Revised version
LU TP 02-08
November 21, 2002

Testing Similarity Measures with Continuous and Discrete Protein Models

Stefan Wallin*

Complex Systems Division, Dept. of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Jochen Farwer†

Free University of Berlin, Dept. of Biology, Chemistry and Pharmacy,
Institute of Chemistry, Takustr. 6, D-14195 Berlin, Germany

Ugo Bastolla‡

Centro de Astrobiología (CSIC-INTA),
Ctra. de Ajalvir km. 4, 28850 Torrejón de Ardoz (Madrid), Spain

to appear in *Proteins Struct. Funct. Genet.*

*E-mail:stefan@thep.lu.se
†E-mail:farwer@chemie.fu-berlin.de
‡E-mail:bastollau@inta.es

Abstract

There are many ways to define the distance between two protein structures, thus assessing their similarity. Here, we investigate and compare the properties of five different distance measures, including the standard root-mean-square deviation (cRMSD) and a new one that we propose, called the power distance. The performance of these measures is studied from different perspectives with two different protein models, one continuous and the other discrete. Using the continuous model, we examine the correlation between energy and native distance, and the ability of the different measures to discriminate between the two possible topologies of a three-helix bundle. Using the discrete model, we perform fits to real protein structures by minimizing different distance measures. The properties of the fitted structures are found to depend strongly on the distance measure used and the scale considered. We find that the cRMSD measure effectively describes long-range features but is less effective with short-range features, and it correlates weakly with energy. A stronger correlation with energy and a better description of short-range properties is obtained when we use measures based on intramolecular distances, such as the power distance.

1 Introduction

Protein structures are complex three-dimensional objects, characterized by a large number of linked atoms and a hierarchy of description levels, from primary to quaternary structure. Several distance measures may be defined for the task of comparing protein structures and assessing their similarity. The root-mean-square deviation (cRMSD) is the one most commonly used, but several others are frequently applied in the literature. Here we set up to compare four of the most widely used distance measures, plus a new one that we propose.

Measuring protein similarity is becoming increasingly important in the field of bioinformatics, both for the sake of evaluating protein structure prediction methods [1] and for the sake of classifying proteins and investigating distant evolutionary relationships [2–4]. To this end, several similarity scores have been developed. In a recent review [5], merits and drawbacks of some of these measures are pointed out. Similarity measures have been assessed for their ability to generate robust and accurate clusters in hierarchical classification of proteins [6], with the conclusion that the problem of protein structure comparison does not have a unique answer [7–9]. In this work, we choose another approach and assess protein structure similarity measures by using statistical mechanical models of protein folding.

A similarity measure is appropriate if objects scored as similar share similar properties. In the present context, the objects to be compared are reduced representations of protein structures, and their most important property is their effective energy, which can be thought of as the free energy of the reduced structure as obtained by integrating out degrees of freedom which are not represented in the model (depending on the model, these can be solvent atoms, side chains, etc.). The effective energy depends on solvent properties like temperature, pH, denaturant concentration and so on. In practice, it is impossible to compute the effective energy from first principles, and one has to postulate some energy function in such a way that the low-energy states of the model resemble real protein structures.

According to the thermodynamic view of protein folding [10], the native state of a protein is the state of minimal free energy of the protein plus solvent system. Since the configurational entropy of the protein chain in its native state is very small, it is customary to identify the native state of the protein as the reduced state of minimal effective energy plus its thermal fluctuations. Theoretical considerations based on spin glass theory [11] and comparisons between minimal models of random heteropolymers [12,13] and models of well designed sequences [14–16] have shown that a necessary condition for a simplified model to be a viable model of protein folding is

that the energy landscape is well correlated. Qualitatively, this means that structures very different from the ground state should have high effective energy, so that they have a negligible weight in the thermodynamic ensemble. A well correlated energy landscape is a prerequisite for fast folding [11,14,17,18], thermodynamic stability with respect to changes in the solvent, and stability with respect to mutations [19, 20] of the model protein. In assessing the shape of an energy landscape, it is important that the effective energy function and the similarity measure used are well correlated.

There are some reasons why one can expect that the cRMSD measure does not correlate very well with the energy (see also Ref. [21]). The cRMSD compares atoms at the same position in two chains. All positions have the same weight in this comparison, but not all positions contribute in the same way to the effective energy. Residues in a loop can sometimes be moved without changing the effective energy much, thus generating structures with a high cRMSD from each other but very similar energies. By contrast, even a small displacement of a residue in the hydrophobic core of the native state is likely to produce an atomic collision and thereby a drastic increase in energy. Thus we can find structures with high similarity but very different energies.

A natural way to try to increase the correlation with energy is to replace the cRMSD by a distance measure based on intramolecular atomic distances. In fact, most energy terms are functions of atom-atom distances. It seems reasonable to expect that such a measure can provide a stronger energy correlation, especially if atom pairs at short distances are given a higher weight.

To quantitatively investigate this issue, we perform two kinds of tests. In the first one we evaluate how different distance measures correlate with the effective energy function of a continuous model for protein folding [22–24]. For this purpose we use configurations generated through Monte Carlo sampling at the folding temperature. At this temperature, folded as well as unfolded structures exist, which makes it possible to evaluate the correlation between energy and distance over a wide range of distances.

In the second test we investigate discrete protein models. These models maintain the simplicity of lattice models in that they have a finite state space, and can, at the same time, reproduce native structures of real proteins very closely, as showed by Park and Levitt [25]. Following their approach, we work with a set of six directions; each residue along the chain selects one of these directions. We find that structures obtained this way can fit all protein structures in a very large database with an average cRMSD of less than 1.6 Å. However, a high similarity in terms of cRMSD does not mean that the fitted structures are similar to the real ones in all respects. In

particular, we find that the distribution of C_α-C_α distances is quite different for fitted and real structures, respectively. The discrete structures that are most similar to the native ones in terms of cRMSD tend to exhibit a large number of atomic collisions, which are of course prohibited in real proteins. Such unphysical collisions can be avoided by using, for example, the power distance introduced in this work.

The third issue that we address is the ability of different distance measures to distinguish structures that are locally similar but globally different. We address this issue using the continuous protein model, which possesses two states that are very similar in terms of energy, entropy and secondary structure, but different in overall topology. In this test, the cRMSD outperforms distance measures based on intramolecular distances which, in the presence of thermal noise, have difficulties in discriminating between the two topologies.

On the other hand, when it comes to energy correlations, our analysis shows that three of the four distance measures that we compare to the cRMSD are indeed better than this measure. In particular, this holds for a contact-based distance measure. Hence, the choice of distance measure depends on the type of problem addressed. Some distance measures perform rather poorly for certain tasks, and it is advisable not to use them in such situations.

The paper is organized as follows. In the next section we introduce the distance measures studied and the continuous and discrete models used to test them. In section 3 the results of the tests are presented. The paper is closed with a short discussion and summary of the results.

2 Materials and Methods

2.1 Distance Measures

From a mathematical point of view, a distance is a function $D(a, b)$ associating to any pair of elements, a and b, of a metric space a non-negative real number. It has the distinctive properties: (1) symmetry, $D(a, b) = D(b, a)$; (2) positivity, $D(a, b) \geq 0$, where equality holds if and only if a and b are identical; and (3) the triangular inequality, that is $D(a, b) \leq D(a, c) + D(b, c)$ for any c.

In this work, we represent a structure Γ as an ordered set of coordinates of N atoms,

$\Gamma = \{\mathbf{r}_1 \cdots \mathbf{r}_N\}$. Each amino acid will be represented either by a single atom, C_α or C_β, or by the three backbone atoms N, C_α and C'. If the two structures to be compared would represent different proteins, one would need a criterion to define an alignment, i.e. a correspondence between the residues of the first protein and those in the other one. Alignment methods aim at putting in correspondence evolutionarily related amino acids (sequence alignment) or structurally related amino acids (structural alignment). In what follows, we assume that the two proteins have already been aligned so that there is a one-to-one correspondence between the atoms of the two structures Γ.

There are several ways to define distance measures in the $3N$-dimensional space of all possible structures Γ. We consider five of them, which are listed below.

2.2 Root-mean-square deviation (cRMSD)

The most natural and most frequently used distance measure is the coordinate root-mean-square deviation (cRMSD), which is the ordinary Euclidean distance in $3N$-dimensional space,

$$D_{\mathrm{cRMSD}}(\Gamma^a, \Gamma^b) = \min \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i^a - \mathbf{r}_i^b)^2} \,. \tag{1}$$

Structures related through rigid rotations or translations of all the N atoms are considered equivalent. The cRMSD between two equivalence classes Γ^a and Γ^b is defined as the minimal cRMSD with respect to all possible translations and rotations of the two structures. This minimization can be performed analytically [26, 27].

It has been noted that the cRMSD tends to increase with the number of atoms, N. Maiorov and Crippen proposed a normalization meant to make the cRMSD effectively independent of N [28]. This correction is important if one compares distances corresponding to different N. However, in the present paper we deal with fixed values of N and will adopt the usual definition, equation (1).

The cRMSD measure gives the same weight to all atoms, whatever their structural role. As discussed in the Introduction, this typically leads to a poor correlation with energy, which makes it interesting to look at alternative distance measures.

2.3 Distance RMSD (dRMSD)

A protein consisting of N atoms has a set of $N(N-1)/2$ (not independent) internal atomic distances r_{ij}. Using distance measures based on the r_{ij}'s is appealing partly because the effective energy usually is a function of these distances. Another advantage is that the minimization needed to obtain the cRMSD can be avoided.

The simplest way to construct a distance measure based on the r_{ij}'s is to use a Euclidean distance. This gives the so-called dRMSD measure [29–31], defined as

$$D_{\mathrm{dRMSD}}(\Gamma^a, \Gamma^b) = \sqrt{\frac{1}{N_{\mathrm{pair}}} \sum_{i<j} \left(r_{ij}^a - r_{ij}^b\right)^2}, \tag{2}$$

where N_{pair} is the number of distances compared. $D_{\mathrm{dRMSD}}(\Gamma^a, \Gamma^b)$ is the root-mean-square deviation between two sets of distances, r_{ij}^a and r_{ij}^b.

