have sufficiently similar homologues with known 3D structures (Topham *et al.*, 1991); in contrast, physical prediction methods and combinatorial modelling calculate structures with r.m.s. errors of approximately 3·5 Å for small proteins (Cohen & Kuntz, 1989; Wilson & Doniach, 1989). On the other hand, comparative modelling is not as accurate as X-ray crystallography and NMR,

[†] Author to whom all correspondence should be addressed. Current address: Dept. of Chemistry, Harvard University, 12 Oxford Street, Cambridge MA 02138, U.S.A.

[‡] Abbreviations used: 3D, three-dimensional; r.m.s., root-mean-square; pdf, probability density function; NMR, nuclear magnetic resonance.

0022-2836/93/230779-37 $08.00/0

which can determine protein structures with an r.m.s. error of approximately 0·3 and 0·5 Å, respectively (Clore & Gronenborn, 1991). It is also restricted to sequences with closely related proteins with known 3D structures. Nevertheless, since 28% of the known sequences have at least a 25% residue identity with one of the known structures (Chothia, 1992), we can estimate that an order of magnitude more sequences can be modelled by comparative modelling than there have been protein structures determined by experiment. This ratio is likely to increase as the fraction of the known structural motifs increases and the gap between numbers of the known sequences and 3D structures widens.

In the early eighties, manual comparative modelling (Browne *et al.*, 1969; Warme *et al.*, 1974) was facilitated by manipulation of protein molecules on the graphics terminal (Greer, 1981) that was made possible by the computer programs such as FRODO (Jones, 1978). The method was later improved by the introduction of largely automated modelling algorithms that can use several known structures to model the unknown member of the family (Sutcliffe *et al.*, 1987*a*,*b*). This approach is based on assembling the model from parts of dissected related structures (Blundell *et al.*, 1986, 1987; Blundell & Sternberg, 1985; Claessens *et al.*, 1989; Greer, 1981, 1990; Robson *et al.*, 1987; Schiffer *et al.*, 1990; Stewart *et al.*, 1987; Unger *et al.*, 1989). Known structures that are homologous to the sequence being modelled are first superposed as rigid bodies using multiple least-squares fitting. The sequence of the unknown is then aligned with the consensus sequence of the known structures. The model is assembled from rigid blocks of structure corresponding to the core regions, loops and side-chains from the aligned protein structures. This modelling procedure is very successful when the known structures cluster around that to be predicted and where the percentage sequence identity to the unknown is greater than 40%. For example, the model of bovine trypsin built using the known structures of four other serine proteinases has the r.m.s. difference from the known structures of only 0·64 Å for the 150 residues in the core of the molecule (Overington, 1991). Similarly, 80% of side-chain conformations are correctly predicted for closely homologous structures. In all cases, the accuracy of the prediction decreases quickly as the sequence identity between the known and unknown decreases.

In addition to the assembly of rigid body fragments, there are other automated and semi-automated methods for comparative modelling. Modelling by satisfaction of spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure was proposed by Šali *et al.* (1990). An elegant distance geometry approach for constructing all-atom models from distance constraints was described by Havel & Snow (1991). A similar method was presented by Srinivasan *et al.* (1993). Another method based on satisfaction of main-chain distance restraints by molecular dynamics was described by Fujiyoshi-Yoneda *et al.* (1991). Neural networks and optimization in Cartesian space were used to calculate a model from a C^α distance plot of a homologous protein (Bohr *et al.*, 1990). Recently, comparative modelling by optimization of a potential function constructed from a sequence alignment with related structures was described (Snow, 1993). Protein structure was predicted by optimizing an associative-memory Hamiltonian whose parameters were obtained by using the random energy model with the data from the related protein structures (Friedrichs *et al.*, 1991). Several methods for constructing full backbone co-ordinates from the positions of the C^α atoms alone were described (Bassolino-Klimas & Bruccoleri, 1992; Correa, 1990; Holm & Sander, 1991; Levitt, 1992; Luo *et al.*, 1992; Payne, 1993; Reid & Thornton, 1989; Rey & Skolnick, 1992). These methods can be applied to comparative modelling when homologous structures are used as the source of the guiding C^α positions and when combined with the loop and side-chain construction algorithms (Holm & Sander, 1991, 1992). And finally, a new class of methods based on recognition of the native fold using database of all known protein structures can be seen as a first step towards modelling sequences that are only distantly related to the known protein structures (Thornton *et al.*, 1991). These methods include template matching with three-dimensional profiles (Bowie *et al.*, 1991), topology fingerprinting (Godzik *et al.*, 1992), optimal threading of a sequence onto a 3D structure (Jones *et al.*, 1992), tertiary structure recognition (Friedrichs *et al.*, 1991), and detection of native-like models for a given sequence (Sippl & Weitckus, 1992).

Numerous other techniques have been described that do not predict the whole structure but only some aspects of it. These methods can often be used in combination with each other. Side-chain conformation has been predicted from similar structures, from proteins in general, and from energy considerations (Desmet *et al.*, 1992; Dunbrack & Karplus, 1993; Holm & Sander, 1992; Lee & Subbiah, 1991; McGregor *et al.*, 1987; Ponder & Richards, 1987; Schiffer *et al.*, 1990; Singh & Thornton, 1990; Summers *et al.*, 1987; Summers & Karplus, 1989; Tuffery *et al.*, 1991; Wilson *et al.*, 1993). Substitutions, insertions and deletions, such as those in loops, have been modelled by regularizing a suitable fragment selected from homologous or other structures, or by conformational search based on minimizing the energy of a segment, or by a combination of the approaches (Bruccoleri & Karplus, 1987; Chothia *et al.*, 1989; Dudek & Scheraga, 1990; Jones & Thirup, 1986; Martin *et al.*, 1989; Mas *et al.*, 1992; Moult & James, 1986; Sibanda *et al.*, 1989; Summers & Karplus, 1990; Topham *et al.*, 1993). The geometry of disulphide bridges has been predicted on the basis of information on disulphide bridges in experimentally determined protein structures (Sowdhamini *et al.*, 1989; Thornton, 1981).

	1	2	3	4	5	6	7
structure A	A	f	s̃	t	l	Ñ̲	t
structure B	A	f	s̃	s	i	Ñ̲	**t̃**
structure C	A	Ỹ	p	s	i	S̲̃	a
secuence X	G	F	D	T	I	T	T
extrapolation				↓			
structure X	G	f	d̃	t	i	T̲	t̃

Figure 1. Comparative protein modelling by satisfaction of spatial restraints. A 3D model of sequence X has to be calculated from the known homologous structures A, B and C. First, the known 3D structures are compared. In order to indicate spatial features of the known structures, residue codes in the resulting alignment are formatted using the convention of the JOY program (Overington *et al.*, 1990): UPPER CASE, solvent inaccessible amino acid residues; lower case, solvent accessible amino acid residues; underline, hydrogen bond to main-chain carbonyl; **bold type**, hydrogen bond to main-chain nitrogen; tilde (˜), side-chain–side-chain H-bond; *italic*, positive main chain dihedral angle Φ. The sequence of the unknown is then aligned with the related structures. Next, the spatial features of the known structures are transferred to the sequence of the unknown; thus, a number of spatial restraints on its structure are obtained. For example, since there is a conserved hydrogen bond to the main-chain carbonyl at position 6 in all 3 known structures, we assume that the equivalent hydrogen bond also occurs in the sequence of the unknown. Finally, these restraints are satisfied as well as possible to obtain the model for the 3D structure of the unknown.

Future improvements of comparative modelling should aim to model proteins with lower homology to known structures, to increase the accuracy of the models, and to make modelling fully automated. In this paper, we attempt to achieve these goals by pursuing the following fundamental question: What is the most probable structure for a certain sequence given its alignment with related structures? Our approach, outlined in Figure 1, follows from the method for comparison of protein structures implemented in the program COMPARER (Šali & Blundell, 1990; Šali *et al.*, 1990; Zhu *et al.*, 1992). The modelling method was developed to use as many different types of data about the unknown as possible. The method consists of three stages: (1) alignment of the sequence to be modelled with related protein structures and segments, (2) extraction of spatial restraints on the sequence using the alignment, and (3) satisfaction of the restraints to obtain a 3D model. This paper describes the procedures involved in the last two stages.

Spatial restraints on the sequence of the unknown are obtained from the statistical analysis of the relationships between various features of protein structure. A database of 17 family alignments including 80 proteins was constructed to obtain the tables quantifying the relationships, such as those between the two equivalent C^{α}–C^{α} distances or between equivalent main-chain dihedral angles from two related proteins. These relationships were described as conditional probability density functions (pdfs) for the features to be predicted. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of a residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins. The calculation of the 3D model by satisfaction of spatial restraints is achieved by optimization of the molecular pdf. This function is a combination of pdfs restraining individual spatial features of the whole molecule. The optimization method is a variable target function method that applies the conjugate gradients method to positions of all non-hydrogen atoms. To illustrate comparative modelling by satisfaction of spatial restraints, we describe the modelling of the 3D structure of trypsin based on its relation to two other serine proteinases, tonin and elastase.

2. Derivation of Spatial Restraints

If a sufficient number of strong restraints is specified, the 3D structure of the protein is well determined. In this section, relatively simple restraints on the protein conformation are defined from the information about related protein structures.

A restraint is most precisely defined in terms of a probability density function, $p(x)$, for the feature x that is restrained. It can be of any form provided that it is non-negative and that it integrates to 1 over the range of all possible values for x. The actual finite probability of an event $x_1 \leq x < x_2$ is obtained by integration of p:

$$p(x_1 \leq x < x_2) = \int_{x_1}^{x_2} p(x)\,\mathrm{d}x.$$

The expression of a restraint in terms of a pdf gives more information on the possible values of the restrained feature than the mean of measurements alone. It is also more complete than the upper and lower bounds on a certain atom–atom distance, such as those used in the distance geometry approach to derivation of protein 3D structure from multi-dimensional NMR data.

(a) *An outline of the derivation of probability density functions*

Pdfs useful in protein modelling can be calculated either analytically using statistical and classical mechanics (section (e), below) or empirically using the database of known protein structures (sections (f) to (i), below). In either case, the pdf suitable for restraining a certain feature x can be written as:

$$p(x/a, b, \ldots, c). \qquad (1)$$

This is a conditional pdf and gives a probability density for x when $a, b, \ldots, c$ are specified. It can be seen as an ordinary pdf for x that also depends on the values of other variables. For example, $p(\chi_1/\text{residue type}, \Phi, \Psi)$ could be used to predict the side-chain dihedral angle χ_1 from the type of a residue and its main-chain angles Φ and Ψ.

For the pdf to be useful in modelling, all the features in the association $(x, a, b, \ldots, c)$, except for x, must be known at the prediction stage. Additionally, x has to be a spatial feature of the sequence to be modelled. The most useful known feature is of the same type as x but associated with the equivalent position(s) in a related known structure. For example, when restraining a certain C^{α}–C^{α} distance, the most useful information is the equivalent distance from a related structure.

In reality, it is not possible to obtain the true function p, but only its approximations:

$$p(x/a, b, \ldots, c) \approx W_{x,a,b,\ldots,c} \approx f(x, a, b, \ldots, c, \mathbf{q}), \quad (2)$$

where $W_{x,a,b,\ldots,c}$ is a table spanned by $x, a, b, \ldots, c$ that contains as its elements the observed relative frequencies for the occurrence of x given $a, b, \ldots, c$, and f is an analytic function fitted to the observed $\mathbf{W}$. As with any pdf, $\mathbf{W}$ and f must satisfy the integration and non-negativity criteria mentioned above. f is a function with parameters $\mathbf{q}$ which are obtained by applying the least-squares principle. The best $\mathbf{q}$ is defined as the $\mathbf{q}$ that minimizes the function:

$$\text{r.m.s.} = \sqrt{\sum_{x,a,b,\ldots,c} [W_{x,a,b,\ldots,c} - f(x, a, b, \ldots, c, \mathbf{q})]^2}. \quad (3)$$

The multidimensional table of relative frequencies $\mathbf{W}$ is calculated from the absolute frequencies $\mathbf{W}'$ using:

$$W_{x,a,b,\ldots,c} = \frac{W'_{x,a,b,\ldots,c}}{\sum_x W'_{x,a,b,\ldots,c}}. \quad (4)$$

The absolute frequencies, $\mathbf{W}'$, are obtained directly by counting the number of occurrences of each combination of $(x, a, b, \ldots, c)$ values in the sample. In this study, the sample is derived from a database of known protein structures and their alignments. Thus, before the restraints can be derived, a database of known protein structures, their features and alignments must be constructed.

(b) *Database of known protein structures, their features and alignments*

(i) *Overview of the local database*

Members of 17 families of related proteins extracted from the Brookhaven Protein Databank (Abola *et al.*, 1987; Bernstein *et al.*, 1977) are listed in Table 1. The composition of this sample is discussed in subsection (ii), below. The co-ordinate files in the local database were edited to delete records for non-protein atoms, excessive stretches where the atomic co-ordinates were not defined, and duplicate atoms of double occupancies. The files containing several related domains or subunits were split so that there was one homologous structure per file. These homologous structures were then aligned by the program COMPARER (Šali & Blundell, 1990; Zhu *et al.*, 1992) to obtain multiple alignments for each of the families in the local database. These alignments were added to the local database. Finally, a number of features of protein structures were also calculated and stored in the database. These features are defined below. Recently, two other databases of structural family alignments have been described (Holm *et al.*, 1992; Pascarella & Argos, 1992).

The program MDT was written to explore the local database and to derive the best pdfs for comparative modelling. The inputs to the program are names of selected features, a list of discrete values for tabulating these features (numerical or symbolic), and the list of alignments. These are then used to calculate various multi-dimensional frequency tables $W'_{x,a,b,\ldots,c}$ by counting the occurrences of all the required combinations of features $x, a, b, \ldots, c$ in the local database. The tables $\mathbf{W}'$ were subsequently used as outlined above and described in detail below to calculate the relative frequency tables $\mathbf{W}$ and sometimes the corresponding pdfs f. For fitting pdf f to the observed relative frequencies $\mathbf{W}$, the Levenberg–Marquardt algorithm for non-constrained least-squares fitting of a non-linear multidimensional model (Press *et al.*, 1986) was implemented in the program LSQ.

(ii) *Composition of the local database*

The local database contains representatives of all four structural classes of proteins (Table 1): α, β, $\alpha+\beta$ and α/β. The frequency of the β-sheet residues is similar to that of the residues in the helical conformation (Fig. 2(a)). The percentage sequence identity for the sequences compared varies from 6% to 98% (Fig. 2(b)). This is also reflected in the distribution of residue neighbourhood differences for all equivalent pairs of residues in the database (Fig. 2(c)). Most of the protein structures in the database were solved at a medium or high resolution, although there are also a few low resolution structures (Fig. 2(d)). Additional compositional characteristics of the database, frequency of amino acid residue types and distribution of fractional side-chain solvent accessibilities, are shown in Figure 2(e) and (f). Figure 2 indicates that the local database of protein structures and their alignments is a representative sample of globular proteins and is therefore suitable for uncovering the general relationships between features of protein structure.

(iii) *Tabulating associations between protein features*

The following is a detailed description of the MDT program used for quantifying associations between protein features.

The classification of features from COMPARER (Šali & Blundell, 1990) also proves useful for describing the MDT program: features can be either properties associated with a single element or relationships between two or more elements.

Table 1

Families of homologous proteins in the local database

Name	PDB code	Residue range	Resolution (Å)
β Proteins:			
Aspartic proteinases, lobes			
E. parasitica endothiapepsin	4ape-n	1–174	2·1
	4ape-c	175–326	
P. penicillum penicillopepsin	2app-n	1–174	1·8
	2app-n	175–323	
R. chinensis rhizopuspepsin	2apr-n	1–178	1·8
	2apr-c	179–325	
porcine pepsin	5pep-n	1–174	2·3
	5pep-c	175–327	
bovine chymosin	3cms-n	1–174	2·2
	3cms-c	175–327	
HIV-protease	4hvp		2·3
RSV-protease	2rspa		2·0
Serine proteinases			
rat tonin	1ton		1·8
porcine kallikrein	2pkaa		2·0
bovine trypsin	2ptn		1·5
bovine chymotrypsin	4chaa		1·7
porcine elastase	3est		1·6
rat mast cell protease-II	3rp2a		1·9
S. griseus trypsin	1sgt		1·7
S. griseus proteinase A	2sga		1·5
L. enzymogenes α-lytic proteinase	2alp		1·7
S. griseus proteinase B	3sgb		1·8
Azurins			
A. denitrificans azurin	2azaa		1·8
P. aeruginosa azurin	1azu		2·7
Immunoglobulins, domains			
FAB (Lambda) KOL	2fb4h	1–117	2·0
	2fb4	123–221	1·9
FAB (Prime) NEW	3fabh	2–116	2·0
	3fabl	114–214	
	3fabl	3–108	
B-J Fragment REI	1reia	1–107	2·0
HyHEL-5 FAB	2hfll	1–105	2·5
B-J Fragment RHE	2rhe	1–111	1·6
FAB (Kappa) J539	1fbjl	111–213	2·6
	1fbjh	123–218	
FC (Human)	1fc1a	238–341	2·9
γ-Crystallins, motifs			
calf γ-II crystallin	1gcr-1	1–39	1·6
	1gcr-2	40–87	
	1gcr-3	88–128	
	1gcr-4	129–174	
bovine γ-IV crystallin	2gcr-1	1–39	2·3
	2gcr-2	40–87	
	2gcr-3	88–128	
	2gcr-4	129–174	
α+β Proteins:			
Cysteine proteinases			
papain	9pap		1·7
actinidin	2act		1·7
Lysozymes			
hen egg white lysozyme	1lzt		2·0
human lysozyme	1lz1		1·5
α Proteins:			
Globins			
human haemoglobin	2hhba		1·7
	2hhbb		
sperm whale myoglobin	3mbn		2·0
C. thummi thummi erythrocruorin	1ecd		1·4
sea lamprey haemoglobin	2lhb		2·0
leghaemoglobin	1lh1		2·0
Phospholipases			
bovine phospholipase A_2	1bp2		1·7
porcine phospholipase A_2	1p2p		2·6
C. atrox phospholipase A_2	1pp2		2·5
Cytochromes (1)			
albacore tuna cytochrome *c*	3cyt		1·8
rice embryo cytochrome *c*	1ccr		1·5
R. rubrum cytochrome c_2	2c2c		2·0
bonito cytochrome *c*	1cyc		2·3
P. denitrificans cytochrome *c*-550	155c		2·5
Cytochromes (2)			
P. aeruginosa cytochrome *c*-551	351c		1·6
A. vinelandii cytochrome c_5	1cc5		2·5
Photosynthetic reaction centres, domains			
R. viridis reaction centre	1prcm	1–323	2·3
	1prcl	1–273	
R. sphaeroides reaction centre	4rcrm	1–305	2·8
	4rcrl	1–275	
α/β Proteins:			
Ferredoxins			
P. aerogenes ferredoxin	1fdx		2·0
A. vinelandii ferredoxin	4fd1		1·9
Flavodoxins			
Clostridium mp flavodoxin	3fxn		1·9
D. vulgaris flavodoxin	1fx1		2·0
Dehydrofolate reductases			
E. coli dihydrofolate reductase	4dfra	1–159	1·7
L. casei dihydrofolate reductase	3dfr	1–162	1·7
Dehydrogenases			
horse alcohol dehydrogenase	8adh	193–318	2·4
porcine malate dehydrogenase	4mdha	4–154	2·5
lobster glyceraldehyde dehydrogenase	1gpd	1–148	2·9
dogfish lactate dehydrogenase	6ldh	22–165	2·0
Oxidoreductases			
human glutathione reductase	3grsn	18–160	1·5
	3grsm	186–294	
P. fluorescens p-hydroxybenzoate hydroxylase	1phh	1–164	2·3

The structures were extracted from the Brookhaven Protein Databank (PDB) (Abola *et al.*, 1987; Bernstein *et al.*, 1977). Chain identifiers are shown as the fourth character in the PDB code. Special designators: aspartic proteinases, -n for N-terminal lobes, -c for C-terminal lobes; immunoglobulins, -v for variable domains, -c for constant domains; γ-crystallins, -1, -2, -3, -4, for motifs 1, 2, 3 and 4.

Features are defined at both the residue and the whole protein level. For example, there are residue–residue relationships such as a C^{α}–C^{α} distance and protein–protein relationships such as the fractional sequence identity. There are also residue properties, such as a residue solvent accessibility and protein properties, such as the resolution of an X-ray analysis. Associations among features of two related proteins are crucial for comparative modelling.

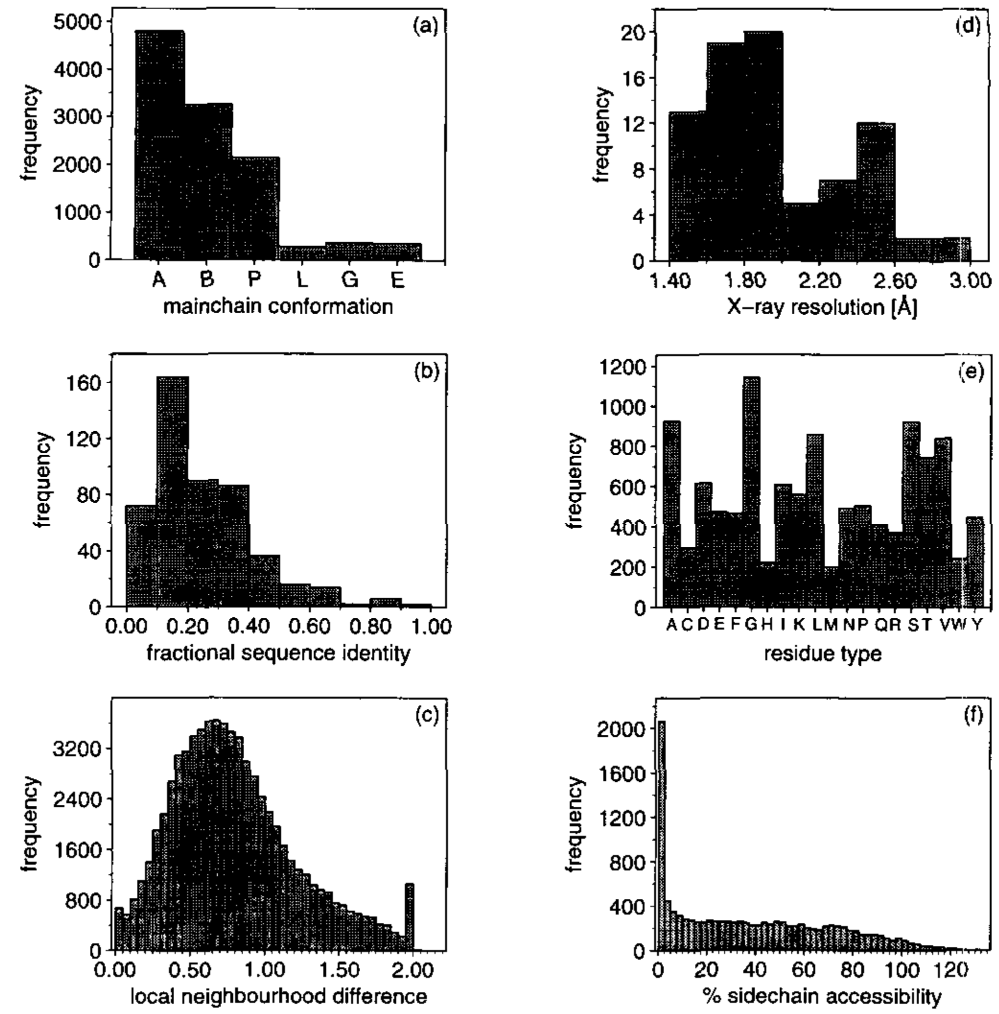

Figure 2. Composition of the local database. Distributions of various features in the local database are shown. (a) Main-chain conformation classes for all residues (Wilmot & Thornton, 1990). (b) Fractional sequence identity score for all pairs of related proteins. (c) Residue neighbourhood difference for all equivalent residue pairs. (d) Resolution of X-ray analysis for all proteins. (e) Residue types for all residues. (f) Fractional side-chain solvent accessibility for all residues.

