Better Bond Angles in the Protein Data Bank

C.J. Robinson and D.B. Skillicorn
School of Computing
Queen's University
{robinson,skill}@cs.queensu.ca

Abstract

The Protein Data Bank (PDB) contains, at least implicitly, empirical information about the bond angles between pairs of amino acids. There is considerable variation in the observed values for a given amino acid pair, and it is not clear whether this variation represents a wide range of conformal possibilities or is due to noise. We show, by applying singular value decompositions to the sets of examples of particular amino acid sequences, that there appear to be relatively few possible conformations for a given amino acid pair, and hence noise is a plausible explanation for the variation in the raw data. This has implications for secondary structure prediction which typically depends on the PDB values.

1 Introduction

The Protein Data Bank (PDB) is a repository of protein structure, mainly gathered using X-ray Crystallography (XC) and Nuclear Magnetic Resonance (NMR). Implicit in the PDB is the conformation of each protein's backbone sequence, that is its primary structure. This can be described using the bond angles, ϕ and ψ, between each pair of amino acids along the backbone. These angles are constrained by the structure and size of the side-chains of the amino acid. The possible bond angles can be displayed using a Ramachandran plot, a two-dimensional representation of ϕ versus ψ showing regions that are admissible on physical or energetic grounds.

The actual bond angles associated with a given pair of amino acids in the PDB show wide variations. In some cases, it is clear that this is due to different conformations; the bond angle values are far apart. In other cases it is not clear whether the variations in values correspond to conformal possibilities or simply to noise. We address this question empirically and show that it is plausible, perhaps likely, that most of the variation in values is due to noise. As a result, we are able to show distinct conformations of amino acid pairs that are not distinguishable in the raw data. By extending these results to chains of amino acids, we suggest that secondary structure can be built from primary structure in a bottom-up fashion. As a side-effect we are able to produce Ramachandran plots with much more tightly constrained admissible regions.

Our strategy is to select all occurrences of a particular amino acid sequence from the PDB, together with the bond angles between its members. Singular Value Decomposition (SVD) is then applied to the resulting dataset; the resulting matrices are truncated at some k components; and then remultiplied to produce a matrix analogous to the original bond angles. This new matrix can be thought of as defining *canonical bond angles* for the amino acid sequence being considered. The different occurrences show strong tendencies to cluster in the transformed space, providing evidence of a limited set of conformations distorted by noise rather than a wide range of possible conformations. Since most secondary structure prediction uses data from the PDB, much improved results would be expected from using canonical, rather than measured, bond angles.

2 Related Work

The effects of neighboring amino acids, *n-1*, *n+1*, on the conformation of amino acid n was studied as early as the 1970's. Advanced, automated analysis of this variety is now possible over the WWW via

the Conformation Angles Database (CADB) [9]. Interfaces to CADB can automatically generate Ramachandran plots of specific amino acids with respect to neighboring residues. This is currently limited to three amino acids and is also restricted by the nature of information contained in the CADB.

Aggregating PDB data to create empirical Ramachandran plots (as opposed to ones based on energy minima, see Section 4.1) has been done by Kleywegt and Jones [5] and later by Hovmoller *et al.* [3]. The plots were created by extracting bond angles for every instance of an amino acid within the PDB. These types of analysis are useful in that they utilize all the information available to provide a detailed plot, but the relative simplicity limits the application of the results.

Clustering algorithms have been applied to length 5 amino acid sequences obtained from the PDB [6]. The clustering was done based on amino acid type and not related to bond angles or sequence conformation. Though similar in concept to our work, the results are fundamentally different as the clustering is not based on bond angles.

A geometric analysis of bond angles for a specific protein structure is available directly at the PDB website. The analysis can show bond angles that deviate from the generally accepted values by more than a given threshold. This is a generally accepted method to identify potential noise and/or errors introduced from the XC or NMR process [1]. Since it is a simple comparison of threshold values, the results obtained do not reflect great confidence and the analysis is limited to one protein structure at a time.

The paper by Rost *et al.* [8] summarizes work on secondary structure prediction. Much of this work is derived from the conformation information implicit in the PDB.