An unwanted feature of $D_{\mathrm{dRMSD}}(\Gamma^a, \Gamma^b)$ is that the pairs of atoms contributing most are those with largest r_{ij}, which are those contributing least to the energy. As a result, one may expect the correlation with energy to be even worse than for cRMSD.

2.4 Contact distance

A simple way to remove this unwanted feature is to introduce the binary quantities C_{ij} which, for a given structure with atomic distances r_{ij}, are defined as

$$C_{ij} = \begin{cases} 1 & \text{if } r_{ij} < r_c \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where r_c is a cutoff distance. The matrix C is called the contact map and represents an equivalence class of structures whose configurational entropy is extensive (of order N), as has been seen using lattice models to represent protein structures [32, 33].

Despite the gross simplification implied by equation (3), models based on the contact map have been intensively studied in the last 15 years (for a recent review, see Ref. [34]). Such models have an energy function of the form $E(\Gamma, \{B_{ij}\})/k_BT = \sum_{ij} C_{ij}(\Gamma) B_{ij}$, where the quantities B_{ij} are effective contact interactions in units of k_BT. For these models, the natural measure of similarity between two structures Γ^a

and Γ^b is their contact overlap, defined as

$$q(\Gamma^a, \Gamma^b) = \frac{\sum_{i<j} C^a_{ij} C^b_{ij}}{\max\left(\sum_{i<j} C^a_{ij}\,, \sum_{i<j} C^b_{ij}\right)}\,, \tag{4}$$

where C^a and C^b are the contact maps of the structures Γ^a and Γ^b, respectively. The numerator is the number of common contacts in the two structures and the denominator is chosen such that the overlap takes the maximum value $q = 1$ if and only if the two contact maps coincide. Alternatively, we could use a denominator of the form $\sqrt{\sum_{i<j} C^a_{ij} \sum_{i<j} C^b_{ij}}$, which is qualitatively very similar and has the advantage that the contact overlap could be interpreted as the "cosine" between the two contact maps. A contact-based distance measure is most easily defined as

$$D_{\text{cont}}(\Gamma^a, \Gamma^b) = 1 - q(\Gamma^a, \Gamma^b)\,. \tag{5}$$

We prove in the Appendix that this distance satisfies the triangular inequality.

2.5 The Holm and Sander score

In the context of structure alignment, Holm and Sander [35,36] proposed a similarity score based on internal atomic distances, which does not have the drawback of the dRMSD of strongly weighting large values of r_{ij}, and which, in contrast to the contact distance, does not require a discretization. The corresponding distance measure can be defined as

$$D^*_{\text{HS}}(\Gamma^a, \Gamma^b) = \sum_{i<j} \frac{|r^a_{ij} - r^b_{ij}|}{r^a_{ij} + r^b_{ij}}\, \mathrm{e}^{-(r^a_{ij}+r^b_{ij})^2/4r_0^2}\,. \tag{6}$$

Only features on scales of order r_0 and smaller contribute significantly to this score (in the original paper, the value $r_0 = 20\,\text{Å}$ was chosen). The exponential weight function is qualitatively similar to a softening of the condition (3). For the task of comparing protein structures, we found it more convenient to use a normalized version of the Holm and Sander score, given by

$$D_{\text{HS}}(\Gamma^a, \Gamma^b) = \frac{D^*_{\text{HS}}(\Gamma^a, \Gamma^b)}{\sum_{i<j} \mathrm{e}^{-(r^a_{ij}+r^b_{ij})^2/4r_0^2}}\,. \tag{7}$$

This score is restricted to the interval $[0, 1]$ regardless of protein length and has the advantage of being less sensitive to details than the original one.

Neither the Holm and Sander score nor its normalized version is a real distance, since they do not strictly satisfy the triangular inequality (see Appendix). However, an advantage of the normalized version D_{HS} is that violations of the triangular inequality seem to become extremely rare. In fact, no violation was found for D_{HS} for the set of protein-like structures obtained with the continuous protein model. For D^*_{HS}, violations were observed for values of r_0 less than 11 Å, but not for the original value $r_0 = 20$ Å used in Ref. [35].

2.6 Power distance

Finally, we propose another distance measure that gives a large weight to pairs of atoms with small r_{ij}. This measure is defined as

$$D^{(0)}_{\mathrm{pow}}(\Gamma^a, \Gamma^b) = \sum_{i<j} \left| (r^a_{ij})^{-m} - (r^b_{ij})^{-m} \right| , \tag{8}$$

where m is a parameter that effectively controls the relative weight of small distances r_{ij}. If m is too small, the correlation between energy and distance to the native state becomes weak. If, on the other hand, m is too large, then local features are weighted too strongly, so that structures that are similar in terms of D_{pow} can have rather different global features. We tested different integer values of m, finding the best overall results for m equal to 2 or 3.

The function $D^{(0)}_{\mathrm{pow}}$ satisfies all properties of a distance, but it is expressed in bizarre units ($\mathrm{Å}^{-m}$), and it depends very strongly on the length and compactness of the structures compared. These problems can be alleviated by normalizing the measure so that it takes values between 0 and 1. One possible normalization is given by

$$D^{(1)}_{\mathrm{pow}}(\Gamma^a, \Gamma^b) = \frac{1}{N_{\mathrm{pair}}} \sum_{i<j} \frac{\left| (r^a_{ij})^{-m} - (r^b_{ij})^{-m} \right|}{(r^a_{ij})^{-m} + (r^b_{ij})^{-m}} , \tag{9}$$

where N_{pair} is the number of distances compared. The function $D^{(1)}_{\mathrm{pow}}$ coincides with the Holm and Sander score if $m = -1$ and $r_0 = \infty$ in equation (7). Unlike the general Holm and Sander score, $D^{(1)}_{\mathrm{pow}}$ fulfills the triangular inequality (see Appendix). The definition (9) has the drawback, however, that it is not necessarily true that small r_{ij}'s are given a higher weight. This can be seen by rescaling r^a_{ij} and r^b_{ij} for one particular pair ij by a common factor λ. Then, as $\lambda \to 0$, the term ij becomes dominant in equation (8), whereas it remains unchanged in equation (9).

An alternative normalization is to use

$$D_{\mathrm{pow}}(\Gamma^a, \Gamma^b) = \frac{\sum_{i<j} \left| (r_{ij}^a)^{-m} - (r_{ij}^b)^{-m} \right|}{\sum_{i<j} \left[(r_{ij}^a)^{-m} + (r_{ij}^b)^{-m} \right]} \,. \tag{10}$$

This distance measure does not strictly fulfill the triangular inequality. However, we found very few violations of this inequality for randomly generated coordinates and none for protein-like structures. In the following, we will use D_{pow}, defined by equation (10), as the power distance.

2.7 Sequence cutoff

Giving a large weight to short distances r_{ij}, as some of the discussed distances do, has the side-effect that amino acid pairs kl with short sequence separation $|k-l|$ get a high weight. This is an unwanted property because the relative position of such pairs provide little information about the overall structure of the chain. To overcome this problem, it is useful to define a sequence cutoff s and only consider pairs such that $|k-l| > s$ in the evaluation of the distances. Since in alpha helices amino acids at separation $|k-l| = 4$ form hydrogen bonds, $s = 2$ seems to be a good choice to get rid of local details without losing important energetic contributions.

2.8 Continuous protein model

The first part of our analysis is performed using a continuous protein model, introduced and studied in Refs. [22] and [23]. The model describes a protein with 54 amino acids which are of three different types (hydrophobic, polar and glycine). Each amino acid is represented by the backbone atoms N, C_α, C', H and O, as well as the C_β atom (except for glycine), which is treated as either hydrophobic or polar, depending on the amino acid type. The degrees of freedom are the Ramachandran torsion angles ϕ_i and ψ_i. The energy function is composed of four terms:

$$E = E_{\mathrm{loc}} + E_{\mathrm{sa}} + E_{\mathrm{hb}} + E_{\mathrm{AA}} \,. \tag{11}$$

The local potential E_{loc} has a standard form with 3-fold symmetry,

$$E_{\mathrm{loc}} = \frac{\epsilon_\phi}{2} \sum_i (1 + \cos 3\phi_i) + \frac{\epsilon_\psi}{2} \sum_i (1 + \cos 3\psi_i) \,. \tag{12}$$

The self-avoidance term $E_{\rm sa}$ is given by a hard-sphere potential of the form

$$E_{\rm sa} = \epsilon_{\rm sa} \sum_{i<j}{}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{13}$$

where the sum runs over all possible atom pairs except those consisting of two hydrophobic C_β. The hydrogen-bond term $E_{\rm hb}$ is given by

$$E_{\rm hb} = \epsilon_{\rm hb} \sum_{ij} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \,, \tag{14}$$

where i and j represent H and O atoms respectively, and

$$u(r_{ij}) = 5\Big(\frac{\sigma_{\rm hb}}{r_{ij}}\Big)^{12} - 6\Big(\frac{\sigma_{\rm hb}}{r_{ij}}\Big)^{10} \tag{15}$$

$$v(\alpha_{ij}, \beta_{ij}) = \begin{cases} \cos^2\alpha_{ij} \cos^2\beta_{ij} & \alpha_{ij}, \beta_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

In these equations, r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC$'$ angle. Finally, the hydrophobicity term $E_{\rm AA}$ has the form

$$E_{\rm AA} = \epsilon_{\rm AA} \sum_{i<j} \left[\Big(\frac{\sigma_{\rm AA}}{r_{ij}}\Big)^{12} - 2\Big(\frac{\sigma_{\rm AA}}{r_{ij}}\Big)^{6} \right], \tag{17}$$

where both i and j represent hydrophobic C_β. Further details of the model, including numerical values of all the parameters, can be found in Ref. [22], where the model was introduced and studied with Monte Carlo simulations.

The model exhibits a first-order-like folding transition from an extended phase to the folded one, where the chain forms a three-helix bundle. A three-helix bundle has two possible topologies; if the first two helices form a U, the third helix can go either in front of or behind this U. The model is unable to distinguish between these two possible ways of arranging the helices. Hence, the model exhibits a twofold topological degeneracy.