Consequently, for each feature type MDT distinguishes at least two variants that are associated with a single protein. The first variant is a feature associated with the first protein in a pairwise alignment and the second variant is the same feature associated with the second protein in the alignment. These two proteins are treated as the template and the target in prediction, but at this stage both the structures are known. MDT can also use triple alignments to correlate features from three proteins (see section 3(a)(i) for an application).

The statistical sample for construction of the frequency table **W**′ depends on the type of features that are correlated among themselves. For example, if only a distribution of residue types is wanted, then the sample consists of all amino acid residues in the local database; if the distribution of protein–protein comparison scores is required, then the sample includes all homologous protein pairs in the local database; if C^α–C^α distances are tabulated, then the sample is all intra-molecular residue–residue pairs in the local database; if C^α–C^α distances in one protein are correlated with C^α–C^α distances in another protein, then the sample includes all pairs of equivalent C^α–C^α distances in all homologous pairs of proteins in the database. The MDT program automatically constructs the correct type of the sample from the nature of the features to be tabulated. Sizes of the samples for various types of features, with the current local database, are listed in Table 2. The sample for the combination of features is always the largest of the samples for the individual features in the analysis.

Table 2
Size of the local database

Alignments	17
Protein structures	80
Homologous protein pairs	488
Residues	11,352
Equivalent residue pairs	92,780
Intramolecular residue pairs	1,888,881
Equivalent intramolecular residue pairs	20,270,936

The following is a description of the protein features that can be selected in MDT and were used in this paper. For a summary of these features and their symbols, see Table 3.

Amino acid residue type. Twenty standard amino acid residue types are distinguished. Asx is classified as Asn and Glx as Gln. All residue types other than these 22 are ignored.

Main-chain dihedral angles Φ *and* Ψ. These are computed by the program DIH, following the IUPAC convention (Kendrew *et al.*, 1970).

Secondary structure class of a residue. This coarse definition of main-chain conformation is based on the secondary structure definition by Kabsch & Sander (1983). If the Φ angle is positive, the main-chain conformation is assigned to class +Φ; otherwise, the secondary structure assignments from SSTRUC (written by David Smith) are used to select one of the three remaining classes: helical (Kabsch and Sander codes: H, α-helix; G, 3_{10}-helix; I, π-helix), extended (E, strand in a β-sheet; B, β-bulge; ' ', extended chain) and other (T, turn; S, bend).

Table 3

Features used in this paper that may be selected in MDT to span multi-dimensional frequency tables **W'**

Index	Variable	Feature
1	r	Amino acid residue type
2	Φ	Main-chain dihedral angle Φ
3	Ψ	Main-chain dihedral angle Ψ
4	t	Secondary structure class of a residue
5	M	Main-chain conformation class of a residue
6	α	Fractional content of residues in the main-chain conformation class A
7	χ_i	Side-chain dihedral angle χ_i, $i = 1, 2, 3, 4$
8	c_i	Side-chain dihedral angle χ_i class, $i = 1, 2, 3, 4$
9	a	Residue solvent accessibility
10	$\bar{a}$	Average accessibility of two residues in one protein
11	s	Residue neighbourhood difference between two proteins
12	$\bar{s}$	Average residue neighbourhood difference between two proteins
13	i	Fractional sequence identity between two proteins
14	d	C^{α}–C^{α} distance
15	Δd	Difference between two C^{α}–C^{α} distances in two proteins
16	h	Main-chain N–O distance
17	Δh	Difference between two main-chain N–O distances in two proteins
18	b	Average residue B_{iso}
19	R	Resolution of X-ray analysis
20	g	Distance of a residue from a gap in alignment
21	$\bar{g}$	Average distance of a residue from a gap

The second column lists the variable names that are used for the features in pdfs and in frequency tables **W'**. Features that are not associated with 2 proteins can be used independently for 2 related proteins in a pairwise alignment or for 3 related proteins in a triple alignment. For example, a 2D table can be constructed that is spanned by a residue type r in one protein and a residue type r' at the equivalent position in a related protein; the prime is generally used to designate that the feature is from the second protein and 2 primes that it is from the third protein. The Δ symbol refers to the difference between features f and f': $\Delta f = f - f'$. O, N, oxygen and nitrogen atoms in the main-chain peptide group; B_{iso}, atomic isotropic temperature factor.

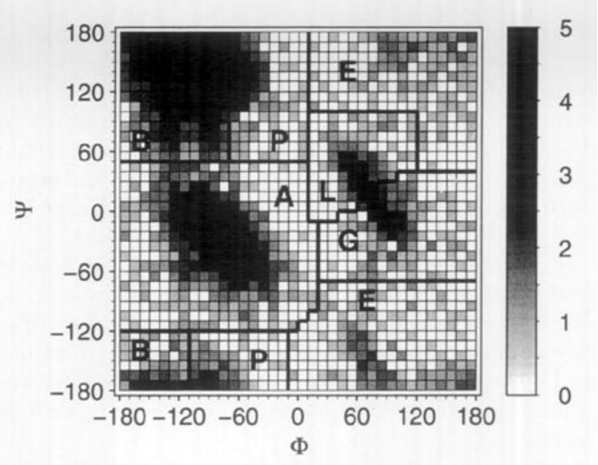

Figure 3. Definition of the main-chain conformation classes. The plot shows $W'(\Phi, \Psi)$ determined from the local database of protein structures. It is divided into 10° × 10° squares. The areas corresponding to the 6 characteristic peaks are delimited by thick lines. These areas define the residue main-chain conformation classes A, B, P, L, G and E. The scale on the right corresponds to $\ln[W'(\Phi, \Psi) + 1]$.

Main-chain conformation class of a residue. A convenient way of describing the residue main-chain conformation is by the pair of the main-chain dihedral angles (Φ, Ψ). For all residues except Pro, those two degrees of freedom cover most of the conformation space allowed to the main-chain. The third main-chain dihedral angle, ω, has a monomodal distribution with a mean of 180° and a small standard deviation of approximately 6°. Pro is the only exception where the ω distribution is bimodal, with the second maximum at 0° corresponding to the *cis* conformation. This *cis* conformation in some Pro residues is ignored in the definition of the main-chain conformation class.

The distribution of (Φ, Ψ) pairs obtained from the protein structures in the Brookhaven Protein Databank shows six different peaks (Fig. 3, Wilmot & Thornton, 1990). They correspond to right-handed α-helix (A), idealized β-strand (B), polyproline conformation (P), the ε region accessible primarily to Gly residues with positive Φ angle (G),

Table 4

Parameters of the main-chain conformation classes

	Mean (°)		Standard deviation (°)	
	$\bar{\Phi}_i$	$\bar{\Psi}_i$	$\sigma_i(\Phi)$	$\sigma_i(\Psi)$
A	−65	−41	15	15
B	−130	135	15	20
P	−65	140	15	15
G	60	40	10	10
L	90	−10	15	10
E	130	180	25	25

Approximate mean values and standard deviations are given for each of the 6 main-chain conformation classes.

left-handed α-helix (L), and extended conformation (E). These six peaks represent six different conformation states of the main-chain of a particular residue. The whole of the plot can therefore be divided into six areas centred at those peaks (Fig. 3). The approximate mean values and standard deviations of the main-chain dihedral angles in each of these six classes are listed in Table 4.

Fractional content of residues in the main-chain conformation class A. This feature is introduced to describe the structural classes of proteins that differ in the relative contents and distribution of helices and strands ($\alpha, \beta, \alpha/\beta, \alpha+\beta$). It is defined as the fractional content of residues in the main-chain conformation class A.

Side-chain dihedral angles χ_1, χ_2, χ_3 and χ_4. The IUPAC definition of side-chain dihedral angles (Kendrew *et al.*, 1970) is adopted with the exception of the χ_2 angle of the residues with C_2 symmetry around the C^β–C^γ bond (Asp, Phe and Tyr). In these cases, the χ_2 dihedral angle is calculated using the atom that happened to be labelled $C^{\delta 1}$ in the Brookhaven file; if the angle is smaller than 0°, it is increased by 180°. Similar considerations apply to χ_3 of Glu. Thus, the absence of any significant difference between the angles $-\alpha$ and $-\alpha+180°$, i.e. the C_2 symmetry, is properly reflected in the equality of the χ angles for the two cases. Pro residues are considered to have a rigid side-chain with χ_1 angle of 29° and are neglected in this analysis.

Classes of side-chain dihedral angles χ_1, χ_2, χ_3 and χ_4. It is well known that the χ_i angles for most residue types follow trimodal distribution with peaks at approximately 60°, 180° and −60° (Janin *et al.*, 1978; Ponder & Richards, 1987). These distributions can be used to assign a particular side-chain dihedral angle to one of the three classes, +, *t* and −, respectively, grouping all the values in one peak to one class (Table 5). There are exceptions that do not follow this trimodal distribution; they include the χ_2 angles with a trigonal C^γ atom that are distributed bimodally (His, Trp) and monomodally (Phe, Tyr, Asp), as well as the trimodal distribution of Asn with unique positions of the maxima (Table 5; Šali, 1991).

Residue solvent accessibility. Absolute or fractional contact areas (Richmond & Richards, 1978) for the side-chain, main-chain or a whole residue can be used. These areas are calculated by the program PSA as described by Hubbard & Blundell (1987). The fractional area is obtained by dividing the contact area of a given residue by the standard contact area of the corresponding residue type X in the extended tripeptide Gly-X-Gly.

Difference between two equivalent residue neighbourhoods in two proteins. The difference between two equivalent residues, one from each protein, is defined as follows. First, all neighbours of the residue in structure *A* are found. Then the equivalent residues in structure *B* are obtained using the alignment between the two proteins. The sum of the residue–residue dissimilarity scores is then calculated for these pairs of equivalent residues. Dissimilarity scores from the residue type matrix of COMPARER were used (Šali & Blundell, 1990). Next, a gap penalty (usually 2) is added to this sum

Table 5

Definition of the χ_1, χ_2, χ_3 and χ_4 side-chain dihedral angle classes

χ type	Residue types	Class	Range (°)	Mean $\bar{\chi}_i$ (°)	Standard deviation $\sigma(\chi_i)$ (°)
χ_1	C, D, E, F, H, I, K, L, M, N, Q, R, S, T, V, W, Y	+	[0, 120]	63	10
		t	[120, 240]	180	10
		−	[−120, 0]	−63	10
χ_2	E, I, K, L, M, Q, R	+	[0, 120]	65	10
		t	[120, 240]	180	10
		−	[−120, 0]	−65	10
	D	1	[0, 180]	0	25
	N	1	[−180, 0]	−50	10
		2	[0, 80]	10	10
		3	[80, 180]	140	10
	H, W	1	[−180, 0]	−75	10
		2	[0, 180]	75	10
	F, Y	1	[0, 180]	75	10
χ_3	K, M, R, Q	+	[0, 120]	65	10
		t	[120, 240]	180	10
		−	[−120, 0]	−65	10
	E	1	[35, 85]	60	15
		2	[85, 395]	180	35
χ_4	K	+	[0, 120]	65	15
		t	[120, 240]	180	15
		−	[−120, 0]	−65	15
	R	+	[0, 70]	45	10
		t	[70, 255]	170	35
		−	[−105, 0]	−80	10

Approximate mean values and standard deviations for each peak are also given. See Šali (1991) for histograms of χ_i for each residue type that result in these definitions of the classes.

where there is a deletion in structure B; this defines the total score. The difference between the residue neighbourhoods of the two proteins is then calculated by dividing the total score by the number of residues in the neighbourhood of the residue in protein A. If there are no neighbours in protein A, the difference score is 0. Three parameters have to be specified to define precisely the neighbours of a residue in protein A: the atoms used to search for contacts, the cutoff for the contact distance, and the number of atom–atom contacts required to have a residue–residue contact. Usually, all the atoms of a residue, a contact distance of 6·0 Å and one atom–atom contact are used.

Average residue neighbourhood difference between two proteins. This quantity measures the difference between two proteins in the neighbourhoods of the two residues that span a certain distance. It is the average of the individual residue neighbourhood differences defined above.

Fractional sequence identity between two proteins. This relationship is defined as the number of pairs of identical residues in the alignment divided by the length of the shorter protein sequence.

Difference in two C^α–C^α distances in two proteins. This is a double relationship: it is defined as the difference between the equivalent distances in two related proteins. The distances are equivalent only if both the first and the second atom defining the distance in the first protein are equivalent to the first and the second atom of the distance in the second protein. A similar difference is defined for the distance between the main-chain N and O atoms.

Average residue isotropic temperature factor, B_{iso}. B_{iso} values are read from the co-ordinate file when present. The average is calculated for side-chain atoms only; C^α atom is included in the side-chain to obtain a value for Gly. If there are no B_{iso} values for a certain residue in the co-ordinate file, the residue is assigned the "undefined" value. This feature separates well-defined parts of a molecule from flexible or undetermined parts.

Resolution of the X-ray analysis. This is obtained from the co-ordinate file as distributed by the Brookhaven Protein Databank.

Distance of a residue from a gap. For each residue in the pairwise alignment, this is defined as the number of positions from this residue to the closest gap. For a residue aligned with a gap, it is 0; for a residue next to a gap, it is 1, etc. Since the structure varies more when closer to a gap in the alignment, this feature can be used to describe the structural variability.

Average distance of an intra-molecular residue pair from a gap. It is defined as the average of the two residue distances from the closest gaps. This feature can be used to describe the variability of the distance spanned by the two residues.

(c) *Calculation of a probability distribution from a sparse data set*

Calculation of probabilities **W** from frequencies **W′** (eqn (4)) is strictly valid only when the frequencies are very large. Unfortunately, the limited size of the database usually results in a sparse matrix **W′** and, consequently, in significant errors in some of the estimated probabilities. This section describes the procedure that can reduce the problem of a sparse data set.

We extend the treatment of sparse data sets that was proposed by Sippl (1990). The original method is briefly described as follows. Suppose we are interested in a discrete probability distribution of a random variable x. We can imagine calculation of the probability distribution in a stepwise process, as an alternative to equation (4). Each step is associated with a single measurement that contributes one data point to **W′**. We start with an *a priori* probability distribution, such as a uniform distribution, and then allow every measurement to perturb this *a priori* distribution so as to make the measured event slightly more likely. The resulting perturbed distribution is used as the *a priori* distribution for the next perturbation step. Exhausting all the measurements leads to the final estimated probability distribution **p**. This iterative procedure can be written as a single operation (Sippl, 1990):

$$p(x_i) = \omega_1 A(x_i) + \omega_2 W(x_i) \qquad (5)$$

with weights

$$\omega_1 = \frac{1}{1+N/(\sigma n_x)} \quad \text{and} \quad \omega_2 = 1-\omega_1, \qquad (6)$$

where N is the number of points in **W′**, n_x is the number of bins, i.e. $i = 1, 2, \ldots, n_x$, and σ is a parameter that determines the relative contributions of the *a priori* distribution **A** and a measured distribution **W** to the final estimate **p**. When N is large, the *a priori* distribution has a negligible effect on the smoothed probability distribution **p**; in this case, **p** is determined by experiment alone and equations (5), (6) are equivalent to equation (4). On the other hand, when N is small, the estimate **p** depends on the *a priori* distribution as well as on the experiment. Thus, an estimate of **p** is robust since the sparseness of data cannot introduce large deviations from the *a priori* distribution; it is also unbiased since, for a small N, the *a priori* distribution is unbiased and, for a large N, **p** is reliably determined by experiment alone. When the average number of data points per bin is σ, the *a priori* and experimental distributions contribute equally to the smoothed estimate **p**. This elegant smoothing procedure was successfully used to derive robust potentials of mean force (Sippl, 1990).

It is possible to extend this idea to multi-dimensional frequency tables used here. For example, suppose we want to obtain probabilities for the occurrence of any of the three classes of the χ_1 dihedral angle $(+, t, -)$ in proteins. Additionally, we want to obtain this distribution for each residue type separately. In other words, we want to estimate the conditional probability distribution $p(\chi_i/r_j)$, where χ_i and r_j stand for the class i of a side-chain dihedral angle (3 values) and residue type j (20 values), respectively. According to equation

(5), the elements i of the smoothed conditional distribution for each residue type j are:

$$p(\chi_i/r_j) = \omega_1^j A(\chi_i/r_j) + \omega_2^j W(\chi_i/r_j). \qquad (7)$$

The only problem now is to find the best *a priori* distribution **A**. One possibility is to use a uniform distribution. However, the distribution of χ_1 angles irrespective of the residue type is, on the average, a better *a priori* distribution than a uniform distribution. So we set:

$$A(\chi_i/r_j) = p(\chi_i) \qquad (8)$$

for all j. $p(\chi_i)$ could be determined directly from the observed frequencies **W′** (eqn (4)), but a more robust estimate is obtained by applying the smoothing operator again:

$$p(\chi_i) = \omega_1 A(\chi_i) + \omega_2 W(\chi_i) \qquad (9)$$

with a proper set of weights ω_i and with the uniform distribution as the *a priori* distribution, i.e. $A(\chi_i) = 1/n_x$. Now we have all the quantities to calculate $p(\chi_i/r_j)$.

The following is a rigorous definition of the smoothing procedure for the relative frequency table **W′** spanned by one dependent and $N-1$ independent variables. There are N stages which can be seen as a build-up of the final $\mathbf{p}^N$ starting with $\mathbf{p}^1$ or, alternatively, as a recursive evaluation of $\mathbf{p}^N$ from $\mathbf{p}^{N-1}$, $\mathbf{p}^{N-2}$, etc. y_i stands for any of the possible values of the variable y, and y_i^j for the ith value of the variable y_j. Therefore, $p(x_i/y_j)$ is a scalar, $p(x/y_j)$ is a vector, and $p(x/y)$ is a matrix.

Each stage n consists of the following operation:

$$\begin{aligned} p^n(x_i/y_j^1, y_k^2, \ldots, y_l^{n-1}) = {} & \omega_1^{jk\ldots l} A^n(x_i/y_j^1, y_k^2, \ldots, y_l^{n-1}) \\ & + \omega_2^{jk\ldots l} W^n(x_i/y_j^1, y_k^2, \ldots, y_l^{n-1}), \end{aligned} \qquad (10)$$

where p^n is a conditional probability for x_i, given that the independent variables assume values $y_j^1, y_k^2, \ldots, y_l^{n-1}$. W^n is the corresponding conditional probability determined from the observed frequencies **W′** using equation (4), more explicitly written here as:

$$W^n(x_i/y_j^1, y_k^2, \ldots, y_l^{n-1}) = \frac{\sum_{y^n, y^{n+1}, \ldots, y^{N-1}} W'^n(x_i, y_j^1, y_k^2, \ldots, y_l^{N-1})}{\sum_{x, y^n, y^{n+1}, \ldots, y^{N-1}} W'^n(x_i, y_j^1, y_k^2, \ldots, y_l^{N-1})}. \qquad (11)$$

The *a priori* distribution $\mathbf{A}^n$ is calculated as a weighted average of all p^{n-1}:

$$A^n(x_i/y_j^1, y_k^2, \ldots, y_l^{n-1}) = \sum_c \rho_c p_c^{n-1}; \quad \sum_c \rho_c = 1, \qquad (12)$$

where the sum runs over all possible combinations of $n-2$ independent variables with their values set to a subset of $y_j^1, y_k^2, \ldots, y_l^{n-1}$. If there are $N-1$ independent variables in total, there are $(N-1)!/\{[(N-1)-(n-2)]!(n-2)!\}$ different combinations of $n-2$ variables, each combination corresponding to one term of the sum in equation (12). The following rationale was used to calculate weights ρ_c. Suppose the independent variables $y_j^1, y_k^2, \ldots, y_l^{n-2}$ in a particular $\mathbf{p}^{n-1}$ do not provide any additional information about the value of x. Then this $\mathbf{p}^{n-1}$ should not be used at all in the calculation of $\mathbf{A}^n$. Conversely, the more x depends on the variables in $\mathbf{p}^{n-1}$, the greater should be the weight ρ_c for this $\mathbf{p}^{n-1}$ in the calculation of $\mathbf{A}^n$. A convenient measure of the dependence of x on the independent variables is the entropy of $\mathbf{p}^{n-1}$, defined as:

$$\begin{aligned} S_c(\mathbf{p}^{n-1}) = {} & -\sum_i p^{n-1}(x_i/y_j^1, y_k^2, \ldots, y_l^{n-2}) \\ & \times \log p^{n-1}(x_i/y_j^1, y_k^2, \ldots, y_l^{n-2}). \end{aligned} \qquad (13)$$

The maximum of entropy, $S_{\max} = \log n_x$, corresponds to a uniform distribution. In such a case, independent variables contain no information about x. Therefore, ρ_c can be defined as:

$$\rho_c = \frac{(S_{\max} - S_c)}{\sum_c (S_{\max} - S_c)}. \qquad (14)$$

The iterative smoothing procedure is started by setting $n = 1$ and $\mathbf{A}^0$ to a uniform distribution, $\mathbf{A}^0(x_i) = 1/n_x$. It is finished when $n = N$.

