Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 44 (2003) 4373–4379
polymer
www.elsevier.com/locate/polymer

Series approximation of protein structure and constructing conformation space

G.M. Crippen*

College of Pharmacy, University of Michigan, 428 Church Street, Ann Arbor, MI 481091065, USA

Received 21 November 2002; received in revised form 8 February 2003; accepted 8 February 2003

Abstract

Series approximations of the three-dimensional structure of protein conformations can provide insightful ways to detect and manipulate global features and those local to contiguous segments of the chain. Discrete cosine transforms have proven to be very useful in the past, and now wavelet transforms appear to have additional advantages. Here the emphasis is on a new generalization of the discrete Haar transform for chains of arbitrary length, as opposed to the customary powers of 2. This can be used to define a true, concrete conformation space, where different conformations correspond to points in the space, and a measure of distance between points corresponds to the customary root-mean-square deviation after optimal pairwise superposition (rmsd). Examples are given of how to do this to high accuracy. The key is to devise a rule for placing individual conformers in a standard position relative to the coordinate system, rather than superimposing them on a pairwise basis.

Keywords: Wavelets; Structural superposition; Conformational similarity

1. Introduction

There are many answers to the question, 'How similar are these two proteins?', depending on the criteria and methods used. If only the primary structure is considered, then each protein is an ordered linear sequence of amino acid types that has an orientation (i.e. direction) from N- to C terminus. Standard sequence alignment algorithms produce a one-to-one matching of some subset of the residues of protein A to an equal number of the residues of protein B that strictly preserves the ordering. Thus if residue a_i is matched to b_j, and a_k to b_l, then it must be true that $i \neq k$, $j \neq l$ and $i < k$ implies $j < l$. Here we are concerned with comparing the three-dimensional structure of proteins largely without regard to amino acid sequence, but these sequence comparison ideas underlie much of the thinking. Proteins are regarded as linear, oriented sequences of labelled points in space, one point per residue, and in comparisons no point may be used twice, nor may the residue–residue matching violate the sequence ordering.

When the structural comparison involves two different conformations of the same protein, then the required one-to-one matching of residue points in space is obvious, and the main problem is to translate and rotate the one set of points onto the other to achieve an optimal superposition that is independent of the starting coordinates. This is very well handled by standard methods as described below. For comparisons between different proteins (still concentrating on single polypeptide chains apiece), the primary concern is devising a suitable matching. If there is substantial sequence similarity, the customary approach is to first align the two sequences without regard to structure and then use the implied matching for the superposition step. Failing strong sequence similarity, the gapped alignment can use instead a combination of secondary structural state of the residues and structural similarity [1]. Alternatively, one might develop a matching strictly on structural criteria. For example, the two interresidue distance matrices can be aligned, allowing for insertions and deletions, but preserving sequence ordering [3]. The degree of similarity so measured then depends on the details of the alignment gap penalties because regions contributing to poor structural similarity tend to be eliminated by sequence deletions. One way to balance these factors is to find the maximal cardinality matching such that after optimal rigid body superposition, the worst

* Tel.: +1-734-763-9722; fax: +1-734-763-2022.
E-mail address: gcrippen@umich.edu (G.M. Crippen).

0032-3861/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0032-3861(03)00131-9

superimposed atom pair is closer than a given cutoff [4]. Relaxing the one-to-one matching requirement has been explored, for example by seeking the rigid body superposition such that a surface running between the two chains has minimal area [5]. Others have proposed relaxing the strict preservation of sequence order by permitting reordering of sequence secondary structure segments or even reversing their orientation [2]. Insisting on a strict rigid body superposition may also not be reasonable when comparing multidomain protein structures [2].

Here we examine two other facets of the general concept. The direct Cartesian coordinates of the points representing the residues may not always be the most suitable way to view conformational features. For some purposes, different coordinate transforms have advantages. The other idea is analogous to going from pairwise sequence alignment to multiple alignments. The algorithm for comparing two protein structures may not readily generalize to a way to view many different structures at once.

