Compressing Protein Conformational Space

Malik Magdon-Ismail[*] Yu Shao [†] Daniel Freedman[‡]
Christopher Bystroff[§]

Abstract

Motivation: A conformational search to determine protein structure is unrealistic for large proteins. We study the extent to which protein conformational space is compressible.

Results: We demonstrate that most representations of protein conformational space are redundant, above and beyond translational and rotational invariance. Using some well known compression techniques, principle components analysis (PCA) and Fourier transforms, we show that the 3D coordinates of protein fragments of length 40 may be compressed from a 3N-3=117 dimensional space into a 10 to 20 dimensional space, and accurately reconstructed back from the compressed space. Compression of the highly non-linear 2N-2=78 dimensional torsion space representation of the same set was unsuccessful. Compression of distance matrices having N(N-1)/2=780 dependent variables yielded interesting results when reduced to 20-40 dimensions. The ability to use a much smaller representation might enable conformational searches that were previously infeasible because of high dimensionality and non-linearity. Applications and improvements are discussed.

Availability: Implementations (some in MATLAB and some in C) may be obtained via email.

Contact: `magdon@rpi.edu; shaoy@rpi.edu;freedd@cs.rpi.edu; bystrc@rpi.edu`

Key words: Folding, compression, principle components analysis (PCA), Fourier.

1 Introduction

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. The final structures representing folded proteins are stored in the Protein Data Bank (PDB). Many energetic and geometric functions exist that potentially explain the sequence-dependent structures that proteins can adopt (Liwo *et al.*, 1999), but these functions

[*] Dept. of Computer Science, RPI, 110 8th Street, Troy, NY 12180. `http://www.cs.rpi.edu/ magdon`

[†] Dept. of Biology, RPI, 110 8th Street, Troy, NY 12180.

[‡] Dept. of Computer Science, RPI, 110 8th Street, Troy, NY 12180. `http://www.cs.rpi.edu/ freedd`

[§] Dept. of Biology, RPI, 110 8th Street, Troy, NY 12180. `http://isites.bio.rpi.edu/bystrc/`

depend on models with atomic detail, and therefore cannot be tested on large polypeptides because the conformational space is too large to be searched in atomic detail.

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 correctly identifies the correct structure, given the sequence. This is conceptually similar to the "threading" approach, (Smith *et al.*, 1997), where the search space is essentially the database of known proteins, and an energy function is determined which correctly assigns sequence to structure. However, threading, or fold recognition, ignores the physical process of folding. Since it is well known that the folding protein does not sample all possible conformations, there is no reason to believe that an energy function exists that solves the descrimination problem.

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 exhautive 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 priciple 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 recontructed 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.

We have found no prior work that addresses the compressibility of conformational space. Work along the lines of compressing the linear structure of proteins has been addressed, see for example (Nevill-Manning & Witten, 1999). The paper is organized as follows. First we discuss protein structure representation and the data that is available. Following this we briefly overview the compression techniques and give the results of our simulations. We conclude with some remarks on improvements and the use of such techniques to perform conformational searches.

2 Protein Structure Representation

There are at least three ways to represent the 3D structure of a protein backbone at atomic resolution: as 3-dimensional coordinates, as a set of inter-residue distances, or as a set of backbone torsion angles. Some of these representations are highly redundant in terms of the information content, while others are compact. The most compact representation will not necessarily be the best for compression, and in fact we will see later that the opposite is true. A representation that is redundant but manages to somewhat linearize the space will be more useful for linear compression techniques.

The most compact representation is the angular one, where the position of the $l+1^{\text{th}}$ amino acid relative to the l^{th} amino acid is specified by the two ϕ, ψ angles. This results in a total of $2N-2$ angles where N is the number of amino acids in the protein. All translational and rotational invariances have already been incorporated into this representation.

The next representation is the set of 3D coordinates for each amino acid, a total of $3N-3$ parameters. But these are highly redundant in their information content. For example, the coordinates of the $l+1^{\text{th}}$ residue are highly dependent on the coordinates of the l^{th} residue.

The last representation we consider is the distance matrix, where one specifies each of the $N(N-1)/2$ unique distances, $d(l,j)$ between amino acid l and amino acid j in the sequence. These numbers can be represented in a symmetric matrix. All translational and rotational invariances are built into this representation, however, there are many other geometric constraints that such a matrix must satisfy, such as the triangle inequality amongst every 3 amino acids and bounded distances between neighboring amino acids.

We will consider the compressibility of protein conformational space in each of these representations.