The thermodynamic behavior of this model protein has been studied in detail before [22] through extensive Monte Carlo simulations at different temperatures. A set of 5000 configurations, separated by at least 10^5 elementary Monte Carlo steps, were recorded at each temperature. In our analysis, we use ensembles of configurations sampled at the folding temperature $T_{\rm f}$, where the folded and unfolded phases coexist, and at a lower temperature $T_{\rm low} = 0.95 T_{\rm f}$, where the chain is folded.

The previous study also determined representative conformations for the two topologies, through a quenching procedure [22]. These two structures, having minimum

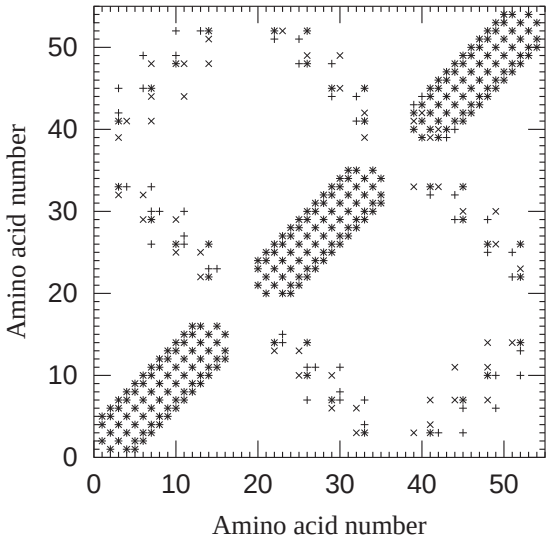

Figure 1: Contact maps of the two minimum-energy conformations BU (+) and FU (×); shared contacts appear as '∗'. Two amino acids with C_β atoms closer than 7 Å are regarded in contact.

energy within their respective topologies, are referred to as FU and BU. In Fig. 1 we compare the contact maps of these two conformations. The contacts can be divided into interhelical and intrahelical ones, the latter ones being located near the diagonal. The two intrahelical contacts sets are found to be essentially the same, as expected. The two interhelical contact sets resemble each other, but many of the contacts are not exactly the same in the two conformations.

In our analysis of the model, all distances except the contact distance are calculated over the backbone atoms N, C_α and C'. In the contact distance, we use the C_β atoms to define contacts.

2.9 Discrete protein models

Continuous models are much more complex to simulate than lattice models where the conformation space is much simpler. Most minimal models of random heteropolymers and designed sequences have been studied on the cubic lattice where each new residue that is added to the chain has at most five possible directions (going backwards is forbidden by hard core repulsion). Although these models have provided valuable qualitative insights into the principles of folding, their application to real proteins is

severely limited. Several groups have then studied models based on more complex lattices (reviewed in Ref. [25]) and discrete off-lattice models where each new residue can choose between a finite number of predetermined dihedral angles [25, 37].

Park and Levitt [25] evaluated the accuracy of discrete models measuring the minimal cRMSD between model structures and protein native states, for a set of 149 proteins. They showed that, for four-state models, it is possible to optimize the set of allowed directions in such a way that the average of the minimal cRMSD is 2.2 Å. Moreover, two six-state models with angles chosen according to the distribution of torsion angles in the PDB by Rooman *et al.* [37] and Park and Levitt [25] can fit the same proteins up to 1.74 Å and 1.90 Å, respectively, on the average. These results suggest that optimized discrete models can reproduce protein structures very accurately, despite their low complexity.

In this work, following Park and Levitt, we optimize sets of six directions. We choose a C_α representation of protein structures, parameterizing directions in terms of the pseudo-bond angle α (the angle formed by three consecutive C_α atoms) and the pseudo-torsion angle τ (the torsion angle formed by four consecutive C_α atoms). Our scoring function is based on distances to a training set of native protein structures. The same analysis is repeated using different distance measures.

2.10 Modified build-up algorithm

The task to determine the discrete structure that minimizes the distance to a given target protein structure is computationally very hard. One approach to this problem is to apply a Monte Carlo algorithm at low temperature, or some form of simulated annealing. However, Park and Levitt [25] showed that a simple deterministic algorithm provides good approximate solutions to the problem.

One such algorithm is the build-up algorithm, where new residues are attached to the growing chain one at a time. Since each residue can be attached in k different configurations, where $k = 6$ for six-state models, the number of possible chain configurations increases exponentially with the number of residues. At every growth stage n, configurations are ranked according to their score and the N_{keep} configurations with the best score are selected. These configurations are then used as building blocks to build the kN_{keep} configurations at stage $n + 1$. The procedure is iterated until the chain is completed. Surprisingly, this very simple algorithm, which is guaranteed to find the lowest energy configuration for $N_{\text{keep}} = k^N$, where N is the chain length,

converges close to the optimal value already at $N_{\rm keep}$ of the order of a few hundreds, apparently independent of N [25].

We found out that the performance of this build-up algorithm can be improved by adding a little bit of randomness. This is accomplished by selecting the $N_{\rm keep}$ conformations at stage n using the scoring function

$$\text{Score} = D(\Gamma^n, \Gamma^{\rm nat}) + T_r \epsilon_n \left(1 - \frac{n}{N}\right) , \tag{18}$$

where $D(\Gamma^n, \Gamma^{\rm nat})$ is the distance between the discrete structure at stage n and the first n residues of the native structure, T_r is a parameter measuring the amount of randomness in the score, and ϵ_n is a random variable uniformly distributed between -1 and 1. For $T_r = 0$, the original, deterministic build-up algorithm is recovered. The factor $1 - n/N$ is chosen such that the random part of the score vanishes for the completed chain, since $D(\Gamma^N, \Gamma^{\rm nat})$ is to be minimized. For moderate but non-vanishing T_r and sizeable $N_{\rm keep}$, the modified algorithm typically finds structures of lower distance than for $T_r = 0$. In fact, for $T_r = 0$, all the selected configurations are rather correlated, and the algorithm explores a relatively small number of directions in a high-dimensional space. After adding some randomness, the algorithm explores a broader spectrum of directions and has a better chance to find low-distance states.

3 Results and Discussion

3.1 Relationships between distances

The first question we address is how the different distance measures are related to each other. Obviously, we expect that they are correlated, but the extent of the correlation will depend on factors such as the relative weight given to local versus global features. We study this question using the conformations obtained by sampling of the continuous system described above at its folding temperature $T_{\rm f}$.

Since the native state of this chain has a twofold degeneracy, we define native distance as

$$D^{\rm n}(\Gamma) \equiv \min\left[D(\text{FU}, \Gamma), D(\text{BU}, \Gamma)\right] , \tag{19}$$

where FU and BU are the representative conformations of the two topologies and D denotes any one of the distances previously introduced. We consider five different native distances: $D^{\rm n}_{\rm cRMSD}$, $D^{\rm n}_{\rm dRMSD}$, $D^{\rm n}_{\rm cont}$, $D^{\rm n}_{\rm HS}$ and $D^{\rm n}_{\rm pow}$. Throughout our analysis of

	$\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$	$D^{\mathrm{n}}_{\mathrm{cont}}$	$D^{\mathrm{n}}_{\mathrm{HS}}$	$D^{\mathrm{n}}_{\mathrm{pow}}$
$\log(D^{\mathrm{n}}_{\mathrm{cRMSD}}+1)$	0.97	0.84	0.87	0.90
$\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$		0.81	0.85	0.89
$D^{\mathrm{n}}_{\mathrm{cont}}$			0.95	0.97
$D^{\mathrm{n}}_{\mathrm{HS}}$				0.98

Table 1: Correlation coefficients between pairs of native distances.

the continuous model, the parameters of $D^{\mathrm{n}}_{\mathrm{cont}}$, $D^{\mathrm{n}}_{\mathrm{HS}}$ and $D^{\mathrm{n}}_{\mathrm{pow}}$ are taken as $r_c = 7\,\text{Å}$, $r_0 = 13\,\text{Å}$ and $m = 3$, respectively. These values were chosen so as to maximize the correlation with energy (see section Energy correlation).

Table 1 shows correlation coefficients between different pairs of native distances. The three measures $D^{\mathrm{n}}_{\mathrm{cont}}$, $D^{\mathrm{n}}_{\mathrm{HS}}$ and $D^{\mathrm{n}}_{\mathrm{pow}}$ are found to be strongly correlated with each other, the correlation being strongest between $D^{\mathrm{n}}_{\mathrm{HS}}$ and $D^{\mathrm{n}}_{\mathrm{pow}}$. Figure 2a shows the correlation between $D^{\mathrm{n}}_{\mathrm{HS}}$ and $D^{\mathrm{n}}_{\mathrm{cont}}$.

These three measures turn out to be approximately exponentially related to the two others, $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ and $D^{\mathrm{n}}_{\mathrm{dRMSD}}$. This is illustrated in Fig. 2b, which shows the $D^{\mathrm{n}}_{\mathrm{pow}}, D^{\mathrm{n}}_{\mathrm{dRMSD}}$ distribution. An exponential relation has been observed previously between $D^{\mathrm{n}}_{\mathrm{cont}}$ and $D^{\mathrm{n}}_{\mathrm{cRMSD}}$, in a study of database structures [38]. A fit of our data to the form $D^{\mathrm{n}}_{\mathrm{cRMSD}} \approx a\exp(b * D^{\mathrm{n}}_{\mathrm{cont}})$ yields an exponent $b = 3.4$. The correlation coefficient is 0.84 between $D^{\mathrm{n}}_{\mathrm{cont}}$ and $\log(D^{\mathrm{n}}_{\mathrm{cRMSD}}+1)$ (see Table 1). We note that the logarithm of a distance can be used to define another distance through the formula $D_{\log}(a,b) = \log((D(a,b)+\epsilon)/\epsilon)$ for any $\epsilon > 0$; it is easy to verify that $D_{\log}$ satisfies the three distance properties if D does so.