2. Series approximation of structures

2.1. Fourier series

Suppose we have a periodic function $f(x)$ with period 2π, i.e. $f(x+2\pi k)=f(x)$ for any integer k. Then there is an infinite set of orthonormal functions $\{\psi_0,\psi_1,\ldots\}=\{(2\pi)^{-1/2},\pi^{-1/2}\sin(kx),\pi^{-1/2}\cos(kx),\ldots\}$ for integer $k\geq 1$ in the sense that $\int_0^{2\pi}\psi_i(x)\psi_j(x)\,dx=\delta_{ij}$, where $\delta_{ij}=1$ if $i=j$ and 0 otherwise. Then except for pathological cases, $f(x)$ can be approximated by a convergent series $\sum_{k=0}c_k\psi_k(x)$, where the coefficients are easily calculated by $c_k=\int_0^{2\pi}f(x)\psi_k(x)\,dx$. The discrete version of all this samples $f(x)$ at evenly spaced values x_i, $i=1,\ldots,n$ over the range 0 to 2π, and the above integrals are converted into sums. Only n terms in the final series are required to reproduce all n samples $f(x_i)$ exactly. For computational efficiency in the fast Fourier transform (FFT), n is chosen to be a power of 2.

If there are discontinuities in $f(x)$, then the series approximation does not converge rapidly. In particular, if $f(x)$ is not really periodic, then the discontinuity $f(0)\neq f(2\pi)$ causes large $|c_k|$ for large k, the well-known Gibbs phenomenon. Worse yet in the discrete case, if n is not a power of 2, then the 'missing' data is customarily supplied as $f(x_i)=0$, a procedure called zero-filling that generally introduces discontinuities.

Nonetheless, the general approach is an attractive way to view polymer chain conformations. Suppose we represent a polypeptide chain as a sequence of n points in space, chosen to be the C^α atoms of the residues running from N- to C terminus. Then with respect to a fixed external coordinate frame, we have three discretely sampled functions, x_i, y_i, z_i for $i=1,\ldots,n$. Each can be independently approximated by a series, and neglecting high frequency terms corresponds to viewing the original chain at low resolution.

2.2. Discrete cosine transform

One way to convert an aperiodic function defined on a finite interval into a periodic one is to append its reflection and rescale the composite to the standard interval $[0,2\pi]$. As a result, the odd sine terms all have zero coefficients, and hence it is called the cosine transform. In the discrete case [6] this works out to be a set of n transform coefficients calculated by

$$\hat{x}_k=\frac{2b_k}{n}\sum_{j=1}^{n}x_j\cos\left[\frac{(2j-1)(k-1)\pi}{2n}\right] \quad (1)$$

for $k=1,\ldots,n$,

from which the precise original values can be recovered by the inverse transform

$$x_j=\sum_{k=1}^{n}b_k\hat{x}_k\cos\left[\frac{(2j-1)(k-1)\pi}{2n}\right] \quad (2)$$

for $j=1,\ldots,n$,

where in either case

$$b_k=\begin{cases}2^{-1/2}, & \text{for } k=1\\ 1, & \text{otherwise}\end{cases}. \quad (3)$$

One way to view this is that the original n points having coordinates (x_j,y_j,z_j) can be converted into n points in the transform space having coordinates $(\hat{x}_k,\hat{y}_k,\hat{z}_k)$, and vice versa. Each transform point is a linear combination of all the original points, and vice versa. There are a number of nice relations between the two sets of points [7] along the general lines of overall properties of the whole set of real points correspond to properties of individual transform points. For example, deleting or moving to the origin some of the high frequency transform points corresponds to smoothing the original chain path. How many transform points are kept governs the flexibility of the backtransform, so that keeping two permits only straight line configurations, keeping three permits one bend, etc. Translating the whole set of original points so that their unweighted center of mass is at the origin corresponds to moving the first transform point to the origin: $(\hat{x}_1,\hat{y}_1,\hat{z}_1)=(0,0,0)$. A subsequent rotation of the original points about the origin corresponds to the same rotation of the transform points, and vice versa. The radius of gyration R is related in the two sets of points by

$$R^2=n^{-1}\sum(x_i^2+y_i^2+z_i^2)=2^{-1}\sum(\hat{x}_i^2+\hat{y}_i^2+\hat{z}_i^2). \quad (4)$$

