Journal Title: J Mol Biol
ISSN: 0022-2836
Article Author(s): Holm L
Article Title: Evaluation of protein models by atomic solvation preference.
Year: 1992 May 5
Volume: 225
Issue: 1
Pages: 93-105
PMID: 1583696
User's Comments: In color, if available

J. Mol. Biol. (1992) 225, 93–105

# Evaluation of Protein Models by Atomic Solvation Preference

**Liisa Holm and Chris Sander**

*European Molecular Biology Laboratory*
*Meyerhofstrasse 1, D-6900 Heidelberg, Germany*

*(Received 29 July 1991; accepted 12 November 1991)*

Important properties of globular proteins, such as the stability of the folded state, depend sensitively on interactions with solvent molecules. An excluded volume approximation to protein–solvent interaction, the solvent contact model, was used to derive atomic solvation preference parameters from a database of known protein structures. The ability of solvation preference to discriminate between correct and incorrect three-dimensional structures for a given sequence, or to identify the correct sequence placement in a given structure, was tested. Backbone co-ordinates were taken from experimentally known structures or hypothetical models and side-chain conformations (in rotamer space) were optimized by an efficient Monte Carlo algorithm using simulated annealing and simple potential functions. Discrimination by solvation preference was very clear between deliberately misfolded and correct globular models as well as between native-like and non-native-like topologies of combinatorially generated myoglobin models. Due to its statistical nature, the evaluation works best on entire protein models, while the identification of incorrect parts of models is more difficult. In one case locally incorrect chain tracing in a crystal structure was identified. The method is computationally fast compared to methods based on surface area calculations and is recommended for use as a diagnostic tool in model building based on sequence similarity, in folding simulations and in protein design.

*Keywords:* excluded volume approximation; hydrophobic effect; model building; Monte Carlo optimization; solvent contact model

## 1. Introduction

With the tremendous flow of new protein sequences compared to the slower pace at which new structures are determined, the prediction of three-dimensional (3-D†) structures of proteins has become a necessity in biochemical studies. Developments in methods of sequence analysis allow the identification of increasingly distant relations for model building. Advances in genetic engineering have made the design of completely new proteins an experimental reality. Unfortunately, as the protein folding problem is as yet essentially unsolved, both hypothetical models and theoretical designs have a less than optimal chance of agreeing with physical reality. The key aspect is the development of criteria with sufficient discriminatory power to tell a good model from a bad one. An example is provided by deliberately misfolded proteins in which the sequence of a protein known to have an all-helical 3-D structure is placed into a known structure of a completely different type, an anti-parallel $\beta$-barrel, and *vice versa*. For the evaluation of the quality of these clearly incorrect hypothetical structures, intramolecular energy, calculated *in vacuo* using standard empirical potentials, is not a sensitive criterion (Novotny *et al.*, 1984, 1988). The free energy difference between the folded and unfolded states would be an ideal criterion, but present theories are not capable of calculating free energy differences to sufficient accuracy.

Faced with the lack of an accurate theory of protein folding, empirical observations of regularities gleaned from the database of solved structures can be very useful. Thermodynamic arguments suggest that the hydrophobic effect is a major driving force of protein folding (Kauzmann, 1959; Sharp, 1991). A variety of statistical criteria that measure the preferential distribution of hydrophobic side-chains in the interior of proteins have been used to discriminate successfully between deliberately misfolded and native structures (Baumann *et al.*, 1989; Bryant & Amzel, 1987; Novotny *et al.*, 1988; Hendlich *et al.*, 1990). Solvent-modified potential energy functions have been generated, for example, by omitting the attractive part of the Lennard–Jones potential of exposed non-polar

† Abbreviations used: 3-D, three-dimensional; r.m.s., root-mean-square; PDB, Protein Data Bank.



0022-2836/92/090093–13 $03.00/0

carbon atoms (Novotny *et al.*, 1988), which effectively gives them a preference for the interior, or by introducing an effective residue pair potential calibrated on hydrophobicity that is attractive for pairs of hydrophobic residues (Levitt, 1976). Empirical solvation free energy functions have been derived using atomic solvent accessible surface areas (Eisenberg & McLauchlan, 1986; Ooi *et al.*, 1987), or the volume of the hydration shell (Kang *et al.*, 1988), and various hydrophobicity scales or observed frequencies of "buried" residues (Janin, 1979). For example, the atomic transfer free energy parameters of Eisenberg & McLachlan (1986) for five atom types, based on accessible surface areas and calibrated on $\Delta G_{\text{octanol/water}}$ (Fauchere & Pliska, 1983), were capable of discriminating between correct and deliberately misfolded conformations (Novotny *et al.*, 1988; Eisenberg & McLauchlan, 1986; Chiche *et al.*, 1990). Recently, Vila *et al.* (1991) tested surface area based models with several empirical free energy scales derived from model compounds, the best of which showed the desired discrimination among the native and a set of near-native conformations of pancreatic trypsin inhibitor.

In this study, we used the database of known protein structures to derive a novel set of atomic solvation preference parameters (for 87 atom types) by characterizing the environment of atoms according to the solvent contact model (Colonna-Cesari & Sander, 1990), and demonstrate the ability of solvation preference to identify the correct fold among models that have been misfolded in various ways.

## 2. Methods

### (a) *The solvent contact model*

The protein folding process can be viewed as competition between protein–protein and protein–solvent contacts. As the conformation of a protein changes, the contacts that protein atoms make with other protein atoms are replaced by contacts with solvent molecules and *vice versa*. The principal difficulty in estimating protein–water interactions lies in the uncertainty of the positions of water molecules. In the solvent contact model (Colonna-Cesari & Sander, 1990), the time-average of the strength of these interactions is assumed to depend only on the average number of water molecules in the 1st hydration shell around protein atoms. This number is quantified as the volume occupied by water in the neighbourhood of an atom, which is taken to be the complement of the volume occupied by protein atoms in the neighbourhood. The basic idea is that an atom makes a constant number of nearest-neighbour contacts that is the sum of contacts with other protein atoms and those with solvent molecules. Empty space between atoms too small to be occupied by solvent molecules is assumed to provide, on average, a constant background per atom that does not depend on protein conformation. A conceptually similar approach, different in detail, is that of Kang *et al.* (1988), in which the volume of the hydration shell around a molecule is computed by exact geometrical methods from the volumes of overlapping hydration spheres of the atoms (Gibson & Scheraga, 1988).

In detail, we calculate the occupancy (*OccAtm*) a a protein atom $i$ as the sum over all volumes $V$ of p atoms $j$ in a shell of 6 Å (1 Å = 0·1 nm) radius we with an envelope function that depends on the distance $r_{ij}$ from the atom:

$$OccAtm(i) = \sum_j V_j \times env(r_{ij}).$$

Here, the envelope function, *env*, is a simple l square well (equal to 1·0 from $r = 0{\cdot}0$ to $r = 3{\cdot}2$ Å decreasing linearly down to 0, which is reache $r = 6{\cdot}0$ Å). Defined in this way occupancy has the d sion of volume and can be thought of as a refined d tion of the volume of the 1st hydration shell. solvation factor (*SolFac*), a dimensionless quanti defined to reflect the 2 extremes "fully occupied" ( packed protein interior, $SolFac = 0{\cdot}0$) and solvated" (only covalent neighbours, protein sur $SolFac = 1{\cdot}0$) and is calculated from the atomic pancy as:

$$SolFac(i) = \frac{MaxOcc(t_i) - OccAtm(i)}{MaxOcc(t_i) - MinOcc(t_i)},$$

where *MinOcc* and *MaxOcc* are the minimal and maxi occupancies of atom type $t_i$ of atom $i$. Intuiti speaking, the solvation factor simply represents the so tion state of an atom, on a scale from 0 to 1. calculation of the solvation factor for a given atom i context of a protein structure depends only on i atomic distances (eqn (1)) and involves only a s number (10s, not 100s) of floating point operations ( (1) and (2)) per atom and is therefore very fast comp to geometrical calculations of surface areas (Le Richards, 1971) or volumes (Kang *et al.*, 1988). The p cal approximation involved, that empty space on ave is occupied by solvent atoms, is reasonable conside the dynamic nature of protein conformation and considerable fluctuations of water molecule position aqueous solution at room temperature.

To calculate solvation factors, the volumes of e atom type in eqn (1) were defined as the fragmental v der Waals' volume constants estimated by Moto Marshall (1985), such that a sum over fragmental ato volumes approximates well the volume of the molec Minimal occupancies (*MinOcc*) for each atom type w defined as the average atomic occupancy in extend GGXGG (G, glycine; X, any residue) peptides (co-or nates available on request), where the average is ta over 3 main rotamers of the side-chain of the cen residue. Maximal occupancies (*MaxOcc*) were define the average atomic occupancy of atoms in residues wit relative solvent accessibility of less than 4% in a data of known structures. Solvent accessibilities were ca lated using the program DSSP (Kabsch & Sander, 19 with the maximum values taken from Baumann *et* (1989).