Performance of the smoothing procedure in one simple application is illustrated in Figure 4. The aim is to find the distribution of a χ_1 side-chain dihedral angles for each residue type; the Cys and Ser residues are used as two typical cases.

First, a small database of only eight proteins (1245 residues with χ_1 dihedral angles) was used to derive the distribution of the χ_1 dihedral angle for the Cys residue (Fig. 4(a)). Since only 11 Cys residues exist in this small database, the distribution is very sparse and large errors are likely to occur. On the other hand, when the whole local database of 80 proteins is used for the same purpose, a dense and reliable distribution consisting of 297 Cys residues is obtained (Fig. 4(c)). These two distributions can be compared with the distribution smoothed as described above (Fig. 4(b), $\sigma = 5$). The comparison clearly shows the improvement due to the smoothing procedure. The improvement was possible because Cys has a typical χ_1 distribution. Consequently, information provided by other residues improves the estimate of the Cys distribution.

The second set of histograms (Fig. 4(d) to (f)) shows the corresponding distributions for the Ser residue. This example shows that even when the distribution to be determined (Fig. 4(f)) is very different from the average distribution (approximated by Fig. 4(c)), the smoothing procedure performs satisfactorily, provided that the original dataset (Fig. 4(d)) is not too small.

The third kind of situation is when an atypical distribution has to be obtained from a sparse data set. Since this would require a creation of information from nothing, as opposed to the best usage of the available information, it is clear that no procedure exists for such a task.

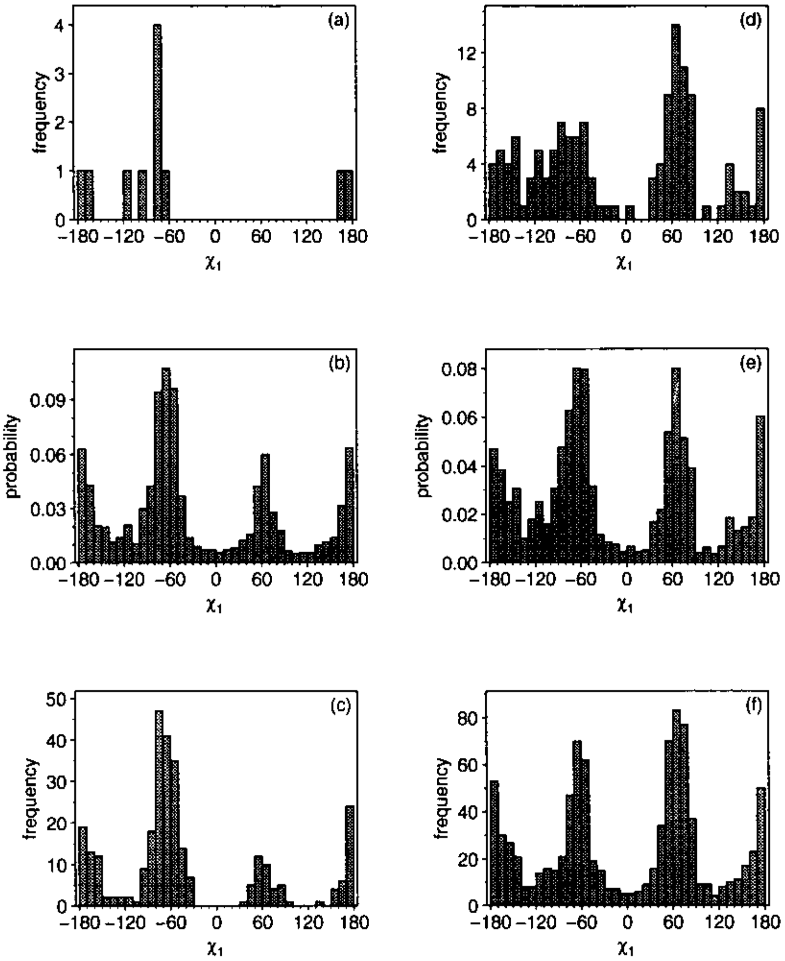

Figure 4. Smoothing sparse distributions of χ_1 side-chain dihedral angles. (a) A distribution of χ_1 angles of 11 Cys residues obtained from a small database of 8 proteins, totalling 1245 residues with χ_1 side-chain dihedral angles. The proteins are 4hvp, 2rspa, 4ape-n, 2app-n, 2apr-n, 3cms-n, 5pep-n, 2apr-c (see Table 1 for the protein names). (b) The smoothed Cys distribution obtained from the small database of 8 structures. The smoothing parameter σ was 5. (c) The "reliable" non-smoothed distribution of Cys χ_1 was obtained from the whole local database of 80 protein structures. This distribution consists of 297 Cys residues. (d) 138 Ser residues in the small database. (e) The smoothed Ser distribution. (f) 923 Ser residues in the large database.

(d) *Strength of associations among the features of protein structure*

We distinguish between the significance of an association between two or more features and the strength of such an association. An association may be significant if it is based on a large amount of data, yet still weak if the values of independent features do not provide strong restraints on the dependent feature. The significance is measured by the χ^2 test, while the strength is measured by the entropy of the conditional pdf (Press *et al.*, 1986). We also distinguish between accuracy and precision. A prediction is precise if the differences between independent realizations of this prediction are small, yet it may still be inaccurate if the mean is very different from the true value.

The most useful pdf for modelling is that which predicts the unknown feature most accurately. Provided that pdf values are not constructed from a sparse and non-representative database, the most precise pdf is on the average also the most accurate pdf; therefore, the most accurate pdf is the pdf with the sharpest shape. A quantitative measure of sharpness is entropy of pdf:

$$S[p(x)] = -\int_{-\infty}^{\infty} p \ln p \, \mathrm{d}x, \qquad (15)$$

and for a discrete probability function:

$$S[p(x)] = -\sum_i p(x_i) \ln p(x_i). \qquad (16)$$

For a discrete conditional probability function, entropy is defined similarly as:

$$S[p(x/a, b, \ldots, c)] = \sum_{a,b,\ldots,c} p(a, b, \ldots, c) S[p(x/a, b, \ldots, c)]. \qquad (17)$$

Thus, to find the best known features $(a, b, \ldots, c)$ for prediction of the unknown feature x, we search for the features that minimize entropy S of a corresponding conditional pdf. A convenient measure of how much the independent features determine the dependent feature is given by the uncertainty coefficient of x (Press *et al.*, 1986):

$$U(x/a, b, \ldots, c) = \frac{S[p(x)]-S[p(x/a, b, \ldots, c)]}{S[p(x)]}. \quad (18)$$

This measure lies between 0 and 1. The value 0 means that x is not associated with $(a, b, \ldots, c)$, and the value 1 implies that $(a, b, \ldots, c)$ completely determines x.

(e) *Stereochemical restraints*

All stereochemical restraints are easily obtained from the amino acid sequence of a protein. Stereochemical restraints used here include bond distances, bond angles, planarity of peptide groups and side-chain rings, chiralities of C^α atoms and side-chains, van der Waals contact distances and the bond lengths, bond angles and dihedral angles of cystine disulphide bridges.

The mean values and standard deviations for bond lengths, bond angles and dihedral angles are obtained from the GROMOS86 IFP37C4 parameter set (Berendsen *et al.*, 1984) which, in turn, are derived from the values found in small molecules from X-ray crystallography, spectroscopic studies and theoretical calculations. The van der Waals radii are also obtained from the atomic radii in the GROMOS parameter set by multiplication with a constant factor (usually 0·82).

(i) *Bond lengths, bond angles and dihedral angles*

The classical harmonic model for the bond length between two atoms gives the vibrational potential energy of the bond as:

$$E(b) = \tfrac{1}{2}c(b-b_o)^2. \quad (19)$$

The probability density function for the bond length is then found, from classical statistical mechanics, to be a Gaussian probability density function (Hill, 1960):

$$p^b(b) = \frac{1}{\sigma_b\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b-\bar{b}}{\sigma_b}\right)^2\right] = N(\bar{b}, \sigma_b). \quad (20)$$

where $\sigma_b = \sqrt{kT/c}$ and $\bar{b} = b_o$. Only two parameters, the mean ($\bar{b}$) and the standard deviation (σ_b), are needed to describe this distribution.

The derivation of the pdf for the bond angle is equivalent to that for the bond length. The final result is again a Gaussian pdf $p^\alpha(\alpha) = N(\bar{\alpha}, \sigma_\alpha)$. Similarly, a monomodal pdf for a torsional or an improper dihedral angle ξ is $N(\bar{\xi}, \sigma_\alpha)$. The monomodal pdf $N(\bar{\xi}, \sigma_\xi)$ is used to restrain several different features of protein structure, following the use of a harmonic potential for restraining improper dihedral angles in GROMOS (Berendsen *et al.*, 1984). They include peptide and ring planarities, and chiralities of C^α atoms, Thr and Ile side-chains.

To allow for the *cis*-peptide conformation, the main-chain dihedral angle ω may be restrained by a sum of two Gaussian functions:

$$p^\omega = \omega_1 \cdot N(180°, 6°) + w_2 . N(0°, 6°); \; w_1 + w_2 = 1. \quad (21)$$

If not specified otherwise, weight w_2 is set to 0, corresponding to the *trans* conformation.

(ii) *van der Waals repulsion*

van der Waals repulsion is the only stereochemical feature which is not described by the harmonic model. Instead, the following pdf is used for two atoms:

$$p^v(d) = c \cdot \begin{cases} N(d_o, \sigma_w); & d \leq d_o \\ \dfrac{1}{\sigma_w\sqrt{2\pi}}; & d_o < d < d_{\max}, \end{cases} \quad (22)$$

where d is the distance between the two atoms, d_o is the sum of their van der Waals radii and σ_w is the standard deviation of the Gaussian part of the whole pdf (usually 0·05 Å). $d_{\max}$ is the maximal possible linear dimension of a protein and constant c is chosen so that $p^v(d)$ integrates to 1. This pdf does not differentiate between contact distances larger than d_o, but it does select against distances smaller than d_o. This is achieved by imposing a repulsive harmonic potential on atoms that are less than d_o apart.

(iii) *Disulphide bonds*

The disulphide bond restraints reflect only the generally allowed stereochemistry of disulphide bridges (Thornton, 1981); they do not automatically take into account the geometry of equivalent disulphide bonds in related proteins.

First, disulphide bonded pairs of Cys residues in the sequence to be modelled have to be specified. Then the geometry of the disulphides is restrained by the mean values and standard deviations of Gaussian pdfs for distances and angles taken from the GROMOS IFP37C4 force field. Additionally, the pdf for the C^β–S–S–C^β dihedral angle, which has been obtained from the disulphide bonds in high resolution protein structures (Thornton, 1981), is used. The distribution of this dihedral angle is bimodal with peaks at −87·1° and 93·9° with standard deviation of 10°. Correspondingly, the dihedral angle is modelled by a weighted sum of the two Gaussians with both weights equal to 0·5 if not specified otherwise.

(f) *Restraining a distance between two C^α atoms*

The unknown feature is defined as the difference between two equivalent C^α–C^α distances, $d-d'$; d' is from the "known" or template structure and d from the "unknown" or target structure. As described in general terms in subsection (b), (iii), above, the local database and the MDT program are used to find the distribution of $d-d'$ as a function of four independent variables: the corresponding C^α–C^α distance in the known structure (d'), the fractional sequence identity of the two aligned sequences (i), the average of the fractional solvent accessibilities of the two residues spanning the distance in the known

Table 6

Derivation of the distributions for the C^α–C^α distance and the main-chain N–O distance

Feature	Symbol	Start	End	Interval	No. of intervals
Average distance from a gap	$\bar{g}$	0	20	[a]	6
Fractional sequence identity	i	0	1	0·2	5
Average residue accessibility (%)	$\bar{a}'$	0	100	20	5
Distance (Å)	d', h'	5	30	5	5
Distance difference (Å)	$d-d'$, $h-h'$	−7·0	7·0	0·5	28

Ranges used for the tabulation of the distributions $W(d-d'/\bar{g}, i, \bar{a}', d')$ and $W(h-h'/\bar{g}, i, \bar{a}', h')$ are shown.

[a] The intervals for $\bar{g}$ are 0 to 2, 3, 4 to 5, 6 to 9, 10 to 16, 17 to 24.

structure ($\bar{a}'$), and the average distance from a gap of the residues spanning the distance ($\bar{g}$). Details about the values of these variables are given in Table 6.

Dependence of the C^α–C^α distance difference on the other four variables is quantified in Table 7. The large decrease in entropy from 4·372 to 2·620 when going from $p(d)$ to $p(d-d')$ is a result of the similarity between related structures (Table 7). A further decrease of entropy to 2·479 when the four independent variables are added to $p(d-d')$ is small but still significant. The strongest influence on the distance difference by a single feature is exerted by the fractional sequence identity and the average gap distance, but the accessibility and the magnitude of the distance are also important.

Examples of the histograms of probability distributions obtained by the MDT program are shown in Figure 5(a) and (b). These histograms demonstrate that the conditional distribution of the distance differences may be approximated by a Gaussian function with a mean of zero and a standard deviation dependent on the values of the independent variables. Therefore, the pdf restraining a C^α–C^α distance in the sequence of an unknown, given an alignment with a single related known structure, can be modelled as:

$$p^d(d/\bar{g}, i, \bar{a}', d') = \frac{1}{\sigma(\bar{g}, i, \bar{a}', d')\sqrt{2\pi}} \times \exp\left[-\frac{1}{2}\left(\frac{d-d'}{\sigma(\bar{g}, i, \bar{a}', d')}\right)^2\right]$$

$$\begin{aligned}\sigma(\bar{g}, i, \bar{a}', d') = {} & \alpha_1+\alpha_2\bar{g}+\alpha_3 i+\alpha_4\bar{a}'+\alpha_5 d' \\ & +\alpha_6\bar{g}^2+\alpha_7\bar{g}i+\alpha_8\bar{g}\bar{a}'+\alpha_9\bar{g}d' \\ & +\alpha_{10}i^2+\alpha_{11}i\bar{a}'+\alpha_{12}id'+\alpha_{13}\bar{a}'^2 \\ & +\alpha_{14}\bar{a}'d'+\alpha_{15}d'^2+\alpha_{16}\bar{g}^3+\alpha_{17}\bar{g}^2 i \\ & +\alpha_{18}\bar{g}^2\bar{a}'+\alpha_{19}\bar{g}^2 d'+\alpha_{20}\bar{g}i^2 \\ & +\alpha_{21}\bar{g}i\bar{a}'+\alpha_{22}\bar{g}id'+\alpha_{23}\bar{g}\bar{a}'^2 \\ & +\alpha_{24}\bar{g}\bar{a}'d'+\alpha_{25}\bar{g}d'^2 \\ & +\alpha_{26}i^3+\alpha_{27}i^2\bar{a}'+\alpha_{28}i^2 d'+\alpha_{29}i\bar{a}'^2 \\ & +\alpha_{30}i\bar{a}'d'+\alpha_{31}id'^2+\alpha_{32}\bar{a}'^3 \\ & +\alpha_{33}\bar{a}'^2 d'+\alpha_{34}\bar{a}'d'^2+\alpha_{35}d'^3. \quad (23)\end{aligned}$$

Table 7

Strength of associations between the dependent and independent features in the distributions of the C^α–C^α and main-chain N–O distances

	$x=d$		$x=h$	
pdf	Entropy S	Conditional entropy U	Entropy S	Conditional entropy U
Uniform $p(x)$[a]	5·298		5·298	
$p(x)$	4·372		4·333	
Uniform $p(x-x')$[b]	3·332		3·332	
$p(x-x')$	2·620	0·0000	2·657	0·0000
$p(x-x'/i)$	2·539	0·0301	2·582	0·0286
$p(x-x'/\bar{g})$	2·568	0·0190	2·605	0·0195
$p(x-x'/\bar{a})$	2·605	0·0047	2·646	0·0041
$p(x-x'/x')$	2·600	0·0067	2·643	0·0054
$p(x-x'/\bar{g}, i, \bar{a}', x')$	2·479	0·0527	2·527	0·0493

[a] A uniform distribution with the same number of bins (200) as the $p(x)$ distributions. These distributions also have the same bin width (0·5 Å) as the $p(x-x'/a, b, \ldots)$ distributions.

[b] A uniform distribution with the same number of bins (28) as the $p(x-x'/a, b, \ldots)$ distributions. The entropies (eqn (17)) and conditional entropies (eqn (18)) for the non-smoothed pdfs are shown. The smoothed pdfs ($\sigma=20$) have similar values; the largest difference occurs for the most sparse W:$S[p(d-d'/\bar{g}, i, \bar{a}', d')]=2·398$ and 2·479 for the smoothed and non-smoothed pdf, respectively. This small difference implies that the pdfs can be determined accurately from the non-smoothed histograms because the database is not sparse.

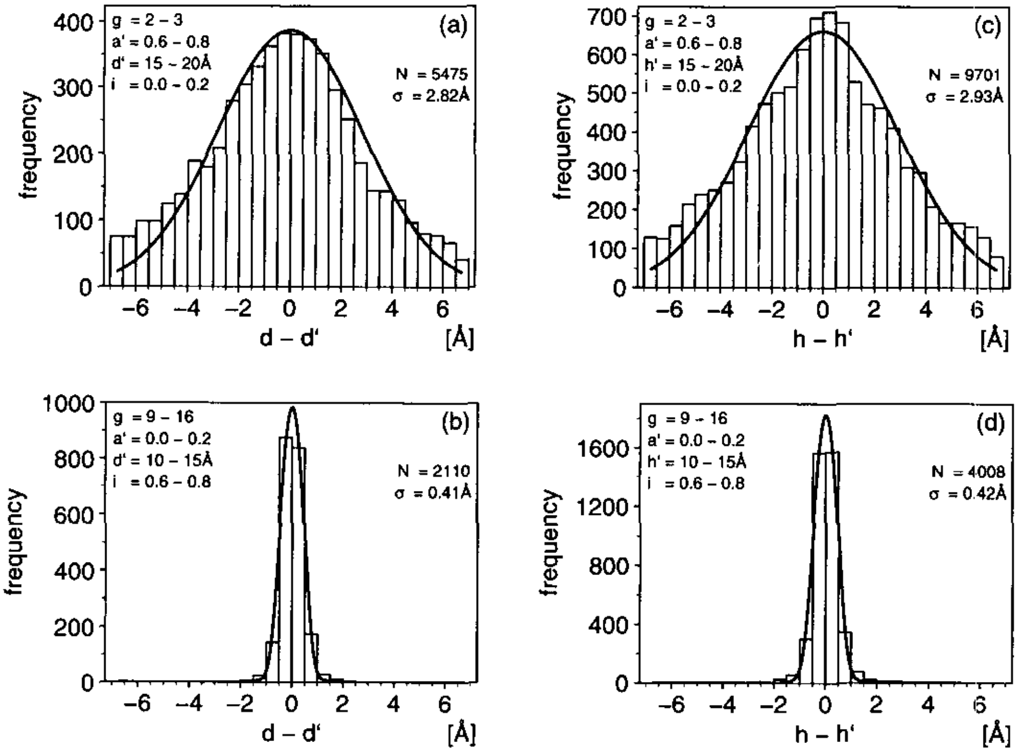

Figure 5. Distribution of the differences between 2 equivalent distances. The histograms show the frequency of the differences between 2 equivalent distances as observed by MDT in the local database. (a) and (b), C^α–C^α distances; (c) and (d), main-chain N–O distances. The curves are fitted Gaussian models (eqns (23) and (24)). The values of the dependent variables, the number of C^α–C^α distances in the database (N) and the standard deviation of the Gaussian model (σ) are shown for each histogram.

If $\bar{g} > 20$, $\bar{g}$ is reset to 20. In relation to equation (23), the four features can be seen as the measure for the degree of transferability of the distance from the known to the unknown structure; the distance in the unknown is more likely to be closer to the equivalent distance in the known when the distance is short, the two residues spanning the distance are buried, the two structures are similar overall, and the residues are distant from the gaps in the alignment.

The remaining problem is to determine the best values of parameters α_i. This is achieved by least-squares fitting (subsection (b)(i), above) the model p^d in equation (23) to the histograms **W** obtained from the database scan (Table 8). The Gaussian conditional pdfs $p^d(d/\bar{g}, i, a, d')$, calculated from the least-squares parameters, are superposed on the experimental histograms in Figure 5(a) and (b). These plots provide additional graphical evidence that the Gaussian model can describe the association between the unknown C^α–C^α distance and the four independent variables included in this analysis.