For our purposes, the main advantage of the discrete cosine transform over the discrete Fourier transform is that chains having arbitrary length n residues can be equally well treated, and there are no side effects due to the fact that single chain proteins are not cyclic, i.e. their coordinates are not periodic. The disadvantage is that the transform in Eq. (1) is relative to the whole chain length, so $\hat{x}_1$ is an average

over the entire chain, $\hat{x}_2$ reflects differences between the first and second halves of the chain, $\hat{x}_3$ is related to the quarters of the chain, etc. Because an α-helix always has 3.6 residues per turn, independent of n, we see from Eq. (1) that the $\hat{x}_k$ most affected by overall helical content are those where $k \approx 1 + 0.645n$. In other words, features involving a fixed scale of a certain number of residues affect different terms of the transform depending on the total chain length. An analogous problem is that conformational features of a small segment of the original chain are reflected in the positions of many of the transform points.

2.3. Haar transform

A different parameterization of chain conformations that seems very attractive is to use wavelets. As with the Fourier and cosine transforms, the general idea is to develop a set of orthonormal functions over some domain in order to approximate arbitrary functions as series expansions. The unusual feature of wavelets is that the individual basis functions are nonzero over only limited portions of the entire domain, which is advantageous when approximating aperiodic functions over a finite interval, especially when certain subintervals contain features of particular interest. This sort of expansion has been used in connection with protein structure mainly to identify conformational features associated with particular parts of the chain, to classify proteins, or to correlate sequence and solvent exposure [8–11]. There are many different kinds of wavelets [12], but for our purposes the simple Haar transform [13] is most suitable.

The standard discrete Haar transform starts with signal values sampled at evenly spaced intervals, x_i for $i = 1, \ldots, n$, representing in our case the x-coordinates of the C^α atoms of a polypeptide chain having length n. When $n \geq 1$ is a power of 2, define a set of L^2 orthonormal functions $\psi_{w,j}(i)$ called Haar wavelets by:

$$\psi_{w,j}(i) = \begin{cases} (2w)^{-1/2}, & \text{for } i = j, \ldots, j + w - 1 \\ -(2w)^{-1/2}, & \text{for } i = j + w, \ldots, j + 2w - 1 \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

$$\psi_{n,1}(i) = n^{-1/2} \text{ for } i = 1, \ldots, n,$$

where the half-width $w = 1, 2, 4, \ldots, n/2$ increases by factors of 2, and the start of the wavelets having that half-width $j = 1, 2w + 1, 4w + 1, \ldots, n - 2w + 1$ increase by $2w$. There are altogether n wavelets, and they constitute a complete orthonormal basis for vectors of data $\mathbf{x} = [x_1, \ldots, x_n]^T$. In other words, associated with each wavelet is a Haar transform coefficient

$$\hat{x}_{w,j} = \sum_{i=1}^{n} x_i \psi_{w,j}(i) \tag{6}$$

and the exact original signal can be recovered from them.

$$x_i = \sum_{w,j} \hat{x}_{w,j} \psi_{w,j}(i). \tag{7}$$

As shown in Fig. 1, the Haar transform gives a hierarchical view of the chain configuration, going from the mean position of all residues, to the difference in mean position of the first half of the chain and the second half, the difference in the first quarter mean position and the second quarter, the third and fourth quarters, and so on, down to differences in position between sequentially adjacent residues. Quantitatively,

$$\hat{x}_{w,j} = \frac{\sqrt{w}}{2} \left(w^{-1} \sum_{i=j}^{j+w-1} x_i - w^{-1} \sum_{i=j+w}^{j+2w-1} x_i \right), \tag{8}$$

for different segments of the chain, while for the whole chain, $(\hat{x}_{n,1}, \hat{y}_{n,1}, \hat{z}_{n,1}) = (0, 0, 0)$ when the unweighted center of mass has been translated to the origin. In addition, the radius of gyration of a segment of the chain having length m starting at residue k is simply related to the sum of the squares of the transform coefficients for the associated wavelets in that segment

$$\sum_{i=k}^{k+m-1} \left(x_i - m^{-1} \sum_{j=k}^{k+m-1} x_j \right)^2 = \sum_{w=1}^{m/2} \left(\sum_{j=k}^{k+m-2w-1} \hat{x}_{w,j}^2 \right) \tag{9}$$

and for the whole chain this reduces to

$$\sum_{i} \left(x_i - n^{-1} \sum_{j} x_j \right)^2 = \sum_{w,j} \hat{x}_{w,j}^2. \tag{10}$$