### (b) *Derivation of atomic solvation preference parameters*

Frequencies of occurrence of side-chain atom types 11 solvation factor (*SolFac*) bins between 0·0 and 1·0 w collected from a database consisting of 63 representati high-resolution protein structures (list from R. Schnei (personal communication), generated as described by Hobohm *et al.* (in the press)). Proteins similar in sequen to the test set of proteins (Figs 2 to 6) were not present the database. There were on average 60 observations/b Five dummy observations were added to all bins throu

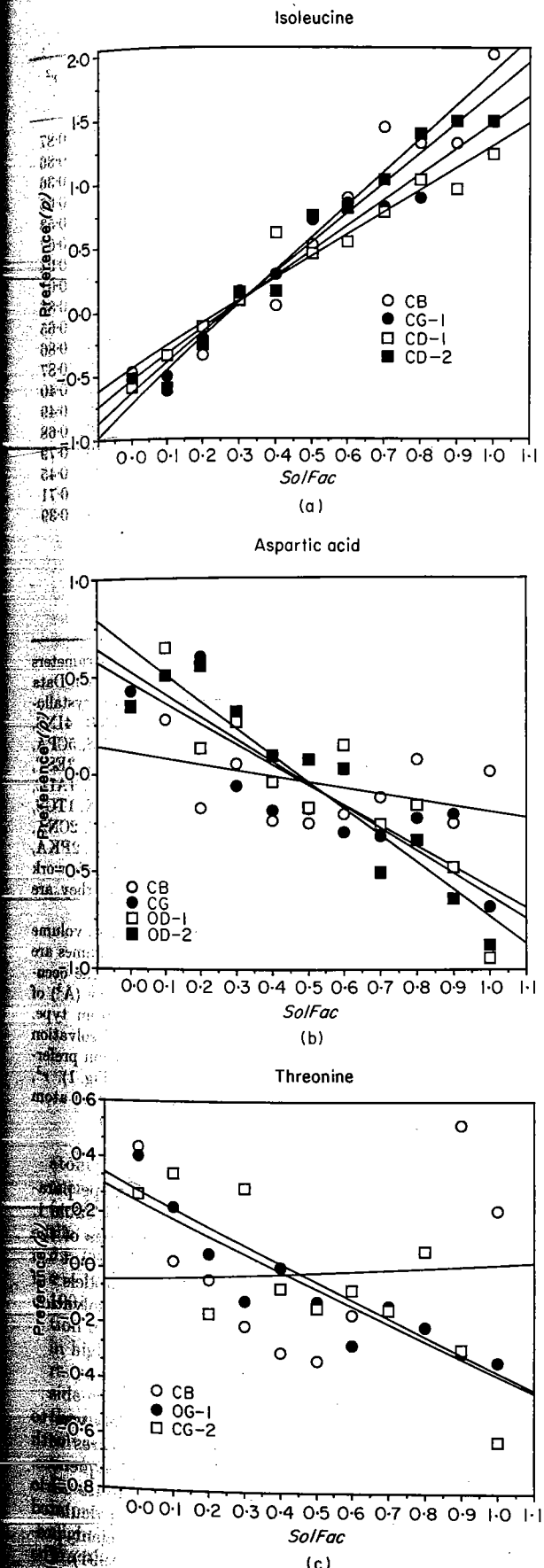



**Figure 1.** Preference parameters for representative side-chain atoms. Preferred states have negative values of *P*. The statistical preference ($p$) of each atom type $t_i$ to occur in a given solvation state is calculated using eqn (3) and plotted as a function of the solvation factor (*SolFac*). Straight lines were fit to these curves to give the intercept ($I$) and slope ($S$) parameters in Table 1. (a) Isoleucine, which clearly prefers interior positions (low solvation factor). (b) Aspartic acid, with an anti-preference for interior positions. (c) Threonine. There is a clear dip for $C^\beta$ around the middle, i.e. neither the maximally nor the minimally solvent exposed states are preferred. Similar dips are observed for some other atom types in polar residues. We do not fully understand the cause of this effect. In the present linear approximation, $C^\beta$ of threonine is neutral, with a slope near zero.

out to help to alleviate the problem of small-number statistics. The preference $p$ of an atom type to occur at a given *SolFac* value is calculated from the "observed : expected" ratio:

$$p_{t,SolFac} = -\ln \frac{N_{t,SolFac} \times N_{\text{total}}}{N_t \times N_{SolFac}}, \tag{3}$$

where $N_{t,SolFac}$ is the number of observations in a *SolFac* bin for atom type $t$, $N_t$ and $N_{SolFac}$ are the marginal sums of $N_{t,SolFac}$, and $N_{\text{total}}$ is the total number of observations. The preferences are made additive by taking the logarithm and the sign is chosen so that low (e.g. negative) values are favourable, in analogy with free-energy scales.

As the discrete preferences $p$ have an overall linear behaviour as a function of *SolFac* (Fig. 1), smoothed preference parameters *pref* for each atom type were derived by linear regression analysis of $p$ values against *SolFac* and are reported as slope $S$ and intercept $I$ (*pref* at *SolFac* = 0) in Table 1. The strongest deviation from linearity occurs for main-chain atoms and $C^\beta$ of some polar side-chains. We think that this is partly explained by the strong tendency of the main-chain to be involved in secondary structure hydrogen bonding in the interior of globular proteins (Baumann *et al.*, 1989), which makes these atom types underrepresented in the outer *SolFac* bins. For this reason no preference parameters were derived for main-chain atoms in the current implementation. Main-chain atoms are taken into account only in calculating the occupancies of side-chain atoms.

The practical consequence of the linearization procedure here is a reduction in the size of Table 1. A linear relationship between the strength of protein–water interaction and solvation state (surface areas or volumes) has also been used in a related context, that of estimating free energy differences between different molecular conformations (Eisenberg & McLachlan, 1986; Ooi *et al.*, 1987; Kang *et al.*, 1988).

### (c) *Application of solvation preference parameters*

Given a model of a protein, solvation factors of side-chain atoms were calculated from their occupancies in the structure. Hydrogen atoms and cofactors, ligands etc. were ignored. If due to fluctuations the solvation factor of an atom fell outside the interval [0·0, 1·0], it was reset to 0·0 or 1·0. The total solvation preference, *Solp*, of a model is defined as the sum of atomic solvation preferences over all side-chain atoms $i$:

$$\begin{aligned} Solp &= \sum_i pref(t_i, SolFac(i)) \\ &= \sum_i [SolFac(i) \times S(t_i) + I(t_i)], \end{aligned} \tag{4}$$

