Protein Structure by Distance Analysis

edited by

H. Bohr and S. Brunak

Center for Biological Sequence Analysis
The Technical University of Denmark
Lyngby, Denmark

1994

IOS Press

Amsterdam • Oxford • Washington DC

Ohmsha

Tokyo • Osaka • Kyoto

Comparative Protein Modeling by Satisfaction of Spatial Restraints

Andrej Šali and Tom Blundell[1]

Department of Chemistry, Harvard University, Cambridge, MA 02138, USA
[1] ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, London WC1E 7HX, England

Abstract

We describe a comparative protein modeling 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 (pdf's) for the features restrained. For example, the probabilities for main chain conformations of a modeled 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 pdf's are obtained from the correlations between structural features in 98 families of homologous proteins which have been aligned on the basis of their 3D structures. The pdf's 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 pdf's 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 modeling of trypsin from two other serine proteases.

1 Introduction

Comparative protein modeling uses experimentally determined protein structures to predict conformation of other proteins with similar amino acid sequences [for reviews see refs. (1) and (2)]. This is possible because a small change in the sequence usually results in a small change in the 3D structure (3,4). The accuracy of protein models obtained by comparative modeling compares favorably with that of models calculated by other theoretical methods. The comparative method produces models with an RMS error as low as 1 Å for sequences that have sufficiently similar homologues with known 3D structures (5). However, comparative modeling is 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 (6), we can estimate that an order of magnitude more sequences

	1	2	3	4	5	6	7
structure A	A	f	s̄	t	l	N̄	t
structure B	A	f	s̄	s	i	N̄	t̄
structure C	A	Ŷ	p	s	i	S̄	a
sequence X	G	F	D	T	I	T	T
extrapolation				|		
structure X	G	f	d̄	t	i	T̲	t

Figure 1: Comparative protein modeling 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 (11): 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 three 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.

can be modeled by comparative modeling than compared to the 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.

Future improvements of comparative modeling should aim to model proteins with lower homology to known structures, to increase the accuracy of the models, to make modeling fully automated, and to allow inclusion of many different types of information. 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 (7,8,9). The modeling method was developed to use as many different types of data about the unknown as possible (2). The method consists of three stages: alignment of the sequence to be modeled with related protein structures and segments, extraction of spatial restraints on the sequence using the alignment, and satisfaction of the restraints to obtain a 3D model. This paper describes the procedures involved in the last two stages and illustrates the approach by application to modeling trypsin from two other serine proteases.

Table 1: Size of the alignments database.

Alignments	98
Protein structures	379
Homologous protein pairs	2,188
Residues	70,996
Equivalent residue pairs	412,910
Intra-molecular residue pairs	21,453,983
Equivalent intra-molecular residue pairs	117,959,350

2 Derivation of spatial Restraints

In this section, relatively simple restraints on the protein conformation are defined from the information about a related protein structure. A restraint is most precisely defined in terms of a pdf, $p(x/a, b, \ldots, c)$, for the feature x that is restrained. This is a conditional pdf and gives a probability density for x when $a, b, \ldots, c$ are specified. 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 dihedral angles Φ and Ψ.

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

$$p(x/a, b, \ldots, c) \approx W_{x,a,b,\ldots,c} \approx f(x, a, b, \ldots, c, \mathbf{q}) \quad (1)$$

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 whose parameters $\mathbf{q}$ are fitted to the observed $\mathbf{W}$. The multidimensional table of relative frequencies $\mathbf{W}$ is calculated from the absolute frequencies $\mathbf{W}'$ using

$$W_{x,a,b,\ldots,c} = \frac{W'_{x,a,b,\ldots,c}}{\sum_x W'_{x,a,b,\ldots,c}}. \quad (2)$$

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.

Initially, a small database of 17 family alignments was built (2,10) by the protein structure comparison program COMPARER (7,8). This small database was then gradually extended and used to obtain environment specific residue substitution tables (11,12), to improve homology modeling of loops (13), and to increase the sensitivity and accuracy of aligning sequences with structures (14). Recently, a large number of alignments was collected and presented (15). Currently, 378 members of 98 families of related proteins were extracted from the Brookhaven Protein Databank (16) and aligned both by COMPARER and MNYFIT (17) to obtain multiple alignments for each of the families (A. Šali & J. Overington, in preparation).[1] The size of the database is illustrated further in Table 2. A number of features of protein structures were also calculated and stored in the database (Table 2).

The program MDT was written to explore the alignments database and to derive the best pdf's for comparative modeling. The inputs to the program are names of features selected

[1]Several results in this paper were derived from the first smaller database of only 17 alignments.

Table 2: Features used in this paper that may be selected in MDT to span multi-dimensional frequency tables $\mathbf{W}'$. The first column lists the variable names that are used for these features. It also indicates whether an intra-molecular average or inter-molecular difference can be calculated. The $\hat{}$ symbol indicates an average of the feature at two residue positions in the same protein, such as an average accessibility of a certain residue pair. Features that are not associated with two proteins can be used independently for two related proteins in a pairwise alignment or for three 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 two primes that it is from the third protein. The Δ symbol refers to the difference between features f and f': $\Delta f = f - f'$.