Another advantage of the Haar transform over the cosine transform is that the subdivisions are not relative to the full chain length, but have a fixed scale of 2, 4, 8, etc. sequentially adjacent residues. One can more easily deal with effects that are local to a short segment of the whole

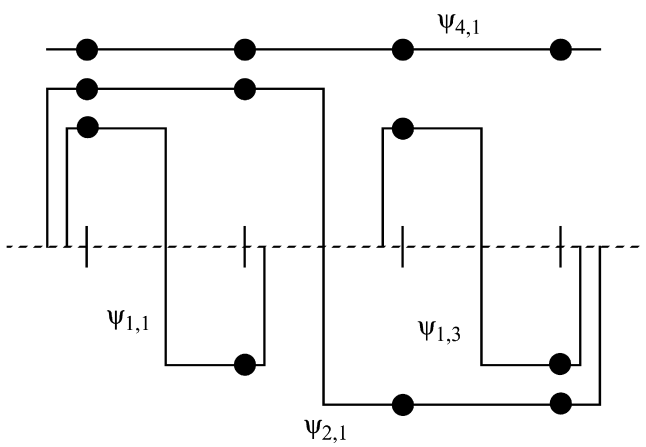

Fig. 1. Schematic drawing of the four discrete Haar wavelets for $n = 4$. Dots indicate the discrete values for which each wavelet is defined. Amplitudes have been altered for clarity.

chain, since narrower wavelets are nonzero only over that segment.

If n is not a power of 2, the conventional approach is to zero-fill, i.e. to add $x_i = 0$ to the end of the vector up to the nearest power of 2 and then calculate the wavelets accordingly, especially for centered data where $\sum_{i=1}^{n} x_i = 0$. The disadvantage is that if x_n differs greatly from zero, then all wavelets (except the first one) having $j + 2w - 1 > n$ will have large $|\hat{x}_{w,j}|$. We have generalized Haar wavelets to arbitrary n by using Eq. (5) when $j + 2w - 1 \leq n$, but otherwise using

$$\psi_{w,j}(i) = \begin{cases} \left(\dfrac{n+1-j-w}{(n+1-j)w}\right)^{1/2}, & \text{for } i = j, \ldots, j+w-1 \\ -\left(\dfrac{w}{(n+1-j)(n+1-j-w)}\right)^{1/2}, & \text{for } i = j+w, \ldots, n \\ 0, & \text{otherwise} \end{cases}, \tag{11}$$

where $w = 1, 2, 4, \ldots, 2^k$ and $k > 0$ is the largest integer such that $2^k < n$. These wavelets are still an orthonormal complete basis, but the positive side may be longer than the negative side, and their amplitudes differ. With this arrangement, there are always exactly n transform coefficients, and their values are proportional to the difference between the mean x_i on the positive side and the mean x_i on the negative side.

Here the emphasis is on low resolution C^{α} trace representations of protein structure. Note that the structural hierarchy inherent in Haar transforms can also include all-atom representations. Let the polypeptide chain be a sequence of points ordered by sequence number i but including, say, 8 points for each residue in some kind of standard ordering such as N_i, C_i^{α}, C_i, O_i, C_i^{β}, C_i^{γ}, etc. Longer sidechains may have only selected atoms included, and shorter ones may have duplication of atoms to bring the total up to 8. Then wavelet $\psi_{8,8i+1}$ is associated with the mean position of the atoms of residue $i + 1$ relative to the mean position of the atoms of residue $i + 2$. Narrower wavelets reflect the relative positions of atoms within one residue.

