Compressing Protein Conformational Space

Malik Magdon-Ismail‡, Yu Shao†, Daniel Freedman‡,
Mohammed Zaki‡, Srinivas Akella‡ & Chris Bystroff†
Department of Biology†, Department of Computer Science‡,
Rensselaer Polytechnic Institute, Troy, NY 12180

{shaoy,bystrc}@rpi.edu, {magdon,freedd,zaki,sakella}@cs.rpi.edu

Introduction

Compressing conformational space is the process of defining a subspace of minimal dimensionality where any point may represent a protein-like structure. This is similar to the problem of image compression, where it is desired to reconstruct an image from a small amount of information. In this case, the similarity (in atomic detail) between a true protein and the protein like structure obtained by projecting the compressed protein back into real space is the measure of the success of the compression algorithm. If proteins may be accurately compressed to a space that is efficiently searchable, and then decompressed back to real space, existing energy functions that use atomic detail may finally be rigorously tested in an exhaustive conformational search. (Note: Unlike in the threading approach, such an exhaustive simulation search may consider the folding pathway, and therefore the folding kinetics.) In this paper, we apply compression techniques to various representations of the proteins of known structure. We apply principle component analysis (PCA) to the coordinate, backbone angle, and distance matrix representations. Additionally, we applied Fourier transform techniques to distance matrix space. The success of the compression was measured by the structural difference between the original and the reconstructed coordinates for proteins that were not used in the development of the compression algorithm. It is found that some representations of the model are more easily compressed that others. We find, unexpectedly, that the models that retain the most atomic detail may be compressed to the smallest subspace.

Methods

The general premise of compression is that the actual space in which the objects of interest (in our case proteins) reside is a lower dimensional manifold of the representation. The benefits of identifying this lower dimensional manifold are many. For example, sampling the manifold can be done more efficiently than sampling the entire space, and discrimination techniques will perform more accurately when highly correlated dimensions are discarded. For both of these reasons, we are interested in compressing protein conformational space, and here we briefly describe two linear techniques, principle component analysis (PCA) and the Fourier transform (FT).

1. Principle Components Analysis (PCA)

The goal of PCA is to identify a lower dimensional linear subspace that contains as much of the variance in the data set as possible (Bishop, 1995). Let $\{\{\mathbf{x}_i\}_{i=1}^N$e the N proteins in the PDB. Suppose that the x_i have a mean of zero (this can always be ensured by subtracting the mean in the event that it is not already zero).

The mathematical formulation of the problem is to find a set of K directions - the PCA directions such that the projection of the vectors x_i onto the space spanned by these K directions is as close to the original x_i as possible in the mean squared error. Letting the K directions be given by the K unit vectors $\{\mathbf{y}_j\}_{j=1}^K$, this reduces to the following optimization problem

$$\underset{\mathbf{y}_i}{\text{maximize}} \sum_{i=1}^{K} \mathbf{y}_i^T \Sigma \mathbf{y}_i \qquad \text{subject to the constraint} \qquad \mathbf{y}_i^T \mathbf{y}_j = \delta_{ij} \tag{1}$$

where $\Sigma = \frac{1}{N}\sum_{i=1}^{M} \mathbf{x}_i \mathbf{x}_i^T$ is the covariance matrix for the x's and δ_{ij} is the Kronoeker δ function that equals 1 when i=j and zero otherwise. This optimization problem can be solved by choosing the y_j to be the eigenvectors of $\sum$ with the K largest eigenvalues. In algorithmic form, the steps are as follows

1. Perform a translation to obtain zero mean vectors: $\mathbf{z}_i = \mathbf{x}_i - \frac{1}{N}\sum_i \mathbf{x}_i$.
2. Compute the covariance matrix: $\Sigma = \frac{1}{N}\sum_i \mathbf{z}_i \mathbf{z}_i^T$.
3. Obtain the eigen-vector matrix of $\sum$ which we label ρ, the columns of which are the eigenvectors of $\sum$. Let the eigenvalue corresponding to eigenvector y_i be given by λ_i and assume that the eigenvalues are sorted in decreasing order. The $\{\mathrm{y}_i\}$ can be chosen to be orthonormal, and for any $0 < i \leq K$, $\mathbf{y}_i{}^T \Sigma \mathbf{y}_i = \lambda_i$.
4. Thus the maximal amount of variance that can be picked up by at most K eigen directions is achieved by taking the first K eigen-directions. Usually one continues to add eigen directions until the percentage of variance accounted for exceeds a threshold.
5. The compressed data point is given by the K projections onto the K PCA directions chosen, and thus is a K dimensional vector. The reconstruction back in the original space based upon these K projections is given by

$$\mathbf{z}_i' = \mathbf{Y}_K \mathbf{Y}_K^T \mathbf{z}_i$$

where Y_K is a matrix whose columns are given by the K PCA eigenvectors. The smaller K is, the greater the compression. Compression is useful only if the discarded dimensions (information) is non-essential.

