Compact Representation of Protein Surface Patches

Mohammed AlQuraishi

Introduction

Prediction of protein structure from amino acid (AA) sequence is one of the grand challenges of modern biochemistry. While *ab initio* structure prediction is difficult, prediction of local structure is more tractable if experimentally determined structures of homologous proteins exist. The prediction of local patches is useful in predicting interaction partners of proteins and nucleic acids.

The high-dimensionality of this problem has so far prevented machine learning methods from making significant inroads. The space of AA sequences grows exponentially (20^n) and the flexibility of proteins affords them an infinite conformational space. In this project we will focus on the problem of reducing the dimensionality of protein structures. We will tackle this problem in the context of predicting local patches of the DNA-binding region of helix-to-helix (HTH) proteins. We have chosen this protein family due to its biological significance and the relative availability of crystallized HTH-DNA complexes.

Algorithm

Our focus is on reducing the problem's dimensionality by encoding HTH surfaces using an efficient basis set, then performing PCA on the entire family of HTH structures to characterize differences amongst individual family members. Once a significant reduction is achieved (we expect ~20 dimensions), it will be possible to apply machine learning methods to predict local structure from the AA sequence.

1. Binding Patch Extraction

The first step is defining and extracting the region of the HTH responsible for DNA binding. All HTH proteins employ an α-helical structure to bind DNA. Picking the helix closest to the DNA molecule is one way of identifying the correct α-helix. Unfortunately, the region within the helix responsible for DNA-binding varies between proteins and so does the angle at which binding occurs (Figure 1). Since we want to represent the protein surface from the DNA's perspective, we will spatially register all HTH-DNA complexes to bring their DNA molecules into superimposition. After alignment, the centroid of the DNA base closest to the binding helix will be used as the center of a rectangular region which will define the DNA-binding surface patch of the protein.

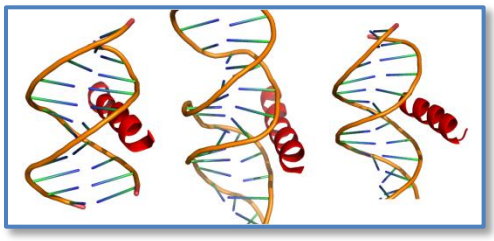

Figure 1: Helices interact with the HTH binding pocket at different angles. (Helices are in red)

1.1 DNA-binding α-helix identification

To identify the DNA-binding helix, the Euclidean distance between every atom in the DNA molecule and every atom in all helices is calculated. The helix with at least n residues that have at least 1 atom within ε of any atom of the DNA molecule is selected as a DNA-binding helix. Letting h_m be 1 if the m^{th} helix binds DNA and 0 if it does not, we have:

$$h_m = 1\left\{ \sum_{\substack{\text{helix}\\ \text{residues}\\ \{i\}}} 1\left\{ \sum_{\substack{\text{residue}\\ \text{atoms}\\ \{j\}}} \sum_{\substack{\text{DNA}\\ \text{atoms}\\ \{k\}}} 1\{\|p_{m\,i\,j} - p_k\| < \epsilon\} \geq 1 \right\} \geq n \right\}$$

Where p_{ijm} is the position of atom j in residue i in helix m, and p_k is the position of atom a in the DNA. Currently n is set to 4 and ε to 400 Å. This successfully identifies the DNA-binding α-helix (Figure 2.)

1.2 3D Registration

The next step is registration of the HTH-DNA complexes, which is equivalent to the pairwise registration of 3D point clouds. Given the complexity of our surface, optimal methods like geometric hashing (1) are too slow. Instead we implemented an algorithm based on local shape descriptors. It works by computing a handful of distinct local features on the surface, and then uses those features to perform the alignment instead of using every point. Many such algorithms exist, and we implemented Gelfand's et al. method (2) for our purposes.

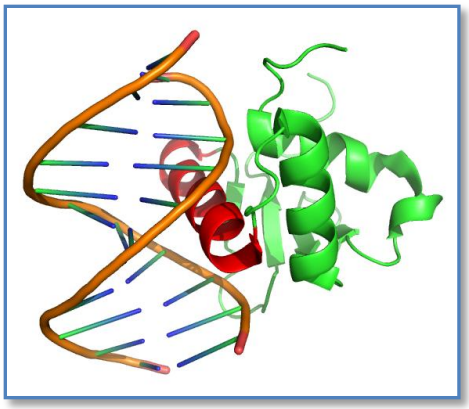

Figure 2: Correctly identified DNA-binding helix is in red, other helices in green.

