Compression of Protein Conformational Space

Yu Shao[1], Malik Magdon-Ismail[2], Daniel Freedman[2], Srinivas Akella[2], Mohammed Zaki[2], and Chris Bystroff[1]

Keywords: Folding, compression, principal components analysis (PCA), Fourier.

1 Abstract

Protein conformational space is large. A folding polypeptide cannot sample all of the possible combinations of the $2N$ backbone angles, but instead explores a small sub-space defined by the energetics of the system. Some final structures representing folded proteins are stored in the Protein Data Bank (PDB). A novel approach to the protein folding problem would be to define a (smaller) space in which the conformational search is possible, then find an energy function in that space that identifies the correct structure, given the sequence. By reducing the system to a minimal set of features, we increase the likelihood that an exhaustive search in feature space would be computationally feasible.

One way to identify this smaller space is to use compression techniques to obtain a subspace of minimal dimensionality where any point may represent a protein-like structure. 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. One such measure is the distance matrix error, dme, which is the root-mean square difference between two distance matrices[3] . The dme correlates with the more familiar root-mean-square distance (rmsd) summed over superimposed coordinates. A good compression will produce a low average dme with as few parameters as possible. 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 tested in an exhaustive conformational search.

For our study, we generated a large number of protein-like decoy structures of length 60 residues using a hidden Markov model for local sequence-structure correlations (HMMSTR[1]) and a Monte Carlo simulation for assembling protein fragments (ROSETTA[2]). Distance matrices were compressed using two approaches: principal component analysis (PCA) and discrete Fourier transforms (FT). We conclude that decoy protein-like structures are compressible. Is the same true of real proteins? Can the compression schemes learned on decoys be applied to true proteins? These issues will be discussed in future research.

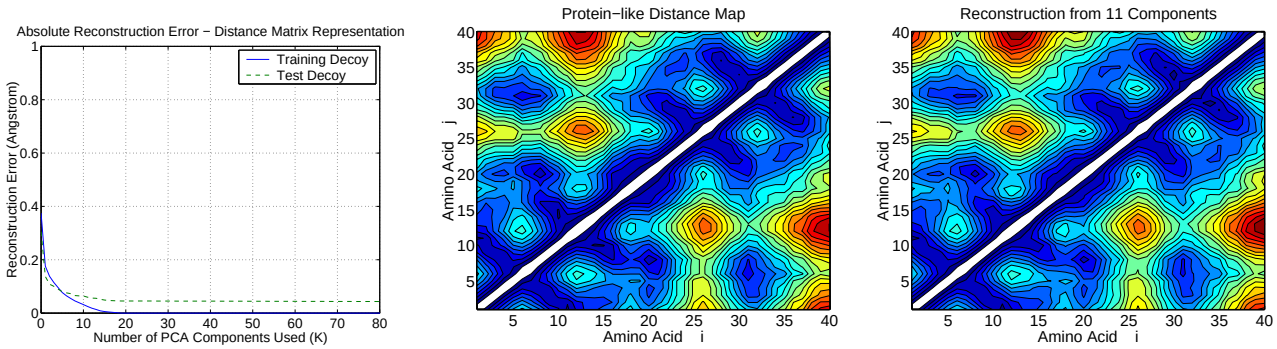
1.1 Principal Component Analysis (PCA)

We constructed a set of 10000 non-overlapping decoy-protein sequences of length 40 amino acids This data set was used to compute the principal component directions and the reconstruction error using the most important (highest variance) directions is illustrated in Figure 1 (a). The test reconstruction error is computed on data that was not used in developing the PCA's. A contour plot of the distance matrix of a typical decoy structure before and after reconstruction using 11 PCA components is also shown (Figure 1(b) and (c)). In this 780-dimensional space, we find that 99.5% of the variance in the distance matrices could be explained in a 11-dimensional linear subspace.

[1]Dept. of Biology, RPI, 110 8th Street, NY 12180. {shaoy,bystrc}@rpi.edu

[2]Dept. of Computer Science, RPI, 110 8th Street, NY 12180. {magdon,freedd,sakella,zaki}@cs.rpi.edu

[3]The distance matrix is a representation of the 3D structure of a protein that uses the set of internal distances.

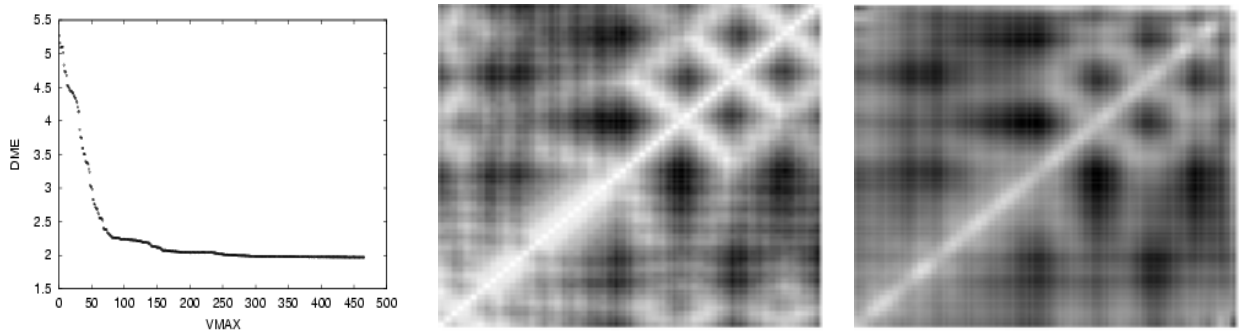


(a) (b) (c)

Figure 1. (a)The dme as the number of principal components increases. (b) Typical distance map before compression, and (c) after decompression using 11 principal components.

1.2 Fourier Transform (FT)

FT-based algorithms, such as JPEG, have been developed primarily for the purposes of image compression. A distance matrix may be thought of as a digital image, in which the distance plays the role of intensity. Preliminary experiments have been performed which compressed distance matrices by Fourier transforming them and then back-transforming using only low-order Fourier coefficients. As few as 200 Fourier coefficients could consistently reconstruct the distance matrices (average dme=2.5Å). In a second experiment, principal Fourier components (PFC's) were chosen as the highest variance Fourier coefficients after transforming 12,000 Rosetta-generated 60 residue structures. Setting the low-variance coefficients to their mean values and varying only 80 PFCs allowed accurate reconstruction of the distance matrices (average dme=2.5Å)



(a) (b) (c)

Figure 2. (a) The dme for a typical protein versus the number of PFCs used in reconstruction (Vmax). (b) The original and (c) the reconstructed distance matrix using 80 PFCs.

References

[1] Bystroff, C., Thorsson, V. & Baker, D. 2000. HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology* 301, pp. 173-90.

[2] 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 Bayesian scoring functions. *J Mol Biol* 268, pp. 209-25.