Variable	Feature
r	amino acid residue type
$\Phi, \Delta\Phi$	main chain dihedral angle Φ
Φ_c	main chain dihedral angle Φ class
$\Psi, \Delta\Psi$	main chain dihedral angle Ψ
Ψ_c	main chain dihedral angle Ψ class
$\omega, \Delta\omega$	main chain dihedral angle ω
ω_c	main chain dihedral angle ω class
χ_i	side chain dihedral angle χ_i, $i = 1,2,3,4,5$
c_i	side chain dihedral angle χ_i class, $i = 1,2,3,4,5$
t	secondary structure class of a residue (positive Φ, α, β, other)
M	main chain conformation class of a residue (24)
a	fractional content of residues in the main chain conformation class A
S	side chain conformation class (χ_1, χ_2)
$a, \bar{a}$	(fractional) contact solvent area of a residue
$s, \bar{s}$	residue neighborhood difference between two proteins
i	fractional sequence identity between two proteins
$d, \Delta d$	distance between two specified atom types
b	average residue isotropic temperature factor
R	resolution of X-ray analysis
n	number of atomic contacts with non-protein non-water atoms per residue
$g, \bar{g}$	distance of a residue from a gap in the alignment
l	number of residues in the protein
G	several residue type groups (*e.g.* hydrophobic/hydrophilic)

from Table 2, 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 alignments database. The tables $\mathbf{W}'$ are subsequently used as outlined above to calculate the relative frequency tables $\mathbf{W}$ and sometimes the corresponding pdf's 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 (18) was implemented in the program LSQ. Since the database is sparse for some pdf's, smoothing of the pdf's was also performed as described in (2); the smoothing method is an

extension of the procedure proposed in ref. (19).

2.1 Stereochemical Restraints

All stereochemical restraints are easily obtained from the amino acid sequence of a protein. Stereochemical restraints include bond distances, bond angles, torsional angles defined by three consecutive bonds, and improper dihedral angles as used in CHARMM (20). The corresponding pdf's are obtained from the CHARMM 22 parameter set (21), based on classical statistical mechanics (22).[2] For example, the pdf for the bond length is a Gaussian probability density function

$$p^b(b) = \frac{1}{\sigma_b\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{b-\bar{b}}{\sigma_b})^2] = N(\bar{b}, \sigma_b), \tag{3}$$

where $\sigma_b = \sqrt{kT/c}$; c and $\bar{b}$ are the CHARMM force constant and mean length, respectively.

The following pdf is used for two atoms restrained by the van der Waals repulsion

$$p^v(d) = c \cdot \begin{cases} N(d_o, \sigma_w); & d \le d_o \\ \frac{1}{\sigma_w\sqrt{2\pi}}; & d_o < d < d_{max} \end{cases} \tag{4}$$

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 0.05Å). 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.

2.2 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. Using the database of alignments described above, the distribution of $d - d'$ was found 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' structure ($\bar{a}'$), and the average number of positions that separate the two residues spanning the distance from the closest gap in the alignment ($\bar{g}$).

Examples of the histograms of probability distributions obtained by the MDT program are shown in Figures 2a–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

[2]Several results in this paper were derived with an earlier version of our program that relied on the GROMOS86 IFP37C4 parameter set (23).

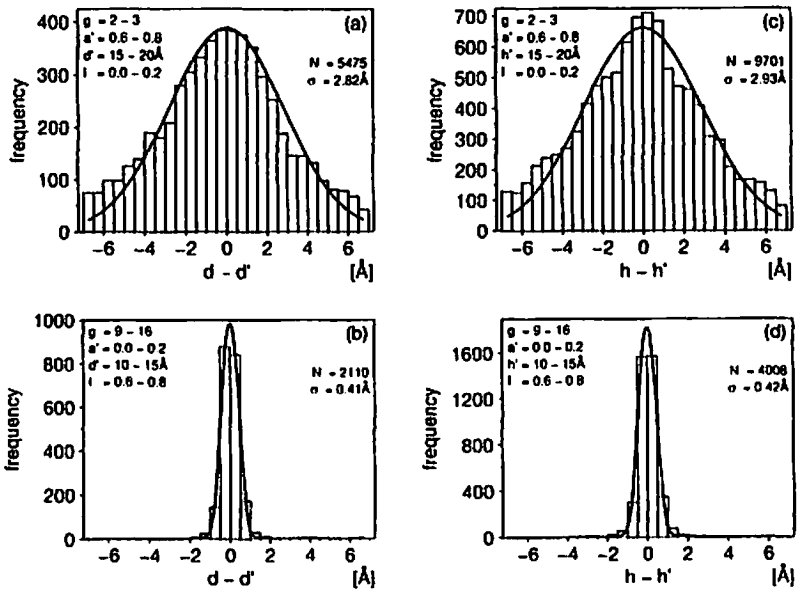

Figure 2: Distribution of the differences between two equivalent distances. The histograms show the frequency of the differences between two equivalent distances as observed by MDT in the alignments database. (a) and (b), C^α–C^α distances; (c) and (d), main chain N–O distances. The curves are fitted Gaussian models [Eqs. (5) and (8)]. 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.

with a single related known structure, can be modeled as

$$\begin{aligned} p^d(d/\bar{g}, i, \bar{a}', d') &= \frac{1}{\sigma(\bar{g},i,\bar{a}',d')\sqrt{2\pi}} \exp[-\tfrac{1}{2}(\tfrac{d-d'}{\sigma(\bar{g},i,\bar{a}',d')})^2] \\ \sigma(\bar{g}, i, \bar{a}', d') &= \alpha_1 + \alpha_2\,\bar{g} + \alpha_3\,i + \alpha_4\,\bar{a}' + \alpha_5\,d' + \\ &\quad \alpha_6\,\bar{g}^2 + \alpha_7\,\bar{g}i + \alpha_8\,\bar{g}\bar{a}' + \alpha_9\,\bar{g}d' + \alpha_{10}\,i^2 + \alpha_{11}\,i\bar{a}' + \\ &\quad \alpha_{12}\,id' + \alpha_{13}\,\bar{a}'^2 + \alpha_{14}\,\bar{a}'d' + \alpha_{15}\,d'^2 + \\ &\quad \alpha_{16}\,\bar{g}^3 + \alpha_{17}\,\bar{g}^2 i + \alpha_{18}\,\bar{g}^2\bar{a}' + \alpha_{19}\,\bar{g}^2 d' + \alpha_{20}\,\bar{g}i^2 + \alpha_{21}\,\bar{g}i\bar{a}' + \\ &\quad \alpha_{22}\,\bar{g}id' + \alpha_{23}\,\bar{g}\bar{a}'^2 + \alpha_{24}\,\bar{g}\bar{a}'d' + \alpha_{25}\,\bar{g}d'^2 + \alpha_{26}\,i^3 + \\ &\quad \alpha_{27}\,i^2\bar{a}' + \alpha_{28}\,i^2 d' + \alpha_{29} i\bar{a}'^2 + \alpha_{30}\,i\bar{a}'d' + \alpha_{31}\,id'^2 + \\ &\quad \alpha_{32}\,\bar{a}'^3 + \alpha_{33}\,\bar{a}'^2 d' + \alpha_{34}\,\bar{a}'d'^2 + \alpha_{35}\,d'^3 \end{aligned} \tag{5}$$