**Table 1**

*Atomic solvation preference parameters*

| Residue N | Atom | *V* | *MinOcc* | *MaxOcc* | *I* | *S* | $r^2$ |
|---|---|---|---|---|---|---|---|
| Ala 1335 | CB | 16·2 | 99 | 250 | −0·25 | 0·65 | 0·68 |
| Pro 681 | CB | 12·8 | 115 | 238 | 0·19 | −0·44 | 0·27 |
| | CG | 12·8 | 107 | 238 | 0·32 | −0·72 | 0·48 |
| | CD | 12·8 | 132 | 244 | 0·02 | 0·01 | 0·00 |
| Ser 1175 | CB | 12·8 | 107 | 256 | 0·32 | −0·79 | 0·79 |
| | OG | 11·0 | 93 | 260 | 0·40 | −0·93 | 0·82 |
| Cys 424 | CB | 12·8 | 115 | 261 | −0·65 | 2·28 | 0·91 |
| | SG | 19·9 | 99 | 260 | −0·57 | 1·89 | 0·93 |
| Thr 993 | CB | 9·4 | 120 | 243 | −0·04 | 0·05 | 0·00 |
| | OG-1 | 11·0 | 111 | 254 | 0·24 | −0·62 | 0·81 |
| | CG-2 | 16·2 | 107 | 243 | 0·28 | −0·65 | 0·58 |
| Val 1218 | CB | 9·4 | 125 | 232 | −0·68 | 2·29 | 0·95 |
| | CG-1 | 16·2 | 111 | 235 | −0·45 | 1·57 | 0·93 |
| | CG-2 | 16·2 | 111 | 237 | −0·51 | 1·84 | 0·96 |
| Ile 843 | CB | 9·4 | 137 | 228 | −0·73 | 2·58 | 0·94 |
| | CG-1 | 12·8 | 123 | 226 | −0·54 | 2·00 | 0·94 |
| | CG-2 | 16·2 | 123 | 234 | −0·45 | 1·73 | 0·94 |
| | CD-1 | 16·2 | 81 | 224 | −0·64 | 2·33 | 0·97 |
| Leu 1173 | CB | 12·8 | 138 | 234 | −0·57 | 2·04 | 0·94 |
| | CG | 9·4 | 121 | 215 | −0·53 | 1·82 | 0·98 |
| | CD-1 | 16·2 | 103 | 220 | −0·28 | 1·12 | 0·66 |
| | CD-2 | 16·2 | 103 | 216 | −0·23 | 1·07 | 0·58 |
| Asp 852 | CB | 12·8 | 122 | 244 | 0·11 | −0·30 | 0·22 |
| | CG | 9·8 | 105 | 244 | 0·47 | −1·05 | 0·76 |
| | OD-1 | 8·2 | 94 | 244 | 0·53 | −1·15 | 0·78 |
| | OD-2 | 8·2 | 94 | 250 | 0·65 | −1·37 | 0·89 |
| Asn 782 | CB | 12·8 | 127 | 252 | 0·16 | −0·40 | 0·13 |
| | CG | 9·8 | 113 | 254 | 0·59 | −1·34 | 0·76 |
| | OD-1 | 8·2 | 79 | 266 | 0·28 | −0·68 | 0·41 |
| | ND-2 | 13·3 | 128 | 257 | 1·01 | −1·91 | 0·74 |
| His 375 | CB | 12·8 | 140 | 258 | −0·35 | 1·00 | 0·73 |
| | CG | 7·3 | 128 | 239 | −0·26 | 0·65 | 0·53 |
| | ND-1 | 9·3 | 116 | 236 | 0·08 | −0·06 | 0·01 |
| | CD-2 | 10·8 | 114 | 246 | −0·09 | 0·20 | 0·14 |
| | CE-1 | 10·8 | 91 | 230 | 0·07 | −0·10 | 0·03 |
| | NE-2 | 9·3 | 90 | 233 | 0·02 | −0·04 | 0·01 |
| Phe 528 | CB | 12·8 | 148 | 252 | −0·49 | 1·71 | 0·89 |
| | CG | 7·3 | 142 | 227 | −0·49 | 1·77 | 0·84 |
| | CD-1 | 10·8 | 131 | 231 | −0·27 | 1·16 | 0·74 |
| | CD-2 | 10·8 | 132 | 229 | −0·36 | 1·26 | 0·81 |
| | CE-1 | 10·8 | 103 | 225 | −0·27 | 1·06 | 0·78 |
| | CE-2 | 10·8 | 102 | 222 | −0·43 | 1·41 | 0·93 |
| | CZ | 10·8 | 92 | 217 | −0·37 | 1·33 | 0·72 |
| Tyr 540 | CB | 12·8 | 147 | 258 | −0·51 | 1·56 | 0·89 |
| | CG | 7·3 | 146 | 239 | −0·41 | 1·25 | 0·89 |
| | CD-1 | 10·8 | 136 | 242 | −0·25 | 0·82 | 0·67 |
| | CD-2 | 10·8 | 137 | 242 | −0·15 | 0·56 | 0·69 |
| | CE-1 | 10·8 | 110 | 234 | −0·22 | 0·62 | 0·53 |
| | CE-2 | 10·8 | 109 | 234 | −0·20 | 0·54 | 0·75 |
| | CZ | 7·3 | 99 | 229 | −0·21 | 0·51 | 0·37 |
| | OH | 10·9 | 70 | 234 | 0·04 | −0·12 | 0·03 |
| Trp 264 | CB | 12·8 | 158 | 257 | −0·42 | 1·37 | 0·82 |
| | CG | 7·3 | 158 | 248 | −0·38 | 1·19 | 0·76 |
| | CZ-3 | 10·8 | 119 | 235 | −0·23 | 1·03 | 0·69 |
| | CD-1 | 10·8 | 138 | 255 | −0·10 | 0·44 | 0·19 |
| | CD-2 | 6·8 | 153 | 231 | −0·44 | 1·45 | 0·80 |
| | NE-1 | 9·0 | 120 | 246 | −0·15 | 0·51 | 0·34 |
| | CE-2 | 6·8 | 127 | 226 | −0·34 | 1·10 | 0·71 |
| | CE-3 | 10·8 | 151 | 243 | −0·27 | 1·20 | 0·71 |
| | CZ-2 | 10·8 | 103 | 226 | −0·30 | 1·07 | 0·89 |
| | CH-2 | 10·8 | 98 | 227 | −0·31 | 1·14 | 0·77 |
| Met 332 | CB | 12·8 | 136 | 246 | −0·45 | 1·36 | 0·86 |
| | CG | 12·8 | 127 | 235 | −0·26 | 0·83 | 0·50 |
| | SD | 16·4 | 80 | 223 | −0·32 | 1·07 | 0·70 |
| | CE | 16·2 | 62 | 227 | −0·24 | 0·85 | 0·75 |
| Glu 781 | CB | 12·8 | 134 | 247 | 0·11 | −0·27 | 0·05 |
| | CG | 12·8 | 121 | 238 | 0·69 | −1·53 | 0·87 |
| | CD | 9·8 | 79 | 230 | 0·79 | −1·63 | 0·68 |
| Glu 781 | OE-1 | 8·2 | 65 | 228 | 0·82 | −1·73 | 0·8 |
| | OE-2 | 8·2 | 65 | 234 | 0·89 | −1·84 | 0·8 |
| Gln 552 | CB | 12·8 | 139 | 244 | −0·19 | 0·49 | 0·3 |
| | CG | 12·8 | 126 | 235 | 0·32 | −0·77 | 0·5 |
| | CD | 9·8 | 84 | 234 | 0·42 | −0·98 | 0·6 |
| | OE-1 | 8·2 | 65 | 235 | 0·41 | −0·93 | 0·6 |
| | NE-2 | 13·3 | 76 | 244 | 0·64 | −1·45 | 0·7 |
| Lys 927 | CB | 12·8 | 136 | 245 | −0·04 | 0·08 | 0·1 |
| | CG | 12·8 | 130 | 223 | 0·42 | −0·99 | 0·6 |
| | CD | 12·8 | 91 | 209 | 0·64 | −1·36 | 0·7 |
| | CE | 12·8 | 71 | 195 | 1·12 | −2·24 | 0·8 |
| | NZ | 13·3 | 53 | 192 | 1·45 | −2·73 | 0·8 |
| Arg 600 | CB | 12·8 | 133 | 254 | −0·26 | 0·61 | 0·2 |
| | CG | 12·8 | 131 | 241 | 0·16 | −0·42 | 0·4 |
| | CD | 12·8 | 98 | 231 | 0·37 | −0·86 | 0·6 |
| | NH-1 | 9·0 | 57 | 239 | 0·68 | −1·49 | 0·7 |
| | CZ | 7·0 | 62 | 225 | 0·57 | −1·21 | 0·6 |
| | NH-2 | 9·0 | 50 | 226 | 0·71 | −1·49 | 0·7 |
| | NE | 9·0 | 81 | 224 | 0·37 | −0·86 | 0·6 |
| Gly 1331 | CA | 12·8 | | | | | |
| | CA | 9·4 | | | | | |
| Main chain | N | 13·3 | | | | | |
| | C | 7·3 | | | | | |
| | O | 8·2 | | | | | |

**Table 1** *(continued)*

The structural database used to derive the parame consisted of the following structures taken from the Protein D Bank (Bernstein *et al.*, 1977) in the order of nominal crystal graphic resolution: 5PTI, 7RSA, 1UTG, 4PTP, 1NXB, 4I 2OVO, 1CRN, 2SGA, 1CCR, 2PRK, 2SNS, 1LZ1, 3GRS, 5C 1PAZ, 1GCR, 451C, 3TLN, 1PCY, 2CPP, 2WRP, 1PSG, 3E 2CCY, 4CHA, 3C2C, 2ALP, 1L01, 8DFR, 1SGT, 1CTF, 1A 2AZA, 3APR, 1CHO, 3RNT, 1GD1, 1UBQ, 2APP, 1TON, 1T 1HNE, 3CPV, 3PR2, 3TPI, 2CA2, 2GCH, 2RSP, 5TNC, 2C 1HOE, 1GOX, 6LDH, 1P01, 1GP1, 1RNS, 1ACX, 2SOD, 2P 1SNC, 3CLA, 2GBP. Full references to the crystallographic w can be found in the headers of the co-ordinate files; they a omitted here for space reasons.

*N*, number of residues of this type in the database. *V*, volu (Å$^3$) according to Motoc & Marshall (1985). Atomic volumes a also given for the main-chain, since it contributes to the oc pancy of side-chain atoms. *MinOcc*, minimal occupancy (Å$^3$) atom type. *MaxOcc*, maximal occupancy (Å$^3$) of atom ty *I*, solvation preference at *SolFac* = 0·0. *S*, slope of solva preference *versus* solvation factor (difference in solvation pre ence going from *SolFac* = 0·0 to *SolFac* = 1·0, see Fig. 1) correlation coefficient of the linear regression fit. Note that at types with bad fits have *I* and *S* parameters close to 0.

where $t_i$ is the atom type of atom $i$, $S$ is the slope pa meter and $I$ is the intercept parameter given in Table The intercept term cancels out if 2 conformations of t same sequence are compared, but is important comparing models of different proteins. Models different proteins can be compared in terms of solvati preference per residue or per side-chain atom.