3 Matrix Decompositions

The singular value decomposition of a matrix A is

$$A = USV'$$

where the dash indicates the transpose. If A is $n \times m$ and has rank r, then U is $n \times r$, S is an $r \times r$ diagonal matrix with decreasing entries $\sigma_1, \sigma_2, \ldots, \sigma_r$ (the singular values), and V is $r \times m$. In addition, both U and V are orthogonal, so that $UU' = I$ and $VV' = I$. In most practical datasets, $r = m$.

The most useful property of SVD for our purposes is that the transformation captures as much variation in the original data as possible in the first transformed dimension, as much as possible of what remains in the second dimension, and so on. Hence, if we truncate the decomposition so that U is $n \times k$, S is $k \times k$, and V is $k \times m$, for some small k, then we have discarded dimensions that have little influence on the correlational structure of A. The dimensions from $k+1$ to r can be considered to represent noise in the original data.

Remultiplying the truncated matrices produces a new matrix that has the same shape and interpretation as A but has been 'denoised'. The truncation parameter, k, should be chosen so that significant information is retained but insignificant discarded. The magnitudes of the singular values are a measure of how important each dimension of the decomposition is. Plotting the values of the diagonal of S and choosing k at the earliest point where these values become small is often a good selection mechanism.

An alternate interpretation of the transformed space produced by an SVD is that points are placed close to other points with which they are correlated. Hence if the original data describes objects of a few different kinds, distorted by noise, we would expect to see tight clusters in the transformed space (which can be plotted and visualized directly if $k = 2$ or 3). On the other hand, if the original data describes data without much similarity, we would not expect to see clusters in the transformed space in any dimensions. SVD reveals the latent cluster structure of data.

4 Background and Methods

This section describes current methods of obtaining protein structure and the inherent problems, databases of acquired protein structure, and the limitations of applying SVD and other data mining algorithms to such data.

4.1 Obtaining structure Current methods for determining the 3-dimensional structure of a protein are slow, error-prone and expensive. The readily available methods employed today are X-ray crystallography (XC) and Nuclear Magnetic Resonance (NMR); both are physical processes. XC and NMR involve determining the 3-dimensional coordinates of every atom in a protein within a known error range. Data collected from both methods are not directly converted to atomic co-ordinates but require human input, interaction and heuristics to refine the data. This is a further possible source of error [4]. The structure of about 28,000 proteins has been determined using these methods but millions of proteins are known to exist.

The Protein Data Bank (PDB) is the worldwide depository for protein structure, almost all of which have been obtained from XC and NMR. The format for protein structure has been carefully designed to be as flexible as possible allowing the data to be utilized in many different fields of study [1]. The PDB consists of individual files for each protein entry, and contains the atomic co-ordinates of atoms within the protein, but does not directly contain bond angles for amino acids in the primary sequence. Unfortunately, data mining applications are not natively supported by this standard format of the PDB.

4.2 Database formats There are many derivatives of the PDB, only a few of which contain bond angles of amino acids in an easily accessible format. The Conformation Angles Database (CADB) [9] and Dihedral Angle Database (DAB) [2] are the newest and most comprehensive sources. CADB is a self-limited database which excludes protein with homologous sequences up to a certain threshold. For data mining purposes, it would be beneficial not to exclude any data *a priori*. CADB also will only supply bond angles for small sequences of amino acids. DAB contains bond angles for every possible set of length 2, 3, 4 and 5 amino acid sequences. However, the database is not currently publicly available. Both of these databases lack the ability to supply a set of bond angles in a matrix format based on a simple query (i.e. for a specific sequence of amino acids, or for every set of 8 amino acids). Since most data mining applications, and in particular SVD, require data to be in matrix form, a new database was required.

4.3 Ramachandran plots Ramachandran plots are a way to display the possible conformation of an amino acid pair by plotting the admissible regions based on energy considerations. Figure 1 shows a typical plot. The region near the top left corresponds to conformations associated with a β sheet, with the two peaks corresponding to parallel and anti-parallel secondary structure. The region midway down and to the left corresponds to conformations associated with α helices. The third region corresponds to conformations of anticlockwise α helices. Not all amino acids are constrained in this way; for example pairs involving glycine can exhibit a much larger range of conformations. We will use Ramachandran plots to compare the conformations possible in the raw PDB data and those suggested after denoising with SVD.

