Recognizing misfolded and distorted protein structures by the assumption-based similarity score

A.P.Golovanov1,2, P.E.Volynsky1, S.B.Ermakova1 and A.S.Arseniev1,3

1Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Ul. Miklukho-Maklaya, 16/10 Moscow V-437, 117871 GSP-7, Russia and 2Université des Sciences et Technologies de Lille, CRESIMM, UFR de Chimie, Bâtiment CS, 59655 Villeneuve d’Ascq, Cedex, France
3To whom correspondence should be addressed

A new similarity score (Σ-score) is proposed which is able to find the correct protein structure among the very close alternatives and to distinguish between correct and deliberately misfolded structures. This score is based on the general principle ‘similar likes similar’, and it favors hydrophobic and hydrophilic contacts, and disfavors hydrophobic-to-hydrophilic contacts in proteins. The values of Σ-scores calculated for the high-resolution protein structures from the representative set are compared with those of alternatives: (i) very close alternatives which are only slightly distorted by conformational energy minimization in vacuo; (ii) alternatives with subsequently growing distortions, generated by molecular dynamics simulations in vacuo; (iii) structures derived by molecular dynamics simulation in solvent at 300 K; (iv) deliberately misfolded protein models. In nearly all tested cases the similarity score can successfully distinguish between experimental structure and its alternatives, even if the root mean square displacement of all heavy atoms is less than 1 Å. The confidence interval of the similarity score was estimated using the high-resolution X-ray structures of domain pairs related by non-crystallographic symmetry. The similarity score can be used for the evaluation of the general quality of the protein models, choosing the correct structures among the very close alternatives, characterization of models simulating folding/unfolding, etc.

Keywords: hydrophilic contacts/hydrophobic contacts/molecular dynamics simulation/protein structure recognition/structure quality

Introduction

The main tasks of protein engineering are the rational change of properties (structure, functional activity, stability, specificity, etc.) of natural proteins and ‘de novo’ protein design. The success of these tasks is strongly limited by the inadequate modeling of interatomic interactions in proteins and protein–protein complexes, mainly due to the ignoring of the hydrophobic effect. Although hydrophobic interactions are considered to be one of the main driving forces of protein folding (Dill, 1990), the correspondent term describing the ‘hydrophobic energy’ is not included in the widely used force fields. As the result, the energy calculation using standard empirical potentials cannot distinguish between the correctly folded and deliberately misfolded structures (Novotny et al., 1984, 1988; Holm and Sander, 1992). In contrast, many successful methods discriminating between native proteins and misfolded models utilize the evaluation of the hydrophobicity-related parameters in various ways (Holm and Sander, 1992; Lüthy et al., 1992; Luthardt and Frommel, 1994; Huang et al., 1995a; Wang et al., 1995a; Miyazawa and Jernigan, 1996; Huang et al., 1996; Park and Levitt, 1996; Park et al., 1997). To describe quantitatively the hydrophobic interactions in proteins, several approaches were used, most of them exploited the tendency of non-polar atoms (or groups of atoms) to avoid contact with polar molecules (solvent) (Eisenberg and McLachlan, 1986; Delarue and Koehl, 1995; Kurochkina and Lee, 1995; Wang et al., 1995a,b; Bahar and Jernigan, 1997). One of the most convenient approaches is that using the molecular hydrophobicity (lipophilicity) potential (MHP), which allows one to calculate quantitatively the surrounding hydrophobicity at any point in space (Fauchère et al., 1988; Furet et al., 1988; Kellogg et al., 1991; Gaillard et al., 1994). This method was used for the detailed characterization of spatial hydrophobicity in membrane and globular proteins (Brasseur, 1991; Efremov et al., 1992; Sansom and Kerr, 1993; Efremov and Alix, 1993; Efremov et al., 1995), and for the detailed characterization and classification (favorable hydrophobic and hydrophilic; or unfavorable hydrophobic-to-hydrophilic) of interresidue contacts in proteins (Golovanov et al., 1995, 1998). The general conclusion from all these studies accords well with the intuitively obvious assumption that in native proteins the polar sets of atoms (e.g. backbone part of the residues) tend to contact the polar sets, the nonpolar sets of atoms (hydrophobic side chains) tend to contact the nonpolar sets, and nonpolar sets avoid contacts with polar sets (but often have to do so because of the polypeptide nature of proteins and the close packing requirement). The assumption that in the protein ‘similar likes similar’ and the MHP approach (which provides the quantitative criteria of hydrophobicity and hydrophilicity) can be used for the construction of a ‘similarity’ score.

In the present work we show that the value of the similarity score is correlated with the quality of the protein model. To prove that, the Σ-score values are calculated for ‘good’ experimental structures of proteins, and these values were compared with that calculated for ‘nearly good’ models as well as for completely wrong models. The confidence interval of the Σ scores was also estimated.