### (d) *Optimization of misfolded models*

The strategy for generating misfolded models was use the backbone co-ordinates of known structures wi side-chains of misaligned or non-native sequence Side-chain conformations were optimized by Monte Ca simulated annealing in rotamer space using precalculate rotamer–rotamer interactions in a highly efficient p cedure (program MaxSprout; Holm & Sander, 1991). T energy function normally used in MaxSprout is a sim 6–9-potential with a minimum of −0·26 energy uni

**Table 2**
*Database of misfolded structures*

| Protein | PDB code | No. of residues | Reference to original structure |
|---|---|---|---|
| A. *Misfolded here* | | | |
| Cellulase tail domain | 1 CBH | | Kraulis *et al.* (1989) |
| Avian pancreatic polypeptide | 1 PPT | 36 | Blundell *et al.* (1981) |
| Ferredoxin | 1 FDX | | Adman *et al.* (1976) |
| Rubredoxin | 5 RXN | 54 | Watenpaugh *et al.* (1980) |
| Staphylococcal nuclease | 1 SN3 | | Almassy *et al.* (1983) |
| Chymotrypsin inhibitor | 2 CI 2 | | McPhalen & James (1987) |
| Cro repressor | 2 CRO | 65 | Mondragon *et al.* (1989) |
| High-potential iron protein | 1 HI P | | Carter *et al.* (1974) |
| Cytochrome *b*5 | 2 B5 C | 85 | Mathews *et al.* (1972) |
| Cytochrome *c*3 | 2 CDV | | Higuchi *et al.* (1984) |
| Subtilisin inhibitor | 2 SS1 | 107 | Satow *et al.* (1980) |
| Phospholipase A2 | 1 BP2 | | Dijkstra *et al.* (1981) |
| Pseudozaurin | 2 PAZ | 123 | Adman *et al.* (1989) |
| Phospholipase A2 | 1 P2 P | | Dijskstra *et al.* (1983) |
| Ribonuclease | 1 RN3 | 124 | Borkakoti *et al.* (1982) |
| Leghaemoglobin | 1 LH1 | | Arutunyan *et al.* (1980) |
| Interleukin 1$\beta$ | 2 I 1 B | 153 | Priestle *et al.* (1989) |
| Bence-Jones protein | 1 REI | | Epp *et al.* (1975) |
| Papain | 5 PAD | 214 | Drenth *et al.* (1976) |
| Rhodanese | 1 RHD | | Ploegman *et al.* (1978) |
| Cytochrome peroxidase | 2 CYP | 293 | Finzel *et al.* (1984) |
| Arabinose-binding protein | 1 ABP | | Gilliland & Quiocho (1981) |
| Myoglobin dimer | 1 PMB | 306 | Dodson *et al.* (1988) |
| Thermolysin | 2 TMN | | Tronrud *et al.* (1986) |
| Tyrosine-tRNA synthetase | 2 TS 1 | 317 | Brick *et al.* (1989) |
| B. *From other sources* | | | |
| Hemerythrin | 1 HMQ | | Stenkamp *et al.* (1983) |
| Immunoglobulin domain | 2 MCP | 113 | E. A. Padlan *et al.* (unpublished results) |
| | | | Misfolded by Novotny *et al.* (1984) |
| Incorrect ferredoxin model | 2 FD1 | | Ghosh *et al.* (1982) |
| Same corrected | 4 FD1 | 106 | Stout (1989) |

To generate 28 misfolded models in section A, the sequences of the pairs (1 triplet) of unrelated proteins of equal chain length listed next to each other were swapped and side-chain conformations optimized, as in Novotny *et al.* (1984).

atom pair. Here, solvation preference was added to the energy function as $\zeta \times Solp$, where $\zeta$ is a unit conversion factor, here set to $\zeta = 1{\cdot}0$. In the Monte Carlo procedure, the simulated annealing protocol had a cooling rate of $\delta = 0{\cdot}0001$, taking the inverse temperature from $\alpha = 0$ to $\alpha = 10$ in 100,000 steps (for details, see Holm & Sander, 1991). The simple 6–9 potential alone, without the solvation term, is sufficient to recreate side-chain conformation in high-resolution X-ray structures with reasonable accuracy: average root-mean-square positional deviation of all side-chain atoms, 1·8 Å; $\chi_1$ dihedral angles within 30° of the native structure for 72% of all residues; and, in the solvent inaccessible core, values of 1·4 Å and 81%, respectively (L. Holm & C. Sander, unpublished results). Reassuringly, adding *Solp* to the cost function further improved the average quality of optimized models slightly (data not shown), although not enough to be of practical value so far. In all Monte Carlo optimizations, the conformational search was restricted to fixed backbones and discrete side-chain rotamer states.

## 3. Results

One approach to the protein folding problem is to try to identify in the database of known structures, or in a repertoire of combinatorially generated models (given secondary structure elements; Cohen *et al.*, 1979), those folds that can accommodate the sequence of interest. We tested the discriminatory power of solvation preference on native and hypothetical sequence–structure pairs, which had been generated by either varying the structure for a given sequence or varying the sequence in a given structure.

### (a) *Deliberately misfolded structures*

Fourteen pairs of proteins with the same number of residues but dissimilar structures were misfolded in the spirit of Novotny *et al.* (1984) by swapping

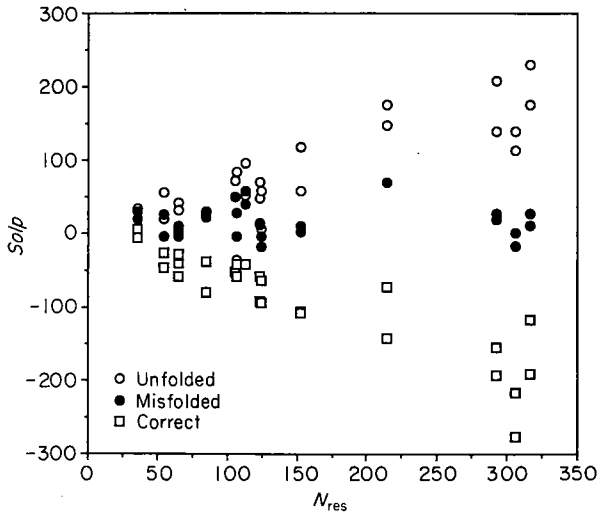



**Figure 2.** Discrimination between deliberately misfolded (Table 2) and correctly folded proteins by solvation preference. Negative *Solp* values indicate more preferred (more stable) structures. (○) Unfolded models (all atoms maximally solvated); (●) misfolded models; (□) parent X-ray structures of the misfolded set. All misfolded models are less favourable than the correct fold. The average stabilization/atom in favour of the correct fold is 0·19 (±0·05 S.D.), range 0·11 to 0·33. The sequence of the multihaem cytochrome 2CDV is exceptional as it gives a negative solvation preference for the unfolded state at $N_{res}=107$. Cofactors were omitted from occupancy calculations. $N_{res}$, number of residues in the protein.

the sequences of each pair (Table 2). Relaxing the co-ordinates with 500 steps of steepest descent energy minimization using the program GROMOS (van Gunsteren & Berendsen, 1987) led to almost indistinguishable potential energy between the misfolded and correct models, consistent with the original report by Novotny *et al.* (1984). So GROMOS potential energy (*in vacuo*) is not a good discriminant of correctness of structure. In contrast, solvation preference values clearly differentiate between correctly folded X-ray structures and deliberately misfolded models (Fig. 2). This is not a trivial achievement, as no proteins which are similar in sequence with the misfolded sequence–structure pairs were present in the database (Table 1) used to derive the preference parameters.

How does solvation preference compare to other empirical models? In Rashin's (1984) empirical free energy model, free energy is a linear function of the total solvent accessible surface area. Would Rashin's model discriminate between the native/misfolded pairs in Table 2? The answer is no, not in all cases, as the difference in solvent-accessible surface area of the misfolded models compared to the correctly folded models varies from −9% to +33% (average +9%). Another diagnostic tool (Baumann *et al.*, 1989), based on database-derived characteristic value ranges for polar fraction (surface area weighted with the absolute values of the partial atomic charge), fails to rank seven of the 28 misfolded models (Table 2) as unusual (data not shown). A third model, the solvation free energy of folding of Eisenberg & McLachlan (1986), works well on three misfolded pairs in Table 2 (Chiche *et al.*, 1990), in that the estimated free energy difference between unfolded and folded states was 25 to 32% larger for the correctly folded models compared to the misfolded ones. Perhaps this model would also perform well in all the cases given in Table 2.