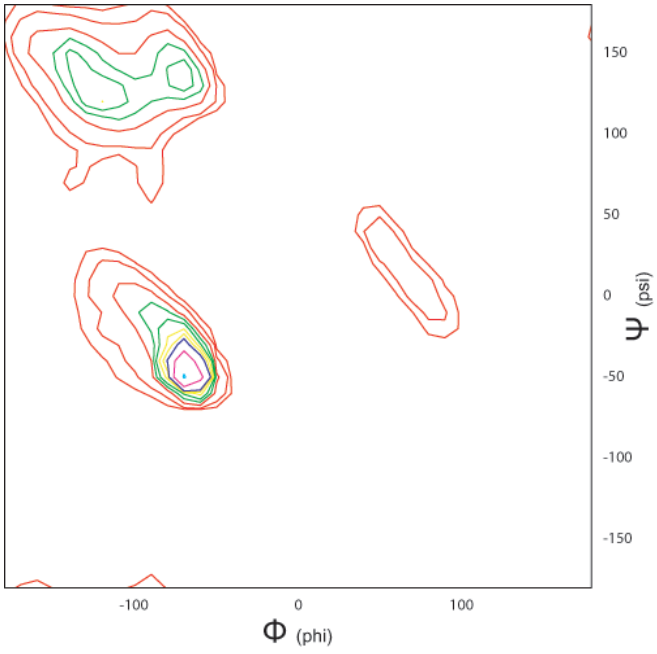

Figure 1: Typical Ramachandran plot (from Kleywegt and Jones (1996) [5]).

4.4 Methods Every protein file listed in the Protein Data Bank was downloaded, converted and appended to a datafile. Bond angles were computed from atomic coordinates for each pair of amino acids in the primary sequence using TORSIONS [7]. The output of TORSIONS was appended to a datafile in the format: ..., amino-acid(location), ϕ, ψ,... (e.g. ..., ALA(150), -150.595, -63.539, SER(151), ...), where each line is the structure of one protein. Due to the time-intensive nature of compiling a database into this format, this intermediate, flexible format was chosen for ease of access for other possible studies. The results presented here are based on 27,544 files which were downloaded from the PDB on November 16, 2004 and transformed into the new format. From this new derivative of the PDB, datasets of specific sets of bond angles can easily be accessed.

The natively implemented svd function in Matlab 6.5r13 was used to perform the decomposition on small datasets. For larger datasets JAMA v5 was used with Java 1.4.2-04 to perform SVD. The resultant decomposed data was geometrically interpreted by plotting the first three dimensions of the U matrix.

The following steps were performed:

- Given an amino acid sequence of length m, the $2m - 2$ internal bond angles associated with each occurrence of this sequence were extracted from the PDB. There are typically 1000–5000 examples of an amino acid sequence of length 3 in the PDB, 100–500 examples of sequences of length 4, and 0–200 examples of sequences of length 5.

- An SVD was performed on the resulting matrix whose rows correspond to examples of the given amino acid sequence and whose columns correspond to a bond angle at a particular position in the sequence.

- The resulting decomposition was truncated at $k = 3$, a value chosen after inspection of a large number of plots of singular values. The first 3 columns of the U matrix was plotted for visualization.

- The truncated matrices were remultiplied to give a matrix of canonical bond angles.

- Each cluster in the transformed space was fitted with a 3-dimensional ellipse and the ellipse mapped back into bond angle space using the SVD 'in reverse'. This ellipse defines a region of the Ramachandran plot corresponding to the cluster.

5 Results

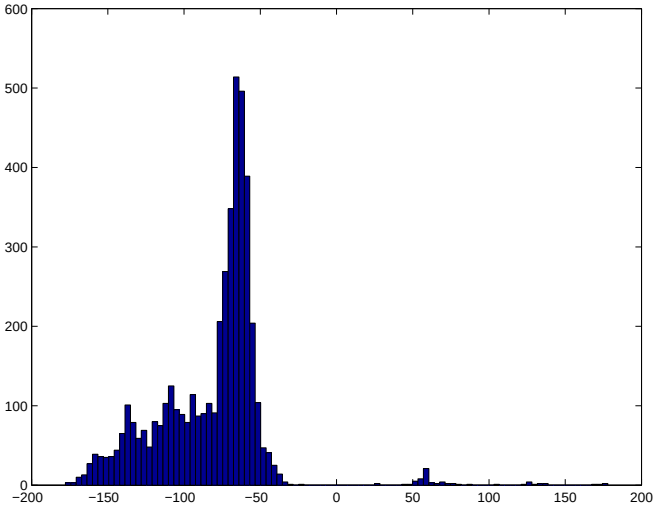

Figure 2: Histogram of the ϕ angles of the LEU-VAL bond in the sequence LEU-VAL-ARG.