Materials and methods

Construction and calculation of the similarity score Σ for the protein models

To calculate the relative strengths and the types of intramolecular contacts (hydrophobic, hydrophilic or unfavorable) between side-chain (S) and backbone (B) segments of amino acid residues in proteins with known spatial structure we used approach described in detail in (Golovanov et al., 1998). The value of MHPX Y i, created by the atoms of segment X (S or B) of amino acid residue i (source residue) in geometrical
center of the segment Y (S or B) of residue j (target residue, \(i \neq j\)) was calculated according to the formula:

\[
M_{ij}^{XY} = \sum_{k=1}^{N_i} f_k \cdot \exp(-r_{kj} / R_{dec})
\] (1)

The atomic hydrophobicity constants \(f_k\) (derived from experimental octanol–water partition coefficients of large set of various chemical compounds) of atom k were taken from Viswanadhan et al. (1989) and complemented as described before (Efremov and Alix, 1993). \(N_i\) is the number of atoms in segment X of residue i, \(R_{dec}\) is the effective radius of decay of the potential (taken to be 1 Å, as shown by Fauchère et al., 1988), and \(r_{kj}\) is the distance (in Å) between atom k and the geometrical center of segment Y of residue j. The exponential distance function in Equation (1) obviously has a sense only when \(r_{kj}\) exceeds a certain value (approximately the sum of two atomic van der Waals radii). This condition is implicitly fulfilled for the protein models which do not have strong steric bumps.

Individual contributions of amino acid residue i to hydrophobic, hydrophilic and unfavorable pairwise interactions were calculated as:

\[
C_{XY}^{\text{phob}} = 10^6 \times \sum_{j=1}^{N_j} M_{ij}^{XY} \times M_{ji}^{XY}, \quad \text{if} \quad M_{ij}^{XY} \times M_{ji}^{XY} > \text{Cutoff}_{MHP}
\]

\[
C_{XY}^{\text{phil}} = 10^6 \times \sum_{j=1}^{N_j} M_{ij}^{XY} \times M_{ji}^{XY}, \quad \text{if} \quad M_{ij}^{XY} \times M_{ji}^{XY} < \text{Cutoff}_{MHP}
\]

\[
C_{XY}^{\text{unf}} = 10^6 \times \sum_{j=1}^{N_j} M_{ij}^{XY} \times M_{ji}^{XY}, \quad \text{if} \quad |M_{ij}^{XY} \times M_{ji}^{XY}| < \text{Cutoff}_{MHP}
\]

(2)

where XY is SS, SB or BB. The value of \(\text{Cutoff}_{MHP} = 0.001\) was chosen (Efremov et al., 1995; Golovanov et al., 1998) to consider only the strongest hydrophobic, hydrophilic and unfavorable contacts and to discard the weak contacts (originating from residue pairs at distances greater ~7 Å). The values of individual contributions (given in arbitrary units) were multiplied by the factor \(10^6\) for convenience.

To characterize the contributions from SS, SB and BB interactions for the whole protein, several parameters (\(S_{\text{phob}}^{SS}, S_{\text{phil}}^{SS}\) and \(S_{\text{unf}}^{SS}\)) were calculated as the sum of correspondent individual contributions (\(C_{XY}^{\text{phob}}, C_{XY}^{\text{phil}}\) and \(C_{XY}^{\text{unf}}\)) over all residues, normalized by the number of residues in the protein. Thus the normalized sums of hydrophobic, hydrophilic and unfavorable contributions were calculated as:

\[
S_{\text{phob}} = S_{\text{phob}}^{SS} + S_{\text{phob}}^{SB} + S_{\text{phob}}^{BB}
\]

\[
S_{\text{phil}} = S_{\text{phil}}^{SS} + S_{\text{phil}}^{SB} + S_{\text{phil}}^{BB}
\]

\[
S_{\text{unf}} = S_{\text{unf}}^{SS} + S_{\text{unf}}^{SB} + S_{\text{unf}}^{BB}
\]

(3)

The similarity score \(\Sigma = S_{\text{phob}} + S_{\text{phil}} + S_{\text{unf}}\) gives the total sum of all favorable (hydrophobic and hydrophilic) and unfavorable (hydrophobic-to-hydrophilic) interactions in the protein, normalized by the number of residues.

The contributions of hydrophobic, hydrophilic and unfavorable interactions for the protein model, as well as similarity score \(\Sigma\) are calculated with our program Hi-EXPLO (Golovanov et al., 1998), which is available from A.P.G.
calculated for Cα atom positions after best fit superposition of structures.

To obtain the large set of alternative protein models close to the experimental ones, 196 proteins from the representative set were subjected to energy minimization using TRIPOS (30 and 100 steps) and CHARMM (30, 150 and 300 steps of minimization) force fields without the electrostatic energy terms. Thus for each of 196 proteins five alternative models were obtained. The small deviations of these alternative models from the experimental structures were characterized by r.m.s.d. calculated for all heavy atoms after best fit superposition of structures.