### (b) *Sequence placement on native backbone*

Is solvation preference (*Solp*) able to find the correct alignment of a sequence in its native backbone structure? This was tested by successively displacing the sequence along its native structure, with cyclic closure, and building an explicit optimized 3-D model for each displacement. Among 124 cyclically permuted models of ribonuclease A (1RN3; Borkakoti *et al.*, 1982), the correct solution

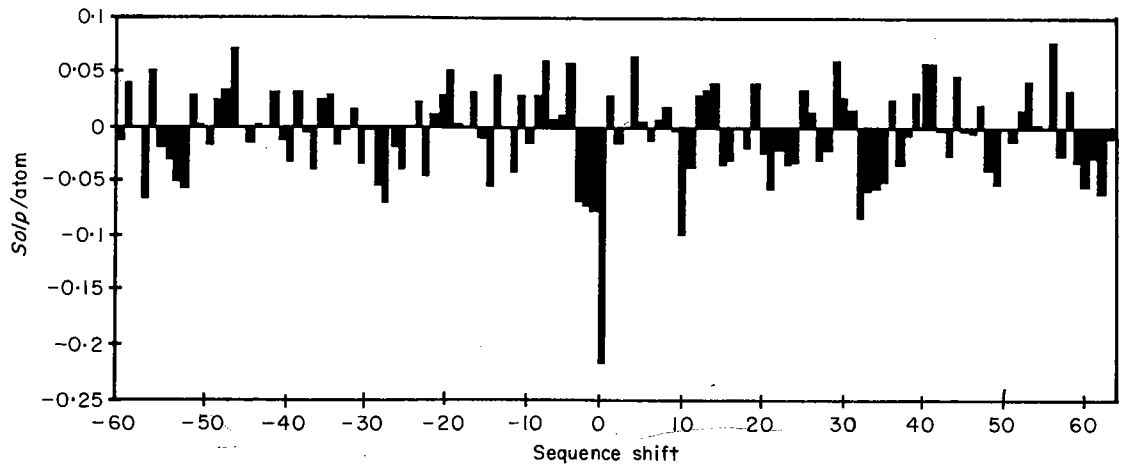
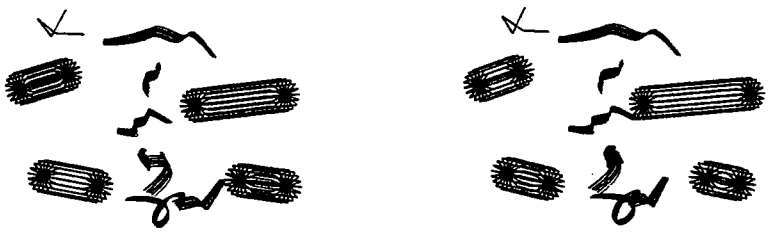


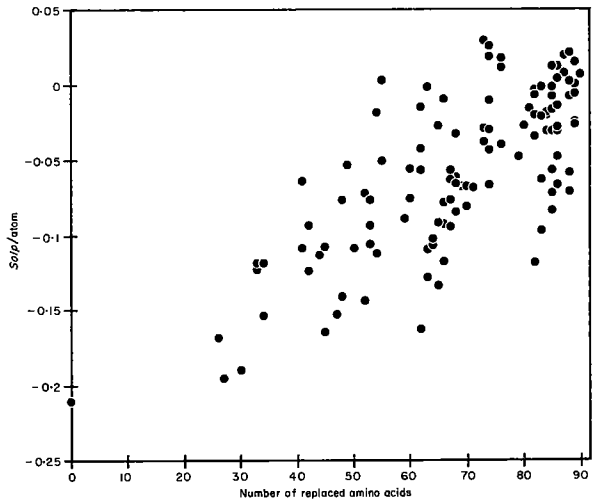

**Figure 3.** Identification of correct placement of a sequence in its native backbone fold. The sequence of ribonuclease A (1RN3) was shifted cyclically through the backbone of the X-ray structure in steps of 1 residue and side-chain orientations were optimized. The native conformation clearly has the best solvation preference.



(a)

| Res. num. | Strand | | Helix | Res. num. | |
|---|---|---|---|---|---|
| 1–5 | ANI**M**K**I**V**Y**WS | ... | G**N**TEK**M**A**E**L**I**AKG**I**IESGK | 10–27 | 1st unit |
| 30–35 | KD**V**N**T**I**N**V**S**D | ... | NVSDV**N**I**D**E**L**L**N**EDILILG | 39–45 | 2nd unit |
| 48–53 | LNED**I**L**I**L**G**C | ... | LEESE**E**F**E**P**F**I**E**E**I**STKIS | 65–73 | 3rd unit |
| 80–88 | G**K**K**V**A**L**F**G**S**Y** | ... | WGD**G**K**W**M**R**D**F**E**E**R**M**NGYGC | 93–104 | 4th unit |
| 110–119 | **V**VETP**L**I**V**ON | ... | QNEPDE**A**E**O**D**C**I**E**F**G**KKIA | 124–136 | 5th unit |

(b)




(c)

**Figure 4.** Identification of correct $\alpha/\beta$ topology by solvation preference. The core of flavodoxin consists of a 5-stranded $\beta$-sheet shielded on both sides by $\alpha$-helices. The strands form a sheet in the order e–d–c–a–b. How does the sequence code for this? The role of hydrophobicity was tested by building models of $5! = 120$ sequence permutations of the $\beta$–$\alpha$-units in the fixed structure. (a) Ribbon diagram (Vriend, 1990) of the core. Loops are omitted. Four helices pack in anti-parallel orientation against the preceding strand. We also include in the core an irregular piece of structure between the 2nd and 3rd $\beta$-strand (top left in (a)). The backbone co-ordinates were extracted from the PDB file 3FXN. (b) Sequence alignment. Each strand and the succeeding helix are treated as 1 unit. The underlined segments indicate which residues/positions were included in the structural models (95 residues). Within each unit, 4 hydrophobic residues from the helix (in diamond formation) pack around a central hydrophobic residue (leucine or isoleucine) on the strand (Cohen *et al.*, 1982). These residues are shown in bold. When shuffling sequence segments, these residues were placed in the corresponding position in the new structural unit. The helices and strands in different units have different lengths but the shuffled models have all the same backbone structure. The sequences corresponding to the strands and helices were truncated or flanking residues included so as to always have the same structural positions occupied, with varying sequence. For example, if the 1st and 2nd unit sequences were interchanged, then the sequence NTINV would be built in positions 1 to 5 of the native structure and the sequence IMKIVY in positions 32 to 37, and similarly for the helices. (c) Solvation preference for the optimized models with shuffled sequences plotted against the number of amino acid substitutions compared to the native sequence-structure fit. It appears that the native topology has the most favourable solvent interactions.

stands out clearly as the best fit with *Solp*/atom = −0·20, while the wrong arrangements had much higher *Solp*/atom values ranging from −0·10 to +0·08 (Fig. 3).

Is it possible to generate misfolded models with a better solvation preference value than the native fit? In general, this seems not to be the case, but we have found at least one interesting exception. The structure of avian pancreatic polypeptide (1PPT; Blundell *et al.*, 1981) consists of a proline-rich tail packing against an α-helix. Shifting the sequence by 16 residues, we did find a *Solp*/atom value slightly lower than that of the native structure, but this particular misfolded model can be rejected on other grounds, as it places the polyproline segment in an α-helix.

### (c) *Alternative α/β topologies*

In a more difficult test, is solvation preference able to detect the correct topological arrangement of sequence fragments in a 3-D template? To test this, the core of flavodoxin (3FXN; Smith *et al.*, 1977) was taken as the 3-D template and divided into five β/α units. Solvation preference was used to evaluate all 120 alternative arrangements of the corresponding sequence fragments in this template, i.e. each strand sequence in each strand position and each helical sequence in each helical position (for details, see Fig. 4). Again, the correct native arrangement stands out with the best solvation preference value (Fig. 4). In part, this is because the side-chains can be accommodated best in the native backbone trace, which was not varied. With increasing deviation from the native arrangement there is a tendency toward less favourable solvation preference values. Apparently the procedure is sensitive to effects such as solvent exposure of the wrong face of a helix or of an interior β-strand, removal from solvent of edge strands and the like. However, its predictive capacity is still relatively weak.

### (d) *Alternative all-α topologies*

Is solvation preference useful in assessing the quality, i.e. the sequence-structure fit, of combinatorially generated models? In a combinatorial approach to protein structure prediction, all possible pairings of preformed secondary structure segments can be generated and evaluated to yield a list of acceptable packing models by applying simple physical and geometrical constraints. Cohen *et al.* (1979) showed that in this way the number of possible models for myoglobin can be reduced from a very large initial number to 20. The models are topologically of two classes. Models with a r.m.s. $C^{\alpha}$ distance of 5 to 9 Å from the native globin structure differ mainly in the position and orientation of the short F helix. Models with $C^{\alpha}$ r.m.s. distance of 11 to 15 Å have a very different mode of packing compared to the native structure. Solvation preference works surprisingly well in this case: the solvation preference values of the models fall into two clusters that coincide with the topological classes and the low-r.m.s. cluster has better solvation preference values close to those of normal proteins (Fig. 5). This partition appears clearer than that obtained by the pair potential for interresidue contacts and packing criteria of Gregoret & Cohen (1990).

