

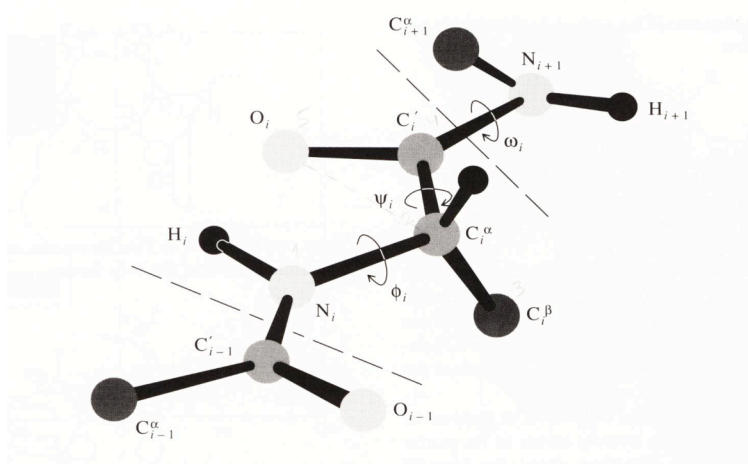
# CLUSTERING CONFORMATIONS of PROTEIN FRAGMENTS

CSE 527

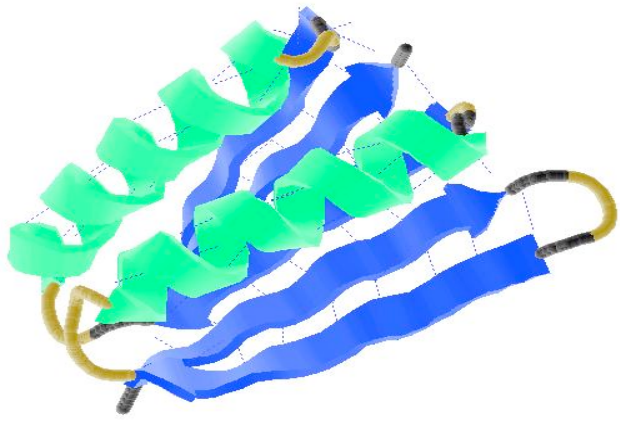
Paul Murphy

12/15/04

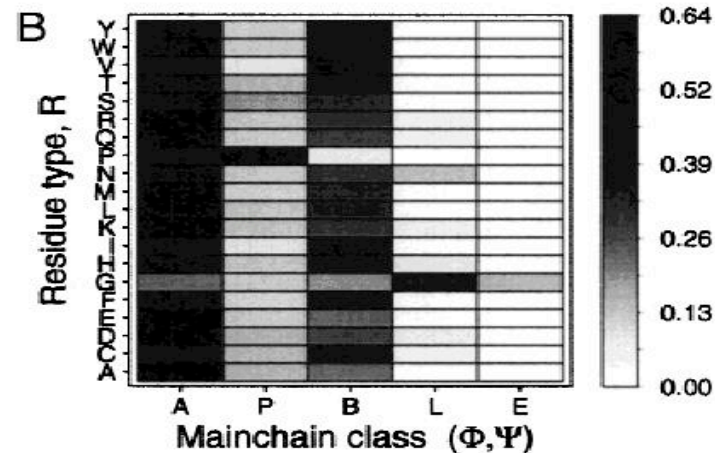
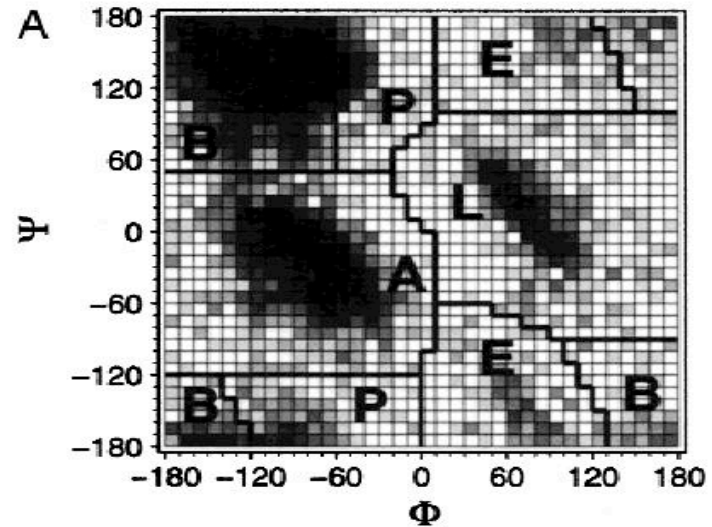
# PROTEIN STRUCTURE