Gelfand's et al. method employs an integral shape descriptor due to its noise-tolerance, in lieu of differential descriptors such as curvature which are sensitive to noise. For each point in the model, a sphere of radius r is formed around the point, and the volume of the surface inside the sphere is calculated (Figure 3.) Volume calculations are performed by discretizing the space around the sphere, and using a ray shooting algorithm (3) to determine whether each discretized point in space is inside or outside the surface. Volumes of the inside points are summed to approximate the total volume.

Since we only want features that define unique points on the surface, we select the most distinct and rare features. This is done by binning features based on volume, and then picking out features from the least populated (most unique) bins. Features within radius R_e of each other are removed.

To improve robustness, we also filter features to insure that they are persistent across multiple scales, and thus unlikely to have arisen by noise. We do so by computing the descriptors at multiple radii, and then picking features deemed rare at multiple consecutive scales.

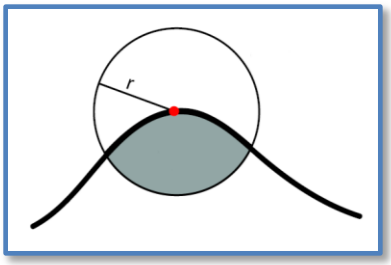

Figure 3: Shaded region represents volume of integral descriptor. Figure adapted from (2).

Descriptors are calculated as described for the model and data surfaces. For each feature in the data surface, we search in the model to find features whose volume is within ε of the data surface feature. This results in a set of putative matches for each point in the data surface. We QT cluster this set (4), and then pick the feature from each cluster that is least different between the model and data surfaces.

For each feature in the data surface we now have a set of putative matches in the model. To perform the registration, we use a branch and bound algorithm to find an optimal correspondence. Assume that k-1 points have already been matched. For each possible kth point, we calculate the global dRMS of the resulting correspondence. Points that result in a dRMS above R_c are pruned from the search. The first point to result in a dRMS below R_c is used to form a new correspondence out of k points. This process is iterated. If all features within a branch are found to increase the dRMS above R_c, we backtrack and consider earlier points that were not previously considered. Once we have run through all features, we test the putative correspondence by transforming the full model and calculating the resulting cRMS over all points. Note that we can also perform partial registration by excluding some points from the correspondence.

After global registration, a post-processing step is necessary to bring the two structures into final alignment. We implemented the iterative closest points (ICP) algorithm to perform this local alignment. The combination of those two methods has proven effective for aligning molecular surfaces (Figure 4.)

2. Surface Representation using Wavelets

After the structures have been registered, a fixed portion of the protein surface involved in DNA binding is projected onto a rectangular 2D plane. Wavelet transforms will then be applied on this 2D representation. The need for a transformation step before PCA is necessitated by the fact that a sufficiently detailed 2D projection of the protein surface is very high-dimensional (65k points per protein), rendering direct application of PCA computationally expensive (65k x 150 matrix).

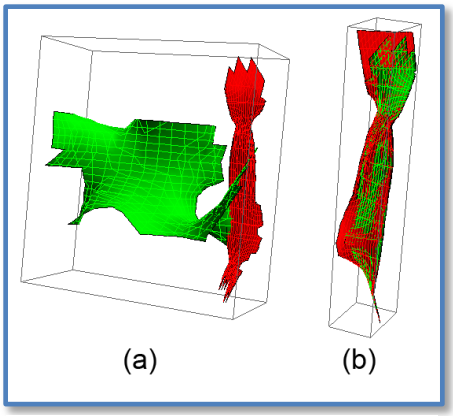

Figure 4: Registration of two molecular surfaces. (a) Pre-registration, and (b) post-registration.

Our choice of wavelets was motivated by their decomposition of signals in both frequency and space domains. Given our long-term objective of using this representation for protein structure prediction, we sought to exploit the localized nature of protein structure. Amino acid variations along the protein chain typically have only localized effects on the protein surface, rendering the effect of AA identity on the wavelet coefficients more direct if the contribution of each wavelet coefficient is itself localized in space.

2.1 Surface Projection

For each structure, we defined a fixed rectangular region in space upon which the surface of the DNA-binding α-helix is projected. The location and orientation of this region were determined as follows. First, the centroid of the DNA base closest to the α-helix was computed as the mean position of all its backbone carbon atoms. The origin was defined to be the centroid, with the y axis running parallel to the long axis of the DNA helix, and the x axis perpendicular to it such that the resulting plane is equidistant from the DNA and the protein. The rectangle was centered at the origin with dimensions 12Å x 8Å, enough to contain the α-helix (Figure 5.)