There are many possible sequences of length 3, 4, and 5, so we only show some typical results here.

We being by considering the sequence LEU-VAL-ARG of length 3. There are 4529 examples of this sequence in the dataset. Figures 2 and 3 show histograms of the bond angles of the LEU-VAL bond taken from these examples. There is obvious structure in both histograms, but it is hard to make use of it. The distribution of ψ angles suggests two clusters; but is the distribution of ϕ angles one big and one small cluster, or two big clusters and one small one? It is not straightforward to build conformations from such information. It is also not clear how much of the visible variation is due to noise and how much represents different possible conformations. Figure 4 is a Ramachandran plot of

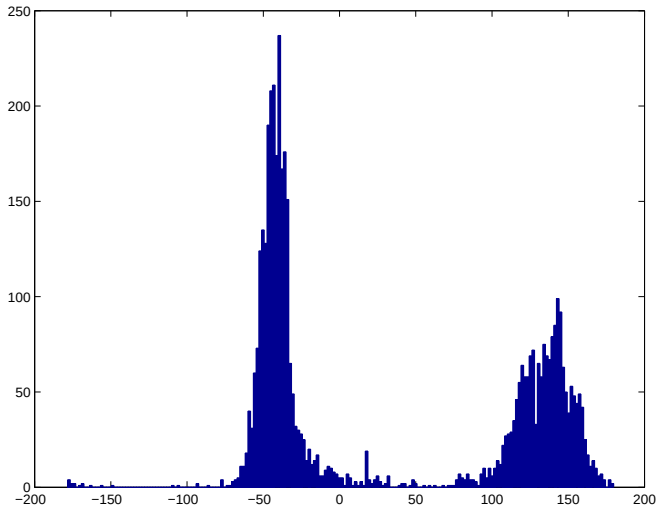

Figure 3: Histogram of the ψ angles of the LEU-VAL bond in the sequence LEU-VAL-ARG.

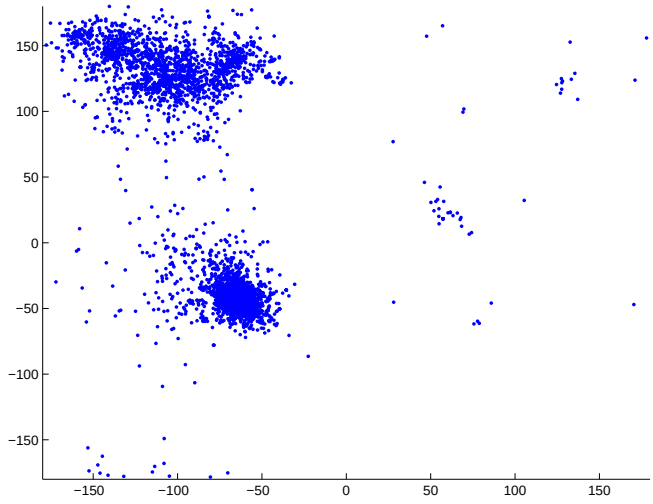

Figure 4: Ramachandran plot for the LEU-VAL bond in the sequence LEU-VAL-ARG.

the bond angles for the LEU-VAL bond. It is clear that there are some α helix conformations, some β sheet conformations, and a few anticlockwise α helix conformations, but the space of possible conformations is apparently not restricted beyond this.

Figure 5 shows the first 3 dimensions of the U matrix obtained from an SVD of the matrix of bond angles for the sequence LEU-VAL-ARG. There are four clusters visible, two large and two much smaller. There are a number of outliers, and perhaps some smaller clusters, but they represent a small fraction of the total number of examples.