In relation to Eq. (5), 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

Table 3: The best parameters for restraining $C^\alpha - C^\alpha$ distances (d) and main chain N – O distances (h). Full expressions for the standard deviations of the Gaussian p models for W [Eqs. (5) and (8)] are given. If $\bar{g} > 20$, $\bar{g}$ is reset to 20. 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 RMS deviations between the p models and W' are 0.0524 and 0.0441 for d and h, respectively.

$$\begin{aligned}\sigma(\bar{g},i,\bar{a}',d') &= 0.849 - 2.033\,\bar{g} - 1.227\,i + 0.971\,\bar{a}' + 1.467\,d' + \\ &\quad 1.382\,\bar{g}^2 + 1.539\,\bar{g}i - 0.504\,\bar{g}\bar{a}' - 0.259\,\bar{g}d' + 2.412\,i^2 - 1.496\,i\bar{a}' - \\ &\quad 3.094\,id' - 0.425\,\bar{a}'^2 + 0.670\,\bar{a}'d' - 0.159\,d'^2 - \\ &\quad 0.307\,\bar{g}^3 - 0.213\,\bar{g}^2 i + 0.088\,\bar{g}^2\bar{a}' + 0.020\,\bar{g}^2 d' - 0.969\,\bar{g}i^2 + 0.453\,\bar{g}i\bar{a}' + \\ &\quad 0.177\,\bar{g}id' - 0.058\,\bar{g}\bar{a}^2 - 0.042\,\bar{g}\bar{a}'d' + 0.020\,\bar{g}d'^2 - 0.847\,i^3 + \\ &\quad 0.055\,i^2\bar{a} + 1.546\,i^2 d' + 0.527 i\bar{a}'^2 - 0.220\,i\bar{a}'d' + 0.254\,id'^2 + \\ &\quad 0.066\,\bar{a}'^3 + 0.153\,\bar{a}^2 d' - 0.153\,\bar{a}'d'^2 - 0.0019\,d'^3\end{aligned} \tag{6}$$

$$\begin{aligned}\sigma(\bar{g},i,\bar{a}',h') &= 0.957 - 2.044\,\bar{g} - 1.078\,i + 0.995\,\bar{a}' + 1.477\,h' + \\ &\quad 1.572\,\bar{g}^2 + 1.148\,\bar{g}i - 0.525\,\bar{g}\bar{a}' - 0.483\,\bar{g}h' + 1.505\,i^2 - 0.655\,i\bar{a}' - \\ &\quad 2.849\,ih' - 0.625\,\bar{a}'^2 + 0.499\,\bar{a}'h' - 0.126\,h'^2 - \\ &\quad 0.369\,\bar{g}^3 - 0.243\,\bar{g}^2 i + 0.121\,\bar{g}^2\bar{a}' + 0.067\,\bar{g}^2 h' - 0.592\,\bar{g}i^2 + 0.346\,\bar{g}i\bar{a}' + \\ &\quad 0.276\,\bar{g}ih' - 0.032\,\bar{g}\bar{a}^2 - 0.061\,\bar{g}\bar{a}'h' + 0.036\,\bar{g}h'^2 - 0.329\,i^3 - \\ &\quad 0.318\,i^2\bar{a} + 1.472\,i^2 h' + 0.284 i\bar{a}'^2 - 0.293\,i\bar{a}'h' + 0.198\,ih'^2 + \\ &\quad 0.382\,\bar{a}'^3 + 0.110\,\bar{a}^2 h' - 0.079\,\bar{a}'h'^2 - 0.0095\,h'^3\end{aligned} \tag{7}$$

achieved by least-squares fitting the model p^d in Eq. (5) to the histograms W obtained from the database scan (Table 3). The Gaussian conditional pdf's $p^d(d/\bar{g},i,a,d')$, calculated from the least-squares parameters, are superposed on the experimental histograms in Figures 2a–b. These plots provide additional graphical evidence that the Gaussian model can describe the association between the unknown $C^\alpha - C^\alpha$ distance and the four independent variables included in this analysis.

2.3 Restraining a distance between main chain N and O atoms

The N – O distance in the target protein was modeled in the same way as the $C^\alpha - C^\alpha$ distance above (Figures 2c-d):

$$p^h(h/\bar{g},i,\bar{a}',h') = \frac{1}{\sigma(\bar{g},i,\bar{a}',h')\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{h-h'}{\sigma(\bar{g},i,\bar{a}',h')})^2]. \tag{8}$$

2.4 Restraining residue main chain conformation

The residue main chain conformation is defined by dividing the Ramachandran plot spanned by the Φ and Ψ main chain dihedral angles into six areas (Figure 3): A, B, P, L, G, and E