3. Comparison of conformations

3.1. RMSD

Given two sets of points A and B, such as the C^{α} traces for two polypeptide chains, the most widely used measure of conformational similarity is the root-mean-square deviation in corresponding coordinates after optimal superposition by rigid body motions (rmsd). We might call this a 'labelled' comparison because it requires a given one-to-one correspondence between at least some subset of the points in A and an equal sized subset of points in B by which the motion and subsequent rmsd are calculated. This does not cause a problem when A and B are different conformations of the same molecule, but it is not obvious how to treat two protein structures having different chain lengths. The advantage is that once given a correspondence, there is a unique optimal rigid body motion for the superposition, and it is readily calculated from the two coordinate sets. Any rigid body motion can be viewed as the composite of one translation and one proper rotation. In one dimension, the translation t applied to set B that minimizes $\sum_{i=1}^{n} (x_{Ai} - x_{Bi} - t)^2$ is easily shown to be $t = n^{-1}(\sum x_{Ai} - \sum x_{Bi})$. This is easily achieved by translating both sets of points so that their respective means are at the origin, and the same holds independently for the y- and z-coordinates. The subsequent optimal rotation about the origin of one set onto the other is more complicated to calculate, but we find Kabsch's approach very reliable [14]. Other algorithms for this same task include iterative methods [15] and the use of quaternions [16].

Rmsd resembles a metric insofar as $\mathrm{rmsd}(A,B) = \mathrm{rmsd}(B,A)$, $\mathrm{rmsd}(A,B) \geq 0$, and $\mathrm{rmsd}(A,B) = 0$ if and only if $A = B$ in the sense that the relative positions of corresponding points in the two sets are identical. For different sets A and B, the magnitude of rmsd(A,B) is governed largely by the radii of gyration of the two sets, which in the case of molecular conformations correlates with molecular weight. Thus a 5 Å rmsd between two conformations of a 500 residue protein indicates great similarity, whereas the same value for a 10 residue peptide indicates great dissimilarity. To a first approximation one can put these on the same scale [17]

$$\rho(A,B) = \frac{2\mathrm{rmsd}(A,B)}{(2R^2(A) + 2R^2(B) - \mathrm{rmsd}^2(A,B))^{1/2}}, \tag{12}$$

where $R(A)$ is the radius of gyration of point set A. More precisely, one first adjusts the two sets to have unit radius of gyration and three equal moments of inertia. Then $\rho = 2$ corresponds to maximal conformational dissimilarity, and $\rho \leq 0.5$ indicates obvious conformational resemblance.

While we can make good sense of pairwise comparisons, for neither rmsd nor ρ can we simply interpret dissimilarity as a true distance measure between many pairs of points corresponding to structures in some abstract space. This is because the optimal superposition of conformer B onto conformer A and C onto A does not generally correspond to the optimal superposition of B and C. The result is a failure to always obey the triangle inequality, $\mathrm{rmsd}(B,C) \leq \mathrm{rmsd}(B,A) + \mathrm{rmsd}(C,A)$, and other such relations. For example, one can easily generate 100 random compact conformations of a freely jointed chain of five steps with a fixed step length of 3.8 Å, simulating (non-self-avoiding) pentapeptide conformers. Then if the rmsd between each pair is supposed to represent the distance between their corresponding points in some Euclidean conformation space, metric matrix embedding [18] can readily calculate coordinates for the 100 points. If coordinates can be found in some $k < 100$ subspace, then an intermediate step in the calculation will determine that the metric matrix has k positive eigenvalues, and the other $100 - k$ eigenvalues will

be zero. As it turns out, roughly half the eigenvalues are negative, a strong indication that the given rmsds did not correspond to interpoint distances in any Euclidean space whatsoever.

3.2. Principal axes standard position

If we slightly alter our measure of conformational similarity, it is possible to construct a consistent conformation space. The trick is to define a standard positioning of each conformation in terms of some rigid body translation and rotation. Then for n residues, the standard position of each conformer is given by $3n$ Cartesian coordinates, the abstract conformation space has $3n$ dimensions, and the distance (i.e. conformational dissimilarity) between conformations is just the usual Euclidean distance between the two sets of standard position coordinates. The suitability of a standard positioning scheme may be judged by the correlation between pairwise rmsd and standard position distance for many pairs of conformers.

