Paul Murphy – CSE 527 Final Project – 12/14/2004

INTRODUCTION:

Computational simulations of protein dynamics have become increasingly important for a number of applications which require understanding of protein behavior at a molecular level. Protein structure prediction and rational drug design are two such applications, which require fine-grained modelling of protein structure in order to evaluate hypotheses about the likelihood that a protein will fold into a particular conformation, or the likelihood that a small modecule will inhibit an enzyme of interest. Computational modelling is required to deal with the vast number of interactions that can occur between proteins, ligands and solvent in the simulation. Functions of the coordinates of each atom in the system have been defined which approximate the free energy of the system, and many programs seek to minimize these functions over the conformation space available to the components of the system.

However, the conformational space available to a polypeptide chain is large, since there may be on the order of thousands of rotable bonds in the system, each of which can adopt a wide range of values. The vastness of this space constitutes a major challenge for simulation, since *a priori* one has only hypothetical knowledge about which region may be pertinent to search. This also provides a motivation for devloping reduced basis representations of protein structure. Such approaches have resulted in discretized sets of conformations of the side chains of proteins known as "rotamers", which have eased the task of side chain structure prediction [Dunbrack 2002]. Similar approaches to representing backbone structure with reduced sets of conformation have also been successful [Deane 2000]. The validity of representing backbone structures with a reduced set of conformations is based on the knowledge that substructures of proteins adopt canonical configurations in some cases [Chothia 1987], [Al-Lazikani 1997] [Al-Lazikani 2000].

In this report, I will describe a clustering based approach to reducing the size of the search space for structure prediction. Using backbone torsion angles of several substructures of proteins, I applied several clustering algorithms and evaluated them with respect to coverage of conformation space and with respect to known relationships between structure and sequence in canonical structures. Hypothetically, cluster centers that emerge from this approach could be used as reduced basis sets for simulations of protein structure in the applications described above.

DATA:

Many protein structures have been elucidated by X-ray crystallography. In this technique, the location of most atoms in a protein can be determined to a certain resolution. In this project, I used 3356 such structures from the Protein Data Bank [Berman 2000] which had been determined to a high-resolution. I extracted the backbone dihedral angles for each polypeptide contained in the set of structures. For comparison with previous results, I extracted from this database, sets of substructures which corresponded to known canonical structures, *ie* 4, 5, and 6-residue hairpin turns. Hairpin turns connect adjacent strands in a beta sheet, so that the first and last residue are hydrogen bonded through both the amide nitrogen and carbonyl oxygen on each residue. The location of the first and last residue is constrained by this hydrogen bonding

requirement, and therefore there are fewer degrees of freedom available to the torsion angles of the interior residues. This should makes clustering on the values of those angles easier than clustering without this constraint, and should therefore be a suitably easy problem for a first attempt, in addition to being able to be compared to previous results for hairpin turns. Henceforth, I will refer to a contiguous subset of a polypeptide as a "fragment".

CLUSTERING:

I used clustering tools available in R to perform clustering on the torsional angles. For hierarchical clustering approaches, I used the function *hclust* defined in standard R. For model based clustering approaches I used the *mclust* package available from: [Fraley 2002a,2002b]. Since torsional angles are periodic, I clustered the vectors formed by concatenating sines and cosines of the dihedral angles as opposed to the dihedral angles themselves. That is, for a set of one residue fragments, I would have clustered not over the vectors $\{\varphi \ \psi\}$, but instead over the vectors $\{\cos(\varphi) \cos(\psi) \sin(\varphi) \sin(\psi)\}$. Although this step doubles dimensionality of the space over which to cluster, it was necessary in order to capture the periodic connectivity of dihedral angles.

In addition, I "clustered" the sequences of the fragments from which I had extracted the torsional angles in an *ad hoc* way. Since the amino acid glycine has no true side chain, it is able to adopt conformations outside the region of the Ramachandran plot accessable to most other amino acids. Based on this, I mapped each sequence to a to a string of the same length comprised of letters from an alphabet that is either G for glycine or O for other. That is, the sequence "AGPV" maps to "OGOO". Given the entire set of fragments which I was clustering, and a number *n* of desired fragments, I defined *n-1* "clusters" representing the *n-1* most common strings in the set, and *1* cluster for all other strings. Clustering sequences was necessary to determine the degree of sequence dependence of structural cluster membership, which has been noted in earlier work [Chothia 1987].