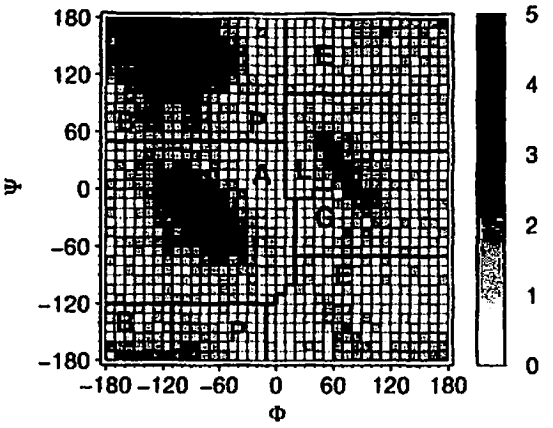

Figure 3: Definition of the main chain conformation classes. The plot shows $W'(\Phi,\Psi)$ determined from the proteins in the alignments database. It is divided into 10° × 10° squares. The areas corresponding to the six characteristic peaks are delimited by thick lines. These areas define the residue main chain conformation classes A, B, P, L, G, and E (24). The scale on the right corresponds to $\ln[W'(\Phi,\Psi)+1]$.

(24). Within each area, the distribution of the two dihedral angles is approximately Gaussian (2). Suppose we can predict the probability ω_i that the restrained residue is in the main chain conformation class i. Then the two pdf's restraining Φ and Ψ dihedral angles can be modeled as a weighted sum of six Gaussian functions, each function corresponding to one of the main chain conformation classes A–E and weighted by a probability that a residue is in the corresponding class:

$$\begin{aligned}p^m(\Phi) &= \textstyle\sum_{i=A,\ldots,E} \omega_i\, N[\bar{\Phi}_i,\sigma_i(\Phi)] \\ p^m(\Psi) &= \textstyle\sum_{i=A,\ldots,E} \omega_i\, N[\bar{\Psi}_i,\sigma_i(\Psi)]\end{aligned} \tag{9}$$

where $N(\alpha,\sigma)$ stands for a Gaussian pdf with mean α and standard deviation σ (2). The remaining problem is to determine the probabilities ω_i of all six main chain conformation classes for each restrained residue. The database of alignments and the program MDT were used to obtain these weights.

The protein features that could correlate with the main chain conformation class of a restrained residue were selected from the list in Table 2. These are the types of the restrained (r) and equivalent (r') 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'). Smoothed and non-smoothed pdf's of the form $p(M/a,b,\ldots,c)$ were derived from the alignments database for the 7249 possible combinations of up to five selected features $(a,b,\ldots,c)$ listed above. Each of the resulting pdf's 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 of seven serine proteases and by comparing these predictions with the

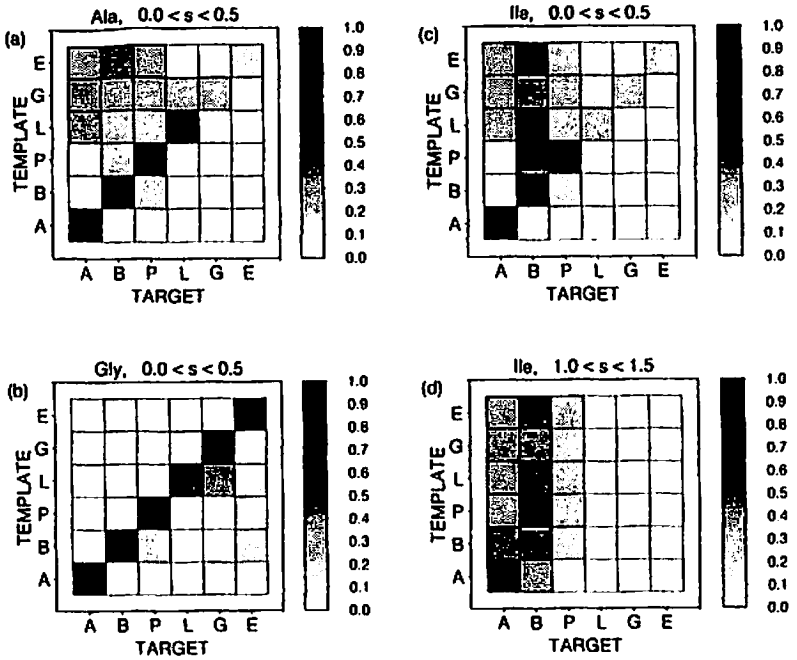

Figure 4: Sample cross-sections through the pdf for prediction of main chain conformation. The probabilities $W(M/M', r, s)$ for conformation classes A–E of a given residue type (horizontal row, M) are shown for each conformation class of an equivalent residue (vertical row, M'). The type of a restrained residue (r), and the residue neighborhood difference (s) are shown above each plot.

actual conformations found in the crystallographic structures.

A cross-section through the best pdf $p(M/r, M', s)$ is shown in Figure 4. Here, s is residue neighborhood difference that measures how different are the types of residues in the two spheres centered on two equivalent central residues; s depends only on the amino acid sequences and can be obtained from the alignment between the target and the template (2). The prediction success of this pdf on the test set of seven serine proteases is listed for the individual residue types in Table 4. The residues that are predicted most accurately (approximately 85%) are Trp, Gln, Pro, Phe, and Cys, whereas the least accurately predicted residues include Gly, Asn, Glu, and Leu (approximately 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

Table 4: Success for the prediction of the main chain conformation class. Total number is the number of residue pairs in the test set that contain 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.

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

the (L, G) classes. When these two pairs are treated only as two classes, the prediction success increases to 87.4%.

2.5 Restraining residue side chain conformation

Side chain restraints are formulated in a similar way to the main chain conformation restraints. Most of the side chain dihedral angles are clustered in up to 3 characteristic intervals that span the range from $-180°$ to $180°$; this results in a small number of side chain rotamers (25,26). Thus, each dihedral angle can be described by a corresponding dihedral angle class within which the distribution of the dihedral angle is Gaussian (2). Similarly to the prediction of the main chain conformation class, the side chain dihedral angles χ_i are

modeled as a weighted sum of Gaussians

$$p^s(\chi_i) = \sum_j \omega_{ij}\ N[\bar{\chi}_{ij}, \sigma_j(\chi_i)] \qquad (10)$$

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 (2). The remaining problem is to determine the probabilities ω_{ij} of all side chain conformation classes for each restrained residue; the same approach is followed as for the derivation of the weights for the main chain conformation classes.