Table 8

The best parameters for restraining C^α–C^α distances (d) *and main-chain N–O distances* (h)

$$\begin{aligned}\sigma(\bar{g}, i, \bar{a}', d') = {} & 0{\cdot}849 - 2{\cdot}033\bar{g} - 1{\cdot}227i + 0{\cdot}971\bar{a}' + 1{\cdot}467d' + 1{\cdot}382\bar{g} + 1{\cdot}539\bar{g}i - 0{\cdot}504\bar{g}\bar{a}' - 0{\cdot}259\bar{g}d' \\ & + 2{\cdot}412i^2 - 1{\cdot}496i\bar{a}' - 3{\cdot}094id' - 0{\cdot}425\bar{a}'^2 + 0{\cdot}670\bar{a}'d' - 0{\cdot}159d'^2 - 0{\cdot}307\bar{g}^3 - 0{\cdot}213\bar{g}^2i \\ & + 0{\cdot}088\bar{g}^2\bar{a}' + 0{\cdot}020\bar{g}^2d' - 0{\cdot}969\bar{g}i^2 + 0{\cdot}453\bar{g}i\bar{a}' + 0{\cdot}177\bar{g}id' - 0{\cdot}058\bar{g}\bar{a}^2 - 0{\cdot}042\bar{g}\bar{a}'d' \\ & + 0{\cdot}020\bar{g}d'^2 - 0{\cdot}847i^3 + 0{\cdot}055i^2\bar{a} + 1{\cdot}546i^2d' + 0{\cdot}527i\bar{a}'^2 - 0{\cdot}220i\bar{a}'d' + 0{\cdot}254id'^2 + 0{\cdot}066\bar{a}'^3 \\ & + 0{\cdot}153\bar{a}^2d' - 0{\cdot}153\bar{a}'d'^2 - 0{\cdot}0019d'^3 \end{aligned} \tag{36}$$

$$\begin{aligned}\sigma(\bar{g}, i, \bar{a}', h') = {} & 0{\cdot}957 - 2{\cdot}044\bar{g} - 1{\cdot}078i + 0{\cdot}995\bar{a}' + 1{\cdot}477h' + 1{\cdot}572\bar{g}^2 + 1{\cdot}148\bar{g}i - 0{\cdot}525\bar{g}\bar{a}' - 0{\cdot}483\bar{g}h' \\ & + 1{\cdot}505i^2 - 0{\cdot}655i\bar{a}' - 2{\cdot}849ih' - 0{\cdot}625\bar{a}'^2 + 0{\cdot}499\bar{a}'h' - 0{\cdot}126h'^2 - 0{\cdot}369\bar{g}^3 - 0{\cdot}243\bar{g}^2i \\ & + 0{\cdot}121\bar{g}^2\bar{a}' + 0{\cdot}067\bar{g}^2h' - 0{\cdot}592\bar{g}i^2 + 0{\cdot}346\bar{g}i\bar{a}' + 0{\cdot}276\bar{g}ih' - 0{\cdot}032\bar{g}\bar{a}^2 - 0{\cdot}061\bar{g}\bar{a}'h' + 0{\cdot}036\bar{g}h'^2 \\ & - 0{\cdot}329i^3 - 0{\cdot}318i^2\bar{a} + 1{\cdot}472i^2h' + 0{\cdot}284i\bar{a}'^2 - 0{\cdot}293i\bar{a}'h' + 0{\cdot}198ih'^2 + 0{\cdot}382\bar{a}'^3 + 0{\cdot}110\bar{a}^2h' \\ & - 0{\cdot}079\bar{a}'h'^2 - 0{\cdot}0095h'^3 \end{aligned} \tag{37}$$

Full expressions for the standard deviations of the Gaussian p models for **W** (eqns (23) and (24)) are given. Before using $\bar{g}$, $\bar{a}'$ and i with the parameters shown, they have to be scaled by 0·1, 0·01 and 0·1, respectively. The r.m.s. deviations between the p models and **W**′ are 0·0524 and 0·0441 for d and h, respectively.

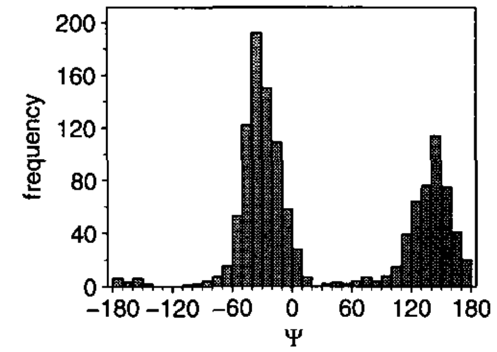

Figure 6. Distribution of the Ψ main-chain dihedral angle at $\Phi = [-80°, -70°]$. The table $W'(\Phi, \Psi)$ was obtained from the local database of protein structures.

(g) *Restraining a distance between main-chain N and O atoms*

The N–O distance in the target protein was modelled in the same way as the C^α–C^α distance above:

$$p^h(h/\bar{g}, i, \bar{a}', h') = \frac{1}{\sigma(\bar{g}, i, \bar{a}', h')\sqrt{2\pi}} \times \exp\left[-\frac{1}{2}\left(\frac{h-h'}{\sigma(\bar{g}, i, \bar{a}', h')}\right)^2\right]. \quad (24)$$

The dependence of the N–O distance on the other four variables is quantified in Table 7. The values of the parameters α_i obtained from the least-squares fitting of the model to the probability distribution histograms are given in Table 8. Examples of experimental histograms overlayed with the analytical curves are shown in Figures 5(c) and (d); they demonstrate that this model is appropriate for restraining N–O distances. The trends observed for the C^α–C^α distance are also valid for the N–O distance. Note that the variability of the main-chain N–O distances is similar to that of the C^α–C^α distances (Fig. 5, eqns (36) and (37)).

(h) *Restraining residue main-chain conformation*

A sample distribution of the Ψ dihedral angle when the Φ dihedral angle is in the range from −80° to −70° is shown in Figure 6. The two peaks correspond to the A and P main-chain conformation classes (Fig. 3). This figure indicates that within each of the six classes, A, B, P, G, L and E (subsection (b)(iii), above), the distribution of the main-chain dihedral angles approximates the Gaussian distribution. The means and standard deviations for the Φ and Ψ angles within each conformation class are estimated in Table 4.

Suppose we can predict the probability ω_i that the restrained residue is in the main-chain conformation class i. Then the two pdfs restraining Φ and Ψ dihedral angles can be modelled as a weighted sum of six Gaussian functions, each function corresponding to one of the main-chain conformation classes A to E and weighted by a probability that a residue is in the corresponding class:

$$p^m(\Phi) = \sum_{i=A,\ldots,E} \omega_i N[\bar{\Phi}_i, \sigma_i(\Phi)]$$
$$p^m(\Psi) = \sum_{i=A,\ldots,E} \omega_i N[\bar{\Psi}_i, \sigma_i(\Psi)], \quad (25)$$

where $N(\alpha, \sigma)$ stands for a Gaussian pdf with mean α and standard deviation σ (Table 4). Note that this scheme takes advantage of the fact that Φ and Ψ angles are highly correlated (Fig. 3). The remaining problem is to determine the probabilities ω_i of all six main-chain conformation classes for each restrained residue. The local database of alignments and the program MDT were used to establish these weights.

The build-up of the pdfs for the main-chain conformation class and for the side-chain conformation class below concentrates only on what is best overall for comparative modelling. No attempt is made to explain the results, to concentrate on individual residue types, or to optimize the pdfs beyond the optimal selection of the currently defined features.

The whole local database was divided into a learning part and a test part. The test set consists of seven serine proteinases and the learning database includes the remaining 16 protein families. The serine proteinases of the test set, tonin, kallikrein, trypsin, chymotrypsin, elastase, rat mast cell protease II, and *S. griseus* trypsin, have had structures determined at a resolution of 2 Å or better (Table 1). This test set contains 21 pairs of related proteins, 1608 residues and 5586 pairs of equivalent residues. Even though serine proteinases beong to the β class of proteins, the content of the six main-chain conformation classes is not significantly different from that of the whole local database; 43% of the residues in the whole database are in the A conformation, as are 28% of the test set residues. The pairwise sequence identities in the test set range from 27 to 55%; the average is 35·4%.

The protein features that may correlate with the main-chain conformation class of a restrained residue were then selected from the list in Table 3. These are the types of the restrained and equivalent residues and the features that can be classified into the following three groups: main-chain conformation of an equivalent residue (M', t', Φ', Ψ', α'), side-chain conformation of an equivalent residue (c_1', c_2', c_3'), and variability measures (s, i, b', a', g, R') (Table 9). The non-smoothed and smoothed ($\sigma = 5$) pdfs of the form $p(M/a, b, \ldots, c)$ were derived from the learning database for the 7249 possible combinations of up to five selected features (a, b, , , c) listed above. Each of the resulting pdfs was evaluated by predicting the most likely main-chain conformation class for each residue in all the 5586 equivalent residue pairs in the test set and by comparing these predictions with the actual conformations found in the crystallographic structures (Table 10).

If only the probabilities $p(M)$ of the six different conformation classes are used in the prediction, the prediction success is 27·9% (Table 10), because that

Table 9

Features used in the derivation of the distributions for the main-chain conformation class M *and side-chain dihedral angle classes* c_i

Type of features	Feature	Symbol	Start	End	Interval	No. of intervals
	Residue types	r, r'	—	—	—	20
Main-chain	Main-chain conformation classes	M, M'	—	—	—	6
	Secondary structure type	t'	—	—	—	4
	Φ Dihedral angle	Φ'	0	360	20	18
	Ψ Dihedral angle	Ψ'	0	360	20	18
	Content of the A class	α'	0	1·0	0·2	5
Side-chain	χ_i' Classes	c_i, c_i'	—	—	—	3
Variability	Distance from a gap	g	0	∞	[a]	5
	Resolution of X-ray analysis	R'	0	∞	[b]	4
	Residue neighbourhood difference	s	0	2·0	0·4	5
	Average side-chain B_{iso}	b'	0	75	15	5
	Fractional sequence identity	i	0	1	0·2	5
	Side-chain accessibility (%)	a'	0	125	25	5

The Table shows the ranges used for the tabulation of the distributions $W(M/a, b, \ldots, c)$ and $W(c_i/a, b, \ldots, c)$. Note that the prime in the feature symbol indicates that the feature is obtained from the template structure, not the target structure.

[a] The intervals for g are 0 to 1, 2, 3 to 4, 5 to 6, 7 to ∞.

[b] The intervals for R are 0 to 1·6, 1·6 to 2·0, 2·0 to 2·5, 2·5 to ∞.

many residues are in the most likely A class. This success is improved marginally by 5·3% when the residue type is taken into account [$p(M/r)$], due to the preference of Val and Tyr for the B class and Pro for the P class. However, the largest improvement due to a single feature is provided by the main-chain conformation class of the equivalent residue [$p(M/M')$]; this increases the prediction success from 27·9% to 70·2%. When the residue type and the equivalent main-chain conformation are combined in $p(M/r, M')$, the prediction success rises to 72·5%. This is then increased by up to 3·6% when a variability measure, such as the fractional sequence identity or local similarity, is also taken into account. The effect of this measure is that the pdf relies more on the residue type when similarity is low and more on the equivalent main-chain conformation when similarity is high (compare Figs 7(c) and (d)). Information about the conformation of an equivalent side-chain does not increase the prediction rate.

Table 10

The build-up of the pdf for prediction of the main-chain conformation

	Entropy		% Correctly predicted	
pdf	Non-smth	Smth	Non-smth	Smth
Uniform pdf	1·792	1·792	16·7	16·7
$p(M)$	1·3169	1·3206	27·9	27·9
$p(M/r)$	1·1549	1·1840	33·2	33·2
$p(M/t')$	1·2124	1·2136	54·0	54·0
$p(M/M')$	1·3330	1·3348	70·2	70·2
$p(M/t', r)$	1·0172	1·0489	58·4	58·4
$p(M/M', r)$	1·0333	1·1466	72·6	72·5
$p(M/M', r, s)$	0·7504	1·2118	73·5	73·3
$p(M/M', r, i)$	0·7172	1·1838	76·3	76·1
$p(M/M', r, \alpha')$	0·7926	1·1419	72·4	72·3
$p(M/M', r, b')$	0·7014	1·1579	71·7	71·6
$p(M/M', r, a')$	0·7332	1·2113	71·4	71·3
$p(M/M', r, R')$	0·8813	1·1848	71·7	71·6
$p(M/M', r, g)$	0·7929	1·1810	74·0	73·7
$p(M/M', r', c_1')$	0·8350	1·2424	72·0	71·9
$p(M/M', r, s, g)$	0·5441	1·2321	72·9	73·8
$p(M/M', r, a', g)$	0·5328	1·2348	72·0	73·3
$p(M/M', r, a', s)$	0·5168	1·2593	72·0	73·2
$p(M/M', r, r', c_1')$	0·4656	1·2035	68·9	71·5

Non-smth, non-smoothed pdf; Smth, smoothed pdf ($\sigma = 5$). A prediction is correct when the most likely conformation obtained from the pdf matches the crystallographically observed conformation. Results for 4 and 5 independent features do not improve the prediction success which remains at around 73%. Only a small sample of the total of 7249 different pdfs is shown. See Table 9 for the description of the variables.

The comparison of prediction successes of the non-smoothed and smoothed pdfs shows that smoothing does not improve the prediction success if three or less independent features are used (Table 10). However, the comparison of entropies for the smoothed and non-smoothed pdfs indicates that whereas the most likely conformation may well be predicted correctly by the non-smoothed pdfs, the errors in the weights for the other less likely conformations are expected to be smaller after smoothing.

The smoothed pdf $p(M/r, M', s)$ was selected for the subsequent calculations of weights ω_i that are needed to restrain the main-chain dihedral angles Φ and Ψ (eqn (25)). Residue neighbourhood difference s is the most appropriate variability measure because it consistently leads to high prediction scores for different test sets (data not shown).

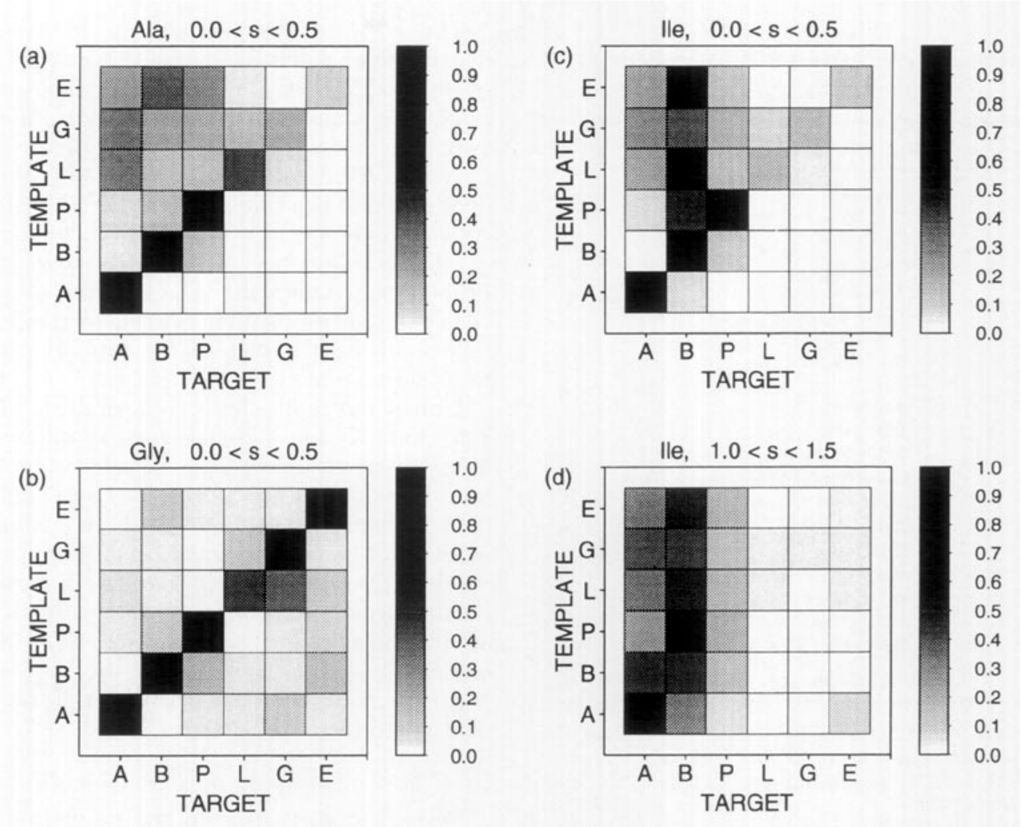

Figure 7. Sample cross-sections through the pdf for prediction of main-chain conformation. The probabilities $W(M/M', r, s)$ for conformation classes A to E of a given residue type (horizontal row, M) are shown for each conformation class of an equivalent residue (vertical row, M'). The type of a restrained residue (r), and the local neighbourhood difference (s) are shown above each plot.

A cross-section through pdf $p(M/r, M', s)$ is shown in Figure 7. The prediction success of this pdf on the test set of seven serine proteinases is listed for the individual residue types in Table 11. The residues that are predicted most accurately (approx. 85%) are Trp, Gln, Pro, Phe and Cys, whereas the least accurately predicted residues include Gly, Asn, Glu and Leu (approx. 63%). This trend probably reflects the distribution of the various residue types in the core and on the surface of the molecule as well as the degree of restraint on the main-chain provided by its side-chain. The conformation of the core residues is expected to be more conserved, and therefore better predicted, than the conformation of the exposed residues. Likewise, the conformationally restrained residues, such as Pro, are predicted better than those that are more flexible, such as Gly. Leu is not predicted reliably because its intrinsic preferences for the A, B and P classes are very similar.

While the 73% prediction rate may seem low, many errors occur because of the swapping between the structurally similar (B, P) classes as well as between the (L, G) classes. When these two pairs are treated only as two classes, the prediction success increases to 87·4%.

(i) *Restraining residue side-chain conformation*

A large amount of information exists about what determines conformation of side-chains. It has been known for a long time that different side-chains have specific preferred values for their dihedral angles (Janin *et al.*, 1978; Ponder & Richards, 1987). More recently, there have been extensive analyses of the dependence of side-chain conformation on the conformation and type of an equivalent side-chain in a related structure (Summers & Karplus, 1989; Sutcliffe *et al.*, 1987*b*), solvent accessibility (Schiffer *et al.*, 1990; Summers & Karplus, 1989), hydrogen bonding (Summers & Karplus, 1989), secondary structure (McGregor *et al.*, 1987; Sutcliffe *et al.*, 1987*b*), main-chain conformation (Dunbrack & Karplus, 1993), and close packing (Desmet *et al.*, 1992; Holm & Sander, 1992; Lee & Subbiah, 1991; Tuffery *et al.*, 1991). The close packing criterion cannot be included in the derivation of restraints formulated as pdfs, but it is satisfied to some degree in the final modelling stage when all restraints are optimized simultaneously (section 3). Likewise, side-chain hydrogen bonds, interactions with water, and salt bridges were not included in the side-chain conformation restraints. However, all the other features listed above and several additional features were tested in a number of different combinations. The aim was to develop the best pdf for restraining side-chain conformation given an alignment with a related structure.

As for the main-chain modelling, the protein features that may correlate with the side-chain dihedral angle classes of a restrained residue, c_i,

Table 11

Success for the prediction of the main-chain conformation class

Residue type	Total number	% Correctly predicted
W	219 (37)	90
Q	391 (67)	89
P	430 (79)	86
F	224 (38)	83
C	360 (62)	83
A	683 (123)	82
V	791 (136)	79
S	751 (139)	78
I	518 (88)	76
H	209 (36)	73
R	262 (46)	73
K	446 (77)	72
T	614 (108)	71
D	382 (69)	70
Y	330 (57)	69
M	137 (23)	68
G	918 (163)	66
N	481 (85)	62
E	319 (57)	62
L	685 (118)	57
	9150 (1608)	73

Total number is the number of residue pairs containing the residue being predicted; the numbers of the predicted residues are shown in parentheses. The smoothed pdf $p(M/M', r, s)$ was used for the prediction of the main-chain conformation class M. The residue types are listed in descending order with respect to the success of the prediction. The bottom line gives the total number of equivalent residue pairs, the total number of residues with a defined main-chain conformation state and the prediction success averaged over all residue types.

were selected from the list in Table 3. These are the types of the restrained and equivalent residues and the features that can be classified into the following three groups: main-chain conformation of an equivalent residue (M', t', Φ', Ψ', α'), side-chain conformation of an equivalent residue (c'_1, c'_2, c'_3, c'_4), and variability measures (s, i, b', a', g, R') (Table 9). For each c_i, the non-smoothed and smoothed ($\sigma = 5$) pdfs of the form $p(c_i/a, b, \ldots, c)$ were derived from the learning database for the 2517 possible combinations of up to four selected features ($a, b, \ldots, c$) listed above. Each of the resulting pdfs was evaluated by predicting the most likely side-chain conformation class for each residue in all 5586 equivalent residue pairs in the test set and by comparing these predictions with the actual conformations found in the crystallographic structures (Tables 12 to 14).

Similarly to the prediction of the main-chain conformation class, the side-chain dihedral angles χ_i are modelled as a weighted sum of Gaussians:

$$p^s(\chi_i) = \sum_j \omega_{ij} N[\bar{\chi}_{ij}, \sigma_j(\chi_i)], \qquad (26)$$

where ω_{ij} are the probabilities that the restrained side-chain dihedral angle i is in class j, and $N(\alpha, \sigma)$ is a Gaussian pdf with mean α and standard deviation σ (Table 5). The remaining problem is to determine the probabilities ω_{ij} of all side-chain conformation classes for each restrained residue.

The c_i class can assume up to three values, depending on i and the residue type (Table 5). Some residues, however, have a smaller number of possible c_i classes; for example, χ_2 in His only has two. The current implementation of the smoothing procedure (subsection (c), above) smooths the data over all three possible χ_i classes for all residue types. As a consequence, the probabilities for non-existing χ_i classes may become greater than zero. This problem is patched after the smoothing by resetting the probabilities for non-existing χ_i classes to 0 and renormalizing the rest of the probabilities so that their sum equals 1.

Unlike the main-chain conformation classes, the side-chain classes c_i are assigned to different residue types in an arbitrary way; i.e. class $c_2 = 1$ of Ser is not the same as class $c_2 = 1$ for Glu (Table 5). As a consequence, whenever a side-chain conformation of a residue being predicted (c_i) or an equivalent residue (c'_i) is included in a pdf, the corresponding residue type (r or r') also has to be added. Thus, the simplest meaningful pdf for prediction of the side-chain conformation is $p(c_i/r)$, not $p(c_i)$. The pdf $p(c_i/r)$ is equivalent to a rotamer library such as that of Ponder & Richards (1987).

(i) *Prediction of the χ_1 conformation class*

When $p(c_1/r)$ is used in the prediction of the χ_1 class, the prediction success is 57·4%, because that many residues are in their most likely classes (Table 12). All residue types prefer the $-\chi_1$ class, except for Val which strongly prefers the t class and Ser which has a weak preference for the + class. When information about the type and χ_1 dihedral angle of an equivalent residue is added to obtain pdf $p(c_1/r, r', c'_1)$, the prediction success increases for 6·4% to 63·8%.

None of the remaining independent variables listed in Table 9 improves the prediction success of $p(c_1/r)$ or $p(c_1/r, r', c'_1)$, irrespective of whether the variables are used on their own, in pairs, or in threes. The prediction successes of the smoothed pdfs remain at around 57% [$p(c_1/r)$] and 63% [$p(c_1/r, r', c'_1)$]. It has been shown that side-chain dihedral angle χ_1 correlates with χ_2 (Janin *et al.*, 1978) and with main-chain conformation (Dunbrack & Karplus, 1993; McGregor *et al.*, 1987; Sutcliffe *et al.*, 1987*b*) of the same residue. In contrast, when χ_2 and main-chain conformation of an equivalent residue are used, these correlations are not statistically strong enough to help the prediction. It is possible, however, that the prediction using these correlations could be improved for sequences that are more similar to the known structure than the present test sequences, for which the average residue identity is 35%. Unfortunately, the current local database is too small to reliably test this possibility.