Since hierarchical clustering and the *ad hoc* sequence clustering algorithm described above both require as input the number of clustered desired in the output, I use model based clustering to generate such a number. That is, since *mclust* computes decomposes the data into clusters with membership defined probabilistically, the BIC score can be used to determine the model which is the best fit for the data without overfitting by tuning too many parameters.

$$BIC = 2 \log p(D \mid \text{theta\_k}, \text{m\_k}) - \text{v\_k} \log(n)$$

Thus the model with the highest BIC score can be considered to be the best model for the data. The number of clusters it contains was the number which I used as input to the hierarchical clustering methods.

RESULTS:

Clusters membership for 4, 5 and 6 residue hairpin turns are shown in the figures below. NB: each figure shows several Ramachandran plots, one for each residue in the fragment, with φ plotted on the x-axis against ψ plotted on the y-axis. Captions belong to the figure above. Mutual information between each clustering method is given in the

table.

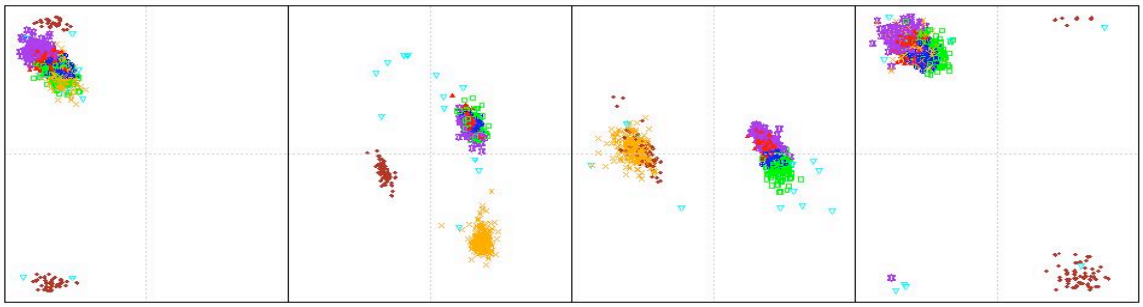

Figure 1 – Ramachandran plot for 1156 x 4-residue hairpin turn, partitioned into 7 clusters based on a VEV model (equal shape, variable size and orientation). This was the model with the highest BIC score for this data set.

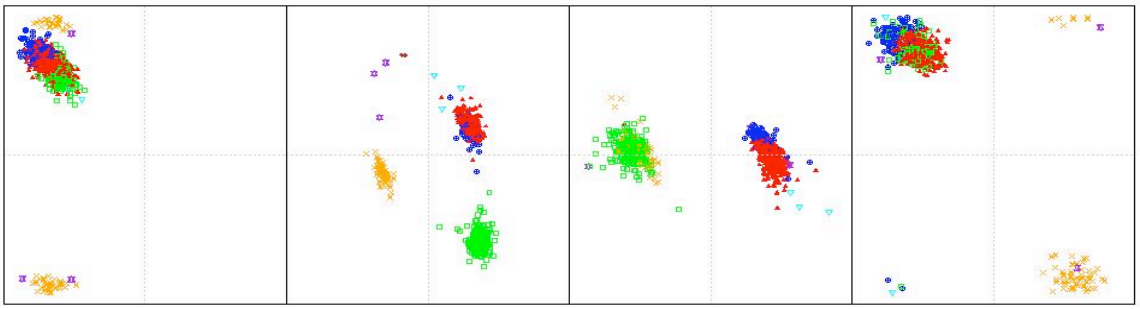

Figure 2 – Ramachandran plot for 1156 x 4-residue hairpin turn, partitioned into 7 clusters using hclust(method="complete") with distances defined as the Euclidean distance between the vectors.