2.1 A Non-Redundant Sample of Constant Length Natural Protein Fragments

The database of known protein structures, PDB (http://www.rcsb.org; (Berman *et al.*, 2000)), contains a wealth of structural information. However, it is highly redundant, having many representatives of a few evolutionary families. A non-redundant set of structures was extracted from this set, PDBselect (Hobohm & Sander, 1994) where the evolutionary similarities have been factored out. After removing membrane proteins, metal-binding proteins and proteins with many disulfide links (these classes of proteins have their own characteristic structures), the list contains 691 globular proteins having a total length of 120,000 amino acid residues.

For the purposes of this work the input data must be of a constant length. Therefore the protein chains were divided into non-overlapping 40 or 60 residue pieces. The compression experiments described below on "proteins" used this non-redundant, non-overlapping set. Polypeptide chains in this set are not necessarily compact nor are they expected to be in a low energy state by themselves.

2.2 HMMSTR and ROSETTA: Protein-Like Decoy Structures

In an attempt to create a large dataset of protein-like structures that is truly non-redundant and representative of a physical model, we built protein structures using a combination of two, well-established statistical models. Unlike natural proteins, these "decoys" cannot be related by evolution. They are stochastically generated representative samples of the physical model used to generate them. They are also of a constant length (60) and have been energy minimized at that length.

HMMSTR, (Bystroff *et al.*, 2000), is a hidden Markov model (HMM) for generalized protein sequence. Generalized HMMs are directed, cyclic graphs where each node is a single symbol emitter. HMMSTR is a "parallel HMM" which emits, from one Markov state, a single amino acid and a symbol for the backbone ϕ and ψ angles. State pathways represent all known local structure motifs, as defined in the I-sites Library (Bystroff & Baker, 1998) They are represented in the model in proportion to the frequency at which they are found in the database of protein structures.

ROSETTA is a folding simulation algorithm that uses the fragment insertion Monte Carlo approach (Simons *et al.*, 1997). The promise of this method has been demonstrated in blind ab initio protein structure predictions as part of the CASP experiments (Moult *et al.*, 1995), correctly predicting protein fragments of up to 107 residues in length with an accuracy of 5Å root-mean-square deviation in superimposed alpha-carbon coordinates (RMSD) (Bonneau *et al.*, 2001). A "fragment insertion" move consists of selecting an insertion point in the target, then selecting a fragment at random from a list associated with that location. The backbone angles of the fragment are inserted, and the new coordinates are computed, then accepted or rejected, depending on ROSETTA's knowledge-based energy function (Simons *et al.*, 1999).

3 Compression 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).

3.1 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. A more detailed discussion can be found in (Bishop, 1995). Let $\{\mathbf{x}_i\}_{i=1}^N$ be the N proteins in the PDB. Suppose that the $\mathbf{x}_i$ have a mean of zero (this can always be ensured by subtracting the mean in the event that it is not already zero). The details do not depend on which specific representation we choose to use.

The mathematical formulation of the problem is to find a set of (K) directions - the PCA directions such that the projection of the vectors $\mathbf{x}_i$ onto the space spanned by these K directions is as close to the original $\mathbf{x}_i$ 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 $\mathbf{x}$'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 $\mathbf{y}_j$ to be the eigenvectors of Σ 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 Σ which we label ρ, the columns of which are the eigenvectors of Σ. Let the eigenvalue corresponding to eigenvector $\mathbf{y}_i$ be given by λ_i and assume that the eigenvalues are sorted in decreasing order. The $\{\mathbf{y}_i\}$ 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 $\mathbf{Y}_K$ 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 (informtion) is non-essential.

3.1.1 Reconstruction Error

To determine whether the compression is successful, we take a data set, compress it according to the compression scheme, and then reconstruct it from the compressed space back into the original space. If the compression was perfect, the resulting data set will be identical to the original one. However, more often than not, some information will be lost in the compression, and so there will be some reconstruction error. We can analyse this error. Suppose that $\mathbf{x}$ is the orriginal data point and $\mathbf{x}'$ is the reconstructed one. Then the squared error can be defined as $(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')$. The average value of this quantity over the entire data set can be used as a measure of the reconstruction error. Taking the square root, one then arrives at the root mean square deviation,

$$RMSD = \sqrt{\frac{1}{N}\sum_{\mathbf{x}}(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')}$$

. A question arises as to what data set we should we use to compute the reconstruction error (RMSD). A data set was used in developing the compression method. This data set could be used to compute the reconstruction error in which case, we have computed what can be called the *training* reconstruction error. However, though the value computed may be a useful measure of the compression, this might be an overly optimistic estimate of the true success of the compression, analogous to trying to predict the weather in the past. One can always learn from past data how to predict rainy from sunny days, however the true measure of success is to be able to predict the weather in the future. Thus a more effective measure of the performance is the reconstruction error on a data set that was never used in the development of the compression scheme. In this case we have computed the *test* reconstruction error.