Some of the pdfs with three or four independent variables are sparse as shown by the difference in the entropies of the non-smoothed and smoothed pdfs, and by a drop by up to 7% in the prediction success of the non-smoothed pdfs compared to the

Table 12

The build-up of the pdf for prediction of the χ_1 class, c$_1$

pd-	Entropy		% Correctly predicted	
	Non-smth	Smth	Non-smth	Smth
Uniform pdf	1·099	1·099	33·3	33·3
$p(c_1/r)$	0·9357	0·9539	57·36	57·36
$p(c_1/r, \alpha')$	0·8778	0·9244	57·72	57·72
$p(c_1/r, t')$	0·8914	0·9230	56·03	56·26
$p(c_1/r, M')$	0·8568	0·9272	56·55	56·48
$p(c_1/r, r', M')$	0·5785	0·9238	52·68	54·21
$p(c_1/r, a', M')$	0·6456	0·9261	54·52	56·65
$p(c_1/r, \Phi')$	0·7753	0·9443	57·47	57·77
$p(c_1/r, \Psi', s)$	0·6087	0·9342	58·01	59·53
$p(c_1/r, \Psi')$	0·8165	0·9303	58·10	59·08
$p(c_1/r, r', \Phi')$	0·5018	0·9399	50·65	55·54
$p(c_1/r, r', \Psi')$	0·4545	0·9283	52·83	58·00
$p(c_1/r, \Phi', \Psi')$	0·4346	0·9399	55·57	59·08
$p(c_1/r, r')$	0·8679	0·9377	54·87	55·61
$p(c_1/r, a')$	0·9190	0·9473	55·27	54·97
$p(c_1/r, s, i)$	0·7496	0·9547	56·96	57·97
$p(c_1/r, r', i)$	0·6567	0·9427	53·52	55·56
$p(c_1/r, c_1', r')$	0·7140	0·9288	62·73	63·77
$p(c_1/r, c_1', r', c_2')$	0·5814	0·9342	61·40	63·87
$p(c_1/r, c_1', r', t')$	0·4912	0·9179	60·26	63·32
$p(c_1/r, c_1', r', a')$	0·4898	0·9312	60·05	64·21
$p(c_1/r, c_1', r', s)$	0·4822	0·9301	60·65	63·87
$p(c_1/r, c_1', r', i)$	0·5008	0·9272	59·53	63·60
$p(c_1/r, c_1', r', M')$	0·4750	0·9217	59·17	63·80
$p(c_1/r, c_1', r', \Phi')$	0·3887	0·9359	58·06	63·15

Conditional pdfs for prediction of c_1 from all combinations of up to 4 independent features were calculated and evaluated. Only a small fraction of the total of 2517 different pdfs are shown. Non-smth, non-smoothed pdf; Smth, smoothed pdf ($\sigma = 5$). See Table 9 for the description of the variables.

Table 13

The build-up of the pdf for prediction of the χ_2 class, c$_2$

pdf	Entropy		% Correctly predicted	
	Non-smth	Smth	Non-smth	Smth
Uniform pdf			51·9	51·9
$p(c_2/r)$	0·6202	0·7749	70·68	70·68
$p(c_2/r, r', c_1')$	0·4806	0·7810	70·15	70·30
$p(c_2/r, r', c_2')$	0·5153	0·7903	71·28	71·65
$p(c_2/r, \alpha')$	0·6095	0·7657	70·76	70·76
$p(c_2/r, M')$	0·5839	0·7608	70·43	70·43
$p(c_2/r, t')$	0·6045	0·7508	70·06	70·06
$p(c_2/r, \Psi')$	0·5734	0·7743	71·24	71·30
$p(c_2/r, \Phi')$	0·5172	0·7850	69·59	69·89
$p(c_2/r, r', M')$	0·3879	0·7891	68·67	69·45
$p(c_2/r, a', M')$	0·4219	0·7815	68·96	70·08
$p(c_2/r, \Psi', s)$	0·4238	0·7943	70·69	71·28
$p(c_2/r, \Phi', \Psi')$	0·2892	0·8127	69·04	70·48
$p(c_2/r, r', \Phi')$	0·3431	0·8048	66·89	69·45
$p(c_2/r, r', \Psi')$	0·3148	0·7992	66·52	69·19
$p(c_2/r, r')$	0·5851	0·7648	69·41	69·54
$p(c_2/r, r', i)$	0·4377	0·7875	69·17	69·76
$p(c_2/r, a')$	0·6108	0·7574	70·24	70·52
$p(c_2/r, s, i)$	0·4930	0·7922	70·22	70·15
$p(c_2/r, c_2', r', t')$	0·3754	0·8017	70·00	71·48
$p(c_2/r, c_2', r', a')$	0·3746	0·8056	68·59	71·37
$p(c_2/r, c_2', r', s)$	0·3578	0·8045	70·02	71·19
$p(c_2/r, c_2', r', i)$	0·3696	0·8024	70·80	71·95
$p(c_2/r, c_2', r', c_1')$	0·3971	0·8005	72·30	72·28
$p(c_2/r, c_2', r', M')$	0·3413	0·8063	70·30	71·06
$p(c_2/r, c_2', r', \Phi')$	0·2924	0·8175	67·17	71·21

See legend to Table 12.

smoothed pdfs. This drop demonstrates the beneficial effect of smoothing: smoothing always increases the prediction success for sparse pdfs and does not significantly decrease the prediction success of dense pdfs.

The results in Table 12 show that $p(c_1/r, c_1', r')$ is the best pdf for prediction of c_1 for a sequence at approximately 35% residue identity with the known structure. However, the pdf $p(c_1/r, c_1', r', s)$ was selected for the subsequent calculations of weights ω_{1j} (eqn (26)) to include the beneficial effect of s. As shown for the prediction of main-chain conformation, pdf with s is capable of relying more on the homologous structure when similarity is high and more on the residue type preference when similarity is low.

The prediction successes of $p(c_1/r)$ and $p(c_1/r, c_1', r', s)$ are listed for the individual residue types in Table 15. The residues that are predicted most reliably (80%) by $p(c_1/r, c_1', r', s)$ tend to be large and buried (Trp, Cys, Leu, Val and Tyr). The residues that are predicted least reliably (50%) tend to be small and exposed (Asn, Met, Arg, Glu and Ser).

The largest improvement as a result of using information about the equivalent side-chain occurs for Trp (30%), His (23%), Asp (17%), Thr (12%), Tyr (10%) and Leu (10%). The amount of information provided by the type and χ_1 of an equivalent residue tends to be large for large or buried residues and small or non-existent for exposed residues. This improvement reflects the degree to which the side-chain conformation of a residue is restrained by its environment.

(ii) *Prediction of the* χ_2 *conformation class*

All residues with the usual trimodal distribution of the χ_2 dihedral angle (Asp, Ile, Lys, Leu, Met, Gln and Arg) prefer the t class, Trp prefers class 1, His and Asn prefer class 2. The overall prediction success for the simplest pdf $p(c_2/r)$ is 70·7% (Table 13). When the residue type and c_2 class of the equivalent residue are added to this pdf, the prediction success is increased marginally to 71·7% $[p(c_2/r, r', c_2')]$. When information about the χ_1 angle of an equivalent residue is added to obtain $p(c_1/r, r', c_1', c_2')$, the overall prediction success is further increased to 72·3%; most of the increase comes from only three residues: His (5%), Leu (2%) and Gln (5%). This small improvement seems to reflect the inter-dependence of the χ_1 and χ_2 angles (Janin *et al.*, 1978). The pdf $p(c_2/r, r', c_1', c_2')$ was selected for the subsequent calculations of weights ω_{2j} that are needed to restrain side-chain dihedral angle χ_2 (eqn (26)). The prediction successes of $p(c_2/r)$ and $p(c_2/r, r', c_1', c_2')$ are listed for the individual residue types in Table 15. Similarly to χ_1, the two residues that are most improved by the information from the equivalent residue are Trp and His.

Table 14

The build-up of the pdf for prediction of the χ_3 class, c_3

pdf	Entropy		% Correctly predicted	
	Non-smth	Smth	Non-smth	Smth
Uniform pdf	—	—	34·2	34·2
$p(c_3/r)$	0·9066	0·9679	58·52	58·52
$p(c_3/r, r', c'_1)$	0·7022	0·9657	56·08	56·78
$p(c_3/r, r', c'_2)$	0·7608	0·9674	59·36	59·42
$p(c_3/r, r', c'_3)$	0·7949	0·9677	60·00	60·26
$p(c_3/r, \alpha')$	0·8773	0·9578	53·83	53·83
$p(c_3/r, M')$	0·8758	0·9801	60·32	60·32
$p(c_3/r, t')$	0·8897	0·9757	61·09	61·09
$p(c_3/r, \Phi')$	0·7770	0·9779	59·74	59·74
$p(c_3/r, \Psi')$	0·8318	0·9698	58·78	58·78
$p(c_3/r, \Phi', \Psi')$	0·4404	0·9698	56·08	57·11
$p(c_3/r, r', M')$	0·5715	0·9696	56·40	59·16
$p(c_3/r, \alpha', M')$	0·7053	0·9615	52·73	53·63
$p(c_3/r, \Psi', s)$	0·5926	0·9709	58·97	58·39
$p(c_3/r, r', \Phi')$	0·5144	0·9694	53·95	55·37
$p(c_3/r, r', \Psi')$	0·4658	0·9620	52·99	58·65
$p(c_3/r, r')$	0·8591	0·9711	59·87	59·55
$p(c_3/r, a')$	0·8847	0·9723	58·59	58·78
$p(c_3/r, s, i)$	0·7264	0·9769	59·68	58·84
$p(c_3/r, r', i)$	0·6111	0·9675	58·20	59·04
$p(c_3/r, c'_3, r', t')$	0·6011	0·9639	58·71	60·64
$p(c_3/r, c'_3, r', a')$	0·5989	0·9607	56·40	58·52
$p(c_3/r, c'_3, r', s)$	0·5670	0·9652	57·75	57·56
$p(c_3/r, c'_3, r', i)$	0·5739	0·9648	56·33	58·71
$p(c_3/r, c'_3, r', c'_1)$	0·6392	0·9641	57·23	58·71
$p(c_3/r, c'_3, r', M')$	0·5444	0·9677	57·04	59·74
$p(c_3/r, c'_3, r', c'_2)$	0·6574	0·9660	57·75	59·68
$p(c_3/r, c'_3, r', \Phi')$	0·4680	0·9671	51·51	57·75

See legend to Table 12.

(iii) *Prediction of the χ_3 conformation class*

Glu residues fall into two χ_3 classes, of which class 2 is dominant with 80% of the Glu residues. The other residues with χ_3 angles (Lys, Met, Arg, Gln) have three classes. For Met and Gln, the three classes are almost equally populated, whereas Arg and Lys have a preference for the t class (47% and 71%, respectively). The overall prediction success for the simplest pdf $p(c_3/r)$ is 58·5% (Table 14). When the residue type and χ_3 dihedral angle of the equivalent residue are added to this pdf to get $p(c_3/r, r', c'_3)$, the prediction success is increased to 60·3%. Examination of the prediction successes of various pdfs $p(c_3/a, b, \ldots, c)$ for different residue types indicated that information about the secondary structure state of the equivalent residue improves the prediction of both Met and Gln by a few percentage points. Therefore, the pdf $p(c_3/r, c'_3, r', t)$ was used in the subsequent calculations of weights ω_{3j} that are needed to restrain side-chain dihedral angle χ_3 (eqn (26)). The prediction success for the individual residues is listed in Table 15. The prediction of Gln is the worst, whereas Lys and Glu are predicted most successfully. Note, however, that Glu has only two classes of which the most probable one spans a large range of degrees (Table 5). Met is the only residue whose prediction is improved significantly by information from a related structure.

Table 15

Success for the prediction of the side-chain χ_i classes, c_i

Residue type	Total number	% Correctly predicted			
		χ_1 class	χ_2 class	χ_3 class	χ_4 class
W	219 (37)	86·8 (56·8)	81·3 (62·2)	—	—
C	360 (62)	81·4 (77·4)	—	—	—
L	685 (118)	74·0 (64·4)	59·3 (55·9)	—	—
V	791 (136)	72·3 (72·1)	—	—	—
Y	330 (57)	69·4 (59·6)	100·0 (100·0)	—	—
I	518 (88)	68·9 (65·9)	73·9 (73·9)	—	—
K	446 (77)	65·2 (66·2)	63·5 (64·9)	76·0 (75·3)	71·4
F	224 (38)	64·7 (57·9)	100·0 (100·0)	—	—
D	382 (69)	64·7 (47·8)	100·0 (100·0)	—	—
H	209 (36)	61·7 (38·9)	62·2 (55·6)	—	—
Q	391 (85)	61·4 (64·2)	66·5 (62·7)	37·3 (35·8)	
T	614 (108)	58·5 (46·3)	—	—	—
N	481 (85)	55·1 (52·9)	55·1 (56·5)	—	-
M	137 (23)	54·7 (52·2)	64·2 (69·6)	54·7 (21·7)	—
R	262 (46)	53·4 (52·2)	72·9 (73·9)	49·2 (54·3)	80·4
E	319 (57)	51·7 (49·1)	64·9 (63·2)	79·6 (80·7)	—
S	751 (139)	45·7 (40·3)		-	—
	7119 (1243)	64·4 (57·4)	72·3 (70·7)	60·6 (58·5)	74·8

Total number is the number of residue pairs containing the residue being predicted; the numbers of predicted residues are listed in parentheses. The smoothed pdfs $p(c_1/t, c'_1, r', s)$, $p(c_2/r, r', c'_1, c'_2)$, $p(c_3/r, c'_3, r', t')$ and $p(c_4/r)$ were used for the prediction of χ_1, χ_2, χ_3 and χ_4 dihedral angle classes, respectively. The prediction successes of the smoothed pdfs $p(c_i/r)$ are shown in parentheses for $i = 1, 2, 3$. The residue types are listed in descending order with respect to the success of the c_1 prediction. The bottom line gives the total number of equivalent residue pairs tested by the pdf, the total number of residues with a defined χ_1 dihedral angle, and the c_i prediction successes averaged over all residue types that have defined χ_i dihedral angles.

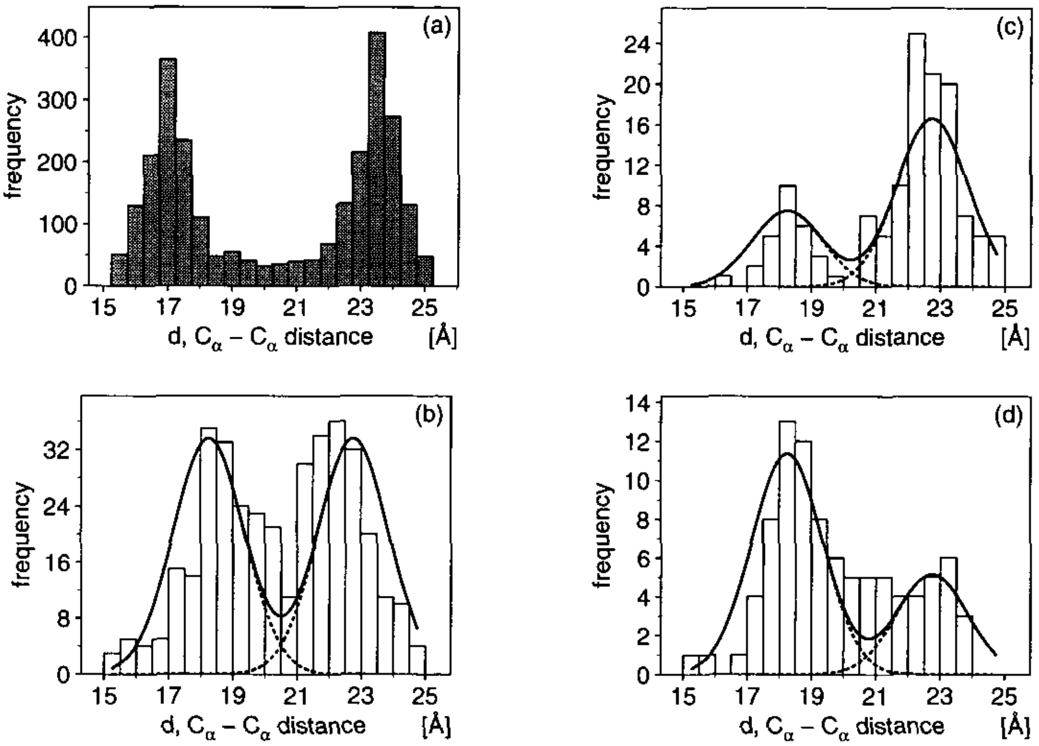

Figure 8. Derivation of a feature pdf from basis pdfs. In all plots, $18{\cdot}0 < d < 18{\cdot}5$ and $22{\cdot}5 < d'' < 23{\cdot}0$. (a) $W'(d/d', d'')$. (b) $W'(d/d', d'', \bar{s}', \bar{s}'')$, where $0{\cdot}2 < \bar{s}' < 0{\cdot}4$ and $0{\cdot}2 < \bar{s}'' < 0{\cdot}4$. (c) $W'(d/d', d'', \bar{s}', \bar{s}'')$, where $0{\cdot}2 < \bar{s}' < 0{\cdot}4$ and $0{\cdot}4 < \bar{s}'' < 0{\cdot}6$. (d) $W'(d/d', d'', \bar{s}', \bar{s}'')$, where $0{\cdot}4 < \bar{s}' < 0{\cdot}6$ and $0{\cdot}2 < \bar{s}'' < 0{\cdot}4$. The histograms are obtained by scanning the local database. The broken lines are the basis pdfs $p^d(d/d', d'', \bar{s}', \bar{s}'')$ calculated from eqn (23). The continuous lines are the feature pdfs $p^D(d/d', d'', \bar{s}', \bar{s}'')$ calculated with eqns (27) and (28).

(iv) *Prediction of the χ_4 conformation class*

Only Lys and Arg residues have a χ_4 side-chain dihedral angle. In the learning database, 64% of Lys residues and 74% of Arg residues are in the most likely t class. The prediction success of the simplest pdf $p(c_4/r)$ is 75·3%. No increase is achieved by addition of any combination of up to three independent variables listed in Table 9. Whatever the values of the independent variables, the most likely conformation of χ_4 remains to be the t class. Smoothed pdf $p(c_4/r)$ is used in the subsequent calculations of weights ω_{4j} (eqn (26)).

3. Satisfaction of Spatial Restraints

It was shown in the previous section how spatial restraints on the sequence to be modelled can be expressed as pdfs. These pdfs were obtained from stereochemical considerations and from a single homologous structure. In this section, we describe how to combine the restraints from several homologous structures and how to use these restraints to derive a 3D model. The aim is to obtain the most probable model for a certain sequence given its alignment with related structures. The 3D model is obtained by an optimization of the molecular pdf which depends on the model and on the restraints. As a result, violations of the given restraints by the model are minimized.

(a) *The molecular probability density function*

The molecular pdf is assembled from feature pdfs which, in turn, are obtained from basis pdfs.

(i) *Derivation of a feature pdf from basis pdfs*

A feature f of a protein structure is defined as any quantity associated with a particular set of atoms $ijk \ldots l$. For example, the distance between atoms i and j is a feature, the distance between atoms j and k is another feature and the angle ijk is yet another feature. The basis pdfs, $p^f(f)$, are the pdfs described in sections 2(e) to (i). In general, every structural feature f can be restrained by several basis pdfs $p^f_k(f)$ for $k = 1, 2, \ldots$. A feature pdf, $p^F(f)$, is a pdf that combines all the information about the possible values that the feature f can assume. The lower-case and upper-case superscripts are used for the basis and feature pdfs, respectively.

The following example clarifies these definitions. The aim is to construct a feature pdf for a particular C^α–C^α distance in a given sequence. Suppose two known related structures with equivalent distances are available; therefore, we have two corresponding basis pdfs for the C^α–C^α distance in the target sequence (eqn (23)). In addition, we also know that each of the two restraints has to comply with the van der Waals criterion, i.e. the distance has to be larger than the sum of the two van der Waals radii (eqn (22)). In order to combine all this information, we have to combine the three basis pdfs into a single

1. Use the MDT program to obtain the frequency table $W'(d, d', d'', \bar{s}', \bar{s}'')$. Distance d ranges from 15 to 25.0Å in steps of 0.5Å, distances d' and d'' from 18 to 22.0Å in steps of 0.5Å and average residue neighbourhood differences from 0 to 1.5 in steps of 0.3.
2. Estimate weights $\omega_1(\bar{s}')$ and $\omega_2(\bar{s}'')$ for all histograms $W(d/d', d'', \bar{s}', \bar{s}'')$ that have more than 200 data points and where d' is at least 1.5Å away from d''. Divide each histogram into two parts at $(d' + d'')/2$. Weights $\omega(\bar{s}')$ and $\omega(\bar{s}'')$ are set proportionally to the number of the data points in the corresponding part of the histogram.
3. Use the LSQ program to fit the model in Eq. 28 to the 570 $\omega_j(\bar{s})$ points from the previous step. The calculated values of the parameters a, b, and c are 0.0331 ± 0.0025, -4.98 ± 0.11, and 1.800 ± 0.079, respectively. The RMS deviation between the data and the model is 0.05.

Figure 9. Determination of the weights of basis pdfs contributing to a feature pdf.

feature pdf. To do that, we can use all possible alignments of three proteins in the local database. An example of the dependence of a C^α–C^α distance on the two equivalent distances from two related structures, $p(d/d', d'')$, is shown in Figure 8(a). The histogram suggests that $p(d/d', d'')$ can be modelled as a weighted sum of the individual pdfs $p(d/d')$ and $p(d/d'')$:

$$p(d/d', d'', \bar{s}', \bar{s}'') = \omega(\bar{s}')\cdot p(d/d') + \omega(\bar{s}'')\cdot p(d/d''). \tag{27}$$

This relation is clearly a better fit to the data than a product of the individual pdfs which would imply that the most likely distance d is an average of d' and d''. The weight ω of each term in this sum is proportional to the average residue neighbourhood difference s between the corresponding structure and the sequence of the unknown. The data can be fitted by the following model for $w(s)$:

$$\omega(s) = \frac{w(s)}{\sum_j w(s_j)},$$

where (28)

$$w(s) = a + \exp(bs^c); \quad \sum_j \omega(s_j) = 1,$$

as described in Figure 9. The result is that the contribution to the 3D model of a related structure falls faster than linearly with the average residue neighbourhood difference between the two sequences. Examples of histograms and analytical curves for the feature pdf corresponding to different weights are shown in Figure 8(b) to (d).