Estimation of the Σ-score confidence interval

The confidence interval of the Σ-score was estimated basing on its calculation for different ‘correct’ models of the same protein, i.e., for the pairs of domains related by non-crystallographic symmetry (resolution better 2.5 Å), and for various models of the same protein in NMR-derived entries. The following X-ray set of domain pairs was used (the PDB code and chain identifiers of two compared chains are given): 1apx (A,B), 1brc (A,B), 1buc (A,B), 1deh (A,B), 1dp (A,B), 1ebg (A,B), 1ebh (A,B), 1gse (A,B), 1ids (A,B), 1les (AB,CD), 1ndp (A,B), 1pvd (A,B), 1pyd (A,B), 1set (A,B), 1smn (A,B), 1tar (A,B), 1wcg (A,B), 2cst (A,B), 2nac (A,B), 2phi (A,B), 2wgc (A,B), 4mdh (A,B), 5p2p (A,B), 7wga (A,B), 8cat (A,B), 9wga (A,B). The set of NMR structures (all NMR entries in PDB searched by keywords ‘protein’ or ‘toxin’ without complexes of short peptides with DNA and entries without explicit prostons) consisted of 277 entries (total 5665 protein models). The Σ-score confidence intervals were estimated as the three mean standard deviations of Σ-score for the pairs of domains (X-ray) or for different models (NMR) of the same protein.

Molecular dynamics simulations in solvent

MD simulations were performed on the NMR structure (Lubieniecki et al., 1994) of barstar (PDB entry 1btr) and the crystal structure (Houset et al., 1994) of scorpion toxin II (PDB entry lptx, resolution 1.3 Å) using the program CHARMM (Brooks et al., 1983). The computational protocols for the two proteins were the same. The proteins were solvated in the spheres of water molecules (50 Å in diameter, ~1800 water molecules). Solvent shells were pre-equilibrated by energy minimization with subsequent short (20 ps, 300 K) MD simulation with the protein atoms fixed. Then the entire systems were minimized, and the corresponded ‘minimized in solvent’ protein structures were taken as the references. The systems were then heated to 300 K and equilibrated. The length of the subsequent MD run was 1600 ps for barstar and 600 ps for toxin II. Stochastic boundary conditions were applied for both systems. The structures were sampled every 1 ps, both for the heating stage and MD simulation. The details of MD simulations will be published elsewhere.

Results and discussion

Design of similarity score

The similarity score characterizes intramolecular contacts in proteins with known three-dimensional structures. It favors both hydrophobic and hydrophilic contacts and penalizes those that are hydrophobic-to-hydrophilic. Each amino acid residue is subdivided into two segments (side-chain and backbone) and pairwise contacts between these segments are summed all over the protein. The type of contact (hydrophobic, hydrophilic or unfavorable) and its relative strength is calculated using the molecular hydrophobicity potential approach as described before (Golovanov et al., 1998). The definitions of various types of contacts are illustrated on Figure 1. If both groups of atoms (segments of amino acid residues) are turned to each other by their hydrophobic sides (i.e., they create positive, hydrophobic potential in the geometrical center of each other) the contact is hydrophobic. If both groups of atoms are turned to each other by their hydrophilic sides (i.e., they create a negative, hydrophilic potential in the geometrical center of each other) the contact is hydrophilic. But if two groups are turned to each other by the sides of different polarity (i.e., they create the potentials of different signs in the geometrical center of each other) the contact is unfavorable. The ‘strength’ of the individual contact is equal to the product $MH{P_i} 	imes MH{P_j}$, which is positive for both favorable hydrophobic and hydrophilic contacts, and negative for unfavorable hydrophobic-to-hydrophilic contacts. The similarity score is the sum of hydrophobic, hydrophilic and unfavorable contacts in the protein, and its value decreases as the favorable contacts decrease, or unfavorable contacts increase. Thus the expansion of the protein model leading to the loss of intramolecular contacts causes a decrease in the similarity score. The form of the similarity score permits one to consider separately the contributions from hydrophobic, hydrophilic and unfavorable interactions, or contributions from side-chain–side-chain (SS, mainly hydrophobic interactions), backbone–backbone (BB, mainly hydrophilic interactions (type and boundary structure)) and side-chain–backbone (SB) interactions. While the similarity score reflects the ‘overall’ quantity and quality of contacts in the protein model, its constituents give additional information about particular types of interactions in this model.

Statistical analysis of representative and misfolded sets of proteins

Two sets of proteins were used to reveal the typical values of Σ-score and its constituents for ‘correct’ and ‘wrong’ models: the representative set of 196 high-resolution protein structures and the set of 26 deliberately misfolded proteins, obtained by threading of amino acid sequence of one protein on the spatial scaffold of another one (see Materials and methods). The calculated values of unnormalized Σ-scores and its unnormalized constituents for two sets of proteins were roughly proportional to the number of amino acid residues (with the correlation coefficients >0.95, data not shown). Therefore for further analysis we used Σ-scores and its constituents (see Materials and methods) normalized by the number of residues in protein.