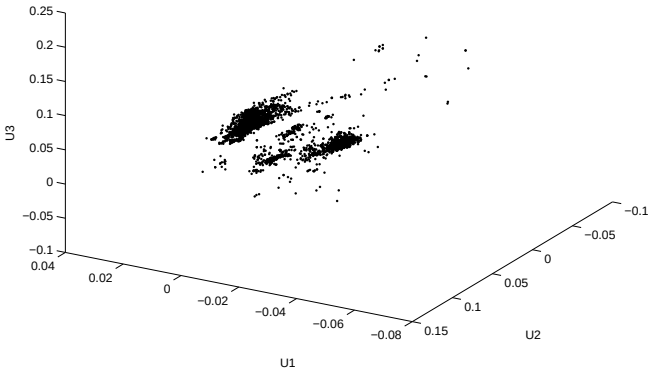

Figure 5: 3-dimensional plot of SVD transformed space for the amino acid sequence LEU-VAL-ARG.

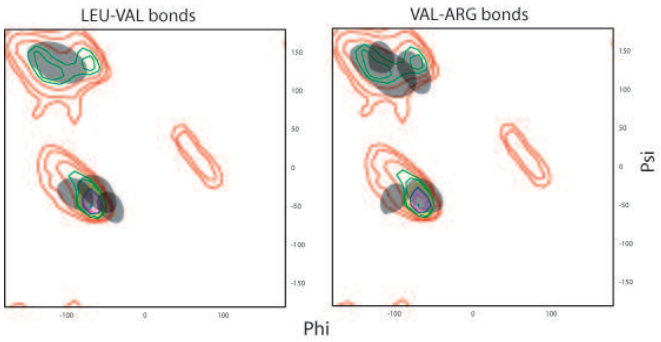

Figure 6: A Ramachandran plot overlayed with regions obtained from mapping clusters from SVD transformed space back to bond angle space for the amino acid sequence LEU-VAL-ARG. Each region is labelled in gray, darker where regions overlap. Compare with Figure 4

Figure 6 shows the location of these clusters when mapped back into bond angle space for each

of the pair bond angles. There are four possible conformations for each of the bond angles: the LEU-VAL bond exists in 4 α helix conformations (with different pitches) and 1 β sheet conformation. These 5 conformations for the VAL-ARG bond split more evenly, with 2 α helix conformations, and 3 β sheet conformations. In other words, 2 of the conformations of the overall sequence are α helices, 1 is a β sheet, and two others exhibit transitions from one shape to another at the VAL amino acid. Note that these transitions are abrupt; there is little evidence that the 'end'of the α helix changes shape because of the conformation of the adjacent bond.

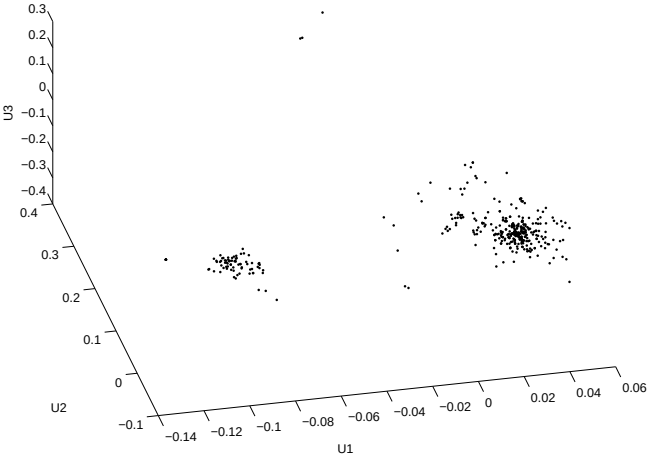

Figure 7: 3-dimensional plot of SVD transformed space for the amino acid sequence LEU-VAL-ARG-ILE.

Figure 7 shows the equivalent SVD transformed space for a sequence that extends the one we have just been considering: LEU-VAL-ARG-ILE. There are 417 examples of this sequence in the dataset. We see that there are two well-separated clusters with hardly any outliers. Figure 8 shows the singular values for the matrix of examples of this amino acid sequence. Almost all of the variation is in the first two dimensions.