The first scheme that comes to mind is just the usual center-of-mass principal axes coordinate system used in classical mechanics. The configuration of points is translated so that $\sum x_i = \sum y_i = \sum z_i = 0$, and it is then rotated so that $\sum x_i^2 > \sum y_i^2 > \sum z_i^2$ and $\sum x_i y_i = \sum x_i z_i = \sum y_i z_i = 0$. Similar distributions of points are similarly positioned so that the scatter is greatest along the x-axis and second greatest along the y-axis. One problem occurs when the variance/covariance matrix has nearly degenerate eigenvalues so that the above inequalities are nearly equalities. Then, for example, a small change in configuration can cause an exchange of the x- and y-axes, i.e. a 90° flip. The other difficulty is that this positioning scheme is most compatible with an unlabeled measure of similarity between point sets A and B, such as the Hausdorff metric

$$H(A,B) = \min[\max_{a\in A} \min_{b\in B} d(a,b), \max_{b\in B} \min_{a\in A} d(b,a)], \tag{13}$$

where $d(a,b)$ is the ordinary Euclidian distance between points a and b. If A and B are extremely similar, there are four equally good positionings of B (as is, or a 180° rotation about the x-, y-, or z-axes), only one of which is compatible with A in the labelled rmsd sense.

3.3. Chain segment standard positioning

One standard positioning scheme for labelled polymer conformers, such as proteins, distinguishes the beginning (N terminus) and end (C terminus) of the chain [19]. As usual, the polypeptide chain is represented as a labelled sequence of C^α atom points. First translate so that the unweighted center of mass is at the origin; then rotate so that the centroid of the first third of the chain is on the positive x-axis and the centroid of the third third is in the xy-plane with positive y coordinate. In other words, $0 = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i = \sum_{i=1}^{n/3} y_i = \sum_{i=1}^{n/3} z_i = \sum_{i=2n/3}^{n} z_i$, and $\sum_{i=1}^{n/3} x_i > 0$, $\sum_{i=2n/3}^{n} y_i > 0$. Obviously there are cases where small perturbations of particular conformations can result in large changes in positioning, such as when any of the three centroids happen to be close to each other, but that appears to be a rare problem for protein-like conformers. After all, the motivation behind focussing on entire thirds of the chain was to keep the positioning insensitive to perturbations of small parts of the chain. For example [19], when a set of 762 conformations of an $n = 35$ residue oligopeptide are so positioned and further simplified by representing each one as a point in a 12-dimensional space ($\mathbb{R}^{12}$), where the coordinates are the discrete cosine transform terms showing the greatest variability over the set, then there is an $r^2 = 0.8$ correlation between interpoint distances in $\mathbb{R}^{12}$ and pairwise superposition ρ values.

3.4. Haar transform standard positioning

An even better standard positioning method uses the generalized discrete Haar transform, Eq. (11). The Cartesian coordinates of the n C^α atoms constitute three signals, **x**, **y**, and **z**. Transforming each signal independently produces $3n$ Haar coefficients. Let the standard translation of the conformer be to place the center of mass at the origin, which ensures that $\hat{x}_{n,1} = \hat{y}_{n,1} = \hat{z}_{n,1} = 0$. Consider the variance/covariance matrix C of the other coefficients.

$$C = \begin{bmatrix} \sum_{w,j} \hat{x}_{wj}^2 & \sum_{w,j} \hat{x}_{w,j}\hat{y}_{w,j} & \sum_{w,j} \hat{x}_{w,j}\hat{z}_{w,j} \\ \sum_{w,j} \hat{x}_{w,j}\hat{y}_{w,j} & \sum_{w,j} \hat{y}_{w,j}^2 & \sum_{w,j} \hat{y}_{w,j}\hat{z}_{w,j} \\ \sum_{w,j} \hat{x}_{w,j}\hat{z}_{w,j} & \sum_{w,j} \hat{y}_{w,j}\hat{z}_{w,j} & \sum_{w,j} \hat{z}_{w,j}^2 \end{bmatrix}. \tag{14}$$