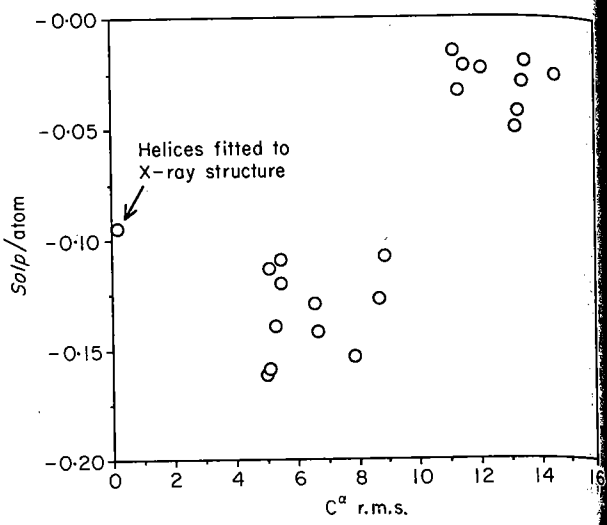



**Figure 5.** Partition of combinatorially generated myoglobin models by solvation preference. The original models (Cohen *et al.*, 1979) have only $C^{\alpha}$ co-ordinates. Side-chains were built to helical backbone models and optimized, including a solvation term in the object function. All models in the low-r.m.s. cluster (bottom left) have native-like topologies, while those in the high-r.m.s. cluster (top right) are quite different from the globin fold. The low-r.m.s. cluster has solvation preferences in the range observed for normal X-ray structures. Loops are not included in the models, which have 103 residues instead of 153 for native myoglobin. A similar model in which the helices (no loops) are packed using the correct X-ray structure (5MBN) as template has a relatively high solvation preference compared to the best combinatorial models, due to an empty haem pocket. Its $C^{\alpha}$-r.m.s. of 0·2 Å is due to proline kinks in the real structure and straight helices used in fitting.

### (e) *Local evaluation of sequence-structure fits*

Given the result that an entire protein structure can be identified as possibly incorrect, can one pinpoint where in the structure errors are located? As solvation preference is evaluated here at the atomic level, one can display the local values in 3-D graphics or plot them sequentially, in a manner similar to that of crystallographic *B*-values. Regions of gross departure from the database average are potential trouble spots. As solvation preference takes no account of electrostatics or specific interactions, such as H-bonds between polar groups or salt bridges, it alone is not capable of unambiguously identifying every incorrectly positioned residue. However, Figure 6(a) suggests that longer stretches of unfavourable solvation preference do

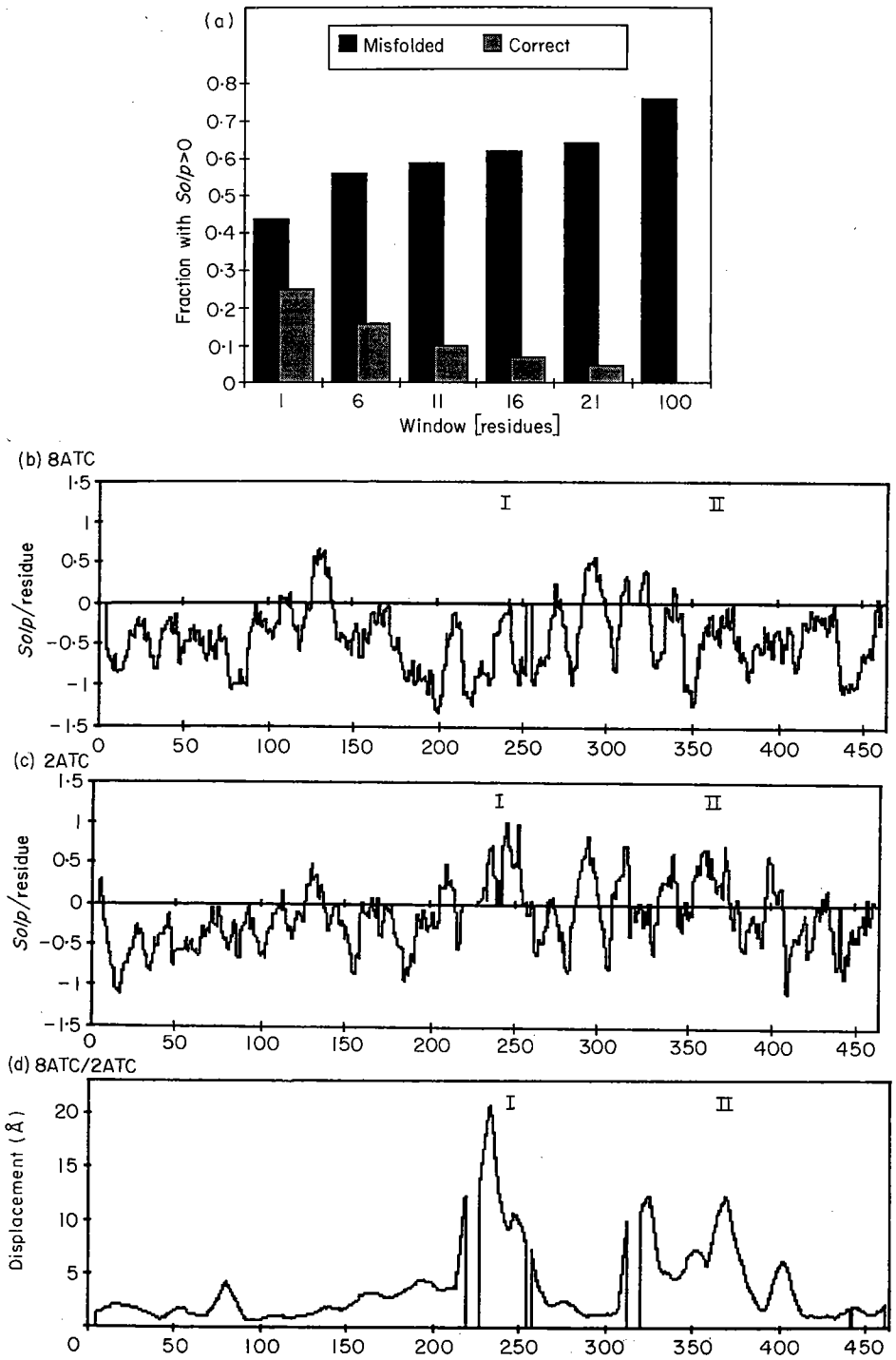



**Figure 6.** Use of solvation preference in locating possible errors. Positive *Solp* values indicate unfavourable regions. (a) *Solp*/residue averaged over different window lengths in misfolded models and their correct parent structures (see Fig. 2). Due to fluctuations, about 1/4 of the residues in correct models have solvation preference values $>0$, but longer stretches of unfavourable averaged solvation preference are increasingly rare. Misfolded models in general have less favourable solvation preference, and the longer fragments one examines, the clearer the discrimination becomes. A window width of 11 residues was used to analyse models of aspartate transcarbamoylase with (b) the correct (PDB dataset 8ATC, 2·5 Å resolution, *R*-value 0·165) and (c) an incorrect chain tracing (PDB dataset 2ATC, 3·0 Å resolution, *R*-value 0·27). The vertical axis is solvation preference per residue. Side-chains were optimized as described in Methods. (d) Displacement of $C^{\alpha}$ atoms (Å) in 2ATC relative to 8ATC in optimal sequence alignment, averaged over 11 residues. The enzyme consists of a catalytic (chain A) and a regulatory subunit (chain B). Chain B starts at position 310. A few gaps were introduced in the graphs to align the sequences. The 1st 7 residues of chain B were not seen in the electron density-map for 8ATC. The 2 widest humps of unfavourable solvation preference in the profile of 2ATC correspond to erroneously traced regions, labelled I and II. Between residues 235 and 250 of chains A, one structure has a loop before a strand and the other has a loop after the strand, so the sequence is misaligned in the strand. The other erroneous region is the switch of 2 $\beta$-strands between residues 40 and 65 in chain B.

indicate possible errors (which warrant checking by other, independent measures). For example, two conspicuously broad positive regions in the solvation preference profile of PDB dataset 2ATC for aspartate transcarbamoylase (Honzatko *et al.*, 1982) correspond to a misaligned edge strand in the catalytic subunit and to the switch of two β-strands in the regulatory subunit compared to the corrected dataset 8ATC (Ke *et al.*, 1988) (Fig. 6(b) to (d)). Not all positive peaks are diagnostic of bad structure however, as there are a number of positive peaks even in the correct structure (Fig. 6(b)).