Amino acids have *two rotatable bonds* along their backbone [Creighton (1993)]



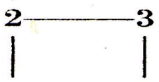
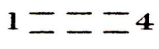
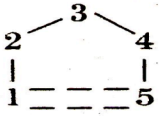
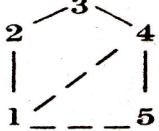
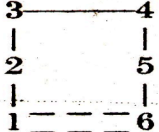
The *conformation of a polypeptide* is a function of the dihedral angles of its backbone. [Kuhlman, Baker (2003)]



The dihedral angles formed by these bonds can be viewed in a *Ramachandran plot* [Fiser, Sali (2000)]

# CANONICAL STRUCTURES of POLYPEPTIDE FRAGMENTS

Table 2  
Conformation of hair-pin turns

Structure	Sequence <sup>a</sup>	Conformation <sup>b</sup> (°)								Frequency <sup>c</sup>
	1 2 3 4 X- G- G- X	$\phi_2$ , +55	$\psi_2$ +35	$\phi_2$ , +85	$\psi_3$ -5 <sup>d</sup>					6/6
	X- G- X- X	+65	-125	-105	+10 <sup>e</sup>					6/7
	X- X- G- X	+50	+45	+85	-20 <sup>d</sup>					7/8
	X- X- X- X	+60	+20	+85	+25 <sup>f</sup>					4/4
	X- X- X- G	$\phi_1$ -135	$\psi_1$ +175	$\phi_2$ -50	$\psi_2$ -35	$\phi_3$ -95	$\psi_3$ -10	$\phi_4$ +145	$\psi_4$ +155	4/4
	1 2 3 4 5 <sup>g</sup> X X X X G	$\phi_2$ -75	$\psi_2$ -10	$\phi_3$ -95	$\psi_3$ -50	$\phi_4$ -105	$\psi_4$ 0	$\phi_5$ +85	$\psi_5$ -160	3/3
	X X X X X	+50	+55	+65	-50	-130	-5	-90	+130	1/1(3/3)
	1 2 3 4 5 <sup>h</sup> X- X- X- N- X G D	$\phi_2$ -60	$\psi_2$ -25	$\phi_3$ -90	$\psi_3$ 0	$\phi_4$ +85	$\psi_4$ +10			13/15
	1 2 3 4 5 6 <sup>i</sup> X- X- X- X- N- X G X	$\phi_2$ -65	$\psi_2$ -30	$\phi_3$ -65	$\psi_3$ -45	$\phi_4$ -95	$\psi_4$ -5	$\phi_5$ +70	$\psi_5$ +35	3/3 2/2 1/1

Some substructures of proteins have a discrete set of typical conformations related to sequence [Chothia, Lesk (1987)]

# MOTIVATION FOR CLUSTERING

- Classification
- Simulation
  - *Protein structure prediction* algorithms require exploration of a large conformation space
  - Clustering can (potentially)...
    - Positively bias the search towards favorable regions of conformation space
    - Address questions regarding the *thoroughness of sampling*

# PROJECT DESCRIPTION

- Input

- **Fragments of proteins**  
derived from a non-redundant database of high resolution crystal structures
- Several different **clustering algorithms**

- Output

- Assess coverage of conformation space by looking at **BIC score** (where applicable)
- Assess classification by looking at **mutual information** classification based on sequence