We use the **reconstruction error** to measure the success of the compression. The reconstruction error is given by

$$RMSD \;=\; \sqrt{trace[(\mathbf{I} - \mathbf{Y}_K \mathbf{Y}_K{}^T)\Sigma_{\mathbf{z}}] + (\mu - \mu_{\mathbf{z}})^T (\mathbf{I} - \mathbf{Y}_K \mathbf{Y}_K{}^T)(\mu - \mu_{\mathbf{z}})} \tag{2}$$

2. Fourier Transform (FT)

The second class of compression algorithms which we implemented was based on the Fourier transform, analogous to many image compression techniques. A distance matrix may be thought of as a digital image, in which the distance plays the role of intensity. As a result, standard schemes from image processing may be brought to bear on the problem. Using FT analysis of distance matrices, we can identify protein-like features that manifest themselves as periodicities in the inter-residue distances. If $d(l,j)$ denotes the entries in the distance matrix between amino acid l and amino acid j, then the discrete Fourier coefficients are given by

$$F(h,k) = \frac{1}{N}\sum_{l,j=1}^{N} d(l,j) e^{\frac{2\pi i h l}{N}} e^{\frac{2\pi i k j}{N}}$$

where $i = \sqrt{-1}$. The reason that these coefficients are useful is because knowing the Fourier coefficients, one can reconstruct the distance matrix as follows

$$d(l,j) = \frac{1}{N}\sum_{h,k=1}^{N} F(h,k) e^{-\frac{2\pi i h l}{N}} e^{-\frac{2\pi i k j}{N}}$$

Thus, given the Fourier coefficients, one can reconstruct the distance matrix and vice versa. Compression can now be achieved by constructing the Fourier coefficients from the distance matrix and then ignoring (i.e., setting to some mean value, usually chosen to be zero) some subset of the Fourier coefficients. The remaining Fourier coefficients can now be used to reconstruct the distance matrix. If the ignored Fourier coefficients are negligible then the resulting reconstructed distance matrix should be close to the original one. We can compute a reconstruction error by taking the RMSD between the reconstructed distance matrix and the original one.
Usually the best compression is obtained by ignoring the highest order Fourier coefficients or the ones with smallest variance, where the variance can be determined on the training set.

Results

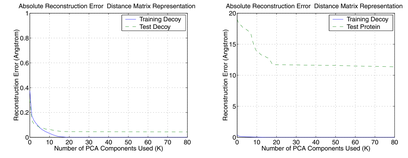
We performed compression experiments on non-overlapping 40 and 60 residue segments of proteins from the PDBselect (Hobohm & Sander, 1994) non-redundant database. We also generated decoy proteins using HMMSTR (Bystroff *et al.*, 2000), a stochastic model for sequence and local structure, and ROSETTA (Simons *et al*, 1997), a program for simulating protein folding by fragment insertion Monte Carlo. The "decoy" proteins may be thought of as sampling from a stochastic model. The compression of decoy data may tell us how to more efficiently sample the same space.

Both data sets were randomly split into training and test sets, the training set was used to develop the compression scheme which was then tested by computing the reconstruction error on various
data sets. We also tested our compression methods with three different structure representations: distance matrix (DM), angular (ANG), and coordinate (XYZ). The quantitative results along with the detailed methodology are presented in what follows.

Training set	Test set	Comp. scheme	Rep.	Experimental Result
Decoy	Decoy	PCA	DM	Compressible conformational space. Compression factor of about 20.
Decoy	PDB	PCA	DM	Compression not very succesful. The first few decoy PCA directions seem to contain significant information about PDB conformational space, but the rest appear to be random directions.
PDB	PDB	PCA	DM	Compressible conformational space. Compression factor of about 8.
PDB	PDB	PCA	XYZ	Compressible conformational space. Compression factor of about 6.5.
PDB	PDB	PCA	ANG	Not compressible.
PDB	PDB	FT	DM	Compressible conformational space. Compression factor of about 2.

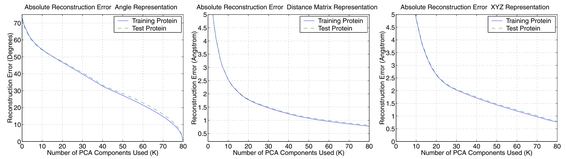
1. Principle Component Analysis (PCA)

A test set of size 50 was sampled from the entire data base, and the remaining data was used to develop the compression method. The reconstruction error is plotted as a function of the number of PCA directions used for reconstruction. The training and test reconstruction errors are both averaged over 1000 runs. Figure 1 shows the result when the decoy structures are used to develop the compression scheme and Figure 2 shows the result when the PDB structures are used to develop the compression scheme.