One can compute a formula for the reconstruction error on an arbitrary data set. For completeness, we provide the formula here, without derivation. Suppose that the (arbitrary) data set for which we wish to compute the reconstruction error is given by $\{\mathbf{z}_i\}_{i=1}^N$, and let the mean of this data set be $\mu_{\mathbf{z}} = \frac{1}{N}\sum_i \mathbf{z}_i$ and let the covariance matrix be $\Sigma_{\mathbf{z}} = \frac{1}{N}\sum_i (\mathbf{z}_i - \mu_{\mathbf{z}})(\mathbf{z}_i - \mu_{\mathbf{z}})^T$. Then the reconstruction error is given by

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

where μ is the mean for the data that the PCA directions were developed from and $\mathbf{Y}_K$ is a matrix that was defined earlier. If the training and test data sets were sampled from the same distribution the two means will be approximately the same and hence the second term is negligible and can be ignored.

3.2 Fourier Transform (FT)

The second class of compression algorithms which we implemented was based on the Fourier transform, analogous to many image compression techniques (for example the JPEG algorithm), see for example (Anderson & Huang, 1971). 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

$$F(h,k) = \frac{1}{N}\sum_{l,j=1}^{N} d(l,j) e^{\frac{2\pi i h l}{N}} e^{\frac{2\pi i k j}{N}} \tag{3}$$

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

$$d(l,j) = \frac{1}{N}\sum_{h,k=1}^{N} F(h,k) e^{-\frac{2\pi i h l}{N}} e^{-\frac{2\pi i k j}{N}} \tag{4}$$

Thus, given the Fourier coefficients, on 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.

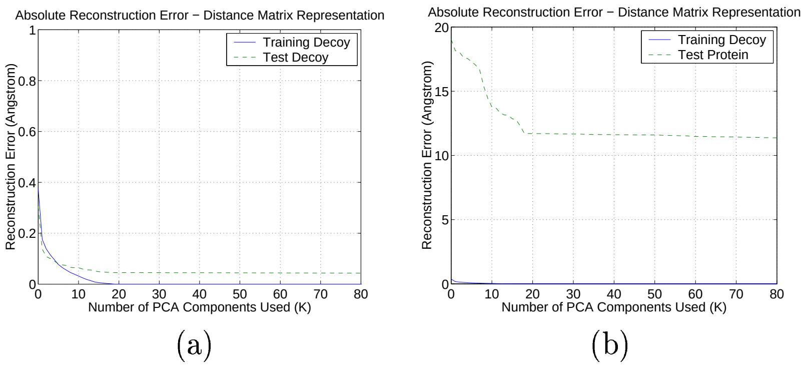


Figure 1: Reconstruction error using PCA derived from decoy structures of length 40 amino acids in the DM representation. (a)Test set: Decoy structures; (b) Test set: PDB structures.

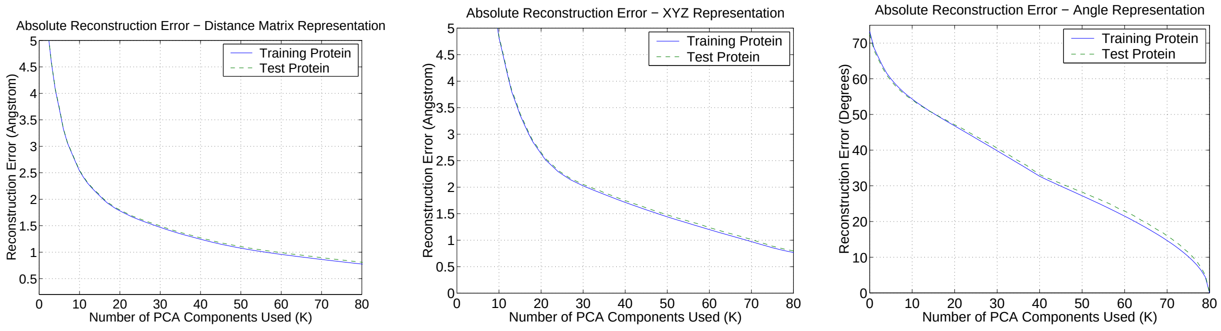
4 Results

We performed experiments on proteins in the PDB as well as the decoy proteins generated using ROSETTA. Two data sets were constructed from the PDB. One containing structures for non-overlapping sequences of length 40 amino acids and one with non-overlapping sequences of length 60 amino acids. We report only the results on the data set with protein sequences of 40 amino acids, since these are more statistically significant.