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. 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 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 $p(c_1/r)$ and $p(c_1/r, c_1', r', s)$ are listed for the individual residue types in Table 5. 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. The restraints for χ_2, χ_3, and χ_4 dihedral angles were derived in a similar way. The prediction successes are summarized in Table 5.

3 Satisfaction of spatial restraints

It was shown in the previous Section how spatial restraints on the sequence to be modeled can be expressed as pdf's. These pdf's 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 3D model is obtained by an optimization of the molecular pdf which depends on the model and on the restraints.

3.1 The molecular probability density function

The molecular pdf is assembled from feature pdf's which in turn are obtained from basis pdf's.

3.1.1 Derivation of a feature pdf from basis pdf's

In general, every structural feature f can be restrained by several basis pdf's $p_k^f(f)$ for $k = 1, 2, \ldots$, such as those described in the preceding Section. A feature pdf, $p^F(f)$, is a pdf that combines all basis pdf's to use all the information about the possible values that the feature f can assume. The lowercase and uppercase superscripts are used for the basis and feature pdf's, respectively. The following example clarifies these definitions. The aim

Table 5: Success for the prediction of the side chain χ_i classes, c_i. Total number is the number of residue pairs in the test set that contain the residue being predicted; the numbers of predicted residues are listed in parentheses. The smoothed pdf's $p(c_1/r, c_1', r', s)$, $p(c_2/r, r', c_1', c_2')$, $p(c_3/r, c_3', r', t')$, 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 pdf's $p(c_i/r)$ 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.

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

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 pdf's for the C^α-C^α distance in the target sequence [Eq. (5)]. 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 [Eq. (4)]. In order to combine all this information we have to combine the three basis pdf's into a single feature pdf. To find how to do that, we can use all possible alignments of three proteins in the alignments 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 5a. The histogram suggests that $p(d/d', d'')$ can be modeled as a weighted sum of the individual pdf's $p(d/d')$ and $p(d/d'')$:

$$p(d/d', d'', \bar{s}', \bar{s}'') = \omega(\bar{s}') \cdot p(d/d') + \omega(\bar{s}'') \cdot p(d/d''). \qquad (11)$$

The weight ω of each term in this sum is proportional to the average residue neighborhood difference s between the corresponding structure and the sequence of the unknown. The

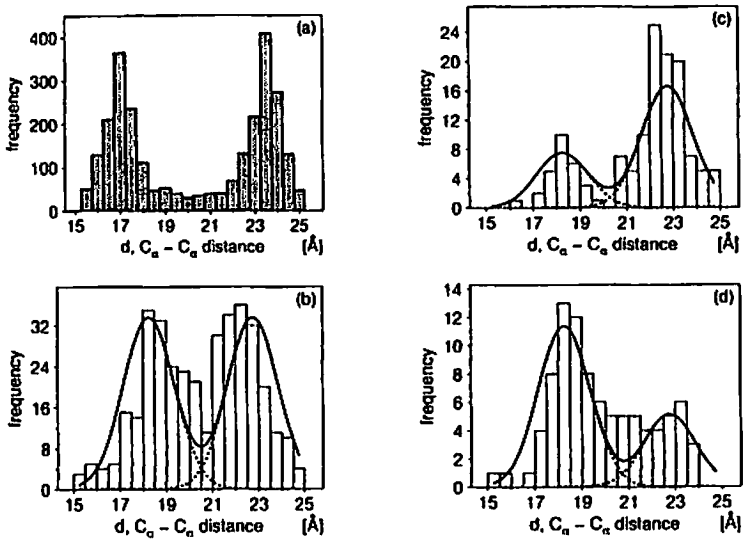

Figure 5: Derivation of a feature pdf from basis pdf's. In all plots, $18.0 < d < 18.5$ and $22.5 < d'' < 23.0$. (a) $W'(d/d', d'')$. (b) $W'(d/d', d'', \bar{s}', \bar{s}'')$, where $0.2 < \bar{s}' < 0.4$ and $0.2 < \bar{s}'' < 0.4$. (c) $W'(d/d', d'', \bar{s}', \bar{s}'')$, where $0.2 < \bar{s}' < 0.4$ and $0.4 < \bar{s}'' < 0.6$. (d) $W'(d/d', d'', \bar{s}', \bar{s}'')$, where $0.4 < \bar{s}' < 0.6$ and $0.2 < \bar{s}'' < 0.4$. The histograms are obtained by scanning the alignments database. The dashed lines are the basis pdf's $p^d(d/d', d'', \bar{s}', \bar{s}'')$ calculated from Eq. (5). The continuous lines are the feature pdf's $p^D(d/d', d'', \bar{s}', \bar{s}')$ calculated with Eqs. (11)–(12).

data can be fitted by the following model for $w(s)$:

$$\omega(s) = \frac{w(s)}{\sum_j w(s_j)} \quad \text{where} \quad w(s) = a + \exp(bs^c)\,; \qquad \sum_j \omega(s_j) = 1\,, \tag{12}$$

where the best values for the parameters, as obtained by the LSQ program, are: $a = 0.0331 \pm 0.0025$, $b = -4.98 \pm 0.11$, and $s = 1.800 \pm 0.079$. The result is that the contribution of a structure to the 3D model of a related structure falls faster than linearly with the average residue neighborhood difference between the two sequences. Examples of histograms and analytical curves for the feature pdf corresponding to different weights are shown in Figures 5b–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 pdf's was used for any number of basis pdf's of the same type that were derived from related structures. When properties such as main chain and side chain conformation are predicted, average residue neighborhood difference is replaced by the residue neighborhood difference.