The values of Σ-score and its constituents calculated for the representative and misfolded sets of protein models are...
Table I. Values of Σ-score and contributions of various interactions to it for misfolded structures and 'correct structures' (representative set)

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Misfolded set</th>
<th>Representative set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>12.9 ± 5.7</td>
<td>17.2 ± 6.6</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>154.7 ± 34.8</td>
<td>179.1 ± 34.0</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;β&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>232.4 ± 49.0</td>
<td>323.4 ± 71.2</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>172.9 ± 66.1</td>
<td>318.6 ± 60.2</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phob&lt;/sup&gt;</td>
<td>1.08 ± 0.92</td>
<td>3.94 ± 1.91</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phob&lt;/sup&gt;</td>
<td>0.27 ± 0.77</td>
<td>0.29 ± 0.51</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>-52.0 ± 20.1</td>
<td>-58.1 ± 14.2</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>-159.9 ± 39.3</td>
<td>-183.2 ± 32.3</td>
</tr>
<tr>
<td>Sum&lt;sub&gt;α&lt;/sub&gt;&lt;sup&gt;phil&lt;/sup&gt;</td>
<td>-139.7 ± 38.2</td>
<td>-99.6 ± 27.3</td>
</tr>
<tr>
<td>φ&lt;sub&gt;phil&lt;/sub&gt;</td>
<td>400.0 ± 58.5</td>
<td>519.7 ± 88.6</td>
</tr>
<tr>
<td>φ&lt;sub&gt;phob&lt;/sub&gt;</td>
<td>174.3 ± 66.8</td>
<td>322.8 ± 60.6</td>
</tr>
<tr>
<td>α&lt;sub&gt;phil&lt;/sub&gt;</td>
<td>-351.5 ± 58.1</td>
<td>-340.9 ± 53.7</td>
</tr>
<tr>
<td>Σ</td>
<td>222.7 ± 82.6</td>
<td>501.6 ± 86.3</td>
</tr>
</tbody>
</table>

Mean values and standard deviation are shown. The strongest contributions in hydrophobic, hydrophilic and unfavorable interactions are shown in bold.

Table I. Values of Σ-score and contributions of various interactions to it for misfolded structures and 'correct structures' (representative set)

Presented in Table I. It can be seen that the values of similarity scores Σ calculated for the misfolded set are generally less than for the representative set, which means that the similarity of packing for the native proteins is higher than that for the misfolded models. As the different terms contributing to the total Σ-score reflect different kinds of intramolecular interaction, it is interesting to compare the typical values of these parameters for the ‘normal’ and misfolded proteins to understand the differences between them. The value Sum<sub>α</sub><sup>phil</sup> contributes mostly to the total favorable hydrophilic interactions φ<sub>phil</sub>. This value generally should not depend on the specific amino acid sequence, but reflects favorable polar backbone–backbone interactions due to formation of secondary structure elements and main chain hydrogen bonds. The values Sum<sub>α</sub><sup>phil</sup> and φ<sub>phil</sub> for the set of misfolded proteins are only slightly lower than for the representative set, but this result is expected as the misfolded proteins conserve the secondary structure of their parent models, although it is distorted. The value of favorable hydrophobic interactions between the side chains Sum<sub>α</sub><sup>phob</sup> contributes to the total hydrophobic interactions φ<sub>phob</sub> most of all (Table I). This value also has the strongest discriminating power between the two sets of proteins. For the misfolded set this value is small, which means that misfolded proteins have a deficit of favorable hydrophobic interactions. The contributions of unfavorable interactions do not differ much between these two sets.

The values of parameters φ<sub>phob</sub>, φ<sub>phil</sub>, α<sub>phil</sub> and Σ, calculated for the individual models from the two test sets are shown in Figure 2A, B, C and D, respectively. It is interesting to analyse which proteins from the representative set the parameters φ<sub>phob</sub>, φ<sub>phil</sub> and Σ-score differ significantly from the average values. Three proteins have very low φ<sub>phob</sub> (see Figure 2A). In all three cases the additional interactions, other than non-covalent interresidue ones, strongly contribute to the structure stabilization. Metallothionein (4mt2, 61 residues) forms complexes with four ions of Cd<sup>2+</sup>, two ions of Zn<sup>2+</sup> and one Na<sup>+</sup> ion. Cytochrome C-3 (2cdv, 107 residues) has four hemes. Lectin (2cwg, 171 residues, 16 disulfide bridges) is the glicoprotein. The values φ<sub>phil</sub> are considerably higher than average [see Figure 2B] for streptavidin (1sri), s-lectin (1slt) and porin (2por, membrane protein). These three proteins consist mainly of β-structure, and hydrophilic backbone–backbone contacts contribute significantly to the φ<sub>phil</sub>.