The last step in the derivation of the feature pdf is to include the van der Waals restraint. Since all stereochemical restraints have to be satisfied in all structures, these restraints are multiplied into the feature pdf and we obtain the final feature pdf:

$$p^D(d) = [\omega_1 p_1^d(d) + \omega_2 p_2^d(d)]p^v(d).$$

This simple approach to combining of two basis pdfs was used for any number of basis pdfs of the same type that were derived from related structures, including the basis pdfs described in sections 2(f) to (i). When properties such as main-chain and side-chain conformation are predicted, average neighbourhood sequence difference is replaced by the neighbourhood sequence difference.

Definitions of all types of feature pdfs follow, with the basis pdfs on the right side of the equations as defined in sections 2(e) to (i). The subscript i in the sum refers to the sequences with known structure that are aligned with the sequence of the unknown. The independent variables $a, b, \ldots$ refer to the features correlated with the restrained feature as described in sections 2(f) to (i). The weights ω_i are determined from equation (28). The derivation of the feature restraints is implemented in the program GETCSR.

1. C^α–C^α distance restraints:

$$p^D(d) = p^v(d) \sum_i \omega_i p_i^d(d/a, b, \ldots) \tag{29}$$

for all pairs of C^α atoms in the sequence of the unknown that satisfy the following three criteria: (1) there is at least one equivalent C^α atom pair in the known structures; (2) there are at least N^α (usually 1) residues between the two residues spanning the distance in the sequence of the unknown; and (3) at least one equivalent distance in the known structures is less than d_d (usually 20 Å). The sum runs over all known structures with an equivalent C^α pair present.

2. Main-chain N–O distance restraints:

$$p^H(h) = p^v(h) \sum_i \omega_i p_i^h(h/a, b, \ldots) \tag{30}$$

for all pairs of main-chain N and O atoms in the sequence of the unknown that satisfy the following criteria: (1) there is at least one equivalent (N, O) pair in the known structures; (2) there are at least N_h (usually 2) residues between the two residues spanning the distance in the sequence of the unknown; and (3) at least one equivalent distance in the known structures is less than d_h (usually 10 Å). The sum runs over all known structures with an equivalent N–O pair present.

3. Stereochemical restraints:

$$p^E(e) = p^e(e). \tag{31}$$

Feature e can be bond length, bond angle, "stereochemical" dihedral angle (section 2(e)), van der Waals contact, or the distance, angles and dihedral angle of a disulphide bond. The feature pdf for van der Waals contacts, $p^V(v)$, restrains only those pairs of atoms that are not already restrained by the feature pdfs for the bond lengths, bond angles, C^α–C^α distances and main-chain N–O distances.

4. Main-chain conformation restraints:

$$p^M(\theta) = \begin{cases} \sum_{i=1}^{n} \omega_i p_i^m(\theta/a, b, \ldots) & n > 0 \\ p^m(\theta/R) & n = 0, \end{cases} \quad (32)$$

where θ stands for either Φ or Ψ main-chain dihedral angle. If there is no equivalent residue in any of the related structures ($n = 0$), the restraint depending only on the residue type in the sequence of the unknown is applied.

5. χ_1, χ_2, χ_3 and χ_4 side-chain dihedral angle restraints:

$$p^S(c) = \begin{cases} \sum_{i=1}^{n} \omega_i p_i^s(c/a, b, \ldots) & n > 0 \\ p^s(c/R) & n = 0, \end{cases} \quad (33)$$

where c stands for either χ_1, χ_2, χ_3 or χ_4 side-chain dihedral angle. A rotamer library based only on the residue type is used when there is no equivalent residue in any of the available structures ($n = 0$).

(ii) *Derivation of a molecular pdf from feature pdfs*

The last stage in the derivation of a molecular pdf is to combine all feature pdfs into a molecular pdf. The 3D structure of a protein is uniquely determined if a sufficiently large number of its spatial features, f_i, are specified. The goal is to find the 3D structure that is consistent with the most probable values of individual features f_i. The molecular pdf should give a probability for occurrence of any combination of these features simultaneously. Then the model for the 3D structure of the unknown would correspond to the maximum of the molecular pdf. Assuming that feature pdfs are independent of each other, the molecular pdf is simply a product of feature pdfs defined in equations (29) to (33):

$$P = \prod_i p^F(f_i). \quad (34)$$

Thus, by maximizing function P we find the most probable model for the 3D structure of the unknown given its alignment with the known structures.

Assuming the independence of feature pdfs is equivalent to the supposition that the total energy of the molecule can be expressed as a sum of individual energy terms. This is clearly incorrect in some cases. For example, the probability of a certain Φ angle depends strongly on the value of the Ψ angle of the same residue (Fig. 3). To model these correlations properly, higher dimensional pdfs of the form $p(x, y, \ldots, z/a', b', \ldots, c')$ would be needed. Unfortunately, the current database of alignments is not large enough to derive such pdfs in general. However, some partial solution of this problem may sometimes still be possible, as demonstrated by our pdfs for the main-chain dihedral angles where the individual pdfs for Φ and Ψ angles were derived from the pdfs for the main-chain conformation classes, which are based on both main-chain dihedral angles. Search for strongly correlated features and their treatment within a single multidimensional pdf will be an important part in future improvements of this approach to protein modelling.

(b) *Optimization of the molecular pdf*

This section describes the tools for optimization of the molecular pdf that are implemented in the program MODELLER†. The latest version of MODELLER consists of about 35,000 lines of code in FORTRAN 77 that was tested to run on IBM RS/6000, Silicon Graphics Iris 4D, SUN Sparcstation, DEC Decstation, DEC Alpha and NeXT workstations. This program has a variety of functions that facilitate modelling, including building of structures from sequence, multiple comparison of proteins, comparison of protein features with restraints, and various graphic routines that output PostScript files (all Figures in this paper, except Figs 1 and 10, were produced by the ASGL program which works in tandem with MODELLER). MODELLER is implemented as an interpreter of a high-level language specialized for dealing with protein structures.

The function that is actually optimized is a transformation of the molecular pdf P:

$$F = -\ln(P), \quad (35)$$

where all the features are expressed in terms of atomic Cartesian co-ordinates. Function F is referred to as the objective function. The same Cartesian co-ordinates that maximize P also minimize F. However, F is computationally better suited for optimization than P, since multiplication of terms in the product of equation (34) is substituted by their addition in equation (35) and since the problem of floating point overflow is smaller for F than for P.

The second transformation of the original molecular pdf that is useful in optimization is scaling of the standard deviations of the basis restraints. This option allows independent scaling of each of the restrained feature types (bond lengths, bond angles, C^α–C^α distances, etc.). By increasing the standard deviation, the restraint becomes less powerful and a larger violation is more likely, similar to the effect that a decrease of a force constant has in energy minimization.

To increase the radius of convergence, the variable target function approach is implemented in MODELLER. This method has been introduced by Braun and Gō in the DISMAN program for calculating protein 3D structures consistent with 2D NMR constraints (Braun & Gō, 1985). The main difference between the original method and the

† MODELLER 0.9 that was used in the present work is available upon request from A.Š. MODELLER 1.1, with several significant improvements, documentation and improved ease of use will be available shortly.

present implementation is that the current optimization proceeds in the Cartesian space, whereas the original procedure optimized the dihedral angles. Following the variable target function method, the optimum of the molecular pdf is found by successive optimizations of increasingly more complex "target" functions, culminating in the true molecular pdf at the end. This series is obtained by starting with sequentially local restraints and then introducing more and more long-range restraints, finally arriving at the true molecular pdf incorporating all the restraints. More precisely, the target function $P(\Delta r)$ is defined as a function of an integer variable $\Delta r = 1, \ldots, N$, where N is the number of residues in the sequence being modelled. The target function $P(\Delta r)$ is obtained in the same way as the molecular pdf, except that only those restraints whose atoms originate from residues not more than Δr residues apart in the sequence are included. The whole calculation consists of a number of conjugate gradient optimizations (Press *et al.*, 1986) of target functions $P(\Delta r)$ with increasing Δr values. The starting conformation for $P(1)$ optimization is either an extended structure or a conformation derived from an extended chain by rotation around the main-chain and side-chain dihedral angles. In the subsequent steps of the variable target function method, the starting conformation is the final model from the previous step. An ensemble of different final models is obtained by using different initial conformations.

4. Modelling of Trypsin

To illustrate the method of comparative modelling by satisfaction of spatial restraints, this section describes the modelling of trypsin from two other serine proteinases, elastase and tonin. The availability of the crystallographic 3D structure of trypsin allowed an evaluation of the model. Two other examples of application of MODELLER include modelling of ferredoxin (Šali, 1991) and of mouse mast cell chymases (Šali *et al.*, 1993).

Serine proteinases are enzymes consisting of two domains and approximately 230 residues, with the active site located in the cleft at the interface between the two domains. Each of the two domains contains a distorted six-stranded β-barrel with a buried structurally conserved core and one or two helices (McLachlan, 1979). The main structural differences between the members of this family are in the length and conformation of the exposed loop segments that connect the conserved strands and helices.

The 3D structures of trypsin (223 residues; Walter *et al.*, 1982), elastase (240 residues; Meyer *et al.*, 1988) and tonin (227 residues; Fujinaga & James, 1982) were compared using the program COMPARER (Šali & Blundell, 1990) (Figs 10 and 11). This program relies on many structural properties and relationships, such as positions of C^{α} atoms, local main-chain conformation, solvent accessibility and main-chain hydrogen bonding patterns. When only those aligned C^{α} atoms that are less than 3·5 Å apart from each other are considered, 196 pairs superpose with the r.m.s. of 0·79 Å in the more similar pair of trypsin and elastase, whereas only 184 pairs superpose with the higher r.m.s. of 0·87 Å in the superposition of trypsin and tonin. This trend is reversed for the sequence comparisons, where the sequence identity between elastase and trypsin is only 38%, and that between tonin and trypsin is 42%.

In a real application of MODELLER, the sequence of trypsin would have to be compared with the alignment of elastase and tonin to obtain the final multiple alignment. Such a sequence alignment, however, would be less suitable for model building than the structure-based comparison. As this modelling example was designed to test the MODELLER program, its performance should not be limited by a suboptimal input alignment. Therefore, the structural alignment was used for extraction of spatial restraints on the sequence of trypsin as described in section 2. The types of restraints and their numbers are listed in Table 16. Note that many more restraints are available than in the refinement of a model using NMR-derived constraints. However, the comparative modelling restraints are generally not as accurate as those obtained from NMR spectroscopy.

Thirty-nine models of trypsin were calculated by optimizing the molecular pdf from 39 different initial conformations. These conformations were obtained by setting the main-chain and side-chain dihedral angles Φ, Ψ and χ_i to random values between 0 and 360°. The progress of modelling was followed by monitoring the average atomic shifts and the value of the objective function. The optimization schedule and a typical progress of optimization are shown in Figure 12. A total of 11 models with low values of the objective function was obtained (10,293±655) (Table 17). These models were close to the correct trypsin structure. The remaining 28 models were the mirror images of either the whole molecule or of a part of it. They all had a significantly higher value of the objective function (>15,000) and were thus easily identified as misfolded models. The model with the lowest value of the objective function (9388) among the 11 successful trials was taken to be the representative trypsin model (the best model). The violations of the restraints by this and other ten models are small (Tables 16 and 17). The stereochemistry of the models is comparable or better than that of the crystallographic trypsin structure refined at a high resolution.

Even though the MODELLER models have good stereochemistry, we used molecular mechanics program CHARMM 22 (Brooks *et al.*, 1983; A. D. MacKerell, Jr, D. Bashford, M. Bellott, R. L. Dunbrack, Jr, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera & M. Karplus, unpublished results) to refine the energy of these models and

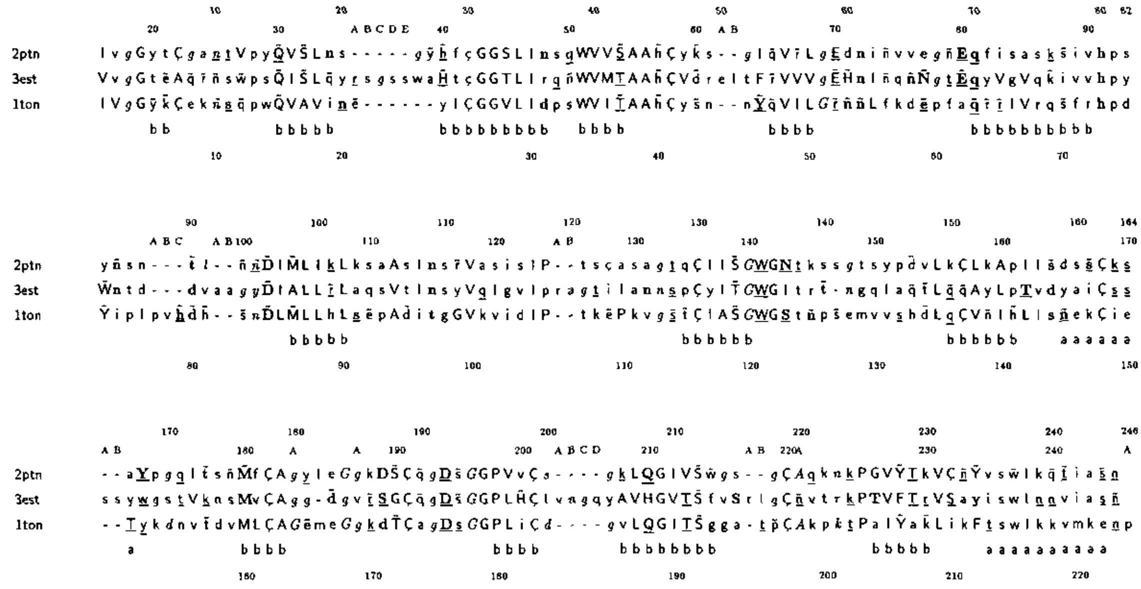

Figure 10. COMPARER alignment of the 3D structures of trypsin, elastase and tonin. The formatting convention of program JOY (Overington *et al.*, 1990) is applied (see legend to Fig. 1). The numbers in the top line refer to the alignment positions. The numbers in the second line are the trypsin residue numbers as specified in the Brookhaven Protein Databank; they were obtained by the alignment of trypsin with chymotrypsinogen. Gaps in the trypsin sequence are indicated by the letters A through E. The numbers in the bottom line count the residues in trypsin. The second line from the bottom gives the secondary structure assignments of individual trypsin residues to the helix (a) and β-strand (b) (Kabsch & Sander, 1983).

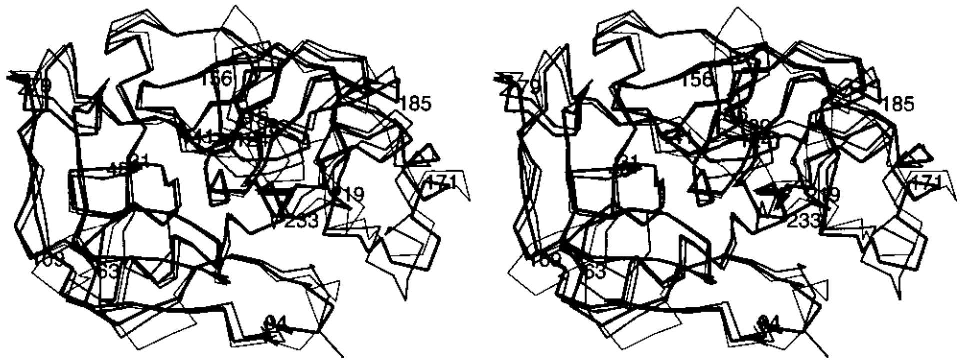

Figure 11. Comparison of trypsin, elastase and tonin. The stereo plot shows the superposition of the C^{α} backbones of elastase (medium line) and tonin (thin line) on that of trypsin (thick line). The pairs of the C^{α} atoms that are aligned in the COMPARER alignment (Fig. 10) were used for the superpositions. Chymotrypsinogen numbering is used.

crystallographic trypsin structure (Table 17). All hydrogen atoms were added to the heavy-atom models and then the structures were refined *in vacuo* by 200 steps of the conjugate gradients minimization of the default CHARMM 22 energy function. This function includes bond length, bond angle, dihedral angle, 6–12 Lennard-Jones, and electrostatic terms. Atomic positions were not restrained during energy minimization because this type of minimization is not capable of large positional changes. The final C^{α} positions had an r.m.s. deviation of approximately 0·75 Å from the initial positions. These shifts resulted in an improvement in energy from approximately 10,000 kcal/mol to −300 kcal/mol, but at a cost of an increase in an r.m.s. from 0·85 Å to 0·95 Å for comparison with the crystallographic trypsin structure. Most of the improvement in energy was due to the relaxation of the structures that were strained in some positions by the addition of the hydrogen atoms (not shown). Additional energy decrease resulted from optimization of a few exposed segments (not shown) that were badly modelled by MODELLER. The value of the objective function also increased significantly as a result of energy minimization from approximately 10,000 to 35,000. The results of the energy minimization of the crystallographic trypsin structure can be used to put these observations into context (Table 17): (1) The crystallographic trypsin structure had as high an initial energy as the 11 MODELLER models. (2) Trypsin structure did not change as much as the MODELLER models during minimization. (3) Refined trypsin structure had a somewhat lower energy than the MODELLER models. We conclude that energy minimization does not significantly improve the overall quality of the

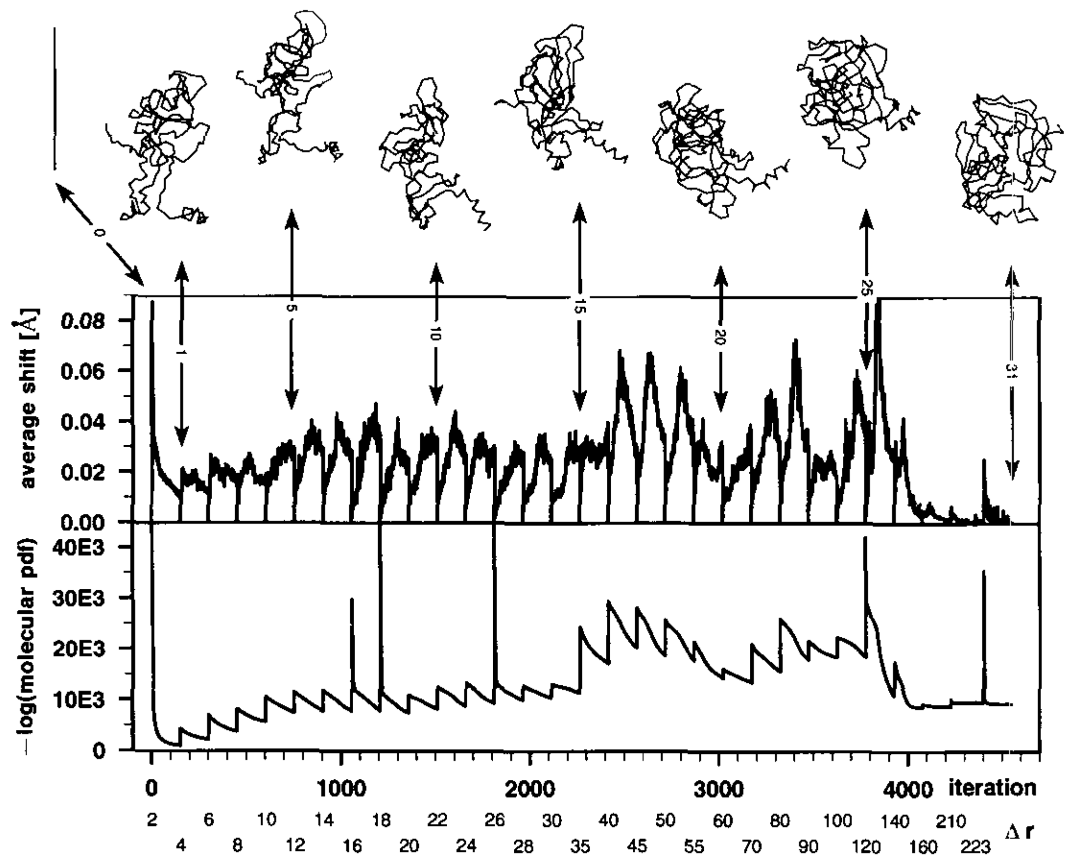

Figure 12. Schedule and progress of optimization. The optimization schedule is specified in the bottom 3 lines. The iteration line counts the conjugate gradient steps. The bottom 2 lines show the changes in Δr: Δr is increased every 150 conjugate gradient steps or when the largest atomic shift is smaller than 0·005 Å. Each change in Δr corresponds to a step in a variable target function method. There are 31 such steps to get one model. The method starts with a few restraints that involve only the atoms from residues at most Δr residues apart and gradually incorporates all restraints (the final Δr equals the length of a sequence). The C^α traces of the evolving model at several stages during the refinement are shown on the top of the Figure. The starting conformation in this case is an extended chain; generally, it is a chain with random Φ, Ψ and χ_i dihedral angles. The van der Waals criterion was gradually introduced in the last 5 steps of the variable target function method by scaling the corresponding standard deviations by 8, 4, 2, 1 and 1. The data for the trial resulting in the model with the lowest value of the molecular pdf are shown. The CPU time needed to calculate one model is 1·5 h on an IBM RS/6000-550 workstation.

models, certainly not in the regions where there are sufficient homology-derived restraints. However, a combination of energy terms and homology-derived restraints may be useful for modelling of exposed parts of the structure where homology-derived restraints are weaker because of structural variability and gaps in the alignment. In the rest of this section, the original MODELLER models are described.