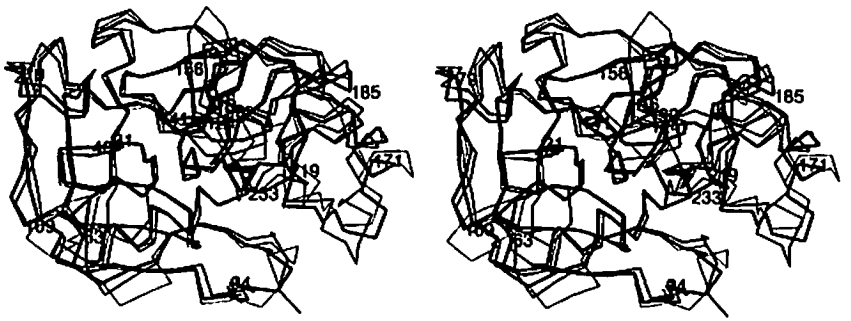

Figure 6: Comparison of trypsin, elastase, and tonin. The stereo plot shows the superposition of the C^α backbones of elastase (medium line) and tonin (thin line) on that of trypsin (thick line). The pairs of the C^α atoms that are aligned in the COMPARER alignment were used for the superpositions. Chymotrypsinogen numbering is used.

Definitions of all types of feature pdf's follow, with the basis pdf's on the right side of the equations as defined in Sections 2.1 – 2.5. 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 2.2 – 2.5. The weights ω_i are determined from Eq. (12).

1. C^α – C^α distance restraints:

$$p^D(d) = p^v(d) \sum_i \omega_i\, p_i^d(d/a, b, \ldots) \tag{13}$$

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:

$$p^H(h) = p^v(h) \sum_i \omega_i\, p_i^h(h/a, b, \ldots) \tag{14}$$

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:

$$p^E(e) = p^e(e)\,. \tag{15}$$

Feature e can be bond length, bond angle, torsion angle, improper dihedral angle, or van der Waals contact (Section 2.1). 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 pdf's for the bond lengths, bond angles, C^α–C^α distances and main chain N–O distances.

4. Main chain conformation restraints:

$$p^M(\theta) = \begin{cases} \sum_{i=1}^{n} \omega_i\, p_i^m(\theta/a, b, \ldots) & n > 0 \\ p^m(\theta/R) & n = 0 \end{cases} \tag{16}$$

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:

$$p^S(c) = \begin{cases} \sum_{i=1}^{n} \omega_i\, p_i^s(c/a, b, \ldots) & n > 0 \\ p^s(c/R) & n = 0 \end{cases} \tag{17}$$

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 ($n = 0$).

3.1.2 Derivation of a molecular pdf from feature pdf's

The last stage in the derivation of a molecular pdf is to combine all feature pdf's 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 pdf's are independent of each other, the molecular pdf is simply a product of feature pdf's defined in Eqs. (13)–(17):

$$P = \prod_i p^F(f_i). \tag{18}$$

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.

3.2 Optimization of the molecular pdf

Derivation of restraints from an alignment and satisfaction of those restraints are implemented in the computer program MODELLER. The protein model may consist of all atoms or any subset such as all heavy atoms, mainchain atoms, or only C^α atoms. The function that is actually optimized is a transformation of the molecular pdf P:

$$F = -\ln(P) \tag{19}$$

where all the features are expressed in terms of atomic Cartesian coordinates. Function F is referred to as the objective function. The same Cartesian coordinates that maximize P also minimize F. 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 (27). The main difference between the original method and the 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 modeled. 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 (18) 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 Modeling of trypsin

To illustrate the method of comparative modeling by satisfaction of spatial restraints, this section describes the modeling of trypsin from two other serine proteases, 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 modeling of ferredoxin (10) and of mouse mast cell chymases (28).

The 3D structures of trypsin [223 residues; (29)], elastase [240 residues; (30)], and tonin [227 residues; (31)] were compared using the program COMPARER (7) (Fig. 6). 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 3.5Å apart from each other are considered, 217 pairs superpose with the RMS of 1.07Å in the more similar pair of trypsin and elastase, whereas only 209 pairs superpose with the higher RMS of 1.18Å 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%. There are only a few short gaps of up to 6 residues in the alignment. 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 6.

39 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 $-180°$ and $180°$. The progress of modeling was followed by monitoring the average atomic shifts and the value

Table 6: Spatial restraints used to model trypsin. [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 pdf's by more than the cutoff in the parentheses is given. These cutoffs generally lie between one and two standard deviations of the corresponding basis pdf's. The best model is defined as the one with the lowest value of the molecular pdf. [d]RMS deviation between the actual values in the best model and the closest optimum. [e]RMS deviation between the actual values in the best model and the most likely optimum. [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 Pro 198 which is aligned with Gly in trypsin. Therefore, no *cis*-peptide bonds were imposed on trypsin.

Type	Basis pdf's[a]	Feature pdf's[b]	Violations[c]	RMS[d]	RMS[e]
bond lengths	1659	1659	0 (0.1Å)	0.005Å	0.005Å
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 (0.2Å)	0.02Å	0.02Å
$C^{\alpha} - C^{\alpha}$ distances	23538	11914	26 (1.5Å)	0.22Å	0.47Å
main chain N – O distances	7480	3832	19 (1.5Å)	0.31Å	0.51Å
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°
disulfide bridge bonds	6	6	0 (0.1°)	0.007Å	0.007Å
disulfide bridge angles	12	12	0 (10°)	3.7°	3.7°
disulfide bridge dihedral angles	6	12	0 (20°)	10.0°	12.9°
cis-peptides[h]	0	0	—	—	—

of the objective function. The optimization schedule and a typical progress of optimization are shown in Figure 7. A total of 11 models with low values of the objective function were obtained (10293 ± 655). 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 (> 15000) 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 10 models are small (Table 6). The stereochemistry of the models is comparable or better than that of the crystallographic trypsin structure refined at a high resolution.