The majority of misfolded and native protein models can be distinguished from each other basing on the similarity score Σ (see Figure 2 D): the values of Σ<sub>≤</sub>350 are typical for misfolded proteins. However, there are few exceptions. Seven natural proteins (cytotoxin 1tgx, metallothionein 4mt2, β-subunit of hydrolase 2kau, DNA-binding regulatory protein 1cm1, ferredoxin 2fd2, and two cytochromes 2cdv and 1ccr) have the Σ-scores less than 350, which are typical for misfolded models. All these proteins are stabilized by additional interactions with ions, hemes, by disulfide bridges, or by the interaction with other subunits. Two misfolded proteins have Σ-scores higher than 350 (models 1lh1lon2i1b and 2ts1lon2tmn, these names include PDB codes of two parent proteins, see Holm and Sander, 1992). The comparison of the Σ-scores for the misfolded protein models with those calculated for their parent experimental structures shows (see Figure 3) that the values of Σ are lower for the misfolded models. That means that although the two misfolded proteins have exceptionally high Σ-scores, their parent structures have larger scores, and hence are ‘more correct’ than the corresponded misfolded child structures. The values of Σ for pancreatic hormone (1ppt) and high-potential iron protein (1hip) are very low and typical for misfolded proteins, and are even lower than for child misfolded models constructed from them (indicated above diagonal on Figure 3). The possible reason is the complex formation with Zn<sup>2+</sup> ion of 1ppt, and the presence of iron-sulfur cluster in 1hip.

In general, the value of similarity score Σ calculated for a spatial model can be used to assess its quality: if this value is considerably lower than the average one, calculated for the representative set (see Table I), the model is probably incorrect, unless it has some additional stabilizing features as heme, disulfide bridges, complexed ions, etc. Similar properties of various energy and scoring functions used for discrimination of protein models were mentioned previously (Casari and Sippl, 1992; Miyazawa and Jernigan, 1996; Huang et al., 1995; Bahar and Jernigan, 1997). The values of parameters φ<sub>phob</sub>, φ<sub>phil</sub> and α<sub>phil</sub> can be utilized additionally to understand better the ‘source’ of problems with the current model, as they correspond to different types of intramolecular interactions.

**Estimation of confidence interval of Σ-score calculation**

Bearing in mind that proteins are not rigid, but possess some flexibility, or in other words, they can have several ‘correct’ conformations, the question arises concerning the variability of Σ-score for these different conformations of the same protein. To get the answer Σ-scores were calculated for the pairs of molecules in the same crystal, related by a non-crystallographic symmetry (as they slightly differ due to crystal packing), and for the sets of models derived from NMR analysis. The values of Σ-scores calculated for the same proteins (26 pairs of X-ray models) in various crystal environments correlate well (correlation coefficient 0.96, while the pairwise r.m.s.d. of C-α atom positions is from 0.1 to 1.77 Å, with the mean value 0.48 Å). The mean standard deviation of Σ in this set σ=11.6. Thus the variability of Σ for high-resolution X-ray structure can be estimated as ±3σ and is equal to ±34.8. This value can be considered as the estimation of Σ-score confidence interval for the individual protein structure revealed by X-ray analysis. The mean standard deviation of Σ-scores calculated for different NMR models of the same protein is much higher (51.8), which reflect higher variability of structures elucidated by this method.
Recognizing distorted structures by similarity score

Analysis of distorted protein structures

The successful potential function (energy or score) should discriminate between the native and near-native structures even if they are very close. In the previous studies the closest to native folds were generated by MD simulations at 298K (Huang et al., 1996; Park et al., 1997), with the average r.m.s.d. (calculated for Cα atoms) of 1.5 Å, and that folds were successfully distinguished from native ones by the score which used a reduced representation for both sequence and structure. The question arises: Is it possible to distinguish between the correct structure and its alternative, if r.m.s.d. is less than 1.5 Å? To monitor the performance of the scoring functions (or energies) in the broad range of r.m.s.d. values we constructed challenging test sets of protein models. For seven high-resolution (better than 1.33 Å) experimental X-ray structures we prepared the sets of distorted structures with the r.m.s.d. (calculated for Cα atoms) from the experimental structure ranging from ca. 0.03 to ca. 20 Å. Small distortions of the experimental models were introduced by conventional unconstrained energy minimization in vacuo and then subsequently larger distortions by MD simulations (see Materials and methods). In the analysis of obtained sets we assumed that: 1) the starting experimental structure is very close to the ideal ‘correct’ structure; 2) the more the model differs from the starting structure (as judged by r.m.s.d. values), the less correct is it. Obviously, the experimental structure differs from the ideal ‘correct’ structure due to experimental errors, usage of non-ideal energy potentials for structure refinement, effect of crystal packing, etc. Although the ideal ‘correct’ structure is unknown, normally it should be in the closest vicinity to the experimental one. So it seems reasonable to assume that some slightly distorted structures can be even ‘better’ by chance than the experimental ones, but the probability to find ‘better’ structure decreases quickly with the increasing r.m.s.d. from the experimental structure.

The dependencies of r.m.s.d. versus structure number in each set of distorted structures (Figure 4A) are very smooth, and are very similar for different sets. Figure 4B displays the dependencies of gyration radii on the structure number. It can be seen that on the initial stage of distortion (corresponding approximately to the first 20 structures which were obtained by energy minimization and MD at low temperatures) the radii of gyration do not increase, showing that the change of protein conformation is due to slight structural rearrangements, but not to the expansion of the protein.