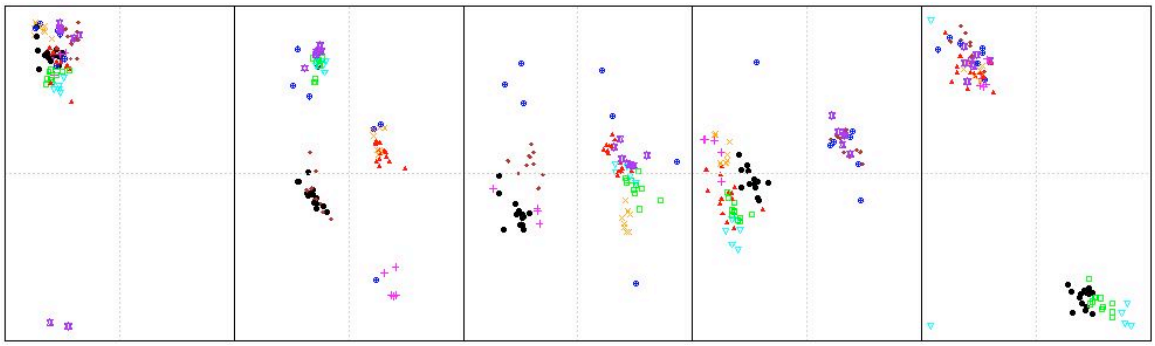

Figure 3 – Ramachandran plots for 88 x 5-residue hairpin turn, partitioned into 11 clusters based on a VVI model (variable shape and size, axially orthogonal orientation)

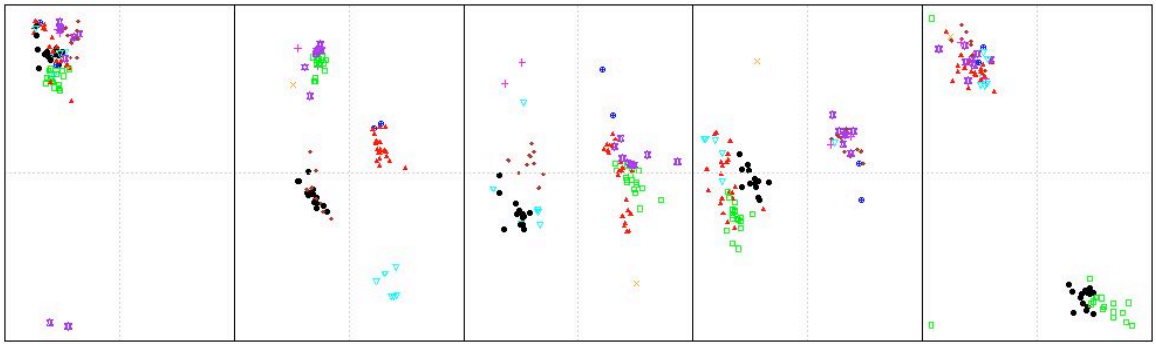

Figure 4 – Ramachandran plots for 88 x 5-residue hairpin turn, partitioned into 9 clusters using hierarchical, complete clustering.

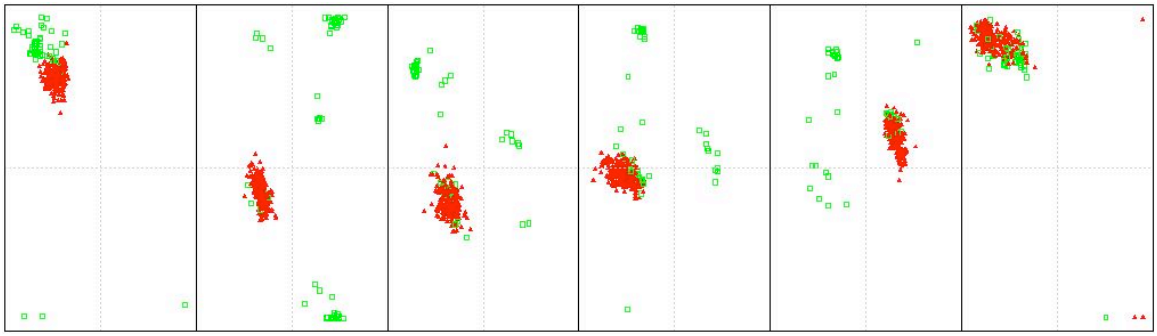

Figure 5 – Ramachandran plots for 311 x 6 residue hairpin turn, partitioned into only 2 clusters using a VVV model (variable size, shape, orientation). This was the highest scoring model, although up to 20 clusters of all model types were tested.