# FRAGMENT LIBRARY

- Using a set of 3526 PDB files, each containing high resolution 3D coordinates of one or more polypeptide chains, built a **database of fragments** containing several pertinent fields

```
mysql> select * from fragment_3 where id < 10;
```

id	phi_0	psi_0	ohm_0	phi_1	psi_1	ohm_1	phi_2	psi_2	ohm_2	seq	dssp	pdbid	chainid	res_num_a	res_num_b
1	1.6747	-2.4298	3.0737	-2.8350	2.8523	3.0105	-1.8506	2.6125	3.0870	GSF	E E	1a12	A	136	138
2	-2.5701	2.7472	-3.0602	1.0333	0.7644	-3.0757	-1.2904	-0.5910	-3.0348	RKQ	E E	1a21	A	185	187
3	-1.2795	2.0514	-3.0934	-1.3064	2.3279	3.1149	-2.0051	2.9817	3.1374	GLG	E E	1a26		297	299
4	-2.2411	2.7711	-3.1325	-1.1222	2.6437	-3.1354	-2.5141	2.2891	-3.1386	YNE	E E	1a26		333	335
5	-1.2686	2.3600	3.1325	-1.9649	-0.4169	-3.0400	-2.1999	2.2437	2.9827	ERI	E E	1a2z	A	80	82
6	-1.8460	2.5746	-3.0573	-1.5097	0.9387	-3.0317	-2.5344	2.3044	3.0166	VNI	ESE	1a2z	A	84	86
7	2.0767	-3.0426	-3.0942	-1.5326	1.2193	-3.1249	-2.5156	2.7913	3.0096	GDI	ESE	1a3k		125	127
8	-1.2955	2.4067	3.1035	-1.2057	-0.6474	3.0700	-1.7423	1.6883	-3.0514	HKD	ESE	1a40		261	263
9	-1.7668	2.2895	-3.1079	-2.4325	2.9845	-3.1146	-2.8288	0.9642	-3.1046	SHE	E E	1a41		82	84

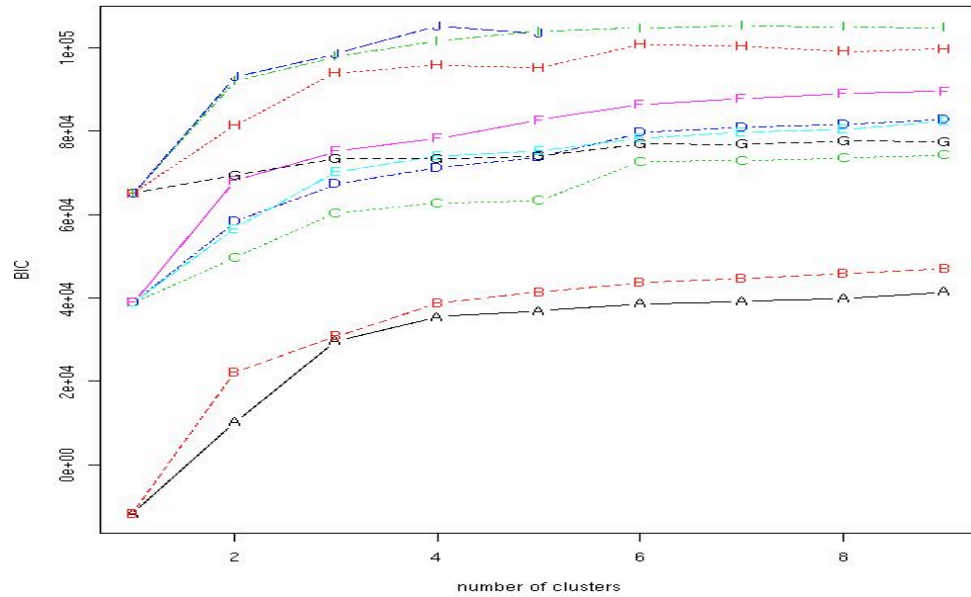
```
9 rows in set (0.04 sec)
```

# CLUSTERING TECHNIQUES

- Hierarchical
  - R function - *hclust*
- Partitional
  - Mclust – R package [Fraley,Raftery (2000)] - *EMclust*
- Clustered  $\text{cbind}(\sin(\theta), \cos(\theta))$  since space is periodic
- Chose  $n$  clusters such that BIC score was maximized, used same  $n$  for hierarchical clustering