Thus prepared sets of distorted protein models were used to check the ability of the proposed assumption-based similarity score to discriminate between the experimental models and...
alternative models. Figure 5A shows that the sums of favorable hydrophobic interactions $S_{\text{phob}}$ generally decrease with the increase of the r.m.s.d. from the experimental structure, although there are some fluctuations on these dependencies. Similar tendencies can be seen for the sums of hydrophilic interactions $S_{\text{phil}}$ (Figure 5B). Still for several structures with r.m.s.d. up to 1.5 Å one can see small increase of favorable hydrophobic interactions (by the expense of decreasing favorable hydrophilic interactions and increasing of unfavorable ones). Thus the parameter $S_{\text{phob}}$ alone can not distinguish between the experimental model and close alternatives. Unlike the favorable hydrophilic and hydrophobic interactions, the unfavorable hydrophobic-to-hydrophilic interactions $S_{\text{unf}}$ (Figure 5C) increase in the absolute value with the increase of r.m.s.d. up to ca. 5 Å, and then tend to decrease in the absolute value. That means that upon the small distortions of the protein part of favorable hydrophobic and hydrophilic interactions are turned into unfavorable ones. Although the parameter $S_{\text{unf}}$ seems to be quite ‘useless’ for distinguishing the experimental structures from deliberately misfolded ones (see Figure 2C), this parameter is useful for detecting the small distortions. The decrease of absolute values of parameters $S_{\text{phob}}, S_{\text{phil}}$, and $S_{\text{unf}}$ with the r.m.s.d. greater 5 Å can be partly attributed to the increase of gyration radii and expanding of the protein structure, with a subsequent weakening of interresidue interactions. Although all the parameters $S_{\text{phob}}, S_{\text{phil}}$, and $S_{\text{unf}}$ manifest some fluctuations, the similarity scores $\Sigma$ which are their sums have rather smooth dependencies for the r.m.s.d. less than 4 Å (Figure 5D). For the largely distorted structures, fluctuations occur with the appearing ‘by chance’ close interresidue contacts.

The unique feature of $\Sigma$-score is that it is dropping considerably even for the very small deviations from the experimental structure, when only unconstrained energy minimization in vacuo distorts the structure. As the gyration radii do not increase for these models, the decrease of $\Sigma$-score can be attributed to the loss of similarity of intramolecular packing rather than to the expansion of the protein. To make an additional check, we calculated the values of $\Sigma$ taking all the atomic hydrophobicities $f_k$ = 1 (see equation (1) in Materials and methods), therefore making $\Sigma_1$ to be sensitive only to the expansion of the protein, but not to the ‘similarity’ of the packing. The value of the parameter $\Sigma_1$ still decreases with the increasing of r.m.s.d. from the experimental structure (data not shown). That means that decrease in $\Sigma$-scores upon deviation from the ‘correct’ structures are caused also by the loss of packing similarity, and not only by the expansion of the protein.

It is also worth to mention that for some proteins in the test sets the first few steps of minimization made structures with even better $\Sigma$-scores than for the experimental ones, while the corresponded r.m.s.d. values (for Cα atoms) were negligibly small (less 0.1 Å). None of structures with r.m.s.d. larger than ca. 0.1 Å have better scores. This agrees with the idea that it is possible to find ‘more correct’ structure (which will be closer to ‘ideal’ than the experimental one) in the close vicinity to the experimental structure.

It is interesting to compare the performance of $\Sigma$-score for the sets of distorted proteins with that of other commonly used scores which are able to recognize the native structure. Figure

Fig. 3. Discrimination between deliberately misfolded models and their correctly folded parent models with similarity score $\Sigma$. The correlation between the values of $\Sigma_{\text{parent}}$ calculated for the ‘parent’ models and corresponded values of $\Sigma_{\text{misp}}$ calculated for the misfolded models is shown. The values of parameters $\Sigma$ are given in arbitrary units. Points marked with arrows are discussed in the text.
Recognizing distorted structures by similarity score

Fig. 5. Dependencies of parameters $S_{phob}$ (A), $S_{phil}$ (B), $S_{unf}$ (C) and $\Sigma$ (D) on the C$^\alpha$-atom r.m.s.d. from the experimental structures for seven sets of distorted protein models. The values of parameters are given in arbitrary units.

Fig. 6. Dependencies of S-scores of 3D-1D profile method (Lüthy et al., 1992) on the C$^\alpha$-atom r.m.s.d. from the experimental structures for seven sets of distorted protein models.