The best trypsin model is very similar to the crystallographic trypsin structure (Figs 13 and 14, Table 18). The accuracy of the model is different for buried and exposed parts; thus, the comparison is done separately for the residues that have fractional side-chain solvent accessibility less than 20% (buried residues) and for the remaining residues (exposed residues). Only six of the 107 buried C^α atoms are more than 3·5 Å away from their correct positions, whereas 22 out of 116 exposed C^α atoms are further than 3·5 Å from their positions in the actual trypsin structure. There is no significant difference between the accuracies of the C^α atoms and all main-chain atoms; the r.m.s. error for buried main-chain atoms is approximately 0·7 Å, and for exposed main-chain atoms, approximately 1·0 Å. Comparison of the inter-molecular distances between the superposed C^α positions in the three pairs of elastase–trypsin, tonin–trypsin and the best trypsin model–trypsin is shown in Figure 14. The majority of the aligned positions are close to each other, but there are several exposed regions where the distance between at least one of the template structures and trypsin is larger than 3 Å. In these cases, one of the template structures is generally significantly closer to the crystallographic trypsin structure than the other template structure. The rule for the combination of basis pdfs (eqns (27) and (28)) takes advantage of this common occurrence. In fact, the largest error in the model (alignment positions 160 to 170) corresponds to the region where the wrong template, tonin, was used. This error was the result of the higher local sequence

Table 16
Spatial restraints used to model trypsin

Type	Basis pdfs[a]	Feature pdfs[b]	Violations[c]	r.m.s.[d]	r.m.s.[e]
Bond lengths	1659	1659	0 (0·1 Å)	0·005 Å	0·005 Å
Bond angles	2250	2250	5 (10°)	2·00°	2·00°
Dihedral angles[f]	919	919	1 (20°)	3·40°	3·40°
van der Waals contacts[g]	531	531	0 (0·2 Å)	0·02 Å	0·02 Å
C^{α}–C^{α} distances	23,538	11,914	26 (1·5 Å)	0·22 Å	0·47 Å
Main-chain N–O distances	7480	3832	19 (1·5 Å)	0·31 Å	0·51 Å
Main-chain Φ dihedral angles	1110	222	2 (20°)	10·8°	21·2°
Main-chain Ψ dihedral angles	1332	222	9 (20°)	10·6°	20·3°
Side-chain χ_1 dihedral angles	528	176	5 (25°)	8·4°	16·8°
Side-chain χ_2 dihedral angles	264	103	3 (25°)	10·2°	13·0°
Side-chain χ_3 dihedral angles	92	32	2 (25°)	11·9°	48·1°
Side-chain χ_4 dihedral angles	48	16	0 (25°)	4·5°	21·9°
Disulphide bridge bonds	6	6	0 (0·1°)	0·007 Å	0·007 Å
Disulphide bridge angles	12	12	0 (10°)	3·7°	3·7°
Disulphide bridge dihedral angles	6	12	0 (20°)	10·0°	12·9°
cis-Peptides[h]	0	0			

[a] Lists a number of basis restraints of a given type that were used to model trypsin.

[b] Lists a number of feature restraints of a given type that were assembled from the basis restraints.

[c] For the best model, a number of the features that differ from the closest optimum in the feature pdfs by more than the cutoff in the parentheses is given. These cutoffs generally lie between 1 and 2 standard deviations of the corresponding basis pdfs. The best model is defined as the one with the lowest value of the molecular pdf.

[d] r.m.s. deviation between the actual values in the best model and the closest optimum in the feature pdfs.

[e] r.m.s. deviation between the actual values in the best model and the most likely optimum in the feature pdfs.

[f] These dihedral angles restrain the planarity of peptide bonds and rings as well as chirality of the chiral carbon atoms.

[g] All pairs of atoms that are not restrained by any of the bond or bond angle terms are restrained by the minimal contact distance. Only the number of pairs that violate this restraint in the final model is listed.

[h] There are no *cis*-peptide bonds in trypsin. The only *cis*-peptide bond in tonin is at Pro198, which is aligned with Gly in trypsin (Fig. 10). Therefore, no *cis*-peptide bonds were imposed on trypsin.

Table 17
Evaluation and energy minimization of the MODELLER models (4–32) and the crystallographic trypsin structure (2ptn)

	MODELLER				Energy minimization				
Model	No. of violations	Objective function	r.m.s. 2ptn (Å)	Energy (kcal/mol)	No. of violations	Objective function	Δ r.m.s. (Å)	r.m.s. 2ptn (Å)	Energy (kcal/mol)
4	6/66	9388	0·852	10,386	31/332	31,712	0·718	0·895	−316
6	23/182	10,652	0·872	10,988	35/490	31,732	0·738	0·957	−249
7	21/143	10,493	0·860	11,145	30/541	37,437	0·790	0·957	−302
11	6/93	9621	0·861	10,620	30/443	31,986	0·762	0·953	−320
15	12/62	9659	0·854	11,569	27/295	32,703	0·664	0·922	−319
17	13/130	10,097	0·831	10,851	26/434	30,505	0·760	0·888	−302
18	8/34	9891	0·837	10,813	30/409	35,019	0·752	0·974	−325
24	18/120	10,167	0·857	11,390	35/547	34,854	0·778	0·967	−300
25	15/218	11,205	0·811	11,370	32/636	37,978	0·815	0·958	−287
30	15/180	10,621	0·850	10,897	31/458	36,757	0·768	0·932	−303
32	18/199	11,431	0·856	11,938	31/709	36,366	0·894	0·931	−283
2ptn	29/450	18,070	0·000	9166	27/667	31,452	0·480	0·480	−405

The MODELLER columns refer to the models as obtained by MODELLER. The energy minimization columns refer to these models as refined by energy minimization in CHARMM 22 (MacKerell, Jr *et al.*, unpublished results). Before energy calculations, hydrogen atoms were added to all the structures so that the energy was minimized without moving the heavy atoms. See the text for description of the energy minimization. The first number in the No. of violations columns is the number of violations of stereochemical restraints (bond lengths, bond angles, dihedral angles, and van der Waals contacts), and the second number is the number of violations of the homology-derived restraints. The violation cut-offs and all restraint types are listed in Table 16. The Objective function columns contain the value of the molecular pdf. r.m.s. 2ptn refers to the comparison with the crystallographic trypsin structure; it is calculated only with pairs of C^{α} atoms that are closer than 3·5 Å. There are between 189 and 194 such pairs. In total, there are 223 C^{α} atoms, 1629 heavy atoms and 1603 hydrogen atoms in the trypsin molecule. The r.m.s. values for all heavy atoms are approximately 25% higher than the r.m.s. deviations for C^{α} atoms only. Δ r.m.s. refers to the shifts of all 223 C^{α} atoms caused by energy minimization. When the steepest descent minimization is used instead of the conjugate gradients method, the r.m.s. shifts are of the order of 0·25 Å, and the energies of the refined models are of the order of 600 kcal/mol (not shown).

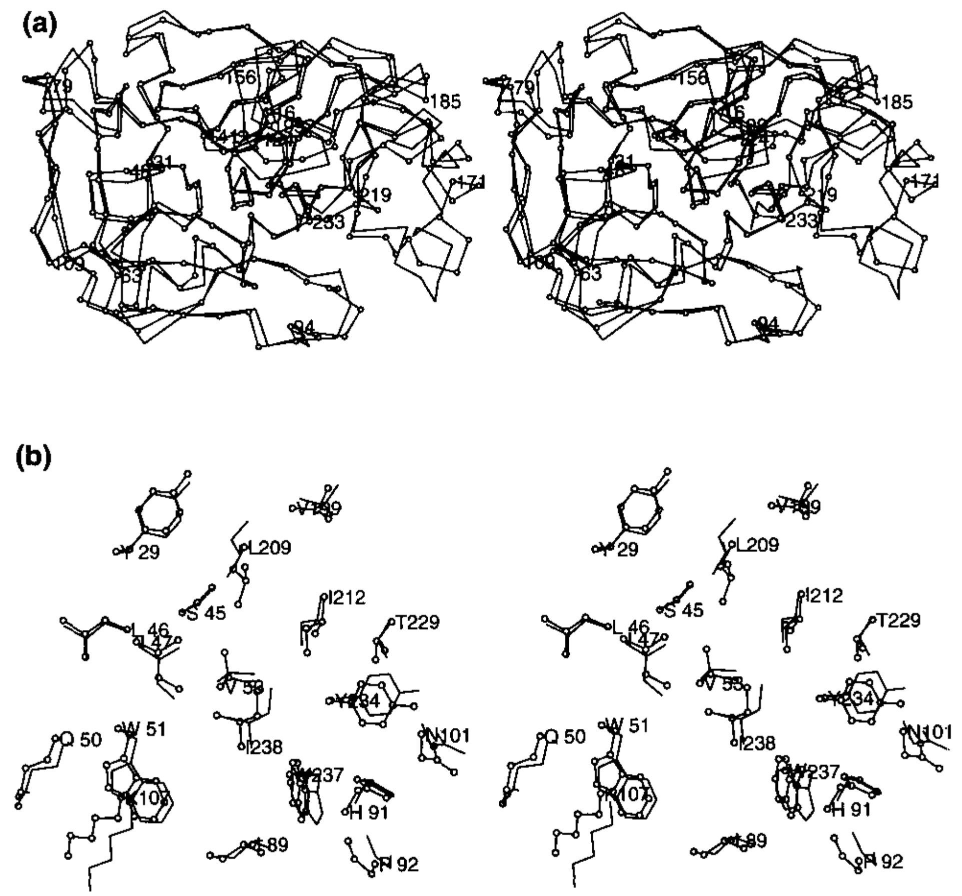

Figure 13. Comparison of the best trypsin model with trypsin. Comparison is obtained by superposing all C^{α} atoms. Chymotrypsinogen numbering is used. Trypsin (bonds with open circles), trypsin model (line). (a) Comparison of the C^{α} traces. (b) Comparison of side-chains in a mostly buried region.

similarity between tonin and trypsin when compared to the local sequence similarity between elastase and trypsin; tonin also has no gaps relative to trypsin, whereas elastase has a two-residue insertion relative to trypsin (Fig. 10). Regardless of these sequence similarities, elastase structure in this

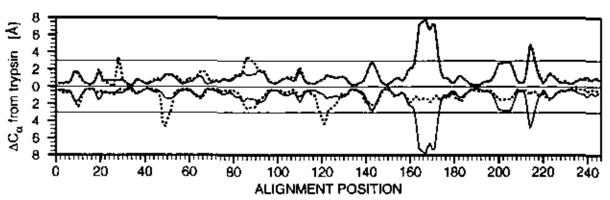

Figure 14. Comparison of the crystallographic trypsin structure with elastase, tonin and the best model of trypsin. Elastase, tonin and the best trypsin model were superposed on trypsin using the C^{α} atoms that are aligned in the COMPARER alignment (Fig. 10). The distances between the aligned C^{α} atoms are plotted for each of the 3 comparisons. The horizontal axis corresponds to the alignment position (line 1 in Fig. 10). Elastase–trypsin and tonin–trypsin comparisons are shown in dotted lines in the top and bottom half of the plot, respectively. The best trypsin model–trypsin comparison is shown in continuous line in both parts of the plot. All 3 curves are smoothed by plotting a value at position i that is an average of the distances at positions $i-1$, i and $i+1$.

region is significantly more similar to trypsin than tonin. Interestingly, one of the 11 models (but not the best one) still selected elastase as the template in this region (Fig. 15). The second largest error in the model (alignment positions 193 to 195) results from a similar situation. Despite these two problems, the optimization of the molecular pdf picks the correct template six times and only misses three times. This success does not appear to be trivial because, in the absence of crystallographic structures, the overall sequence similarity as well as the number and size of the gaps in the alignment would suggest incorrectly that tonin is more similar to trypsin than elastase.

Similarly to the main-chain, buried side-chains were modelled more accurately than exposed side-chains. Eighty-two percent of the buried χ_1 classes and 69% of the exposed classes were predicted correctly. For the χ_2 class, 79% of the buried residues and 80% of the exposed residues were modelled successfully. The average χ_3 prediction score for all χ_3 classes is 68%. There are no buried Arg and Lys residues; they are all exposed and predicted with 75% accuracy.

Four of the six disulphide bridges have the disulphide bridge dihedral angle χ_3 in the correct class (either in the +90° or −90° region). The two bridges that are not modelled correctly are

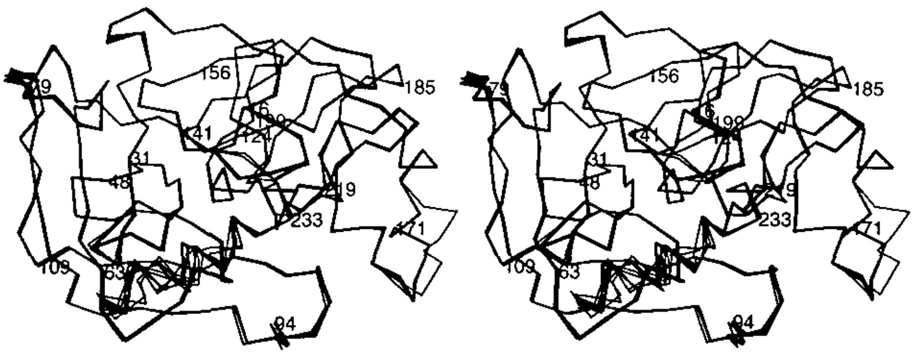

Figure 15. Comparison of the 11 trypsin models. The models are superposed using all the C^α atoms. Chymotrypsinogen numbering is used.

C128–C232, which does not have an equivalent bridge in either tonin or elastase, and C168–C182, which connects the two parts of the main-chain that are modelled least accurately.

The C^α atoms of the 11 models of trypsin are superposed in Figure 15. The variability in the C^α positions is shown more quantitatively in Figure 16. There are five regions where the r.m.s. deviation between the 11 models is larger than 1 Å. These five regions include three segments with the largest errors in the best model; the first and the last variable region do not correspond to the errors in the best model. Thus, the variability among the models can be used as a conservative estimate of the regions in the best model that are most likely to be in error when the experimentally determined structure is not available. The same conclusions are valid for the main-chain N–O distance, and main-chain and side-chain dihedral angles (data not shown).

The modelling example described in this section is not a particularly difficult problem because of a relatively high similarity between the target sequence and the two template structures. There is no region in the target sequence that does not have aligned residues in at least one of the templates. If no equivalent residues in the template structures

Table 18

Comparison of the best trypsin model with trypsin

A. *Main-chains*

		Cut-off = 3·5 Å			No cut-off		
Type of atom	Accessibility	No.	r.m.s.	d.r.m.s.	No.	r.m.s.	d.r.m.s.
C^α	Buried	101	0·689	0·668	107	1·403	0·949
	Exposed	94	1·004	1·073	116	2·041	1·516
	All	195	0·852	0·889	223	1·773	1·265
Main-chain	Buried	400	0·693	0·684	428	1·445	0·971
	Exposed	378	1·017	1·075	464	2·030	1·499
	All	774	0·854	0·906	892	1·781	1·266

B. *Side-chains*

Class	Accessibility	No.	% Correct
χ_1	Buried	82	81·7
	Exposed	94	69·1
	All	176	75·0
χ_2	Buried	47	78·7
	Exposed	56	80·4
	All	103	79·6
χ_3	Buried	5	40·0
	Exposed	26	73·1
	All	31	67·8
χ_4	Buried	0	
	Exposed	16	75·0
	All	16	75·0

A, The 2 main-chains are compared in terms of r.m.s. and distance r.m.s. (d.r.m.s.; Levitt, 1983) for C^α atoms and for all 4 main-chain atoms (N, C^α, C and O). Additionally, one comparison includes only the pairs of aligned atoms that are less than 3·5 Å (cut-off = 3·5 Å) apart, whereas the other comparison (no cut-off) includes all pairs of aligned atoms. The numbers of the equivalent pairs are listed for each case. B, The 2 sets of side-chains are compared in terms of the fractional identity of the side-chain dihedral angle classes for each of the 4 types. The numbers of the angles in each class are also given.

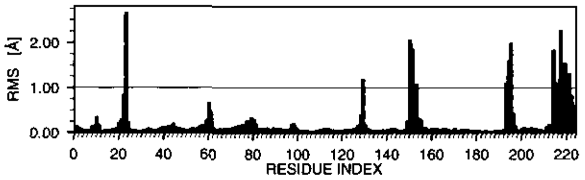

Figure 16. Variability in the positions of the C^{α} atoms in the 11 trypsin models. At each position in the sequence of trypsin, variability is calculated as the r.m.s. deviation among the inter-molecular distances between all pairs of equivalent C^{α} atoms in the 11 models. There are $55 = 11 \times 10/2$ such distances at each sequence position. Before the inter-molecular distance is calculated, the 2 models are superposed using all C^{α} atoms.

were available, MODELLER would use only the main-chain dihedral angle restraints based on the residue type alone. We would not expect such weak restraints to result in an accurate model. Thus, structurally similar segments from the database of all known protein structures would have to be found and added to the alignment. In principle, filtering methods based on the distances between the gap flanking regions (Jones & Thirup, 1986) could be used for this task, but general applicability of this approach is questionable (Tramontano & Lesk, 1992). Another possibility may be an exhaustive conformational search employing energy criteria (Bruccoleri & Karplus, 1987; Moult & James, 1986).

When the starting structure for MODELLER was the actual trypsin structure, the final value of the objective function was never lower than that of the best model calculated with random starting conformations. The value of the objective function for the crystallographic trypsin structure is significantly higher than those for the 11 MODELLER models (Table 17). This suggests that the optimizer is suitable for the problem at hand and that further increase in accuracy of the model will have to come from more accurate restraints, not from a better optimizer.

It is interesting to compare MODELLER to other comparative modelling methods, even if a rigorous comparison is not possible due to, among others, the differences in protein data sets, in proteins modelled, and in measures employed to evaluate the success of the prediction.

For example, Overington (1991) combined careful manual modelling with computer program COMPOSER (Sutcliffe *et al.*, 1987*a*,*b*) to calculate a model of trypsin on the basis of four other serine proteinases. The r.m.s. for the 150 core C^{α} positions was 0·64 Å for his model; it is 0·60 Å for our model, even though only two templates were used by MODELLER.

Recently, Dunbrack & Karplus (1993) developed an elegant automated side-chain prediction method that compares favourably with other side-chain prediction methods. For comparison, we applied the first stage of the method of Dunbrack & Karplus, which is based on the backbone-dependent rotamer library using the template backbone, to the seven serine proteinases in our test set. The prediction success of this procedure was 59% for χ_1 class. The *second refinement stage of the method, which optimizes* packing and energy, improves the accuracy of a model on the average by 6·5% (Dunbrack & Karplus, 1993). Thus, it is reasonable to expect that the method of Dunbrack & Karplus would yield an approximately 65% success rate for the serine proteinases in our test set. This is similar to the 64% success rate obtained with the use of our approach that employed the basis pdf $p(c_1/r, r', c_1', s)$ from a single template (Table 15). When several templates and optimization of the molecular pdf are used, our prediction success is expected to improve, depending on the similarities between the target sequence and the template structures.

The segment match modelling of Levitt (1992) is guided by the positions of some atoms (usually C^{α} atoms) to find the matching segments in the representative database of all known protein structures. Since the matching segments contain the initially missing atoms, the method determines the full atomic model for the initial trace of the guiding atoms. This method can be used for comparative modelling if the C^{α} atoms of a homologous protein are used to guide the segment search. The method appears to be one of the best methods in its class, judging by the accuracy of the constructed models and by the sensitivity of this accuracy to the errors in and the number of the guiding atoms. When an r.m.s. error of less than 1·0 Å is introduced into the guiding positions and all C^{α} positions are used to guide the segment search, the resulting r.m.s. error in the main-chain atoms is 0·8 to 0·9 Å. This should be compared with an r.m.s. error of 0·69 Å for the buried main-chain atoms in the trypsin model (Table 18). For side-chain χ_1 and χ_2 dihedral angles, the average prediction successes of the segment matching method are 72% and 59%, respectively, when the exact C^{α} positions from the target are used to guide the search. The dihedral angle was considered correct when it was within 30° of the exactly correct value. Since the average standard deviation of the angles within each class is approximately 10° (Table 5), the prediction successes must increase by 3·5% for comparison with our criterion. The prediction successes for all χ_1 and χ_2 classes in the MODELLER trypsin model are 75% and 79%, respectively. However, these comparisons are approximate because the MODELLER number is based on only one model, because using the homologous backbone is likely to decrease the accuracy of the side-chain prediction in the segment modelling, and because it is not clear how the sequence similarity between tonin, elastase and trypsin compares with the similarities between the database and the test set of Levitt (1992).

In conclusion, MODELLER appears to be at least as accurate as the other manual or automated knowledge-based methods.

5. Discussion

(a) *Applicability of frequency tables*

Frequency tables and related matrices are commonly used to analyse or predict attributes of protein structure. Dayhoff's MDM250 mutation matrix (Dayhoff *et al.*, 1978) is a close analogue of a 2D table **W**, where one axis represents the residue type at any position in the first protein and the other axis, the residue type at the possibly equivalent position in the second protein. The Dayhoff matrix element measures the likelihood that the two residues are actually equivalent. Overington and co-workers have used multi-dimensional forms of the probability tables **W** and their transformations to search for combinations of protein features that are conserved in evolution (Overington *et al.*, 1990, 1992). Similar matrices were used to detect distantly related sequences (Lüthy *et al.*, 1991), to identify sequences that fold into a known 3D structure (Bowie *et al.*, 1991) and to assess protein 3D models (Lüthy *et al.*, 1992). Other examples of frequency tables and closely related matrices include the Ramachandran plot obtained from a database of known protein structures (Wilmot & Thornton, 1990), various parameter sets for secondary structure prediction (Chou & Fasman, 1974), side-chain rotamer libraries (Dunbrack & Karplus, 1993; Janin *et al.*, 1978; Ponder & Richards, 1987) and hydrophobicity scales found by analysing the known protein structures (Manavalan & Ponnuswamy, 1978).

We describe a systematic and quantitative approach to searching for significant associations between the features of protein structure. This approach exploits the database of known protein structures and their alignments. It is based on expressing the association between selected features as a conditional pdf and on quantifying the strength of the association by entropy, conditional entropy and, where possible, by the prediction success of the tested pdf. To facilitate the derivation, analysis and use of these pdfs, the smoothing procedure of Sippl (1990) is extended to multidimensional tables.

The usefulness of this approach is illustrated by using such associations to model the structure of a protein given its alignment with related structures. It is shown that a certain C^α–C^α distance has a Gaussian distribution around the equivalent distance in a homologous structure. The standard deviation of this distribution is determined as a function of the local environment. When more than one known structure is aligned with the given sequence, the most likely distance in the sequence is not the average of the equivalent distances from the known structures, but the distance from the related structure that has the most similar local environment. The same conclusions are also valid for main-chain N–O distances.

Pdfs are also used to model main-chain dihedral angles Φ and Ψ of a residue in a sequence, given its alignment with a related structure. Main-chain conformation is described as one of the six classes corresponding to the six populated areas in the Ramachandran plot (Wilmot & Thornton, 1990), similar, but not identical, to secondary structure types. The best of the 7249 different pdfs tested takes into account the main-chain conformation class of an equivalent residue, as well as the type of the modelled residue and the sequence similarity of the two equivalent local environments. When pairs of homologous serine proteinases with the average sequence identity of 35% are used as the test case, approximately 73% of residues are predicted correctly. Note that this prediction success is achieved using only one related structure and, without further refinement of the initial prediction in the context of the whole structure.