The two Euclidean distances $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ and $D^{\mathrm{n}}_{\mathrm{dRMSD}}$ correlate well and linearly with each other. Consistent with a previous study [30] we find that $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ tends to be slightly larger than $D^{\mathrm{n}}_{\mathrm{dRMSD}}$ at low values ($< 10\,\text{Å}$), and that the situation is the opposite at high values.

3.2 Energy correlation

We now turn to the correlation between native distance, defined in equation (19), and energy. Before describing our results, we should remark that they are of course dependent on the specific choice of energy function and its parameters. In partic-

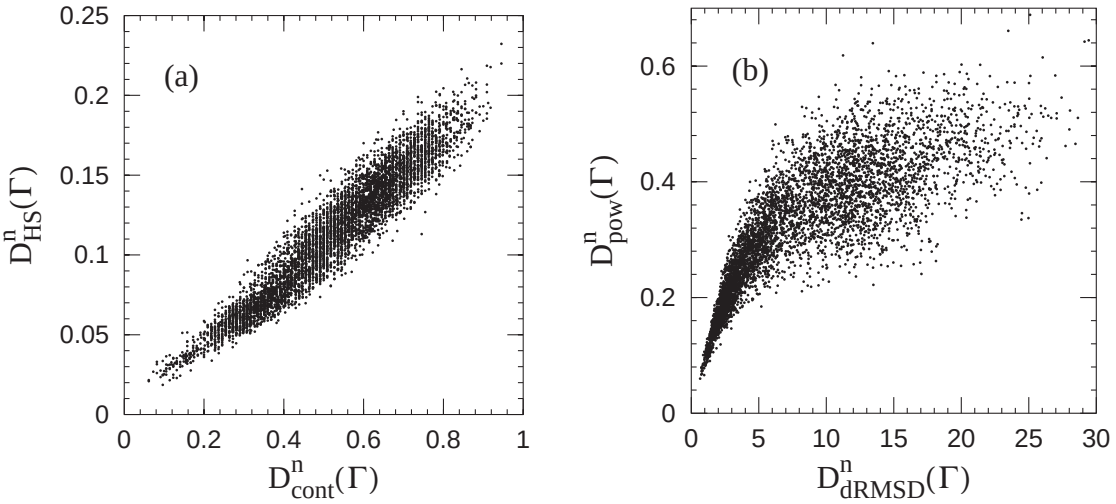

Figure 2: (a) $D^{\rm n}_{\rm HS}$, $D^{\rm n}_{\rm cont}$ and (b) $D^{\rm n}_{\rm pow}$, $D^{\rm n}_{\rm dRMSD}$ scatter plots for conformations at $T_{\rm f}$.

ular, we note that the folding properties of the model used here depend strongly on the hydrogen bond and hydrophobicity strengths, $\epsilon_{\rm hb}$ and $\epsilon_{\rm AA}$, respectively [see equations (14,17)]. For our choice of these parameters, it turns out that folding and collapse occur at the same temperature, and that the transition is first-order-like [22]. Even a moderate change of the relative strength of these parameters leads to a very different behavior. If $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ is made too large, the ground state becomes one long helix rather than a helical bundle. If, on the other hand, $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ is made too small, chain collapse occurs before folding [23]. In these situations, we expect the effective energy gap between the native state and other states to be smaller, which may result in a weaker correlation between native distance and energy.

The main justification for the present form of the model is that it does show two-state folding, as observed for many small proteins. Furthermore, the fact that collapse and helix formation occur simultaneously is in accord with recent experiments [39] on small helical proteins, which found that hydrophobic association and helix formation cannot be separated in the transition state.

Let us first look at the behavior of the cRMSD measure. Figure 3 shows a scatter plot of the energy against the native distance $D^{\rm n}_{\rm cRMSD}$, for conformations taken at the folding temperature $T_{\rm f}$. Low $D^{\rm n}_{\rm cRMSD}$ conformations have a fairly wide range of energies so that, as expected, the correlation is not very strong.

Figure 4 shows scatter plots of the energy against native distance for the other four

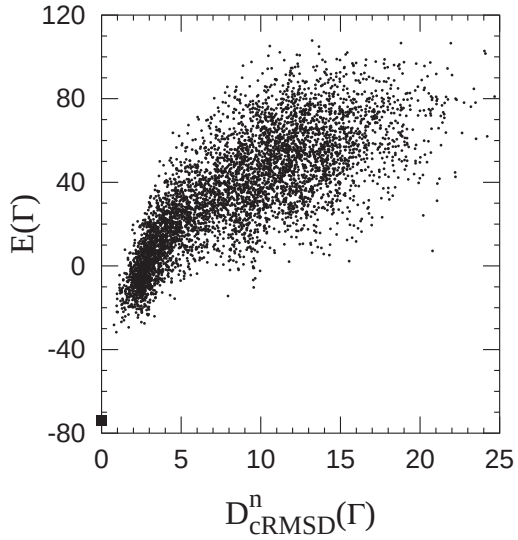

Figure 3: Energy $E(\Gamma)$ (in dimensionless units) against $D^{\rm n}_{\rm cRMSD}$ (in Å), as obtained at the folding temperature $T_{\rm f}$. Also indicated are the representative conformations FU and BU (filled box); both $E(\mathrm{FU})$ and $E(\mathrm{BU})$ are found within the size of the plot symbol.

distance measures. The correlation coefficient R between energy and all the five different distance measures can be found in Table 2. From this table we see that the energy correlation is strongest for $D^{\rm n}_{\rm cont}$ and $D^{\rm n}_{\rm pow}$, strong also for $D^{\rm n}_{\rm HS}$, and significantly weaker for $D^{\rm n}_{\rm cRMSD}$ and $D^{\rm n}_{\rm dRMSD}$. From Table 2 we also see that $\log(D^{\rm n}_{\rm cRMSD}+1)$ and $\log(D^{\rm n}_{\rm dRMSD}+1)$ are more strongly correlated with energy than $D^{\rm n}_{\rm cRMSD}$ and $D^{\rm n}_{\rm dRMSD}$, respectively. This is not surprising since we have seen that $D^{\rm n}_{\rm cRMSD}$ is approximately exponentially related to $D^{\rm n}_{\rm cont}$, and that $D^{\rm n}_{\rm cont}$ is very strongly correlated with energy.

The parameters r_c, r_0 and m for $D^{\rm n}_{\rm cont}$, $D^{\rm n}_{\rm HS}$ and $D^{\rm n}_{\rm pow}$, respectively, are important for the properties of these measures and therefore also for the relationship between energy and distance. The plots in Fig. 4b–d were obtained for the choices $r_c = 7\,$Å, $r_0 = 13\,$Å and $m = 3$, respectively, for which the correlation coefficient R is close to maximal. It turns out, however, that the energy correlation remains strong for fairly wide ranges of these parameters. For example, for the correlation between $D^{\rm n}_{\rm cont}$ and energy we find that $R > 0.90$ if $6\,\text{Å} \le r_c \le 11\,\text{Å}$. The corresponding parameter intervals for $D^{\rm n}_{\rm HS}$ and $D^{\rm n}_{\rm pow}$ are $4\,\text{Å} \le r_0 \le 18\,\text{Å}$ and $2 \le m \le 11$, respectively. Finally, we note that $R = 0.83$ for $D^{\rm n}_{\rm HS}$ and $r_0 = \infty$, and that $R = 0.79$ for $D^{\rm n}_{\rm pow}$ and $m = -1$.

It should be remembered that the above results were obtained for a model whose

Distance	R	parameter
$D^{\mathrm{n}}_{\mathrm{cRMSD}}$	0.78	
$\log(D^{\mathrm{n}}_{\mathrm{cRMSD}}+1)$	0.81	
$D^{\mathrm{n}}_{\mathrm{dRMSD}}$	0.73	
$\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$	0.80	
$D^{\mathrm{n}}_{\mathrm{HS}}$	0.91	$r_0 = 13\,\text{Å}$
$D^{\mathrm{n}}_{\mathrm{pow}}$	0.95	$m = 3$
$D^{\mathrm{n}}_{\mathrm{cont}}$	0.95	$r_c = 7\,\text{Å}$

Table 2: Correlation coefficient R between the energy and different native distances.

native state is a three-helix bundle, and an important question is, of course, to what extent they hold for general proteins. In particular, it would be very interesting to perform the same analysis for proteins with a large beta sheet content. We did repeat the analysis for a 16-amino acid beta hairpin. In this case, the energy correlations were somewhat lower (between 0.73 and 0.78 for all the five distance measures). This is not unexpected because the beta hairpin is smaller and less stable than the three-helix-bundle protein studied. To see whether there is a systematic difference between alpha and beta proteins, it is clear that data for larger beta proteins are needed. Performing such simulations is, however, a very difficult problem. In fact, it takes a lot of effort to develop continuous protein models with suitable folding properties, and it is very hard to obtain models for beta proteins.

3.3 Topology discrimination

An interesting issue is the ability of different distance measures to discriminate structures that are locally similar but globally different. Our three-helix-bundle protein is challenging in this respect, since the FU and BU conformations, in a global sense, are related by an approximate mirror symmetry. It is easy to see that any distance measure based on intramolecular distances is unable to distinguish between a structure and its mirror image. On the other hand, FU and BU are not exact mirror images of each other since there are no left-handed helices, so one may still hope that some of the four measures based on intramolecular distances are able to solve this task.