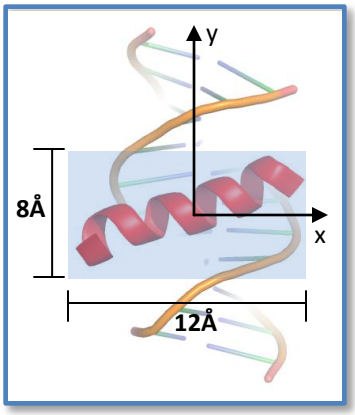

Figure 5: Projection of DNA-binding helix onto 2D plane.

This rectangular region was discretized into a 256x256 grid, and the distance along the normal axis from each point in the grid to the protein surface was calculated. The resulting 2D array of real numbers is in effect the native representation of a protein surface patch in our algorithm.

2.2 Wavelet Transform

Wavelet transforms were applied on the projected surface patch of every HTH protein to transform it into spatially localized frequencies at multiple resolutions. The 2D least asymmetric Daubechies filter of order 4 was used for the transformation. No analytic form exists for these transformations, and so they were numerically computed using Mathematica.

Daubechies filters are the only locally supported, continuous, and orthogonal wavelet bases, which makes them ideal for most wavelet analyses. Different orders were tried and the 4^{th} order appeared to give best results (Figure 6.)

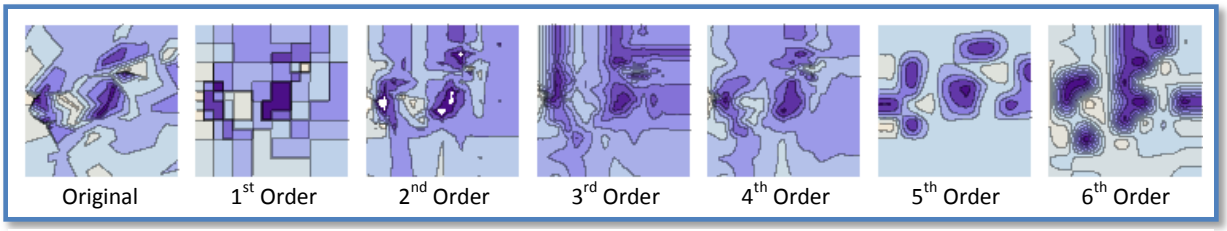

Figure 6: Representation with largest 50 coefficients using least asymmetric Daubechies wavelets of varying orders.

3. PCA of Protein Surfaces in Wavelet Space

Since our HTH proteins of interest come from the same protein family, we expect there to be little variation at the DNA-binding surface patch. More specifically, we expect the variations to be localized in space, and to depend on the underlying amino acid sequence.

After the surface patches have been wavelet transformed, we applied PCA to find the major axes of variation within the HTH protein family. We expect these variations to depend in a localized fashion on the identity of the amino acids that give rise to the structure in the first place. Thus, by transforming the protein surface patches into their principal vectors, we reduced their representation into the primary components that are affected by the identity of the primary sequence.

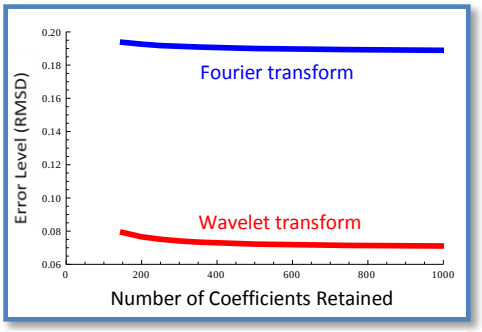

Figure 7: Wavelets vs. Fourier compression.

Since the protein surface patches are all at the same scale, variance normalization was not performed.

Data

A database of HTH-DNA complexes has been manually built and curated. The complexes were obtained from the protein databank (PDB) and filtered to include only structures that contain a DNA molecule bound to one of the following motifs: DNA/RNA-binding 3-helical bundle, λ-repressor-like DNA-binding domain, and HMG-box. This resulted in an initial database of 160 complexes. Subsequent filtering for complexes that are only bound to double-stranded DNA, as well as those that meet homology criteria, reduced the final set to 84 complexes. We tested our compression algorithm on those complexes.

Results

We wavelet transformed the 84 HTH-DNA complexes as previously described, and compared our results to transformations using 2D Fourier transform (Figure 7.) On average, the compression ratio achieved with wavelets was twice as high for the same error level.