A number of pdfs (10,068), or rotamer libraries, are constructed to find the best pdf for comparative modelling of side-chain conformations. The side-chain conformation of a residue is described by a small number (1 to 3) of side-chain conformation classes for each side-chain dihedral angle that exists in the given residue. Even though the main-chain conformation of a residue being modelled strongly determines the conformation of its side-chain (Dunbrack & Karplus, 1993), the information provided by the main-chain conformation of an equivalent residue in a homologous structure is small. The best pdf for modelling side-chain conformation takes into account the side-chain conformation of the equivalent residue and the similarity between the two local environments. The overall prediction successes of the pdfs for the χ_1, χ_2, χ_3 and χ_4 dihedral angles are 64%, 72%, 61% and 75%, respectively, for the pairs of serine proteinases in our test set. The pdfs used here appear to give at least as good a prediction as one of the best of the published automated methods (Dunbrack & Karplus, 1993).

(b) *Modelling by satisfaction of spatial restraints*

We describe the use of an alignment of a target sequence with several related template structures to extract many spatial restraints on the structure of the target sequence. These pdfs constrain stereochemistry, main-chain and side-chain conformation, C^α–C^α and main-chain N–O distances. The aim is to find the 3D structure of the target that is consistent with most probable values of the constrained features. The solution to this problem is achieved by combining all the pdfs into a single molecular pdf such that an optimization of this function leads to the most probable model given the alignment. The molecular pdf is a product of pdfs that constrain individual distances and angles, which themselves may be sums of pdfs obtained from the individual homologous proteins. By optimizing the molecular pdf, violations of the restraints by the model are minimized. The optimization is implemented in the MODELLER program. This program applies the variable target function approach (Braun & Gō, 1985) with the conjugate gradients algorithm to the positions of all non-hydrogen atoms. In this approach, the optimization starts from a random initial conformation and initially uses only the sequentially local restraints. It then proceeds in a number of steps to increase the number of restraints until, finally, all the restraints are included and their violations minimized. Both the extraction of restraints from the alignment and the optimization of the molecular pdf are fully automated.

The method was illustrated by applying it to modelling of trypsin from elastase and tonin, relying on the approximately 40% sequence identity. Eleven models with small violations of restraints were calculated using different initial conformations in MODELLER. The model with the lowest value of the molecular pdf was compared with trypsin. The r.m.s. error for all backbone atoms was found to be approximately 0·7 Å for the 195 equivalent residues that had their C^α atoms less than 3·5 Å apart; 28 residues on the periphery of the trypsin fold were modelled less accurately. Approximately 80% of buried and 73% of exposed side-chain dihedral angle classes χ_1 to χ_4 were modelled correctly. The variability among the 11 models can be used to indicate the errors in the best model. The accuracy of the trypsin model is similar to that of the structures from medium resolution crystallography and NMR experiments (Clore & Gronenborn, 1991). Even though no rigorous comparison of MODELLER with other modelling methods is possible, it appears that MODELLER, in its current form, is at least as accurate as any of the other manual or automated methods based on the homologous or all protein structures.

In the future, MODELLER may be improved by using a larger database to get more accurate restraints, especially for side-chain and main-chain dihedral angles; using additional restraint types, for example, distance restraints involving C^β atoms; analysing the side-chain errors to improve side-chain modelling, possibly by including energy criteria such as hydrogen bonds and solvation terms; using a multiple alignment of the target sequence with many shorter segments corresponding to the variable regions with gaps and possibly finding and applying pdfs specific for these regions; refining the rule for the combination of basis pdfs into feature pdfs.

Recently, MODELLER was applied to calculate 3D models of four mouse mast cell chymases. These models were examined to propose site-directed mutagenesis experiments for identification of proteoglycan-binding regions and to suggest the amino acid segments for raising the protease-specific antigenic epitopes (Šali *et al.*, 1993).

(c) *Comparison of MODELLER with other approaches to protein structure prediction*

This section compares MODELLER with prediction methods based on distance geometry and energy. The comparison focuses on the type of information employed by the methods, not on the techniques using this information or on their performance.

(i) *Distance geometry*

The aim of the metric method of distance geometry is the derivation of the atomic positions consistent with protein stereochemistry and with a limited number of lower and upper bounds on the distances between these atoms. Distance geometry has been used with NMR-derived constraints (Braun & Gõ, 1985; Havel & Wüthrich, 1985) as well as with distance constraints obtained from homologous structures (Havel & Snow, 1991).

MODELLER is similar to distance geometry because it also derives the Cartesian co-ordinates from restraints on spatial features of the sequence to be modelled. However, MODELLER is an extension of distance geometry in a sense that pdfs also restrain features other than distances and that a restraint expressed as a pdf contains more information than a mere specification of lower and upper bounds. From the point of view of information used, MODELLER would emulate distance geometry if feature pdfs for distances had the simple form of the uniform distribution between the bounds, and zero elsewhere.

(ii) *Energy minimization*

MODELLER calculates the model as the most probable structure for a certain sequence given its alignment with related structures, whereas energy minimization methods calculate the model as the lowest energy structure given a force field (Berendsen *et al.*, 1984; Brooks *et al.*, 1983). It is tempting to equate the logarithm of the molecular pdf to the total energy of the system and the logarithms of basis pdfs to the individual energy terms. The analogy is exact for the basis pdfs that constrain the stereochemical features (section 2(e)). However, all other basis pdfs in MODELLER have a purely empirical origin. Thus, a MODELLER optimization does not have any physical significance such as that of the individual energy terms, total energy and molecular dynamics in molecular mechanics. This difference extends the amount of information that can be used in the derivation of the MODELLER model. When the only aim is to predict the native structure, it is more useful to process the information about protein structure at the "probability level", not the "energy level". For example, from the point of view of energy, it would be natural to constrain a certain distance by a sum of quadratic terms each term corresponding to the equivalent distance from one homologous structure; this would force the distance in the model to be between the distances in the known structures. From the point of view of probability, however, we are naturally led to the correct form of the restraint which is a weighted sum of the Gaussian distributions, not their product. This is consistent with the view that crystallographic structures in the family populate most of the minima of the energy landscape for the whole ensemble of sequences belonging to the given family; then, any other sequence in the family is likely to be at an already occupied minimum, resulting in the observed rule for the distribution of the triplets of equivalent distances.

Intermediate between the pdfs developed in this paper and the atomic energy potentials are the potentials of mean force for protein structure prediction from sequence alone (Avbelj, 1992; Casari & Sippl, 1992; Jones *et al.*, 1992; Sippl, 1990).

(d) *Future directions*

The challenge is to unify all the techniques for determination and prediction of protein structure into a single protocol, making the best use of all available information about the structure of a given protein, regardless of whether it is directly based on experiment, on the broader knowledge base, on empirical force potentials, or intuition. The methods that combine molecular dynamics and energy potentials with NMR-derived constraints (Brünger *et al.*, 1987*a*; Clore *et al.*, 1986) and X-ray data (Brünger *et al.*, 1987*a*,*b*) to refine the initial models can be seen as the first step in this direction. Recently, the advantages of a joint crystallographic and NMR refinement were demonstrated (Shaanan *et al.*, 1992).

The following argument illustrates the benefits of a combined method. Before we start prediction of the 3D structure of a protein, we know nothing about positions of the atoms. In the terminology of classical mechanics, the actual structure could be a point anywhere in the phase space spanned by the axes for the positions of all atoms. We can then imagine modelling as a process of reducing the volume of phase space in which we know the actual structure is located. This is achieved by using various kinds of information. First, stereochemical restraints derived from the chemical connectivities can be used to remove some of the *a priori* available phase space. This can be pursued further by inclusion of experimental data, such as that from X-ray crystallography and NMR techniques. We can also add additional theoretical restraints originating from empirical energy potentials and known protein structures. Each of these kinds of information allows the model to be in a different area of the phase space with a different probability. The goal is to find the most probable conformation or a set of most probable conformations according to all types of information. All the information pooled together results in a smaller allowed volume of phase space than any of the methods can locate on their own.

The most useful representation of information is a pdf for the feature that is constrained. The present modelling method uses pdfs in a relatively general way. Thus, the method, even though it has so far been applied only to comparative modelling, could possibly be extended to include other types of information, such as NMR-derived constraints.

6. Summary and Conclusions

(1) A database of family alignments for proteins with known structures was constructed.

(2) It was shown how to use pdfs and other tools to explore quantitatively various relationships between features in individual proteins and in families of proteins.

(3) A method for minimizing the problems of a sparse data set was described and shown to improve the usefulness of the pdfs.

(4) Using these tools and the current database, the best pdfs for comparative modelling of a side-chain conformation of a given residue were constructed. They relied mainly on its type, on the side-chain conformation of the equivalent residue and on the similarity between the two local environments.

(5) The best possible pdf for modelling the main-chain conformation from the main-chain of a homologue was found. It was based on the main-chain conformation of the equivalent residue and on the similarity between the two local environments.

(6) The pdfs for restraining the C^α–C^α distances and the main-chain N–O distances on the basis of homologous structures were calculated. It was shown that the most likely distance corresponded to that in one of the related structures, not to the average of the equivalent distances in the related structures.

(7) A method was developed for calculating the most probable structure for a certain sequence, given its alignment with one or more related structures and the general rules of protein structure.

(8) Once the alignment is determined, the method is completely automated. It can provide a 3D model equivalent to a medium resolution X-ray structure when homologues with at least 40% sequence identity are known. This means that an order of magnitude more sequences can be modelled at a medium resolution that there are entries in the Brookhaven Protein Databank.

We are indebted to John Overington and Mark Johnson for discussions about protein modelling, and to Roland Dunbrack and Martin Karplus for making their manuscript available before publication. We are also grateful to Daša Šali for critical reading of the manuscript and to Igor Podboj of Neo-Systems, Salt Lake City, U.S.A., for advice on the PC-i860 computer system. We also thank David Smith for the SSTRUC program. A.Š. was funded by ORS Awards Scheme, Research Council of Slovenia, Merck Sharp & Dohme, J. Stefan Institute (Slovenia) and Imperial Cancer Research Fund.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases – Information, Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds), pp. 107–132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Avbelj, F. (1992). Use of a potential of mean force to analyze free energy contributions in protein folding. *Biochemistry,* **31**, 6290–6297.

Bassolino-Klimas, D. & Bruccoleri, R. E. (1992). Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. *Proteins,* **14**, 465–474.

Bax, A. (1989). Two-dimensional NMR and protein structure. *Annu. Rev. Biochem.* **58**, 223–256.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.

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

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography,* Academic Press, New York.

Blundell, T. L. & Sternberg, M. J. E. (1985). Computer-aided design in protein engineering. *Trends Biotechnol.* **3**, 228–235.

Blundell, T. L., Barlow, D., Sibanda, B. L., Thornton, J. M., Taylor, W. R., Tickle, I. J., Sterberg, M. J. E., Pitts, J. E., Haneef, I. & Hemmings, A. M. (1986). Three-dimensional structural aspects of the design of new protein molecules. *Phil. Trans. Roy. Soc. Lond. ser. A,* **317**, 333–344.

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

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. & Petersen, S. B. (1990). A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Letters,* **261**, 43–46.

Bowie, J. U., Lütthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164–170.

Braun, W. & Gō, N. (1985). Calculation of protein conformations by proton–proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **186**, 611–626.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.

Brooks III, C. L., Karplus, M. & Pettit, B. M. (1988). *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics,* John Wiley & Sons, New York.

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

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

Brünger, A. T., Campbell, R. L., Clore, G. M., Gronenborn, A. M., Karplus, M., Petsko, G. A. & Teeter, M. M. (1987*a*). Solution of a protein crystal structure with a model obtained from NMR interproton distance restraints. *Science,* **235**, 1049–1053.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987*b*). Crystallographic R-factor refinement by molecular dynamics. *Science,* **235**, 458–460.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential: hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature (London),* **360**, 543–544.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). Conformation of immunoglobulin hypervariable regions. *Nature (London),* **342**, 877–883.

Chou, P. Y. & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry,* **13**, 222–245.

Claessens, M., Cutsem, E. V., Lasters, I. & Wodak, S. (1989). Modelling the polypeptide backbone with "spare parts" from known protein structures. *Protein Eng.* **4**, 335–345.

Clore, G. M. & Gronenborn, A. M. (1991). Two-, three-, and four-dimensional NMR methods for obtaining larger and more precise three-dimensional structures of proteins in solution. *Annu. Rev. Biophys. Chem.* **20**, 29–63.

Clore, G. M., Brünger, A. T., Karplus, M. & Gronenborn, A. M. (1986). Application of molecular dynamics with interproton distance restraints to 3D protein structure determination. *J. Mol. Biol.* **191**, 523–551.

Cohen, F. E. & Kuntz, I. D. (1989). Tertiary structure prediction. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), pp. 647–705, Plenum Press, New York.

Correa, P. E. (1990). The building of protein structures from α-carbon coordinates. *Proteins,* **7**, 366–377.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345–352, National Biomedical Research Foundation, Washington, DC.

Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature (London),* **356**, 539–542.

Dudek, M. J. & Scheraga, H. A. (1990). Protein structure prediction using a combination of sequence homology and global energy minimization. I. Global energy minimization of surface loops. *J. Comp. Chem.* **11**, 121–151.

Dunbrack, R. L. & Karplus, M. (1993). Prediction of protein side-chain conformations from a backbone conformation dependent rotamer library. *J. Mol. Biol.* **230**, 543–571.

Fasman, G. D. (1989). *Prediction of Protein Structure and the Principles of Protein Conformation,* Plenum Press, New York.

Friedrichs, M. S., Goldstein, R. A. & Wolynes, P. G. (1991). Generalized protein tertiary structure recognition using associative memory Hamiltonians. *J. Mol. Biol.* **222**, 1013–1034.

Fujinaga, M. & James, M. N. G. (1987). Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1·8 Å resolution. *J. Mol. Biol.* **195**, 373–396.

Fujiyoshi-Yoneda, T., Yoneda, S., Kitamura, K., Amisaki, T., Ikeda, K., Inoue, M. & Ishida, T. (1991). Adaptability of restrained molecular dynamics for tertiary structure prediction: application to *Crotalus* *atrox* venom phospholipase A_2. *Protein Eng.* **4**, 443–450.

Godzik, A., Kolinski, A. & Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.* **153**, 1027–1042.

Greer, J. (1990). Comparative modelling methods: application to the family of the mammalian serine proteases. *Proteins*, **7**, 317–334.

Havel, T. F. & Snow, M. E. (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.* **217**, 1–7.

Havel, T. & Wüthrich, K. (1985). An evaluation of the combined use of NMR and distance geometry for the determination of protein conformations in solution. *J. Mol. Biol.* **182**, 281–294.

Hill, T. L. (1960). *An Introduction to Statistical Thermodynamics.* Addison-Wesley Publishing Company, Reading, MA.

Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from C^α trace: application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **218**, 183–194.

Holm, L. & Sander, C. (1992). Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins*, **14**, 213–223.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1968.

Hubbard, T. J. P. & Blundell, T. L. (1987). Comparison of solvent inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125**, 357–386.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature (London)*, **358**, 86–89.

Jones, T. A. (1978). A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11**, 268–272.

Jones, T. H. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kendrew, J. C., Klyne, W., Lifson, S., Miyazawa, T., Némethy, G., Phillips, D. C., Ramachandran, G. N. & Scheraga, H. (1970). IUPAC-IUB commission on biochemical nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. *J. Mol. Biol.* **52**, 1–17.

Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217**, 373–388.

Lesk, A. M. & Chothia, C. H. (1986). The response of protein structures to amino-acid sequence changes. *Phil. Trans. Roy. Soc. Lond.* **317**, 345–356.

Levitt, M. (1983). Molecular dynamics of native protein. II. Analysis and nature of motion. *J. Mol. Biol.* **168**, 621–657.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533.

Luo, Y., Jiang, X., Lai, L., Qu, C., Xu, X. & Tang, Y. (1992). Building protein backbones from C_α coordinates. *Protein Eng.* 5, 147–150.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature (London)*, **356**, 83–85.

Lüthy, R., McLachlan, A. D. & Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, **10**, 229–239.

Manavalan, P. & Ponnuswamy, P. K. (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature (London)*, **275**, 673–674.

Martin, A. C. R., Cheetham, J. C. & Rees, A. R. (1989). Modeling antibody hypervariable loops: a combined algorithm. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 9268–9272.

Mas, M. T., Smith, K. C., Yarmush, D. L., Aisaka, K. & Fine, R. M. (1992). Modeling the anti-CEA antibody combining site by homology and conformational search. *Proteins*, **14**, 483–498.

McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295–310.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.

Meyer, E., Cole, G., Radakrishnan, R. & Epp, O. (1988). Structure of native porcine pancreatic elastase at 1·65 Å resolution. *Acta Crystallogr. sect. B*, **44**, 26–38.

Moult, J. & James, M. N. G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins*, **1**, 146–163.

Overington, J., Johnson, M. S., Šali, A. & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity; templates, key residues and structure prediction. *Proc. Roy. Soc. Lond. sect. B*, **241**, 132–145.

Overington, J., Donnelly, D., Johnson, M. S., Šali, A. & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226.

Overington, J. P. (1991). Knowledge-based protein modelling. PhD thesis, University of London, London.

Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng.* **5**, 121–137.

Payne, P. W. (1993). Reconstruction of protein conformations from estimated positions of the C_α coordinates. *Protein Sci.* **2**, 315–324.

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

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes*, Cambridge University Press, Cambridge.

Reid, L. S. & Thornton, J. M. (1989). Rebuilding flavodoxin from C_α coordinates: a test study. *Proteins*, **5**, 170–182.

Rey, A. & Skolnick, J. (1992). Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. *J. Comp. Chem.* **13**, 443–456.

Richmond, T. J. & Richards, F. M. (1978). Packing of α-helices. Geometrical constraints and contact areas. *J. Mol. Biol.* **119**, 537–555.

Robson, B., Platt, E., Fishleigh, R. V., Marsden, A. & Millard, P. (1987). Expert system for protein engineering: its application in the study of chloramphenicol acetyltransferase and avian pancreatic polypeptide. *J. Mol. Graph.* **5**, 5–17.

Šali, A. (1991). Modelling three-dimensional structure of proteins from their sequence of amino acid residues. PhD thesis, University of London, London.

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

Šali, A., Overington, J. P., Johnson, M. S. & Blundell, T. L. (1990). From comparisons of protein sequences and structures to protein modelling and design. *TIBS*, **15**, 235–240.

Šali, A., Matsumoto, R., McNeil, H. P., Karplus, M. & Stevens, R. L. (1993). Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan-binding regions and protease-specific antigenic epitopes. *J. Biol. Chem.* **268**, 9023–9034.

Schiffer, C. A., Caldweml, J. W., Kollman, P. A. & Stroud, R. M. (1990). Prediction of homologous protein structures based on conformational searches and energetics. *Proteins*, **8**, 30–43.

Shaanan, B., Gronenborn, A. M., Cohen, G. H., Gilliland, G. L., Veerapandian, B., Davies, D. R. & Clore, G. M. (1992). Combining experimental information from crystal and solution studies: joint X-ray and NMR refinement. *Science*, **257**, 961–964.

Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). 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**, 759–777.

Singh, J. & Thornton, J. M. (1990). SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. *J. Mol. Biol.* **17**, 195–225.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

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

Snow, M. E. (1993). A novel parameterization scheme for energy equations and its use to calculate the structure of protein molecules. *Proteins*, **15**, 183–193.

Sowdhamini, R., Srinivasan, N., Shoichet, B., Santi, D. V., Ramakrishnan, C. & Balaram, P. (1989). Stereochemical modeling of disulphide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng.* **3**, 95–103.

Srinivasan, S., March, C. J. & Sudarsanam, S. (1993). An automated method for modeling proteins on known templates using distance geometry. *Protein Sci.* **2**, 227–289.

Stewart, D. E., Weiner, P. K. & Wampler, J. E. (1987). Prediction of the structure of proteins using related structures, energy minimisation and computer graphics. *J. Mol. Graph.* **5**, 133–140.

Summers, N. L. & Karplus, M. (1989). Construction of side-chains in homology modelling. Application to the C-terminal lobe of rhizopuspepsin. *J. Mol. Biol.* **210**, 785–811.

Summers, N. L. & Karplus, M. (1990). Modeling of globular proteins. A distance-based search procedure for the construction of insertion/deletion regions and Pro→non-Pro mutations. *J. Mol. Biol.* **216**, 991–1016.

Summers, N. L., Carson, W. D. & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol.* **196**, 175–198.

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

Sutcliffe, M. J., Hayes, F. R. F. & Blundell, T. L. (1987*b*). Knowledge based modeling of homologous proteins, Part II: rules for the conformation of substituted sidechains. *Protein Eng.* **1**, 385–392.

Swindells, M. B. & Thornton, J. M. (1991). Modelling by homology. *Curr. Opin. Struct. Biol.* **1**, 219–223.

Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287.

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

Topham, C. M., Thomas, P., Overington, J. P., Johnson, M. S., Eisenmenger, F. & Blundell, T. L. (1991). An assessment of COMPOSER: a rule-based approach to modelling protein structure. *Biochem. Soc. Symp.* **57**, 1–9.

Topham, C. M., McLeod, A., Eisenmenger, F., Overington, J. P., Johnson, M. S. & Blundell, T. L. (1993). Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* **229**, 194–220.

Tramontano, A. & Lesk, A. M. (1992). Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins*, **13**, 231–245.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. A. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267–1289.

Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989). A 3-D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, **5**, 355–373.

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W. & Huber, R. (1982). On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography. *Acta Crystallogr. sect. B*, **38**, 1462–1472.

Warme, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W. & Scheraga, H. A. (1974). Computation of structures of homologous proteins: α-lactalbumin from lysozyme. *Biochemistry*, **13**, 768–782.

Wilmot, C. M. & Thornton, J. M. (1990). β-Turns and their distortions: a proposed new nomenclature. *Protein Eng.* **3**, 479–493.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with Crambin. *Proteins*, **6**, 193–209.

Wilson, C., Gregoret, L. M. & Agard, D. A. (1993). Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996–1006.

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

Edited by F. Cohen