(a) (b)

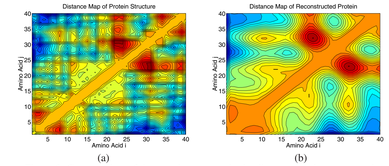
Figure 1: Reconstruction error using PCA derived from decoy structures of length 40 amino acids in the DM representation. (a)Test set: Decoy structures; (b) Test set: PDB structures.



(a) (b) (c)

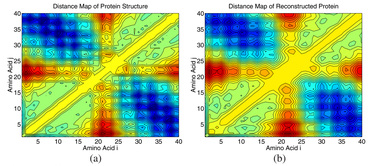
Figure 2: Reconstruction error using PCA derived from PDB structures of length 40 amino acids. (a)Distance matrix representation. (b) Coordinate representation. (c) Angular representation.

A more intuitive picture of what the compression techniques are doing can be seen by comparing the original and reconstructed test structures (Figures 3, 4 and 5). From the figures it is apparent that with fewer PCA directions used in the reconstruction, the reconstructed structures are ``smoother'' than the original. What appears to be happening is that the lower order, or base structure is picked up by the first few PCA directions and the higher order detail gets filled in gradually by the higher PCA directions. Thus, the compression successfully finds a sub manifold that ``represents'' the structure although it does not pick up all of the detail.



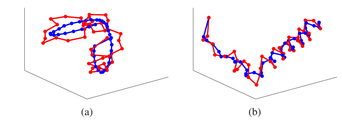
(a) (b)

Figure 3: Contour plots of the protein structure in the distance matrix representation. (a) Original distance matrix. (b) Reconstruction based on 20 PCA directions.



(a) (b)

Figure 4: Contour plots of the protein structure in the distance matrix representation. (a) Original distance matrix. (b) Reconstruction based on 40 PCA directions.

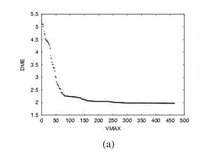


(a) (b)

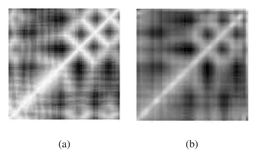
Figure 5: 3D structures in the coordinate representation. The original (red) and reconstructed (blue) are shown on the same axes. (a) Reconstruction based on 20 PCA directions. (b) Reconstruction based on 40 PCA directions.

2. Fourier Transform (FT)

In a preliminary experiment we compressed distance matrices for 60 residue decoys by Fourier transforming them and then removing all but the N low-order Fourier coefficients. Comparing the back-transformed image to the original, it was found that the original distance matrix could be faithfully reconstructed (2.5A) using as few as N=200 Fourier coefficients, regardless of the shape of the original structure. Fourier termination errors were minimized by arranging the distance matrix image in a p4mm 2-D space group (Figure 7). All Fourier coefficients are real.



(a)



(b)

Figure 6: The dme for a protein versus the number of PFC's used in reconstruction(Vmax). (b) The original distance matrix. (c) The reconstructed distance matrix using 80 PFC's.

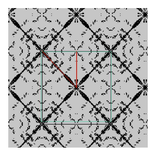


Figure 7: p4mm group.

In a second experiment, the variance in each Fourier coefficient (F) was summed over 12,000 transformed 60-residue decoys. The low-variance F's represented relatively invariant periodicities in distance. A higher degree of compression was obtained by back-transforming using only the N most variable F's (the principle Fourier coefficients, PFC's), and fixing the others to their mean values (Figure 6). Using $N = 80$, an average reconstruction error of 2.5A was obtained. Absolute, rather than relative, variance was the best measure of relative importance in reconstruction. The approach of extracting distance periodicities by FT was thought to capture the overall size of the molecule and the $|d(l,j) - d(l, j+1)| < 3.8$ distance constraint. The characteristic size of compact protein-like 60-mers would have a corresponding characteristic reverse turn frequency, which would manifest itself in invariant low-order F's. The persistence length of polypeptides also restricts the allowable high-frequency periodicities, since $d(l, j+2)$ would be relatively invariant, therefore $|d(l,j) - d(l, j+2)|$ would also be bounded, and high order F's would be small. In fact, low order invariant F's were not found, but of the high order F's, those directed along the diagonal had relatively higher variance, reflecting (in reciprocal space) the predominantly diagonal features of protein contact maps (Figure 6 (b)) caused by beta sheets. Along the axes, the F's with periodicities of 7 to 10 residues were the most variant, perhaps reflecting the vertical and horizontal stripes caused by alpha helices.

References

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Bystroff, C., Thorsson, V. & Baker, D (2000) Hmmstr: a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol*, **301** (1), 173-90.

Hobohm, U. & Sander, C. (1994) Enlarged representative set of protein structures. *Protein Sci*, **3** (3), 522-4.

Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and baysian scoring functions. *J Mol Biol*, **268** (1), 209-25.