### (f) *Robustness of the method*

How robust is the evaluation procedure with respect to errors in side-chain conformation, as a result of a particular model building procedure? To demonstrate the robustness of the method, we have compared solvation preference values for various ways of producing model structures. On average, the correct models, modulo changes in side-chain conformation, had the following average *Solp*/residue values: Monte Carlo optimized models, −0·55; original X-ray structures, −0·39; energy minimized X-ray models, −0·33 and unoptimized models in which side-chains are added in standard conformation, −0·29. Incorrect (misfolded) models, on the other hand, had consistently higher values: Monte Carlo optimized, +0·20; energy minimized, +0·47, and using standard side-chain conformations, +0·34. No matter how the side-chains arrangement was produced, solvation preference evaluation identified misfolded models as incorrect.

This result can be interpreted in terms of the value range of solvation preference for a given backbone model as side-chain conformation varies. A rough estimate gives a fluctuation of ±0·13 per residue. So, in practice, *Solp* values carry this amount of uncertainty when comparing models produced by different side-chain optimization procedures. Note, however, that all comparisons in this study are made between models that have been similarly optimized in rotamer space and with solvation preference in the object function.

How much does the success of the method depend on the fact that side-chains are best accommodated in the native backbone conformation? To test this, we built models (with very approximate loop conformations) of each of six globin sequences, using the native backbone itself, and of each of the five other homologous proteins. The sequence–structure alignments were taken from Bashford *et al.* (1987). Does the solvation preference look less good when a homologous backbone is used? By how much? The *Solp*/residue values for the optimized models were all negative, indicating reasonable structures, in the following value range: haemoglobin α-chains, −0·31 to −0·17; haemoglobin β-chains, −0·51 to −0·32; myoglobin, −0·68 to −0·54; erythrocruorin, −0·56 to −0·32; sea lamprey haemoglobin −0·59 to −0·39; and leghaemoglobin, −0·58 to −0·24. The native backbone had the most favourable *Solp* in four cases out of six. The oligomeric haemoglobin has the most hydrophobic sequence and least favourable *Solp*. In a similar test between a variable and constant immunoglobulin domain (1REI *versus* 1FC1, C r.m.s. deviation 1·8 Å for the core), the native fit had *Solp*/residue values of −0·51 and −0·49, and the exchanged pairs −0·26 and −0·28. These examples demonstrate once more that solvation preference identifies correct backbone folds, and that the placement of side-chains is not overly sensitive to the details of backbone co-ordinates.

## 4. Discussion

The ability of solvation preference to identify correct sequence–structure pairs was tested on misfolded models generated in three different ways: (1) given a native fold, shift the sequence along the structure, (2) given a native fold, replace the sequence by an unrelated one (and *vice versa*), or (3) given the sequence and secondary structure, generate alternative packings of the helices and sheets. In each case, solvation preference was able to identify the correct structure for a given sequence or the correct sequence placement among possible alternatives. The most impressive example was the partition of combinatorially generated myoglobin models, known to be a difficult problem (Cohen *et al.*, 1979; Gregoret & Cohen, 1990).

These examples can be generalized to the problem of 3-D structure prediction for any sequence, based on the fact that only a rather limited set of secondary structure assemblies are observed in proteins (Richardson, 1981; Chothia & Finkelstein, 1990). Therefore, the question of predicting an unknown structure from the sequence can be restated as trying to identify the correct alignment of the sequence in a sufficiently large set of trial structures. In practice, the vast multitude of possible alignments (with gaps) is best scanned using potentials or preference parameters that do not require knowledge of side-chain conformations (Scharf, 1989; M. Scharf & C. Sander, unpublished results; Bowie *et al.*, 1990; Hendlich *et al.*, 1990; Sippl & Weitckus, 1991; Lüthy *et al.*, 1991; Finkelstein & Reva, 1991; Bowie *et al.*, 1991). Solvation preference, which explicitly takes side-chain packing into account, could be used as a second, more sensitive filter applied to the top 100 or 1000 sequence–structure alignments.

The principal limitation of the present method is that it provides only a rough approximation to energetics. All covalent, enthalpic and entropic effects are only indirectly included and averaged out in the distribution of atom types in solvation factor bins. Specific polar and electrostatic interactions are completely ignored in the present implementation, especially in the core. The linear representation of solvation preference as a function of solvation state (Fig. 1), used here for simplicity, is an approximation that could be removed in the future. The

current implementation is only applicable to "good" backbone models as main-chain atoms are not evaluated. Our analysis by solvation preference quite successfully tests an entire sequence–structure pair, but due to statistical fluctuations the identification of incorrect *parts* of structures is more difficult.

The atomic hydrophobicity scale implied here (Table 1) is roughly consistent with residue hydrophobicity scales derived from chemical analogue studies or statistics (Eisenberg & McLachlan, 1986; Janin, 1979; Vila *et al.*, 1991). Conceptually, the parameters derived here are different from several empirical free energy scales based on surface area (e.g. Eisenberg & McLachlan, 1986; Ooi *et al.*, 1987; Vila *et al.*, 1991) in that they are based on an excluded volume model (note that excluded volumes in the 1st hydration shell are strongly correlated with surface areas) and are derived from statistical observations on protein structures. Kang *et al.* (1988) developed an excluded volume model conceptually similar to ours, calibrated on experimental hydration free energies of small organic molecules, but they used a different approach to the calculation of solvent accessible volumes. Their approach is geometrically more precise but more complicated than the one used here (Gibson & Scheraga, 1988). To our knowledge, this method has not been applied to the evaluation of protein models. Our use of solvation preferences is also different from criteria used in previous statistical studies (e.g. Hendlich *et al.*, 1990; Gregoret & Cohen, 1990; Janin, 1979; Baumann *et al.*, 1989; Bryant & Amzel, 1987; Novotny *et al.*, 1988).

A limitation of the misfolded test cases is that the side-chains can be optimally fit into the rigid backbone structure only for the sequence from which the backbone structure was taken. Mutations in homologous proteins tend to lead to small shifts in backbone positions (while retaining the same fold). Clearly, in order to produce "better" misfolded models, it is necessary to incorporate the solvation term into an energy optimization protocol that allows the relaxation of backbone degrees of freedom. Such misfolded models would be more severe competition for correctly folded models.

The technical advantage of the solvent contact model over surface area calculations is that the degree of solvation of an atom becomes a particularly simple function of interatomic distances (eqns (1) and (2)), allowing rapid calculation of solvation-related quantities. Overall, the present results show that solvation preference parameters are a remarkably powerful discriminator between incorrectly and correctly folded globular protein models. We think that this approach will prove to be a useful tool in the screening of *de novo* designed structures, of 3-D structures modelled by sequence similarity and perhaps even of experimental structures. It may also prove useful in the prediction of 3-D protein structure based on aligning a new protein sequence to a representative set of template structures (Bowie *et al.*, 1990; Sippl & Weitckus, 1991; Lüthy *et al.*, 1991; Scharf, 1989; C. Sander & M. Scharf, unpublished results; Bowie *et al.*, 1991).

We thank Fred Cohen for the co-ordinates of myoglobin models; Reinhard Schneider for a non-redundant list of high-resolution PDB proteins; Cornelius Frömmel and Pieter Stouten for discussion. We are grateful to the crystallographers who have made protein co-ordinates available through the Protein Data Bank. An EMBO fellowship (L.H.) is gratefully acknowledged.

## References

Adman, E. T., Sieker, L. C. & Jensen, L. H. (1976). Structure of *Peptococcus aerogenes* ferredoxin, refinement at 2 Ångströms resolution. *J. Biol. Chem.* **251**, 3801–3806.

Adman, E. T., Turley, S., Bramson, R., Petratos, K., Banner, D., Tsernoglou, D., Beppu, T. & Watanabe, H. (1989). A 2·0-Ångströms structure of the blue copper protein (cupredoxin) from *Alcaligenes faecalis* S-6. *J. Biol. Chem.* **264**, 87–99.

Almassy, R. J., Fontecilla-Camps, J. C., Suddath, F. L. & Bugg, C. E. (1983). Structure of variant-3 scorpion neurotoxin from *Centruroides sculpturatus Ewing*, refined at 1·8 Å resolution. *J. Mol. Biol.* **170**, 497–527.

Arutynyan, E. G., Kuranova, P., Vainshtein, B. K. & Steigemann, W. (1980). X-ray structural investigation of leghemoglobin. VI. Structure of acetate ferrileghemoglobin at a resolution of 2·0 Å. (In Russian.) *Kristallografiya*, **25**, 80.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Baumann, G., Frömmel, C. & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng.* **2**, 329–334.

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., Pitts, J. E., Tickle, I. J., Woods, S. P. & Wu, C.-W. (1981). X-ray analysis (1·4 Ångströms resolution) of avian pancreatic polypeptide. Small globular protein hormone. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4175–4179.