A relevant parameter for monitoring the ability of a distance measure to discriminate between the two topologies is

$$\Delta(\Gamma) = \frac{D(\mathrm{FU},\Gamma) - D(\mathrm{BU},\Gamma)}{D(\mathrm{FU},\mathrm{BU})}\,. \tag{20}$$

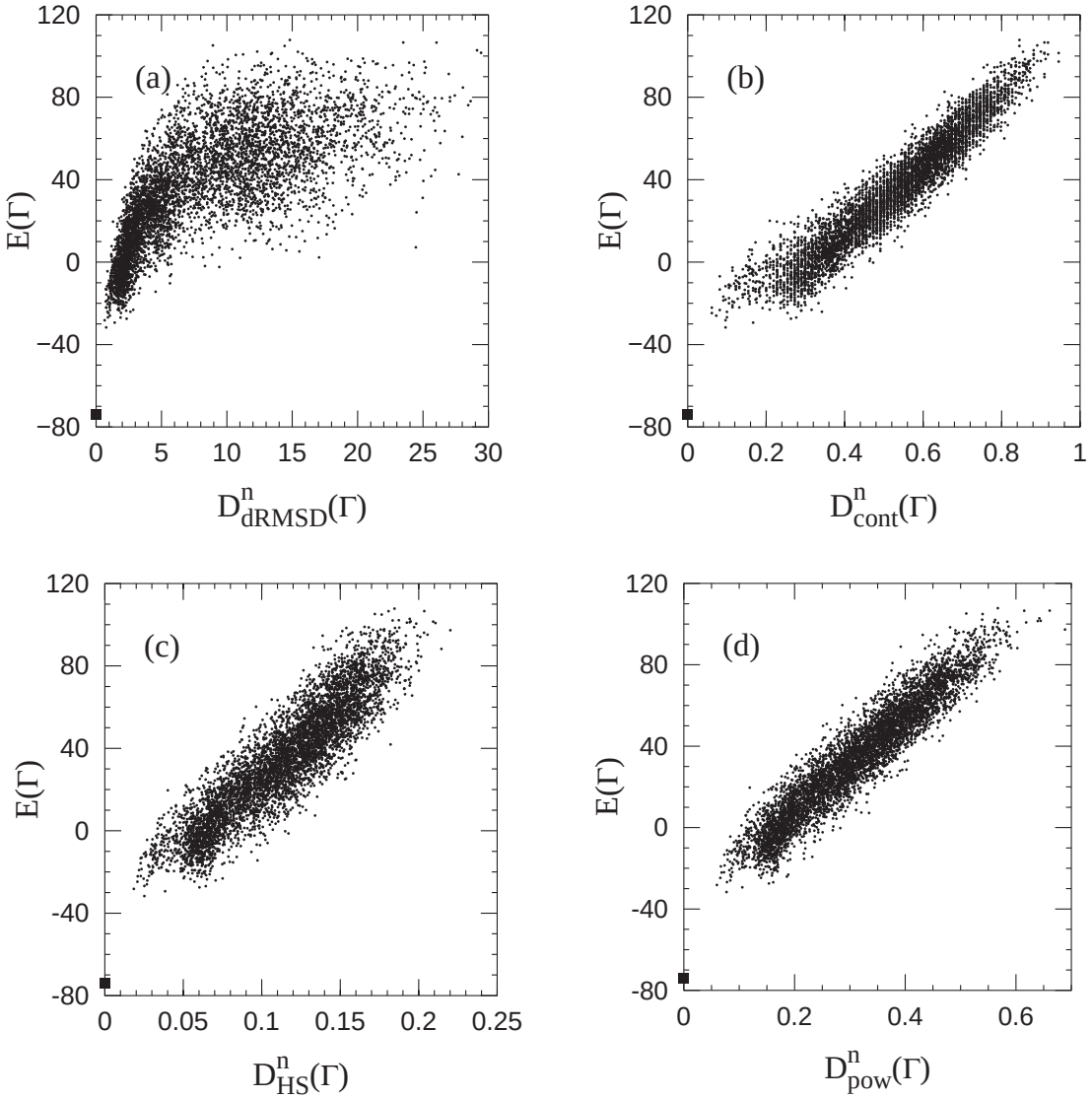

Figure 4: Scatter plots of the energy against different native distances, as measured by (a) $D^{\rm n}_{\rm dRMSD}$ (in Å), (b) $D^{\rm n}_{\rm cont}$ with $r_c = 7\,$Å (see equation (5)), (c) $D^{\rm n}_{\rm HS}$ with $r_0 = 13\,$Å (see equation (7)) and (d) $D^{\rm n}_{\rm pow}$ with $m = 3$ (see equation (10)). The plots are based on conformations taken at the folding temperature $T_{\rm f}$. FU and BU are indicated using a filled box, and native distance is defined through equation (19).

Using the triangular inequality, it is easy to see that $-1 \leq \Delta(\Gamma) \leq 1$, with equality only if Γ coincides with either FU or BU. The behavior of the parameter $\Delta(\Gamma)$ is studied at the temperature $T_{\rm low}$= $0.95T_{\rm f}$, where the unfolded population is very small.

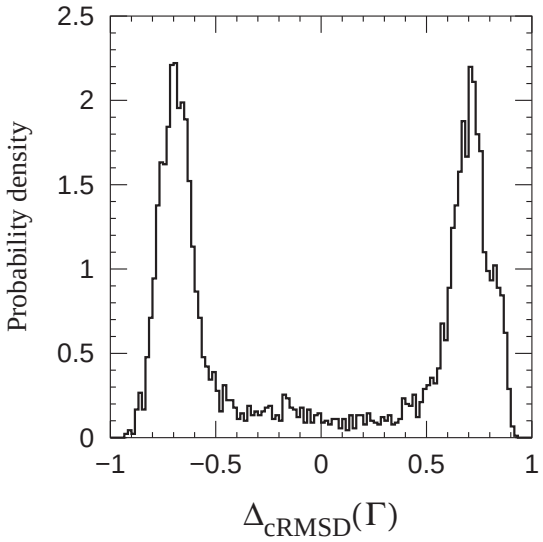

Figure 5: Probability distribution of Δ_{cRMSD} at T_{low}, showing that the cRMSD measure clearly discriminates between the two topologies of the three-helix-bundle protein.

Figure 5 shows the probability distribution of Δ_{cRMSD}. We see that most conformations indeed belong to the basin of attraction of either FU or BU. These two basins of attraction are separated by a large energy barrier, since changing topology requires one of the end helices to cross the U formed by the other two. The fact that the Δ_{cRMSD} distribution has a bimodal shape implies that the cRMSD measure is able to discriminate efficiently between the two topologies.

In the corresponding analysis of the four measures based on intramolecular distances, we make the measures more sensitive to global differences by increasing the sequence cutoff to $s = 4$ (see section 2.7). This is useful because the FU and BU conformations are very similar locally, as we saw in Fig. 1. The D_{dRMSD}, D_{cont}, D_{HS} and D_{pow} measures obtained this way will be denoted by D'_{dRMSD}, D'_{cont}, D'_{HS} and D'_{pow}, respectively.

Figure 6 shows the Δ distributions obtained using the D'_{dRMSD}, D'_{cont}, D'_{HS} and D'_{pow} measures. For D'_{dRMSD}, we see that the distribution is unimodal, in sharp contrast to the clear bimodal shape seen in Fig. 5. As opposed to D_{cRMSD}, the D'_{dRMSD} measure is unable to discriminate between the two topologies of the three-helix bundle. The situation is less clear for D'_{HS}, D'_{pow} and D'_{cont}. These distributions show signs of bimodality, but the separation of the two peaks is much weaker than in Fig. 5.

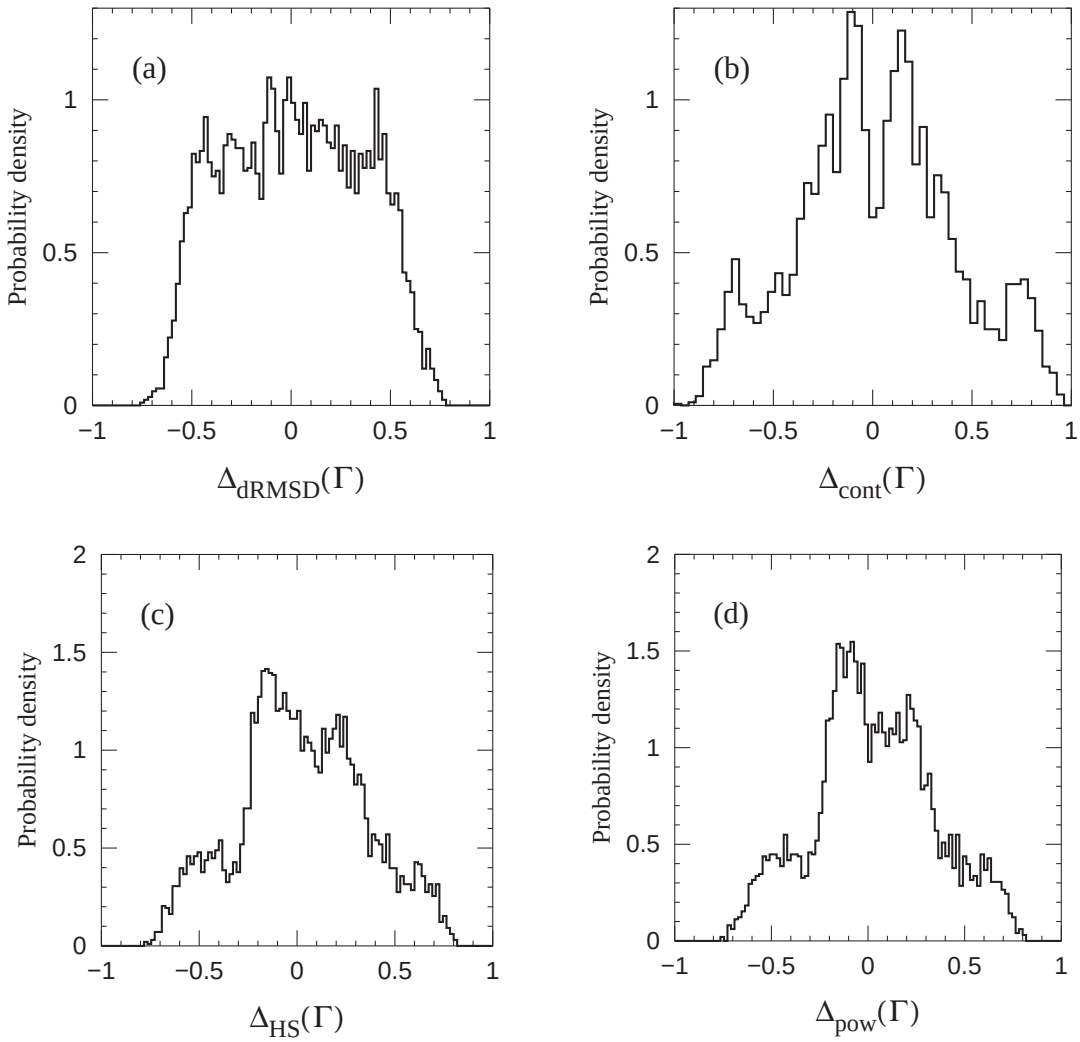

Figure 6: Probability distributions of the Δ parameter (see equation (20)), as constructed using the distance measures (a) D'_{dRMSD}, (b) D'_{cont}, (c) D'_{HS} and (d) D'_{pow}, at T_{low}. The parameters $r_0 = 13\,\text{Å}$, $m = 3$ and $r_c = 7\,\text{Å}$ of D'_{HS}, D'_{pow} and D'_{cont} respectively, are the same as in Fig. 4.