These 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. The following table illustrates the extent of the results that we report.

Training set composed of:	**Test set composed of:**	**Desired Goal:**
Decoy Proteins	Decoy Proteins	Conformational space of decoy proteins is compressible.
Decoy Proteins	PDB Proteins	Compression methods developed on decoy proteins can be used to compress true proteins.
PDB Proteins	PDB Proteins	Conformational space of PDB proteins is compressible.

In addition to the above three types of experiments, a further bifurcation occurs depending on the type of representation used for the proteins - the three representations for which we report results are the *distance matrix (DM)*, *angular (ANG)*, and *coordinate (XYZ)* representations. We present a qualitative survey of the results in the table below. The quantitative results along with the detailed methodology are presented in what follows.



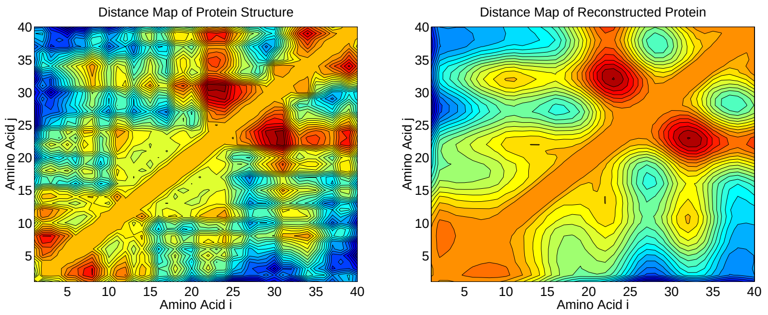
(a) (b) (c)

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.

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.

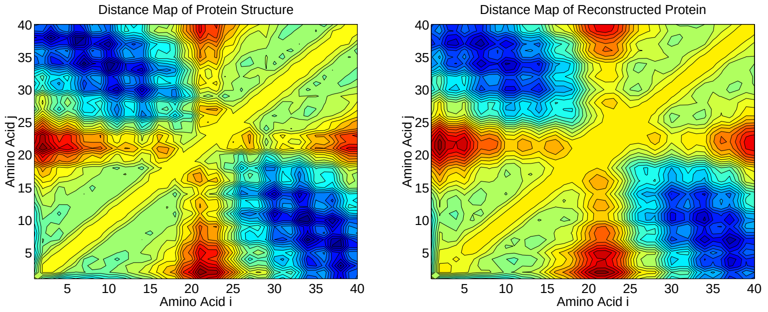
4.1 Principle Component Analysis (PCA)

A test set of size 50 was sampled from the entire data base, and the remaining data was used to develop the compression method. Hence, the compression scheme was independent of the test set, thus, the test reconstruction error is an unbiased estimate of the compression performance. The quantitative results are represented in Figures 1 and 2. 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. For example from Figure 2 (a) we can read off that using 10 PCA components to do the reconstruction, the average error incurred in each distance matrix entry is 2.5 Angstrom.



(a) (b)

Figure 3: Contour plots of the protein structure in the distance matrix representatoin. (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 representatoin. (a) Original distance matrix. (b) Reconstruction based on 40 PCA directions.

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 succesfully finds a sub manifold that "represents" the structure although it does not pick up all of the detail.

4.2 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 (2.5Å) using as few as N=200 Fourier coefficients, regardless of the shape of the

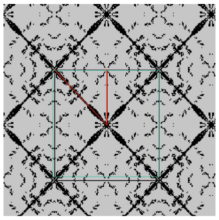
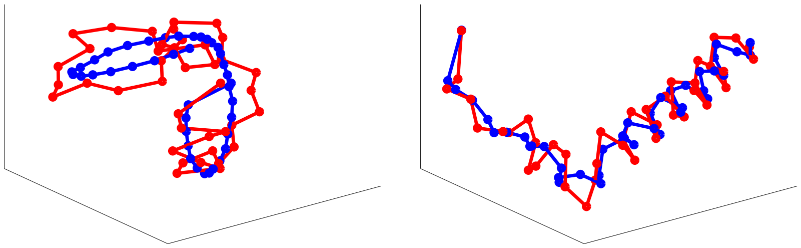
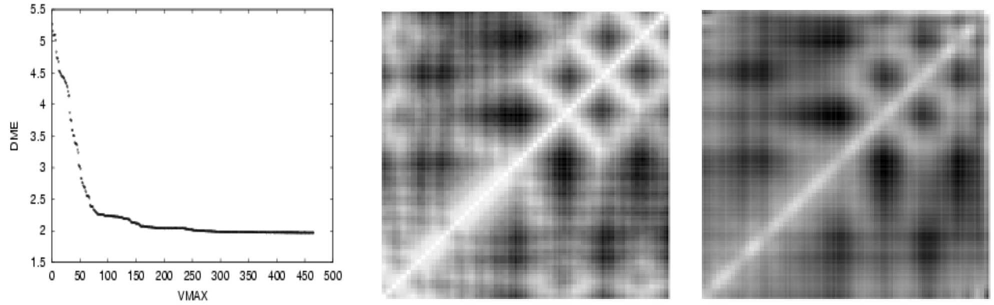


Figure 7: p4mm group.