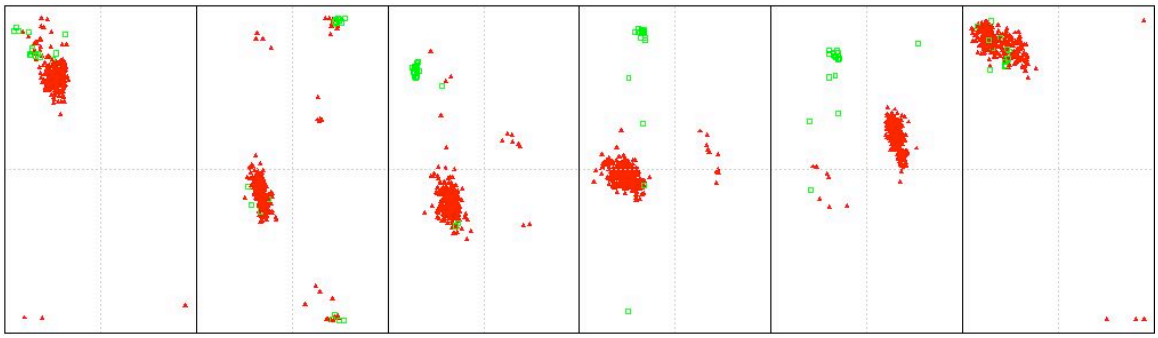

Figure 6 – Ramachandran plot for 311 x 6 residue hairpin turn, partitioned into 2 clusters using hierarchical clustering.

Mutual information quantifies how much information about the value of one random variable is revealed by knowing the value of another random variable. However, to account for sampling bias, excess mutual information must be computed by permuting the data, which reveals independent information. These quantities can be computed as in [Cline 2002].

Table 1 – Excess mutual information (bits) between clustering techniques. S, H and M represent sequence, hierarchical and model based, respectively.

Length	***S vs S***	***H vs H***	***M vs M***	***S vs H***	***S vs M***	***H vs M***
4	1.86618	1.52267	2.49141	0.71023	0.707979	1.1741
5	2.17502	2.28682	2.43372	0.962235	0.845682	2.02843
6	1.07323	0.303199	0.633808	0.0365629	0.16466	0.15463

DISCUSSION:

Hierarchical and model-based clustering in this application seem to differ in

multiple ways primarily in the representation of outliers, *ie* data points unlike any other, in the data. Hierarchical clustering more eagerly assigns outliers to their own cluster. The model based clustering seems to have a set, *ie* the cluster represented by cyan triangles in Figure 1, which represents most or all outliers. Whereas hierarchical clustering produces some tight clusters which may have only a few members, model based clustering seems more likely to partition data into more evenly balanced clusters, at least over this data set. This is represented in the self-information of hierarchical clustering, which is lower in the three examples above than the self-information of model based clustering, indicating a more splayed distribution of the hierarchical clusters. Speculating on the question of which method is superior for coverage of conformation space, it may be superior to use hierarchical clustering for completeness but model based clustering if coverage of densely populated centers is desired.

The question of whether the potentially useful attributes of model based clustering are actually helpful in this circumstance also requires some consideration. At least over these three examples, the BIC score seems to be a somewhat non-intuitive indicator of what the true number of clusters should be. For the 5 residue hairpin turn, 11 clusters are found whereas for the 6 residue hairpin turn only 2 clusters are found. It would have been difficult to predict these outcomes beforehand.

It is likely that the ability of model based clustering to account for varying shape and size is a useful feature. In this situation, not all dihedral angles residues will vary the same amount in each cluster, making the ability to vary cluster size with respect to the cluster and the dimension a useful feature. In addition, given the requirement that these fragments have to position the first and last residue in the same position, a change in a dihedral angle in one residue will need to be offset by a change in the dihedral angle in another residue. For this reason, the ability of model based clustering to pick out covariances between different axes seems like a useful feature.

Correlation between sequence and angular clustering is present and significant, but is not perfect. Mutual information is less than self-information in all cases. This is in line with the general notion that that sequence and structure are loosely related, even in structures as constrained as hairpin turns.

REFERENCES:

Al-Lazikani B, Lesk AM, Chothia C. Canonical structures for the hypervariable regions of T cell alphabeta receptors. J Mol Biol. 2000 Jan 28;295(4):979-95.

Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. J Mol Biol. 1997 Nov 7;273(4):927-48.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov HN, Bourne PE: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol. 1987 Aug 20;196(4):901-17.

Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG Jr, Haussler D. Information-

theoretic dissection of pairwise contact potentials. Proteins. 2002 Oct 1; 49(1):7-14.

Deane CM, Blundell DL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins. 2000 Jul 1;40(1):135-44.

Dunbrack RL. Rotamer libraries in the 21st century. Curr Opin Struct Biol. 2002 Aug;12(4):431-40.

Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97:611:631 (2002a)

Fraley C and Raftery AE (2002b). MCLUST:Software for model-based clustering, density estimation and discriminant analysis. Technical Report, Department of Statistics, University of Washington. See http://www.stat.washington.edu/mclust.