The orthonormal eigenvectors of C ordered according to decreasing (positive) eigenvalues form the rows of a standard positioning rotation matrix. The effect of this rotation on the original coordinates is to make subsequent transform coefficients largest in the first (x) axis: $\sum_{w,j} \hat{x}_{w,j}^2 \geq \sum_{w,j} \hat{y}_{w,j}^2 \geq \sum_{w,j} \hat{z}_{w,j}^2$. Visually, the strands of β-sheets or bundles of α-helices tend to run back and forth mostly in the x-direction, while there is less variation in the y-direction, and less yet in the z-direction. This is because large changes in coordinates along various segments of the chain are reflected in large magnitudes of the corresponding transform coefficients. In order to completely specify the rotation matrix, some of the rows may need to be multiplied by -1 so that the conformation is not reflected, the rotated $\sum_{i=1}^{10} x_i > 0$, and the rotated $\sum_{i=21}^{30} y_i > 0$. The basis for the choice is that the position of residue 5 is relatively uncorrelated with that of residue 25 in a survey over many native protein structures.

Nevertheless, there are still cases where small changes in

the conformation can lead to large shifts in standard position, either because of nearly degenerate eigenvalues of the covariance matrix, or when the chain near residue 5 is near the *yz*-plane, or the chain near residue 25 is near the *xz*-plane. Problems of this nature may be an inevitable consequence of a possible fundamental incompatibility between the rmsd of labelled points and Euclidean space. The obvious conformational similarity measure between point sets A and B in their standard positions is $S(A,B) = [n^{-1}\sum_i d^2(a_i,b_i)]^{-1/2}$, in order to be directly comparable to rmsd(A,B). Because rmsd involves the same comparison of corresponding points after the optimal pairwise superposition, rather than just using their standard position coordinates, it is always true that $S(A,B) \geq \text{rmsd}(A,B)$. Frequently the two measures are nearly equal. Most cases where $S(A,B)$ is substantially larger than rmsd(A,B) are due to the four-way ambiguity of 180° flips described above. In order to make our measure of conformational similarity in standard position more compatible with the realities of rmsd, let $F(A,B)$ be the minimum value of $S(A,B)$ over all four alternatively flipped positions of B, holding A constant. This ensures that $S(A,B) \geq F(A,B) \geq \text{rmsd}(A,B)$. To test the relation between pairwise rmsd and standard position distance, we compared all pairs of a set of 1893 protein chains found in the Protein Data Bank (PDB), truncating the longer of each pair to give the same chain lengths. Pairs of conformations ranged from very similar (nearly 0 Å rmsd) to very dissimilar (30 Å rmsd). The standard positioning distance is always greater than or equal to that by pairwise superposition, and the correlation coefficient between $F(A,B)$ and rmsd(A,B) is an excellent $r^2 = 0.94$, as shown in Fig. 2. The correlation between $S(A,B)$ and rmsd(A,B) is a still good $r^2 = 0.85$, indicating that problems with flips in protein conformations are rather rare.

In addition, standard positioning is robust with respect to small residue insertions and deletions. For example, PDB

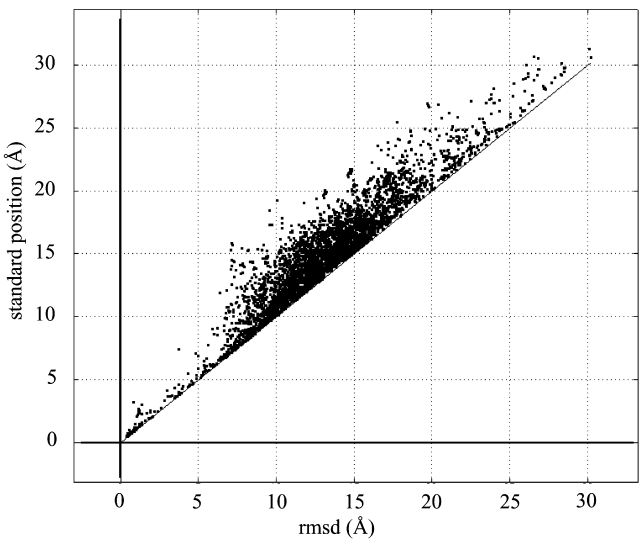

Fig. 2. Rmsd (Å) of pairwise superpositions vs. standard position $F(A,B)$. Solid diagonal line indicates equality of the two measures.