We experimented with varying levels of compression during the wavelet transformation step and the PCA step (Table 1.) In general we found that increasing the number of coefficients retained after the PCA step decreases the error rate, with a high of 17% error when retaining only 10 coefficients, and a low of 12% error when retaining 150 coefficients, averaged across different numbers of wavelet coefficients retained (Figure 8a.) Such a consistent trend was not found however when we averaged across the number of PCA coefficients retained, and varied the number of

Number of PCA Coefficients	Number of Wavelet Transform Coefficients							
	150	200	250	300	350	500	750	1000
10	0.15	0.22	0.16	0.15	0.18	0.19	0.15	0.15
20	0.13	0.13	0.13	0.13	0.21	0.14	0.13	0.13
30	0.12	0.12	0.13	0.12	0.15	0.17	0.12	0.12
50	0.11	0.11	0.13	0.11	0.15	0.15	0.1	0.1
100	0.1	0.098	0.2	0.098	0.095	0.2	0.093	0.093
150	0.1	0.098	0.18	0.11	0.17	0.15	0.093	0.093

Table 1: Compression Error Rates (RMSD)

wavelet coefficients retained (Figure 8b.) Increasing the number of coefficients retained during the wavelet transformation may have resulted in instability in the behavior of PCA due to increased variance in the data.

Given the experimental inaccuracies inherent in protein structure determination, a resolution of 3Å is sufficient for our purposes. Our final compression consisted of keeping 150 wavelet coefficients followed by a PCA step where only the top 30 principal vectors were retained. This represents approximately 2000 fold compression, with an error level of around 10%. At this error rate features greater than 3Å were generally indistinguishable from the uncompressed case (Figure 9.)

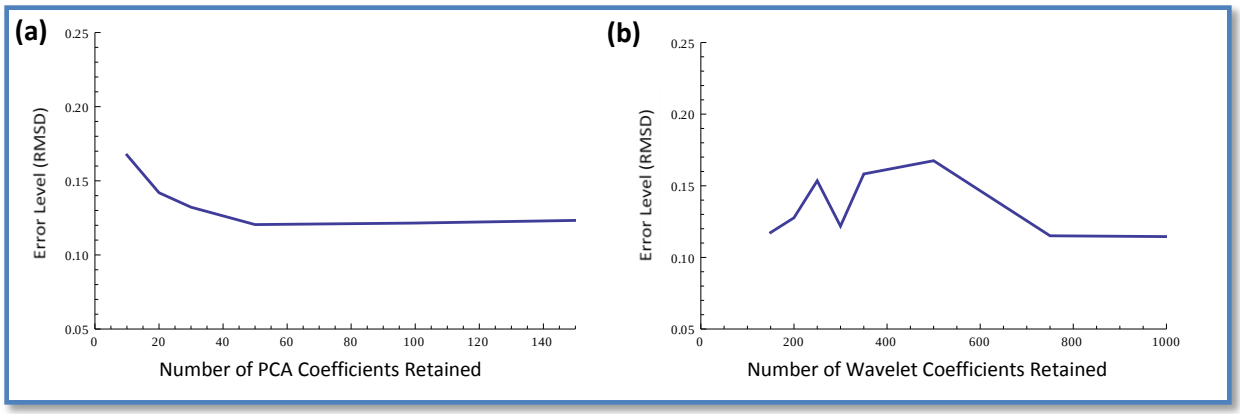

Figure 8: Average error level across (a) different PCA compression levels (averaged across all wavelet compression levels) and (b) different wavelet compression levels (averaged across all PCA compression levels.)

Future Work

We believe that achieving a 30-dimensional representation for local surface patches represents a significant reduction in the complexity of protein structure. Most importantly, such a small number of parameters render the problem highly amenable to machine learning methods, where the relationship between amino acid sequence and protein structure may be automatically inferred. We will explore this possibility in the future, as well as the reduction of DNA surface patches and the prediction of DNA-binding partners of HTH proteins.

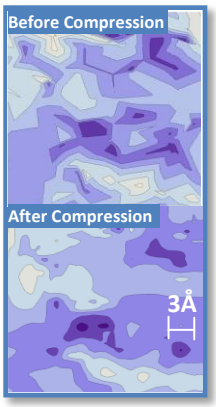

Figure 9: Quality of Final Compression

References

1. *Geometric Hashing: An Overview.* **Wolfson, Haim J. and Rigoutsos, Isidore.** 9924, s.l. : IEEE Computational Science and Engineering, Vol. 1070, pp. 10-21.

2. *Robust Global Registration.* **Gelfand, Natasha, et al.** s.l. : Eurographics Symposium on Geometry Processing, 2005.

3. *Ray Tracing Point Set Surfaces.* **Adamson, Anders and Alexa, Marc.** s.l. : International Conference on Shape Modeling and Applications, 2003. pp. 272-279.

4. *Exploring Expression Data: Identification and Analysis of Coexpressed Genes.* **Heyer, Laurie J., Kruglyak, Semyon and Yooseph, Shibu.** [ed.] 1106-1115. 11, Genome Research, Vol. 9.