The accuracy of the model is different for buried and exposed parts; thus, we will evaluate the model separately for the residues that have fractional side chain solvent accessibility

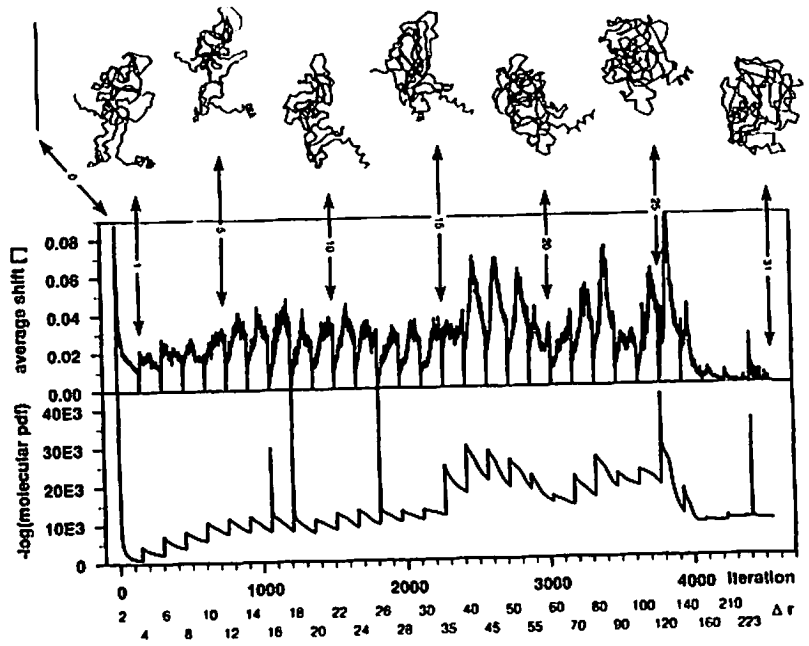

Figure 7: Schedule and progress of optimization. The optimization schedule is specified in the bottom three lines. The 'iteration' line counts the conjugate gradient steps. The bottom two lines show the changes in Δr: Δr is increased every 150 conjugate gradient steps or when the largest atomic shift 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_1 dihedral angles. The van der Waals criterion was gradually introduced in the last five 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 30 minutes on a DEC Alphastation workstation.

less than 20% (buried residues) and for the remaining residues (exposed residues). Only 4 of the 107 buried C^{α} atoms are more than 3.5Å away from their correct positions whereas 6 out of 116 exposed C^{α} atoms are further than 3.5Å 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 RMS error for buried main chain atoms is approximately 0.75Å, and for exposed main chain atoms, approximately 1.3Å.

Similarly to the main chain, buried side chains were modeled more accurately than exposed side chains. 82% 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 modeled 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.

The modeling 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 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 (32) could be used for this task, but general applicability of this approach is questionable (33). Another possibility may be an exhaustive conformational search employing energy criteria (34,35).

5 Discussion

The challenge for the future 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 (9). The methods that combine molecular dynamics and energy potentials with NMR derived constraints (36, 37) and X-ray data (37,38) 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 (39).

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 modeling as a process of reducing the volume of the 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 restrained. The present modeling method uses pdf's in a relatively general way. Thus, the method, even though it has so far been applied only to comparative modeling, could possibly be extended to include other types of information, such as NMR-derived constraints and coarse-grained potentials of mean-force describing residue–residue interactions.

6 Conclusions

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

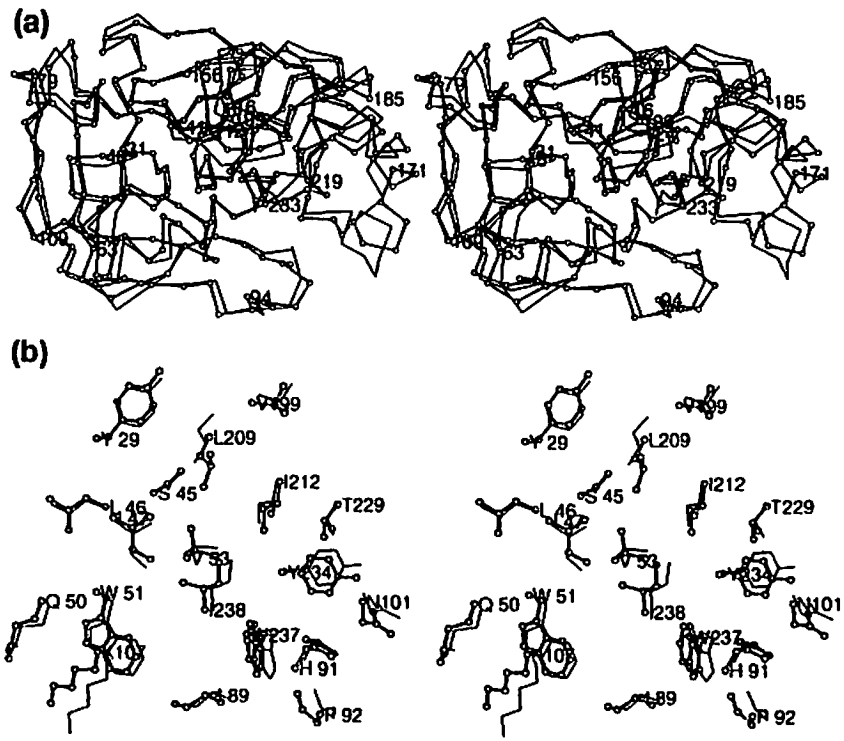

Figure 8: Comparison of the best trypsin model with trypsin. Comparison is obtained by superposing all C^{α} atoms. Chymotrypsinogen numbering is used. Trypsin (open bonds with circles), trypsin model (line). (a) Comparison of the C^{α} traces. (b) Comparison of side chains in a mostly buried region.