(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.



(a) (b) c

Figure 6: (a) 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.

original structure. Fourier termination errors were minimized by arranging the distance matrix image in a p4mm 2-D space group (Figure 7), thus ensuring that the Fourier coefficients are real[1]

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 2.5Å was obtained. Accurate 3D structures may be readily recovered from these distance matrices using distance geometry methods (Aszodi *et al.*, 1997). 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 $|d(l,j)-d(l,j+1)| \leq 3.8$Å distance contraint. 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

[1] The Fourier equations (3 and 4) will then have an internal summation over the 8 space group symmetry operators.

persistence length of polypeptides also restricts the allowable high-frequency periodicities, since $d(l, l+2)$ 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.

5 Discussion and Future Work

Our work addressed the redundancy in conformational space that may explain the apparent dilemma regarding how nature appears to search a huge conformational space in order to fold proteins in real time. The way in which this may be resolved is that proteins have characteristic stable substructures that are recurrent in nature. The geometric constraints imposed by these substructures define a lower dimensional manifold in conformational space, in which all proteins reside. Thus, nature only needs to sample this manifold. Our approach was to assume the existence of such a manifold and attempt to identify it in a data driven manner, rather than from a first principles approach. We used both PCA and FT, which are equivalent in that they are both linear. However, they exploit different properties of the data. PCA detects high variance patterns, while the FT attempts to extract periodicities in structure.

Our results indicate that in certain representations, a significant compression to a linear submanifold can be achieved, indicating for example that while a 40 amino acid protein lives in an 80 dimensional space, we can represent it in a 10-20 dimensional space. By this we mean that, in this significantly lower dimensional representation, we can recover within an acceptable error the original structure of the protein fragment.

One of the many challenges remaining is to assign a sequence-dependent energy to a point in the compressed space. This may be possible if the reconstructed chain is at or close to atomic resolution, where energy calculations are most likely to be accurate. The most successful compression converts a polypeptide to a smooth trace through the chain. The reconstruction violates certain stereochemical constraints of peptides. By enforcing these constraints upon reconstruction (for example that the l to $l+1$ distance is a known constant, 3.8A), it may be possible to recover the structure even more accurately.

We have shown that given a test protein structure, we can recover its base structure from knowledge only of the compressed structure. In a blind test case, we would not know the structure, however we do know that the compressed space is faithful to the true space, and hopefully representative of that space, and therefore believe that the test structure lives in that space. Hence, by sampling in the compressed space, a considerably more feasible operation, we may be able to simulate the search through conformational space more efficiently and perhaps even develop folding pathways in this compressed space. This is the topic of our future research.

References

Anderson, G. & Huang, T. (1971) Piecewise fourier transformation for picture bandwidth compression. *IEEE Transactions on Communications Technology,* **19**, 133–140.

Aszodi, A., Munro, R. E. & Taylor, W. R. (1997) Distance geometry based comparative modelling. *Fold Des,* **2** (3), S3–6.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H. & Westbrook, J. (2000) The protein data bank and the challenge of structural genomics. *Nat Struct Biol,* **7 Suppl**, 957–9.

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

Bonneau, R., Strauss, C. E. & Baker, D. (2001) Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins,* **43** (1), 1–11.

Bystroff, C. & Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol,* **281** (3), 565–77.

Bystroff, C., Thorsson, V. & Baker, D. (2000) Hmmstr: a hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology,* **301** (1), 173–90.

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

Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. (1999) Protein structure prediction by global optimization of a potential energy function. *Proc Natl Acad Sci U S A,* **96** (10), 5482–5.

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins,* **23** (3), ii–v.

Nevill-Manning, C. G. & Witten, I. H. (1999) Protein is incompressible. *Proceedings, Data Compression Conference,* , 257–266.

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** (1), 209–25.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins,* **34** (1), 82–95.

Smith, T. F., Lo Conte, L., Bienkowska, J., Gaitatzes, C., Rogers, R. G., J. & Lathrop, R. (1997) Current limitations to protein threading approaches. *J Comput Biol,* **4** (3), 217–25.