Borkakoti, N., Moss, D. S. & Palmer, R. A. (1982). Ribonuclease-A. Least-squares refinement of the structure at 1·45 Ångström resolution. *Acta Crystallogr. sect. B*, **38**, 2210–2217.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, **7**, 257–264.

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

Brick, P., Bhat, T. N. & Blow, D. M. (1989). Structure of tyrosyl-tRNA synthetase refined at 2·3 Å resolution. Interaction of the enzyme with tyrosyl adenylate intermediate. *J. Mol. Biol.* **208**, 83–98.

Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Peptide Protein Res.* **29**, 46–52.

Carter, C. W., Jr, Kraut, J., Freer, S. T., Xuong, N.-H., Alden, R. A. & Bartsch, R. G. (1974). Two-Ångström crystal structure of oxidized chromatium high potential iron protein. *J. Biol. Chem.* **249**, 4212–4225.

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

Chothia, C. & Finkelstein, A. V. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039.

Cohen, F. E., Richmond, T. J. & Richards, F. M. (1979). Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J. Mol. Biol.* **132**, 275–288.

Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. (1982). Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156**, 821–862.

Colonna-Cesari, F. & Sander, C. (1990). Excluded volume approximation to protein–solvent interaction. The solvent contact model. *Biophys. J.* **57**, 1103–1107.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1981). Structure of bovine pancreatic phospholipase A2 at 1·7 Å resolution. *J. Mol. Biol.* **147**, 97–123.

Dijkstra, B. W., Renetseder, R., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1983). Structure of porcine pancreatic phospholipase A2 at 2·6 Å resolution and comparison with bovine phospholipase A2. *J. Mol. Biol.* **168**, 163–179.

Dodson, G., Hubbard, R. E., Oldfield, T. J., Smerdon, S. J. & Wilkinson, A. J. (1988). Apomyoglobin as a molecular recognition surface. Expression, reconstitution and crystallization of recombinant porcine myoglobin in *Escherichia coli*. *Protein Eng.* **2**, 233–237.

Drenth, J., Kalk, K. H. & Swen, H. M. (1976). Binding of chloromethyl ketone substrate analogues to crystalline papain. *Biochemistry*, **15**, 3731–3738.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature (London)*, **319**, 199–203.

Epp, O., Lattman, E. E., Schiffer, M., Huber, R. & Palm, W. (1975). The molecular structure of a dimer composed of the variable portions of the Bence-Jones protein REI refined at 2·0 Ångströms resolution. *Biochemistry*, **14**, 4943–4952.

Fauchere, J.-L. & Pliska, V. (1983). Hydrophobic parameters π of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369–375.

Finkelstein, A. V. & Reva, B. A. (1991). A search for the most stable folds of protein chain. *Nature (London)*, **351**, 497–499.

Finzel, B. C., Poulos, T. L. & Kraut, J. (1984). Crystal structure of yeast cytochrome c peroxidase refined at 1·7-Ångströms resolution. *J. Biol. Chem.* **259**, 13027–13036.

Ghosh, D., O'Donnell, S., Furey, W., Jr, Robbins, A. H. & Stout, C. D. (1982). Iron-sulfur clusters and protein structure of *Azotobacter* ferredoxin at 2·0 Å resolution. *J. Mol. Biol.* **158**, 73–109.

Gibson, K. D. & Scheraga, H. A. (1988). Volume of intersection of three spheres of unequal size: a simplified formula. *Mol. Phys.* **64**, 641–644.

Gilliland, G. L. & Quiocho, F. A. (1981). Structure of the L-arabinose binding protein from *Escherichia coli* at 2·4 Å resolution. *J. Mol. Biol.* **146**, 341–362.

Gregoret, L. M. & Cohen, F. E. (1990). Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* **211**, 959–974.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180.

Higuchi, Y., Kusunoki, M., Matsuura, Y., Yasuoka, N., Kakudo, M. (1984). Refined structure of cytochrome *c*3 at 1·8 Å resolution. *J. Mol. Biol.* **172**, 109–139.

Hobohm, U., Sander, C., Scharf, M. & Schneider, R. (1992). Selection of representative protein data sets. *Protein Science*, in the press.

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

Honzatko, R. B., Crawford, J. L., Monaco, H. L., Ladner, J. E., Edwards, B. F. P., Evans, D. R., Warren, S. G., Wiley, D. C., Ladner, R. C. & Lipscomb, W. N. (1982). Crystal and molecular structures of native and CTP-liganded aspartate transcarbamoylase from *Escherichia coli*. *J. Mol. Biol.* **160**, 219–263.

Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature (London)*, **277**, 491–492.

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

Kang, Y. K., Gibson, K. D., Nemethy, G. & Scheraga, H. A. (1988). Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell model. *J. Phys. Chem.* **92**, 4739–4742.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1–63.

Ke, H., Lipscomb, W. N., Cho, Y. & Honzatko, R. B. (1988). Complex of *N*-phosphoacetyl-L-aspartate with aspartate transcarbamoylase. X-ray refinement analysis of conformational changes and catalytic and allosteric mechanisms. *J. Mol. Biol.* **204**, 725–747.

Kraulis, P. J., Clore, G. M., Nilges, M., Jones, T. A., Pettersson, G., Knowles, J. K. C. & Gronenborn, A. M. (1989). Determination of the three-dimensional structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry*, **28**, 7241–7257.

Lee, B. & Richards, F. M. (1971). The interpretation of protein strctures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Levitt, M. J. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

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.

Mathews, F. S., Argos, P. & Levine, M. (1972). The structure of cytochrome 65 at 2·0 Ångströms resolu-

tion. *Cold Spring Harbor Symp. Quant. Biol.* **36**, 387–395.

McPhalen, C. A. & James, M. N. G. (1987). Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry*, **26**, 261.

Mondragon, A. Wolberger, C. & Harrison, S. C. (1989). Structure of phage 434 Cro protein at 2·35 Å resolution. *J. Mol. Biol.* **205**, 179–188.

Motoc, I. & Marshall, G. R. (1985). van der Waals' volume fragmental constants. *Chem. Phys. Letters*, **116**, 415–419.

Novotny, J., Bruccoleri, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure prediction. *J. Mol. Biol.* **177**, 787–818.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, **4**, 19–30.

Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 3086–3090.

Ploegman, J. H., Drent, G., Kalk, K. H. & Hol, W. G. J. (1978). Structure of bovine liver rhodanase. I. Structure determination at 2·5 Å resolution and a comparison of the conformation and sequence of its two domains. *J. Mol. Biol.* **123**, 557–594.

Priestle, J. P., Schaer, H.-P. & Gruetter, M. G. (1989). Crystallographic refinement of interleukin-1$\beta$ at 2·0 Ångströms resolution. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 9667–9671.

Rashin, A. A. (1984). Buried surface area, conformational entropy, and protein stability. *Biopolymers*, **13**, 1605–1630.

Richardson, J. S. (1981). Anatomy and taxonomy of protein structures. *Advan. Protein Chem.* **34**, 167–339.

Satow, Y., Watanabe, Y. & Mitsui, Y. (1980). Solvent accessibility and microenvironment in a bacterial protein proteinase inhibitor SSI (*Streptomyces* subtilisin inhibitor). *J. Biochem.* **88**, 1739–1755.

Scharf, M. (1989). Diploma thesis, University of Heidelberg.

Sharp, K. A. (1991). The hydrophobic effect. *Curr. Opin. Struct. Biol.* **1**, 171–174.

Sippl, M. J. & Weitckus, S. (1991). Detection of native like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, in the press.

Smith, W. W., Burnett, R. M., Darling, G. D. & Ludwig, M. L. (1977). Structure of the semiquinone form of flavodoxin from *Clostridium* mp. Extension of 1·8 Å resolution and some comparisons with the oxidized state. *J. Mol. Biol.* **117**, 195–225.

Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1983). Adjustment of restraints in the refinement of methemerythrin and azidomethemerythrin at 2·0 Ångströms resolution. *Acta Crystallogr. sect. B.* **39**, 697–703.

Stout, C. D. (1989). Refinement of the 7 Fe ferredoxin from *Azotobacter* at 1·9 Å resolution. *J. Mol. Biol.* **205**, 545–555.

Tronrud, D. E., Monzingo, A. F. & Matthews, B. W. (1986). Crystallographic structural analysis of phosphoramidates as inhibitors and transition-state analogs of thermolysin. *Eur. J. Biochem.* **157**, 261–268.

van Gunsteren, W. F. & Berendsen, H. J. C. (1987). *GROMOS, Groningen Molecular Simulation Computer Program Package*, University of Groningen, The Netherlands.

Vila, J., Williams, R. L., Vasquez, M. & Scheraga, H. A. (1991). Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins*, **10**, 199–218.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56.

Watenpaugh, K. D., Siebar, L. C. & Jensen, L. H. (1980). Crystallographic refinement of rubredoxin at 1·2 Ångströms resolution. *J. Mol. Biol.* **138**, 615–633.

*Edited by F. Cohen*