2. It was shown how pdf's and other tools can be used to explore quantitatively various relationships between features in individual proteins and in families of proteins.

3. The usefulness of the pdf's was improved by a new smoothing procedure that minimized the problems of a sparse data set.

4. Using these tools and the alignments database, the best pdf's for comparative modeling 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 modeling 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 pdf's 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 modeled at a medium resolution than there are entries in the Brookhaven Protein Databank.

Acknowledgments

We are indebted to Martin Karplus and Alex MacKerell for providing CHARMM 22 parameters. We thank John Overington and Mark Johnson for discussions about protein modeling. We are also grateful to Daša Šali for critical reading of the manuscript. A.Š. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research.

References

1. Blundell, T. L., B. L. Sibanda, M. J. E. Sternberg & J. M. Thornton, "Knowledge-based prediction of protein structures and the design of novel molecules," *Nature* **326**, 347–352 (1987).
2. Šali, A. & T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *J. Mol. Biol.* **234**, 779–815 (1993).
3. Lesk, A. M. & C. H. Chothia, "The response of protein structures to amino-acid sequence changes," *Phil. Trans. Roy. Soc.* **317**, 345–356 (1986).
4. Hubbard, T. J. P. & T. L. Blundell, "Comparison of solvent inaccessible cores of homologous proteins: Definitions useful for protein modelling," *Prot. Eng.* **1**, 159–171 (1987).
5. Srinivasan, N. & T. L. Blundell, "An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure," *Prot. Eng.* **6**, 501–512 (1993).
6. Chothia, C., "One thousand families for the molecular biologist," *Nature* **360**, 543–544 (1992).
7. Šali, A. & T. L. Blundell, "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 (1990).
8. Zhu, Z. -Y., A. Šali & T. L. Blundell, "A variable gap penalty function and feature weights for protein 3-D structure comparisons," *Prot. Eng.* **5**, 43–51 (1992).
9. Šali, A., J. P. Overington, M. S. Johnson & T. L. Blundell, "From comparisons of protein sequences and structures to protein modelling and design," *TIBS* **15**, 235–240 (1990).

10. Šali, A., "Modelling three-dimensional structure of proteins from their sequence of amino acid residues," in *PhD Thesis*, University of London, London, 1991.
11. Overington, J., M. S. Johnson, A. Šali & T. L. Blundell, "Tertiary structural constraints on protein evolutionary diversity; Templates, key residues and structure prediction," *Proc. Roy. Soc. Lond.* **B 241**, 132–145 (1990).
12. Overington, J., D. Donnelly, M. S. Johnson, A. Šali & T. L. Blundell, "Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds," *Protein Sci.* **1**, 216–226 (1992).
13. Topham, C. M., A. McLeod, F. Eisenmenger, J. P. Overington, M. S. Johnson & T. L. Blundell, "Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables," *J. Mol. Biol.* **229**, 194–220 (1993).
14. Johnson, M. S., J. P. Overington & T. L. Blundell, "Alignment and searching for common protein folds using a data bank of structural templates," *J. Mol. Biol.* **231**, 735–752 (1993).
15. Overington, J. P., Z. -Y. Zhu, A. Šali, M. S. Johnson, R. Sowdhamini, G. V. Louie & T. L. Blundell, "Molecular recognition in protein families: A database of aligned three-dimensional structures of related proteins," *Biochem. Soc. Trans.* **21**, 597–604 (1993).
16. Abola, E. E., F. C. Bernstein, S. H. Bryant, T. F. Koetzle & J. Weng, "Protein Data Bank," in *Crystallographic databases — Information, content, software systems, scientific applications*, F.H. Allen, G. Bergerhoff & R. Sievers, eds., Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987, 107–132.
17. Sutcliffe, M. J., I. Haneef, D. Carney & T. L. Blundell, "Knowledge based modelling of homologous proteins, Part I: Three dimensional frameworks derived from the simultaneous superposition of multiple structures," *Prot. Eng.* **1**, 377–384 (1987).
18. Press, W. H., B. P. Flannery, S. A. Teukolsky & W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge, 1986.
19. Sippl, M. J., "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 (1990).
20. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan & M. Karplus, "CHARMM: A program for macromolecular energy minimization and dynamics calculations," *J. Comp. Chem.* **4**, 187–217 (1983).
21. MacKerell Jr., A. D., 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, *in preparation*.
22. Hill, T. L., *An introduction to statistical thermodynamics*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1960.
23. Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola & J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.* **81**, 3684–3690 (1984).
24. Wilmot, C. M. & J. M. Thornton, "β-turns and their distortions: a proposed new nomenclature," *Prot. Eng.* **3**, 479–493 (1990).
25. Janin, J., S. Wodak, M. Levitt & B. Maigret, "Conformation of amino acid side-chains in proteins," *J. Mol. Biol.* **125**, 357–386 (1978).

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

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

28. Šali, A., R. Matsumoto, H. P. McNeil, M. Karplus & R. L. Stevens, "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 (1993).

29. Walter, J., W. Steigemann, T. P. Singh, H. Bartunik, W. Bode & R. Huber, "On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography," *Acta Crystallogr.* **B38**, 1462–1472 (1982).

30. Meyer, E., G. Cole, R. Radakrishnan & O. Epp, "Structure of native porcine pancreatic elastase at 1.65 Å resolution," *Acta Crystallogr.* **B44**, 26–38 (1988).

31. Fujinaga, M. & M. N. G. James, "Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1.8 Å resolution," *J. Mol. Biol.* **195**, 373–396 (1987).

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

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

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

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

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

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

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

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