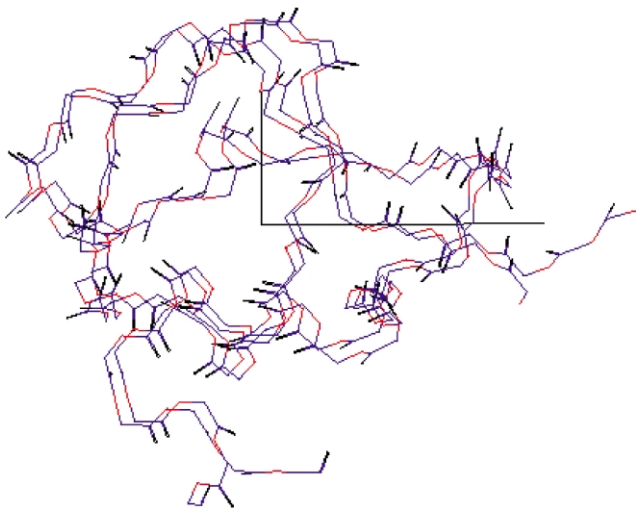

Fig. 3. Standard position of 1OPS and 6MSI. Positive x and y axes drawn as long and short lines, respectively. Note the two extra N-terminal residues of 6MSI relative to 1OPS, located at far right just beyond the horizontal line representing the x axis.

entry 1OPS is clearly recognizable as having conformationally similar residues in similar standard positions compared to 6MSI in spite of missing two N-terminal residues and having only 56% sequence identity (Fig. 3).

4. Conclusions

The standard positioning scheme based on the Haar transform has a remarkably good correspondence to the usual rmsd based on pairwise comparison, at least for a large sample of typical globular protein structures. While this is not a good way to detect similarity between a short chain structure and some small subset of a much larger one, it deals well with small insertions and deletions. Moreover, it produces a useful conformation space for a single protein's different conformations that is otherwise discussed only vaguely in protein folding studies.

References

[1] Yang A-S, Honig B. J Mol Biol 2000;301:665–78.
[2] Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. Protein Sci 1992;1(3):1691–8.
[3] Holm L, Sander C. J Mol Biol 1993;233:123–38.
[4] Akutsu T. IEICE Trans Inf Syst 1996;E79-D:1–8.
[5] Falicov A, Cohen FE. J Mol Biol 1996;258:871–92.
[6] Rao KR. Discrete cosine transform: algorithms, advantages, and applications. Boston: Harcourt Brace Jovanovich; 1990.
[7] Crippen GM, Maiorov VN. J Mol Biol 1995;252:144–51.
[8] Carson AM. J Comp-Aided Mol Des 1996;10:273–83.
[9] Mandell AJ, Selz KA, Shlesinger MF. Physica A 1997;244:254–62.
[10] Hirakawa H, Muta S, Kuhara S. Bioinformatics 1999;15(2):141–8.
[11] Murray KB, Gorse D, Thornton JM. J Mol Biol 2002;316:341–63.
[12] Daubechies I. Ten lectures on wavelets. Philadelphia: SIAM; 1992.
[13] Haar A. Math Ann 1910;69:331–71.
[14] Kabsch W. Acta Crystallogr 1978;A34:827–8.

[15] Diamond R. Protein Sci 1992;1:1279–87.
[16] Zucker M, Somorjai RL. Bull Math Biol 1989;51:55–78.
[17] Maiorov VN, Crippen GM. Proteins: Struct Funct Genet 1995;22: 273–83.
[18] Crippen GM, Havel TF. Distance geometry and molecular conformation. New York: Research Studies Press (Wiley); 1988.
[19] Crippen GM, Ohkubo YZ. Proteins: Struct Funct Genet 1998;32: 425–37.