6 illustrates the dependencies of S-score of 3D-1D profile method (Lüthy et al., 1992), shown to be very useful in protein threading, on the r.m.s.d. of C$^\alpha$ atoms from the experimental structures for the same sets of proteins. In the majority of cases the values of S-scores are not sensitive to small distortions (r.m.s.d. less than 2.5 Å) of protein structure (see Figure 6). S-scores decrease considerably only with r.m.s.d. greater 3 Å, showing that the structures become ‘less correct’. These dependencies differ from those for $\Sigma$-scores, which are very sensitive to the small deviations from the experimental structure, and are rather insensitive to the deviations greater 5 Å (see Figure 5D). This means that both types of scores work well in different r.m.s.d. ranges, and proposed $\Sigma$-score probably will be inappropriate for crude threading, but rather can be useful for the ‘fine tuning’ of the structure in the close vicinity to the ‘correct’ solution.

The process of distorting protein structures utilized in the current work at least at the initial stage resembles MD simulations of protein unfolding (Duggett and Levitt, 1993; Tirado-Rives and Jorgensen, 1993; Hünenberger et al., 1995; Caflisch and Karplus, 1995). The demonstrated sensitivity of $\Sigma$-score and its constituents $S_{phob}$, $S_{phil}$ and $S_{unf}$ to the variation of the structure and the clear physical meanings of these parameters can be used for the analysis of folding/unfolding pathways revealed during such simulations, and for the characterization of various intermediates. The value of $S_{phob}$ is the measure of hydrophobic contacts, $S_{phil}$ reflects the content of secondary structure, mainly contributing to favorable hydrophilic interactions between the backbone segments of amino acid residues, and $S_{unf}$ reflects formation of unfavorable hydrophobic-to-hydrophilic contacts, e.g., due to appearance of unsatisfied hydrogen bond donors or acceptors. We also infer
that the transition of the proteins to the molten globules state (Ptitsyn, 1992), which is characterized by compactness, conservation of large amounts of secondary structure and nonspecificity of tertiary contacts can be successfully monitored by these parameters.

Another test was made with the representative set of high-resolution protein models, which were supposed to be very close to ‘correct’. For this set of structures several runs of unconstrained energy minimization in vacuo with different number of steps were made, using two different force fields (TRIPOS and CHARMM) without an electrostatic energy term. Each time the respective all-heavy-atom r.m.s.d. from the experimental structure was calculated, as well as the difference between similarity scores of minimized and experimental structures \( \Sigma_{\text{min}} - \Sigma_{\exp} \). The results of these calculations are presented on Figure 7. Although in most cases when the minimization caused the increase of similarity score the difference \( \Sigma_{\text{min}} - \Sigma_{\exp} \) falls into the estimated confidence interval (calculated for this difference as \( \pm 34.2 \sqrt{2} \pm 49.2 \)), we cannot exclude that some minimized structures are ‘more correct’ than the experimental ones. For the majority of analysed models, the deviation from the experimental structure (heavy atom r.m.s.d. from 0 to 1.0 Å) caused the decrease of similarity score \( \Sigma \). That means that, in the majority of cases, for the representative set of proteins unconstrained energy minimization in vacuo distorts the structures and leads to the noticeable loss of similarity of intramolecular packing: the overall number and strength of favorable hydrophobic and hydrophilic contacts decreased, and the number and strength of unfavorable hydrophobic-to-hydrophilic contacts increased. This finding cannot be attributed to the properties of the specific force field used for minimization, as two different force fields (TRIPOS and CHARMM) demonstrate similar results. The parameter \( \Sigma \) appeared to be very sensitive even to very small distortions (heavy atom r.m.s.d. less than 1 Å) of the structures relative to the experimental ‘correct’ ones.

Analysis of decoy structures obtained by MD simulations

The ability of \( \Sigma \)-score to discriminate the ‘correct’ structures from decoys generated by MD at 300 K was demonstrated as previously proposed by Huang et al. (1996). We used MD simulations of NMR structure of barstar and X-ray structure of scorpion toxin II in water. To minimize the effect of the differences in computational protocols and force field parameters used for refinement of experimental structures and that used during current MD simulations we took as reference the experimental structures after additional energy minimization in solvent (see Materials and methods). These reference structures were assumed to be ‘native’. The dependencies of \( \Sigma \)-scores on the heavy atom r.m.s.d. from the reference structures for barstar (Figure 8A) and scorpion toxin II (Figure 8B) reveal that the reference structures have nearly the greatest \( \Sigma \)-scores, and only a few alternative models have slightly larger similarity score values. The similarity scores calculated for the structures after energy minimization in vacuo are considerably smaller, although the models are very similar (heavy atom r.m.s.d. \(~1\) Å). The alternative structures obtained during the heating stage and subsequent MD simulation reveal a clear correlation between r.m.s.d. and \( \Sigma \)-score (Figure 8). This funnel-like distribution that has a wide dispersion of scores for conformations far from the correct structure and approaches a linear relationship between the score and r.m.s.d. as the structure approaches to correct one is very similar to...
that predicted earlier for 'good' energy function (Park and Levitt, 1996). Figures 5 and 7 reveal similar tendencies in distributions of Σ-scores versus r.m.s.d. for different test cases. All these data demonstrate the remarkable sensitivity of the similarity score to very small deviations of the structure from the correct one.