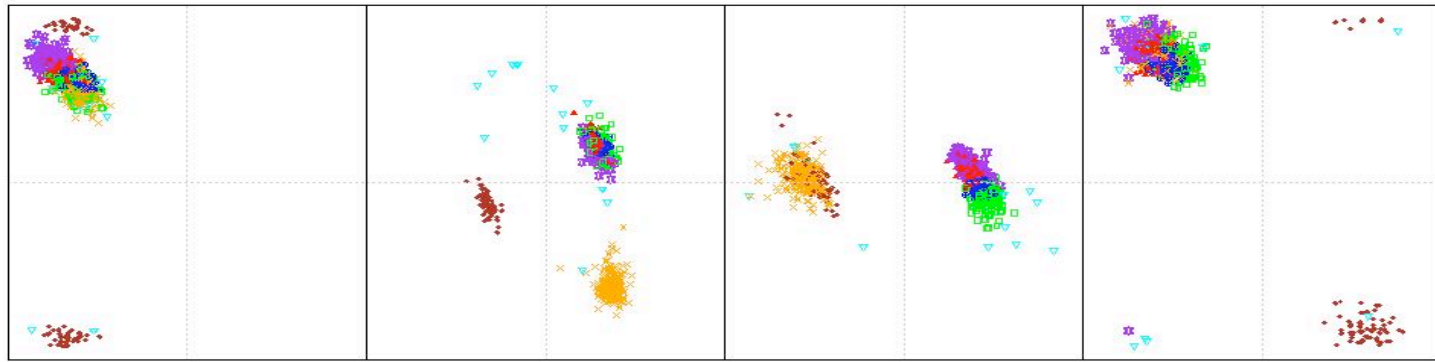
# RESULTS

• Input: 1156 x *4-residue hairpin turns*



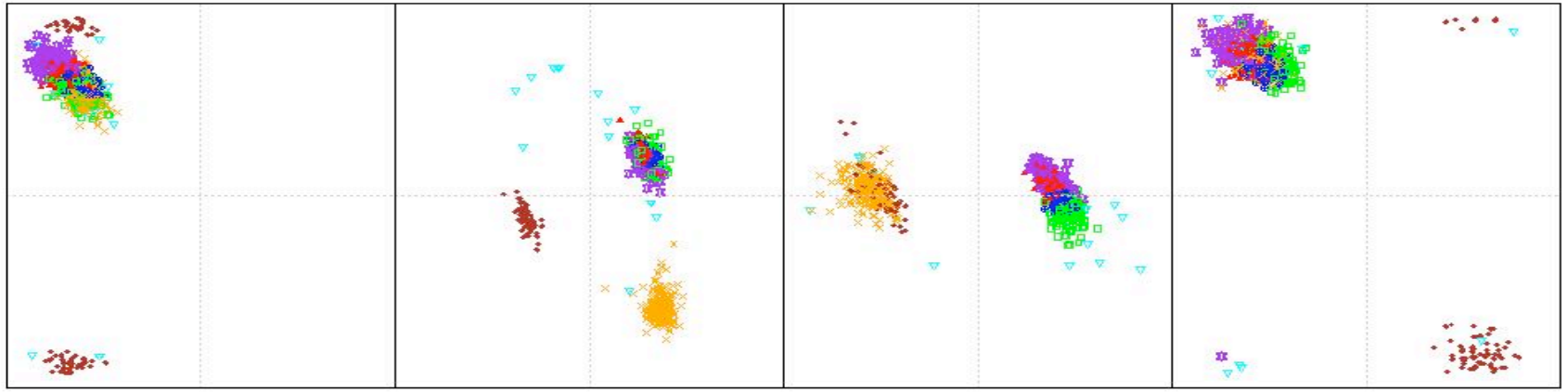
Model	Model
A	EII
B	VII
C	EEI
D	VEI
E	EVI
F	VVI
G	EEE
H	EEV
I	VEV
J	VVV