It is interesting to note that the distance between the "ideal" conformations FU and BU, as measured by the contact measure, is quite large. More precisely, one finds that $D'_{\mathrm{cont}}(\mathrm{FU}, \mathrm{BU}) = 0.62$. The difficulty in discriminating between the two topologies arises when thermal fluctuations are added.

To quantify the ability of the different distance measures to discriminate between the two topologies, we introduce three structural classes: (I) close to FU, (II) close

Class	I	II	III
D_{cRMSD}	0.39	0.39	0.07
D'_{dRMSD}	0.00	0.01	0.69
D'_{cont}	0.07	0.09	0.33
D'_{HS}	0.01	0.02	0.62
D'_{pow}	0.01	0.02	0.61

Table 3: Observed frequencies of class I, II and III at T_{low}, as obtained for the different distance measures, with $\kappa = 0.45$ and $\gamma = 1.0$ (see text).

to BU and (III) far from both FU and BU. Given a measure $D(\Gamma^a, \Gamma^b)$, we assign a conformation Γ to class I if

$$D(\mathrm{FU}, \Gamma) < \kappa D(\mathrm{FU}, \mathrm{BU}) \tag{21}$$

where κ is a parameter, and analogously for class II. Choosing $\kappa < 0.5$ assures that the classes I and II are disjoint, as can be seen by using the triangular inequality. Conformations Γ such that $D(*, \Gamma) > \gamma D(\mathrm{FU}, \mathrm{BU})$ for both $*$=FU and $*$=BU, where we choose $\gamma = 1.0$, are assigned to class III. Conformations with native distances in between $\kappa D(\mathrm{FU}, \mathrm{BU})$ and $\gamma D(\mathrm{FU}, \mathrm{BU})$ are not assigned to any class.

Table 3 shows the result of this classification for D_{cRMSD}, D'_{HS}, D'_{pow}, D'_{cont} and D'_{dRMSD}. We see that the D_{cRMSD} measure classifies a large fraction of the conformations as belonging to either class I or II. By contrast, the other measures classify a large fraction as class III, and leave many conformations unclassified. This tendency is least strong for the contact measure, which does identify significant populations of FU- and BU-like conformations.

From Table 3 it can be seen that there must be many conformations belonging to class I or II with the cRMSD measure that belong to class III with other measures. Notice that for a conformation in class III, the distances to both FU and BU are larger than the distance between FU and BU.

3.4 Discrete models for real protein structures

We now turn to the problem of approximating real protein structures. For this purpose, we use discrete C_α models where each virtual C_α-C_α bond has six possible directions, parameterized by the pseudo-bond angle α and the pseudo-torsion angle τ.

The discrete values of (α, τ) are determined by a Monte Carlo minimization procedure in the twelve-dimensional space of all possible (α, τ) angles. We follow Park and Levitt [25] and minimize a score which, for a given set of (α, τ) angles, is the average distance between a training set of 21–38 protein structures and the corresponding best model fits. The model fits are not strictly optimal since we use a stochastic algorithm (see section 2.10), but the distance found this way approximate well the minimal distance.

The optimized set of angles is then tested on a much larger set of proteins, consisting of 774 non-redundant protein structures without gaps in the crystal structure. The same analysis is carried out for three different distance measures, namely the cRMSD, the power distance with $m = 3$ (see equation (10)) and the contact distance with $r_c = 11\,\text{Å}$ (see equation (3)). The sequence cutoff s (see section 2.7) is taken to be $s = 2$ for the power and the contact distances.

Model fits obtained minimizing the cRMSD and the contact distance may exhibit C_α-C_α distances that are very small. Such structures are unphysical, since they contain atomic collisions. Therefore, we also optimized discrete models with hard core repulsion, rejecting all structures with two atoms closer than a cutoff distance R_c. The problem of atomic collisions does not arise for the power distance, since very short atomic distances are heavily penalized by this distance measure (see equation (8)). Below we present results obtained both with and without hard core repulsion.

3.5 Models without hard core repulsion

The optimized discrete model obtained using cRMSD as distance measure can fit every structure up to a cRMSD of 2.1 Å (it is larger than 2.0 Å for two structures only), with an average cRMSD of 1.57 Å. This is at least 10% better than previous results with the same number of angles. The minimal cRMSD tends to increase with protein length up to around 150 residues and then stay constant, as observed in a previous study [25]. Thus our results show that the similarity between real protein structures and structures built using only six allowed C_α directions is very high in terms of cRMSD.

We performed the same analysis for the contact and the power distances. The contact distance between real and fitted model structures was found to be 0.46 at most and 0.23 on average, the latter value corresponding to 77% common contacts. For the power distance, the maximal and average distances are 0.33 and 0.19, respectively.

	D_{cRMSD}	D_{cont}	D_{pow}
(α,τ) angles	$71.5, 71.2$	$83.3, 72.4$	$85.5, -62.4$
	$87.9, 55.6$	$92.0, 39.8$	$94.9, 96.0$
	$104.2, -111.0$	$115.0, -163.4$	$103.6, 163.0$
	$104.6, 36.6$	$118.0, -64.6$	$115.8, -152.2$
	$124.0, -160.0$	$128.5, 111.5$	$119.6, -22.0$
	$129.5, 128.8$	$129.7, -119.2$	$125.2, 126.8$
average score	1.57 Å	0.23	0.19

Table 4: Optimized (α,τ) angles, obtained without hard core repulsion.

The optimal sets of (α, τ) angles obtained with the three similarity measures are reported in Table 4.

These fits to real structures are based on distance minimization. In the following, we investigate how well model structures that are similar to real structures in terms of the distance measures can reproduce other geometrical features of real proteins. We focus our attention on the distribution of distances between C_α atoms. In Fig. 7 we plot the normalized number of pairs of C_α atoms at distance r, $f(r)$, defined as

$$f(r) = \frac{N(r)}{r^2 \Delta r}\,, \tag{22}$$

where $N(r)$ is the fraction of pairs with C_α-C_α distance in the interval $[r, r + \Delta r]$, and Δr is taken to be 0.1 Å. The figure shows this distribution for real structures and for the three sets of fitted structures discussed above, corresponding to different distance measures. All these distributions are generated using a sequence cutoff of $s = 2$.

The upper left panel of Fig. 7 shows the C_α-C_α distance distribution for crystal structures of real proteins. Several details are worth noting. First, distances smaller than 3.5 Å are absent, which is due to atomic collisions. Second, there is a double-peak at 5.2–6.2 Å, which can be attributed to favorable inter-residue interactions. Third, there is a deep valley at 7.5 Å, probably due to the excluded volume effect of residues in contact with the observed residue. The tail of the distribution decays slowly with r for small r, and then approximately exponentially. In fact, the large-r behavior, $r > 13$ Å, is well fitted by an exponential function $f(r) \propto \exp(-r/\xi)$, with $\xi = 8.30 \pm 0.01$ Å. The average C_α-C_α distance is 28.75 Å.

To what extent the fitted structures reproduce these features depends on the distance measure used. For the cRMSD measure (upper right panel in Fig. 7), it can be seen

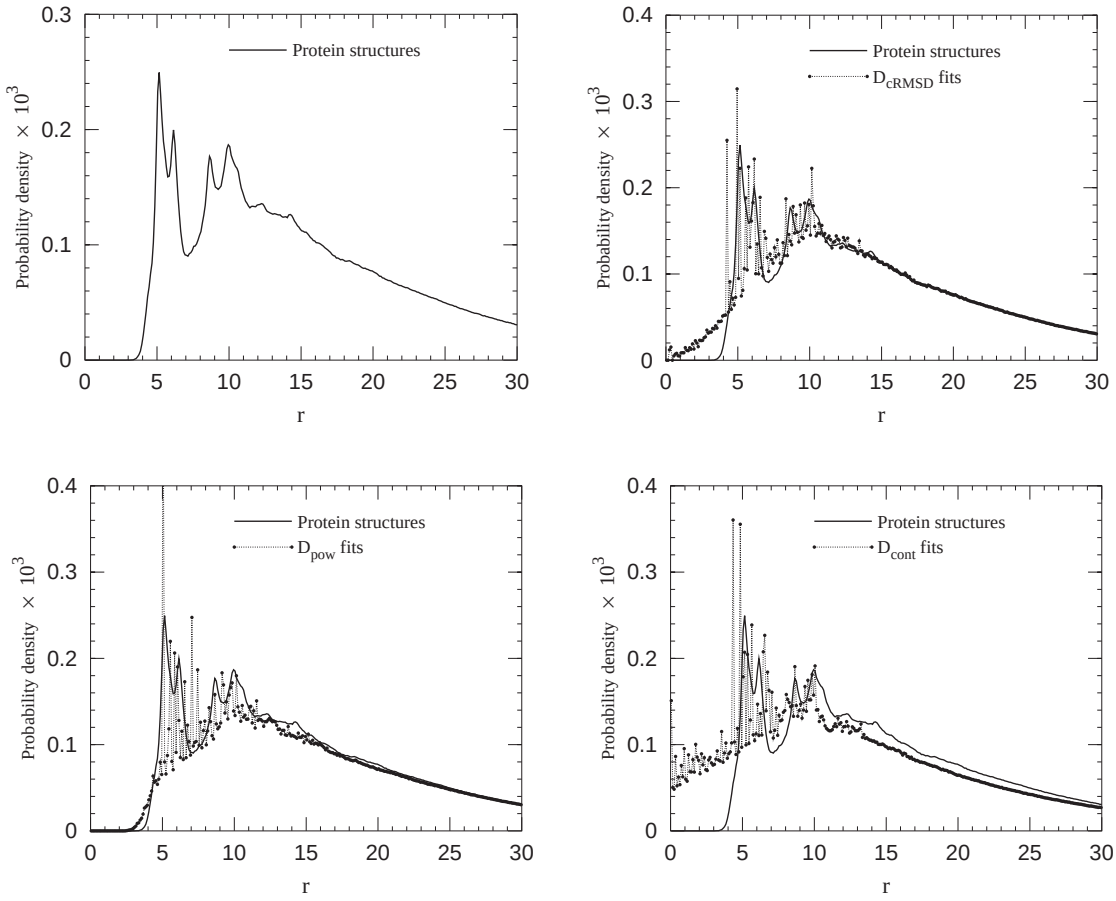

Figure 7: Distributions of C_α-C_α distances (in Å) for real proteins (upper left panel and solid lines in the other panels) and the corresponding models fits (dots). The upper right panel is for cRMSD and the lower panels are for the power distance (left) and the contact distance (right). In the lower left panel, the highest peak is outside the figure. Its value is $f(r) \times 10^3 = 0.53$ at $r = 5.0$.