The advantage of the current approach is that despite using an all-atom representation (including hydrogen atoms) and large diversification of atomic hydrophobicity parameters (f_k values for 29 types of atoms), the computational algorithm is still very simple and fast. Such 'high resolution' representation of the structure permits one to detect the small changes in side-chain orientations, even if they do not result in a conformational change of the backbone. This is different from many other approaches using simplified protein representations (Casari and Sippl, 1992; Huang et al., 1995, 1996; Miyazawa and Jernigan, 1996; Park and Levitt, 1996;), and probably this is the reason why the Σ-score is so sensitive to very small changes in protein conformation.

What can be wrong with the similarity score?

Although the proposed similarity score can easily penalize the expansion of the protein (directly, due to exponential dependencies on interresidue distances, and indirectly, due to mutual cancellation of hydrophobic and hydrophilic contributions of various atoms at larger distances), this score cannot penalize contraction of the protein. If the protein model is more compact than the normal one (e.g., obtained via calculation with smaller van der Waals radii), then the higher similarity score will not mean necessarily that this structure is better. Fortunately, such situations can be recognized by other available techniques (like estimation of energy using 'standard' potentials). In most cases the protein models are obtained using molecular modelling (with or without experimental constraints) with the standard parameters for bond lengths, angles and van der Waals radii, and in the vast majority of cases the protein models are not too contracted (otherwise it would cause strong steric bumps). Using the Σ-scores we implicitly assume that the protein structures were obtained by conventional modelling techniques, and thus the models are not too compact. As the similarity score is the 'attracting' term, its incorporation into conventional force fields cannot be straightforward. The similarity score ignores the electrostatic interactions, effect of crystal packing and protein–solvent contacts.

Another problem is the validation of Σ-score performance for a greater number of examples. In the present work we limited our consideration to the high-resolution crystal structures, assuming that they are close to 'ideal correct' structures. As far as the similarity score works close to the current limits of accuracy of experimental structures, additional studies of the influence of structure refinement protocols on the quality of spatial models are needed. Up to now, the most reasonable algorithm of structure refinement (energy minimization in solvent with the experimental restraints) is the most reasonable algorithm of structure refinement (energy minimization in solvent with the experimental restraints) is the most reasonable algorithm of structure refinement (energy minimization in solvent with the experimental restraints) is the most reasonable algorithm of structure refinement (energy minimization in solvent with the experimental restraints).

Concluding remarks

The assumption-based similarity score Σ proposed in the current work is a measure of similarity of intramolecular packing and predominance of favorable contacts over unfavorable ones in the protein models. A model lacking favorable interactions will have a lower score. Unlike other approaches, the current approach takes into account both hydrophobic and hydrophilic contacts (including hydrogen bonds) in the protein in the same fashion and in the same terms, enabling simultaneously to favour formation of hydrophobic contacts and hydrogen bonds (and hence, secondary structure) and to penalize unfavorable hydrophobic-to-hydrophilic contacts (e.g., formation of unsatisfied hydrogen bond donors or acceptors). The Σ-score can successfully discriminate between the native and deliberately misfolded (i.e., with completely wrong chain topology), and between the native and distorted (i.e., with the similar chain topology) protein models. Proteins additionally stabilized by a large number of disulfide bridges, prosthetic groups and/or complexed ions have lower similarity score values than the proteins mainly stabilized by hydrophobic interactions. The remarkable feature of Σ-score is its sensitivity to the small distortions (heavy atom r.m.s.d. less than 1 Å) of the experimental (‘correct’) structure.

It should be noted that the current Σ-score is not based on the analysis of any statistical preferences, where some training set is used to obtain parameters, which are further statistically tested with other sets. On the contrary, this approach is based on a simple assumption (similar likes similar), and is tested by statistical analysis. The atomic hydrophobicities used as the quantitative criterion of ‘similarity’ are derived from the experimental octanol–water partition coefficients, which indirectly include all the information about non-covalent interactions, including electrostatics, hydrophobic effect and the change of entropy.

The possible application of the proposed Σ-score can be the quality assessment of the protein spatial models and recognition of the ‘correct’ structure(s) among very close alternatives. As parameters Σ_hydro, Σ_hydrophobic and Σ-solvent are very sensitive measures of the intramolecular contacts, strength and similarity of these parameters can also be useful for characterization of models simulating protein folding/unfolding. The further direction of the work could be also including of Σ-score as the additional term in the conventional force fields for the energy minimization and molecular dynamics simulations.

Acknowledgements

We wish to thank Prof. G.Vergoten from CRESIMM, USTL for making available the computational facilities in his laboratory. A.S.A. is grateful for the Research Fellowship from the Region Nord Pas de Calais (France) for six months position at CRESIMM. The work was supported by grants RFBR 95-04-1264, INTAS-RFBR 95-1068 and 03.0002H-326.

References

A.P. Golovanov et al.


Received April 28, 1998; revised June 22, 1998; accepted September 18, 1998