Figure 9 shows the Ramachandran plot derived from these two clusters. As expected, the plots of each of the bond shows well-separated conformal

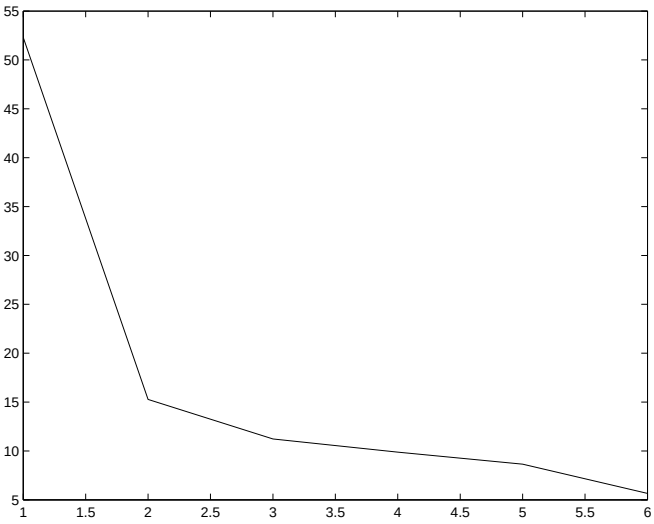

Figure 8: Plot of the singular values of SVD for the amino acid sequence LEU-VAL-ARG-ILE.

possibilities, an α helix and a β sheet; and the conformations for the first two bonds are subsets of those computed from the sequence of length three. It seems likely that each of the possibilities for the bonds match, so that there are only two conformations for this entire length four sequence (and notice the tightening towards the right hand end for both conformations). We have looked at many sequences of length 3 and 4, and the results are similar for most of them.

These results suggest that a much improved version of the PDB's bond angles could be produced by systematically denoising the existing data. However, this is difficult in practice. There are a large number of amino acid sequences of short length, so a great deal of computation is required even to apply the techniques described here to all of their bond angles. It then remains to check whether the canonical bond angles derived from sequences of length 3 agree with (or can be fitted to) those obtained from sequences of length 4 for the same amino acid pair. However, this suggests the possibility of a dynamic programming style algorithm for determining conformations for a sequence of length l from the conformations of overlapping sequences of length $l-1$. We are exploring this

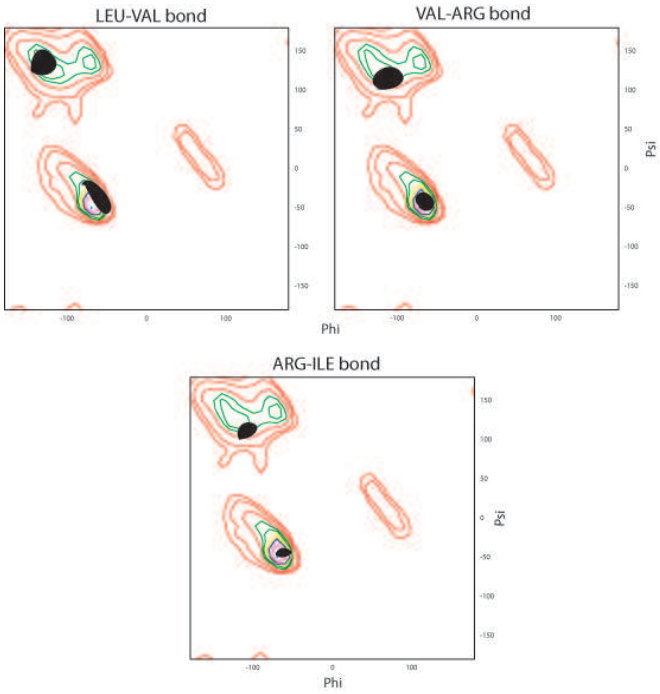

Figure 9: A Ramachandran plot overlayed with areas obtained from mapping cluster from SVD transformed space back to bond angle space for the amino acid sequence LEU-VAL-ARG-ILE.

possibility further.

5.1 Larger structure A random sampling of 1,557,072 length 5 amino acid sequences (approximately a 10% sample of all possible length 5 sequences) was extracted from the PDB. Figure 10 shows the 3-dimensional plot of the SVD transformed space that results. The overall structure is a diamond of clusters. The leftmost cluster corresponds to sequences whose basic conformation is straight (i.e. they are part of β sheets) while the rightmost cluster corresponds to α-helices for the entire sequence. The clusters along the edge of the diamond appear to be sequences in transition between these two basic conformations. For example, the first amino acid in a sequence can be part of an α helix, while the remaining amino acids form a straight segment. The intermediate clusters capture these different conformation possibilities (each conformation appears as two different clusters because it depends on which end we consider first). Clusters in the middle appear to capture conformations with more than one transition, for example from α helix to straight segment to α helix. This figure shows how an SVD analysis of bond angle data can help to elucidate average structure in the PDB. These experiments have been repeated for longer amino acid chains with similar results. One conclusion that can be drawn from this figure is how rare conformations other than α helices and β sheets are.