that there is a significant population of atomic pairs with distance $r < 3.5\,\text{Å}$. Such distances lead to atomic collisions and are therefore unphysical. On the other hand, the large-r behavior of this distribution is in good agreement with the results for real structures. An exponential fit for large r yields $\xi = 8.30 \pm 0.01\,\text{Å}$ and the average C_α-C_α distance is 28.72 Å. These numbers are very close to the corresponding ones for real proteins.

For the power distance (lower left panel in Fig. 7), the situation is the opposite. Small distances are in this case absent, as they should, due to a strong penalty for atomic

collisions. However, this repulsion has the disadvantage that the model structures become less compact than the real ones. This is reflected in the values of the fitted exponent ξ, 8.50 ± 0.01 Å, and the average C_α-C_α distance, 29.82 Å, which are larger than for real proteins. Thus short range features are well reproduced but long range features are not.

Finally, for the contact distance (lower right panel of Fig. 7), we find that neither short range features nor long range ones are well reproduced. The density at $r < 3.5$ Å is even larger than in the cRMSD case; small distances are in fact favored if they increase the number of common contacts. As for long range features, one finds $\xi = 10.69 \pm 0.01$ Å and an average C_α-C_α distance of 32.8 Å. This implies that the contact distance leads to an effective long-range repulsion even stronger than for the power distance.

This effective long-range repulsion is at first sight surprising, but it can be explained by considering the definition of the contact distance. This distance measures the number of common contacts between two structures divided by the maximal number of contacts of the two structures. In order to minimize this distance, one first has to maximize the number of common contacts. Having maximized this number, the contact distance can still be further decreased if the total number of contacts in the model structure is made smaller than the corresponding number for the native structure. Because of this, one finds that the model fits tend to have fewer contacts and be less compact than real structures.

Summarizing, we have seen that discrete structures obtained minimizing the cRMSD reproduce global but not local features of real protein structures, and that those obtained minimizing the power distance, by contrast, reproduce local but not global features. For the contact overlap measure, neither local nor global features are well reproduced.

	D_{cRMSD}	D_{cont}
(α,τ) angles	$83.7, 62.0$	$82.7, 63.9$
	$95.9, 41.6$	$104.8, 169.4$
	$109.7, -151.9$	$106.9, 25.9$
	$110.8, -104.4$	$113.3, -137.3$
	$129.3, 186.7$	$113.8, -70.6$
	$134.0, 120.9$	$128.0, 109.9$
average score	1.54 Å	0.23

Table 5: Optimized (α,τ) angles, obtained with hard core repulsion.

3.6 Models with hard core repulsion

We now turn to the calculations where atomic collisions are avoided by rejecting all model structures having two atoms closer than a cutoff distance $R_c = 2.6$ Å. This value was chosen because it is the minimal C_α–C_α distance among the 774 native structures in our dataset. New optimal angles for the cRMSD and the contact distance were determined, and are shown in Table 5. The model fits obtained with the power distance satisfy the hard core repulsion constraint without rejecting any structures.

The discrete model obtained minimizing the cRMSD with hard core repulsion performs very similarly to the one reported in the previous section: every protein structure can be fitted to less than 2.1 Å cRMSD and the average cRMSD is 1.54 Å. The distribution of the C_α-C_α distances for these fits is shown in Fig. 8 and is very similar to the one obtained without repulsion, except that no distance smaller than R_c is present. The average distance is 28.76 Å and the large-r exponential fit to the probability distribution yields a characteristic length $\xi = 8.29 \pm 0.01$ Å.

The inclusion of hard core repulsion does not modify the performance much for the contact distance either. The average value of the contact distance is 0.23, the same as before, and the largest contact distance is 0.42, slightly better than before. The average C_α-C_α distance for the model fits are 32.50 Å, and the fitted characteristic length is $\xi = 10.62 \pm 0.01$ Å. Overall, the distribution is very similar to the one without repulsion, except for the zero probability density for $r < R_c$.

In conclusion, atomic collisions can be removed from the discrete models by introducing explicit hard core repulsion, without significantly changing the average quality of the fits. This hard core repulsion does, of course, affect the small-r part of the

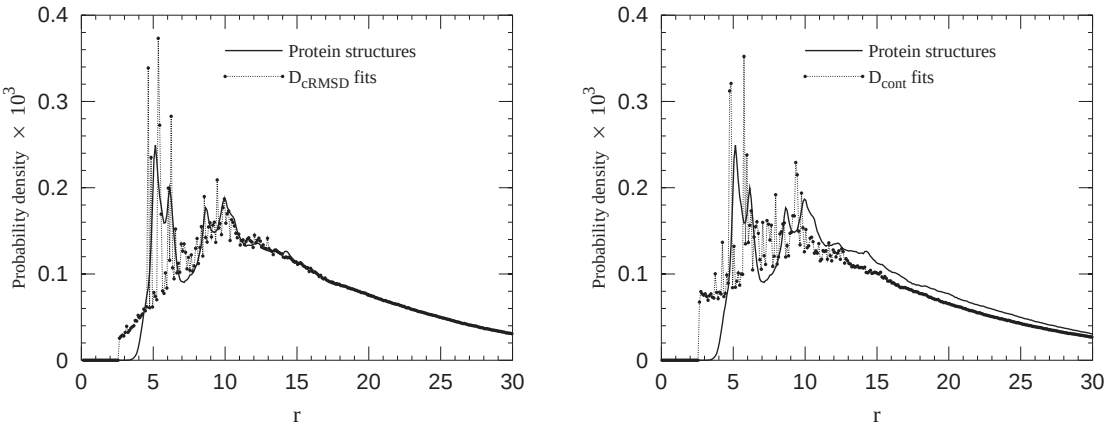

Figure 8: Distributions of C_α-C_α distances (in Å) for fits to real proteins (dots) obtained with hard core repulsion. Left and right panels are for cRMSD and the contact distance, respectively.

distribution of C_α-C_α distances for the fitted structures. However, the overall shape of this distribution, above R_c, changes very little when introducing the hard core repulsion.

4 Conclusions

We have investigated the properties of five different distance measures: the standard cRMSD measure and four measures based on intramolecular distances. We recall that the Holm and Sander measure is not strictly speaking a distance, since it does not fulfill the triangular inequality, but violations of this inequality are very rare. In fact, with our normalized version of this measure, D_{HS}, there were no such violations in our data set.

Using a continuous three-helix-bundle model, the correlation between native distance and energy was studied. We find that this correlation is significantly stronger for the measures D_{HS}, D_{pow} and D_{cont} than for D_{cRMSD} and D_{dRMSD}. This suggests that the former distance measures are more suitable to investigate the shape of the energy landscape. It must be remembered, however, that these results were obtained for one particular protein. This three-helix-bundle protein was chosen because the model has a relatively realistic chain representation and exhibits two-state folding. How general our conclusions are remains to be seen.

On the other hand, the ability to discriminate between the two different topologies of our three-helix-bundle protein is found to be quite limited for all the four measures based on intramolecular distances, while this task is easily solved by the cRMSD measure. This topology problem concerns structures in which not all contacts have been formed, while the two "ideal" minimum-energy structures can be distinguished without difficulties, at least with the contact distance. Therefore, one can still expect that native protein structures determined by X-ray crystallography can be assigned the correct topology. However, our results suggest that caution should be taken with the distance measures based on intramolecular distances. Among these, the best discrimination ability is exhibited by the contact distance and the worst one by the dRMSD.

Using different similarity measures as scoring functions, we performed fits of discrete C_α models with six directions per amino acid to real protein structures. The properties of the fitted structures turn out to depend quite strongly on the similarity measure used. Structures obtained using the cRMSD measure provide a good representation of large scale properties, but fail to reproduce small scale properties. The problem here is that the fitted structures tend to contain unphysical atomic collisions, which are not penalized by the cRMSD measure. Such collisions can be avoided by introducing a hard core repulsion that explicitly rejects structures containing atomic distances below a cutoff value. For the power distance, we find the opposite behavior. Here, the fitted structures reproduce local features very well and global features much worse. The contact measure is bad in both respects; the fits contain atomic collisions and are, at the same time, not as compact as they should. The latter problem reflects the fact that the fits have fewer contacts than the real structures. These results imply that a straightforward application of standard energy functions to fitted discrete structures can be misleading, due to artifacts at small or large scale.