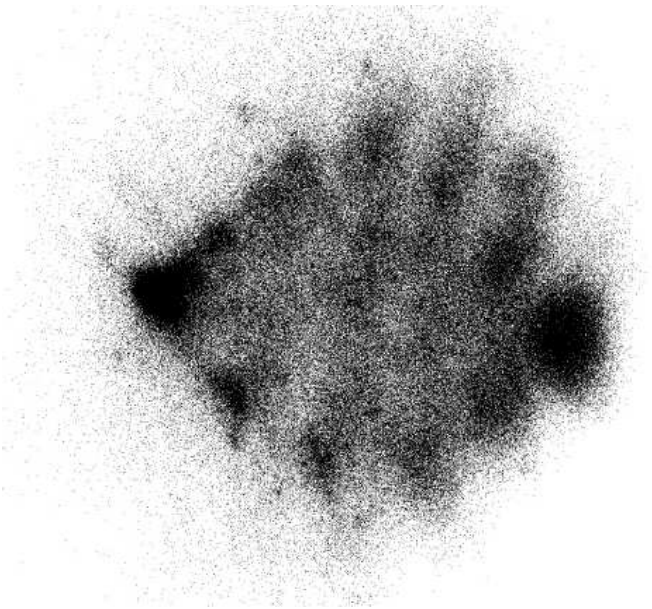

Figure 10: 3-dimensional plot SVD transformed space for a large set of arbitrary amino acid sequences of length 5.

6 Conclusions

We have applied singular value decomposition to datasets of bond angles for particular short sequences of amino acids, using the values from the PDB. The raw data contains a great deal of variability, and it is not clear to what extent this represents noise (or errors) or different conformal possibilities. By examining the clustering structure in the space to which SVD transforms the data, we have provided some evidence that the main source of variation in the raw data is noise. Relatively few

conformal possibilities are revealed in the transformed space.

Mapping the clusters back to the original bond angle space produces a version of each dataset containing 'canonical' bond angles, that is values that have been denoised. These bond angles are constrained to much smaller regions of Ramachandran plots, and exhibit coherence when the same cluster is followed along a sequence of bond angle pairs.

Many secondary structure prediction algorithms and functional studies are based on the PDB. Our results suggest that some effort should be spent on denoising the data before drawing conclusions about more complex structure from it. The techniques described here cannot be applied directly to the entire PDB. Although a single SVD on the entire PDB is (just) possible, new entries are being added all the time, and it is not a computation that is attractive to repeat regularly at this time. Replacement of bond angle data piecemeal using our techniques on sequences of medium length should be straightforward, but requires further analysis of the variation in results for an amino acid pair considered as part of a sequence of length 3, of length 4, and so on.

The elicitation of more robust conformations for short amino acid sequences suggests a method for discovering the conformations of longer sequences by assembling the short conformations into longer ones. We are pursuing this direction.

References

[1] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The Protein Data Bank. *Nucleic Acids Research*, pages 235–242, 200.

[2] S. Dayalan, S. Bevinakoppa, and H. Schröder. A dihedral angle database of short sub-sequences for protein structure prediction. In *2nd Asia-Pacific Bioinformatics Conference*, Dunedin, New Zealand, 2004. Conferences in Research and Practice in Information Technology.

[3] S. Hovmoller, T. Zhou, and T. Ohlson. Conformations of amino acids in proteins. *Biological Crystallography*, D58:768–776, 2002.

[4] G. Kleywegt. Validation of protein crystal structures. *Biological Crystallography*, D56:249–265, 2000.

[5] G. Kleywegt and T. Jones. Phi/psi-chology: Ramachandran revisited. *Structure*, 4:1395–1400, 1996.

[6] W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17:282–283, 2001.

[7] A. Martin. Torsions. acrmwww.biochem.ucl.ac.uk/programs/torsions/, 2004.

[8] B. Rost. Review: Protein secondary structure prediction continues to rise. *Journal Structural Biology*, 134:204,218, 2001.

[9] S. Sheik, P. Ananthalakshmi, G. Ramya Bhargavi, and K. Sekar. Cadb: Conformation angles database of proteins. *Nucleic Acids Research*, 31:448–451, 2003.