Comparing these results to those obtained using the continuous model, a consistent picture seems to emerge. The cRMSD is by far the best for reproducing long range properties of protein structures, but fails to reproduce short range properties, as reflected in a poor energy correlation and the appearance of unphysical C_α-C_α collisions in the fitted structures. The power distance introduced here and the Holm and Sander measure are, by contrast, good at reproducing small scale properties, but the overall size of the fitted structures is too large and the ability to discriminate between the three-helix-bundle topologies is quite poor. For the contact distance the picture is less simple. This measure gives a good energy correlation, a short range property, and is not too bad at topology discrimination, a long range property. However, the fitted structures are relatively poor at both small and large scale. To conclude, it seems that there is no distance measure good for all purposes; a distance

measure to be used in the complex space of protein conformations should be chosen in consideration of the application in question.

Acknowledgements

We are indebted to Anders Irbäck and Ernst-Walter Knapp for valuable discussions and suggestions and to an anonymous referee for suggesting the inclusion of hard core repulsion in the discrete models.

Appendix

In this Appendix, we discuss the triangular inequality. We prove that it holds for the contact distance and find that violations of it are very rare for the power and Holm and Sander distances, although these ones do not strictly fulfill the inequality. The cRMSD and dRMSD measures are guaranteed to satisfy the triangular inequality since they are proportional to the ordinary Euclidean distance in $3N$- and $N(N-1)/2$-dimensional space, respectively.

Contact distance

Let us denote by M_a, M_b and M_c the total number of contacts in structures a, b and c, respectively, and by M_{ab} the number of shared contacts between a and b, and so on. We assume that $M_a \geq M_b \geq M_c$ and consider here only the side ab of the triangle abc. For the other two sides, the triangular inequality can be proven analogously (and more easily). The inequality to be proven reads

$$1 - \frac{M_{ab}}{M_a} \leq 1 - \frac{M_{ac}}{M_a} + 1 - \frac{M_{bc}}{M_b} \,. \tag{23}$$

This can be readily transformed into

$$M_a \left(M_{bc|a} + M_{abc} \right) + M_b M_{ac|b} \leq M_a M_b + M_b M_{ab|c} \,, \tag{24}$$

where M_{abc} denotes the number of contacts present in all the three structures, and $M_{ab|c}$ denotes the number of contacts present in structures a and b but not in c. M_{abc} has been eliminated from both the left- and right-hand sides. Since $M_a \geq M_b$, the l.h.s. is not larger than $M_a(M_{ac|b} + M_{bc|a} + M_{abc})$, which in turn is not larger than $M_a M_c \leq M_a M_b$. This, finally, is not larger than the r.h.s., so the inequality must hold.

Power distance

It is evident that the distance defined in equation (8) fulfills the triangular inequality. The normalized measure $D^{(1)}_{\mathrm{pow}}$ in equation (9) does so too, as can be seen by applying the inequality

$$\frac{|a-b|}{a+b} \leq \frac{|a-c|}{a+c} + \frac{|b-c|}{b+c} \tag{25}$$

to each term, where a, b and c are positive numbers. To prove this inequality, we assume that $a \geq b \geq c$ (the remaining two cases can be treated analogously). In this case, the inequality (25) is equivalent to

$$(a+c)(b+c)(a-b) \leq 2(a+b)\left(ab-c^2\right) . \tag{26}$$

Here, since $b \geq c$, the r.h.s. is not smaller than $2(a+b)(ab-b^2)$. At the same time, the l.h.s. is not larger than $(a+b)(b+b)(a-b) = 2(a+b)(ab-b^2)$, since $a \geq b$ and $b \geq c$. Therefore the inequality is proven.

The triangular inequality does not hold strictly for the normalized version D_{pow} in equation (10) that we use in this paper. However, as mentioned earlier, no violation was detected for the (large) set of protein-like structures that we use.

The Holm and Sander score

For the original version of the Holm and Sander distance, D^*_{HS} in equation (6), and for $N = 2$, the triangular inequality has the form

$$\frac{|a-b|}{a+b}\mathrm{e}^{-(a+b)^2} \leq \frac{|a-c|}{a+c}\mathrm{e}^{-(a+c)^2} + \frac{|b-c|}{b+c}\mathrm{e}^{-(b+c)^2} , \tag{27}$$

which is clearly violated if $c \gg a, b$. For the normalized version D_{HS} in equation (7) and $N = 2$, the triangular inequality is equivalent to the inequality (25), which, as we have seen, is satisfied. The triangular inequality for D_{HS} and $N = 3$ takes the form

$$\frac{\frac{|a_1-b_1|}{a_1+b_1}\mathrm{e}^{-(a_1+b_1)^2} + \frac{|a_2-b_2|}{a_2+b_2}\mathrm{e}^{-(a_2+b_2)^2} + \frac{|a_3-b_3|}{a_3+b_3}\mathrm{e}^{-(a_3+b_3)^2}}{\mathrm{e}^{-(a_1+b_1)^2} + \mathrm{e}^{-(a_2+b_2)^2} + \mathrm{e}^{-(a_3+b_3)^2}} \leq \frac{|a_1-c_1|}{a_1+c_1} + \frac{|b_1-c_1|}{b_1+c_1} \tag{28}$$

in the limit of one distant point (two distances $c_2, c_3 \to \infty$). This inequality can not be satisfied since the r.h.s becomes zero when $a_1 = b_1 = c_1$, and the l.h.s. depends on a_2, a_3, b_2 and b_3. For large N, however, D_{HS} consists of many terms and violations of the triangular inequality become very rare.

References

[1] Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP Experiments. *Proteins Struct Funct Genetics, Suppl.* 2001; 5: 163-170.

[2] Holm L, Sander C. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res* 1994; 22: 3600-3609.

[3] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: A hierarchic classification of protein domain structures. *Structure* 1997; 5: 1093-1108.

[4] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995; 247: 536-540.

[5] Koehl P. Protein Structure similarities. *Curr Opin Struct Biol* 2001; 11: 348-53.

[6] May ACW. Towards more meaningful hierarchical classification of amino acid scoring matrices. *Protein Eng* 1999; 12: 707-712.

[7] Orengo CA, Swindells MB, Michie AD, Zvelebil MJ, Driscoll PC, Waterfield MD, Thornton JM. Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity. *Protein Sci* 1995; 4: 1977-1983.

[8] Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. *Fold Des* 1996; 1: 123-132.

[9] Godzik A. The structural alignment between two proteins: is there a unique answer? *Protein Sci* 1996; 5: 1325-1338.

[10] Anfinsen CB. Principles that govern the folding of protein chains. *Science* 1973; 181: 223–230.

[11] Bryngelson JD, Wolynes PG. Spin-glasses and the statistical-mechanics of protein folding. *Proc Natl Acad Sci USA* 1987; 84: 7524-7528.

[12] Garel T, Orland H. Mean-field model for protein folding. *Europhys Lett* 1988; 6: 307–310.

[13] Shakhnovich EI, Gutin AM. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys Chem* 1989; 34; 187–199.

[14] Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA* 1993; 90: 7195–7199.

[15] Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett* 1994; 24: 3907–3910.

[16] Abkevich VI, Gutin AM, Shakhnovich EI. Free energy landscapes for protein folding kinetics - intermediates, traps and multiple pathways in theory and lattice model simulations. *J Chem Phys* 1994; 101: 6052–6062.

[17] Klimov DK, Thirumalai D. Factors governing the foldability of proteins. *Proteins Struct Funct Genet* 1996; 26: 411–441.

[18] Bastolla U, Frauenkron H, Gerstner E, Grassberger P, Nadler W. Testing a new Monte Carlo algorithm for protein folding. *Proteins Struct Funct Genet* 1998; 32: 52–66.

[19] Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI. Folding and misfolding of designed protein-like chains with mutations. *J Chem Phys* 1998; 108: 757–761.

[20] Bastolla U, Roman HE, Vendruscolo M. Neutral evolution of model proteins: Diffusion in sequence space and overdispersion. *J Theor Biol* 1999; 200: 49–64.

[21] Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J Mol Biol* 1996; 258: 367–392.

[22] Irbäck A, Sjunnesson F, Wallin S. Three-helix-bundle protein in a Ramachandran model. *Proc Natl Acad Sci USA* 2000; 97: 13614–13618.

[23] Irbäck A, Sjunnesson F, Wallin S. Hydrogen bonds, hydrophobicity forces and the character of the collapse transition. *J Biol Phys* 2001; 27: 169–179.

[24] Favrin G, Irbäck A, Wallin S. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins Struct Funct Genet* 2002; 47: 99–105.

[25] Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol* 1995; 249: 493–507.

[26] von Neumann J. Some matrix-inequalities and metrization of matric-space. *Tomsk Univ Rev* 1937; 1: 286–300.

[27] Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog sect A* 1978; 34: 827–828.

[28] Maiorov VN, Crippen GM. Size-independent comparison of protein three-dimensional structures. *Proteins Struct Funct Genet* 1995; 22: 273–283.

[29] Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* 1976; 104: 59–107.

[30] Cohen FE, Sternberg JE. On the prediction of protein structure: the significance of the root-mean-square deviation. *J Mol Biol* 1980; 138: 321–333.

[31] Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol* 1994; 235: 625–634.

[32] Vendruscolo M, Subramanian B, Kanter I, Domany E, Lebowitz J. Statistical properties of contact maps. *Phys Rev E* 1999; 59: 977–984.

[33] Bastolla U, Frauenkron H, Grassberger P. Phase diagram of random heteropolymers: replica approach and application of a new Monte Carlo algorithm. *Jour Mol Liq* 2000; 84: 111–129.

[34] Chan HS, Kaya H, Shimizu S. Computational methods for protein folding: scaling a hierarchy of complexities. In: Jiang T, Xu Y, Zhang MQ, editors. Current topics in computational molecular biology. Cambridge, Massachusetts: MIT Press; 2002. p 403–447.

[35] Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993; 233: 123–138.

[36] Holm L, Sander C. Mapping the protein universe. *Science* 1996; 273: 595–602.

[37] Rooman MJ, Kocher JA, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. *J Mol Biol* 1991; 221: 961–979.

[38] Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins Struct Funct Genet* 2001; 44: 79–96.

[39] Krantz BA, Srivastava AK, Nauli S, Baker D, Sauer RT, Sosnick TR. Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects. *Nature Struct Biol* 2002; 9: 458–463.