Output: Highest BIC: 7 clusters, VEV model (Equal shape, Variable volume & orientation)

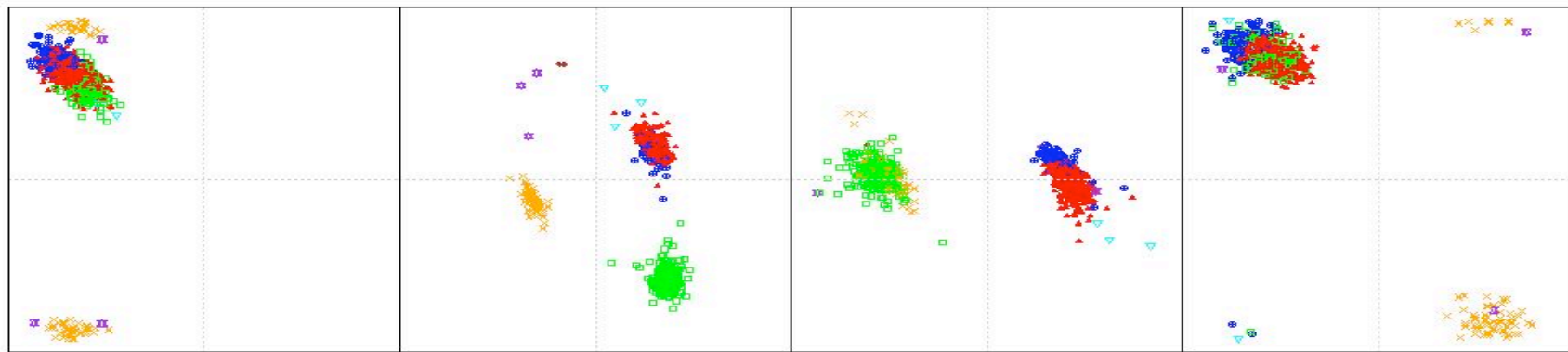




# RESULTS

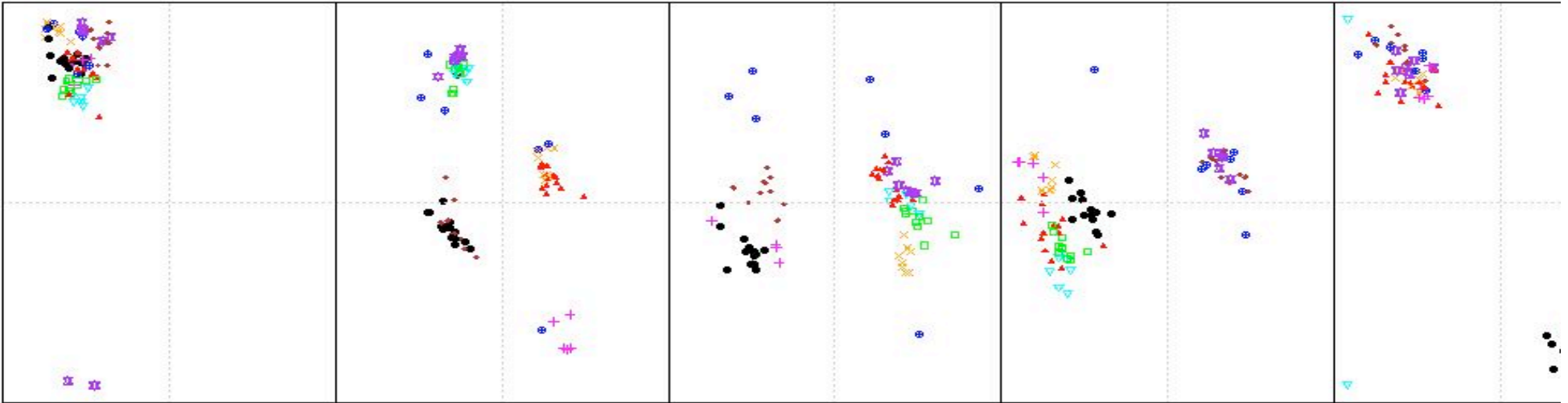


- Model based clustering (again)



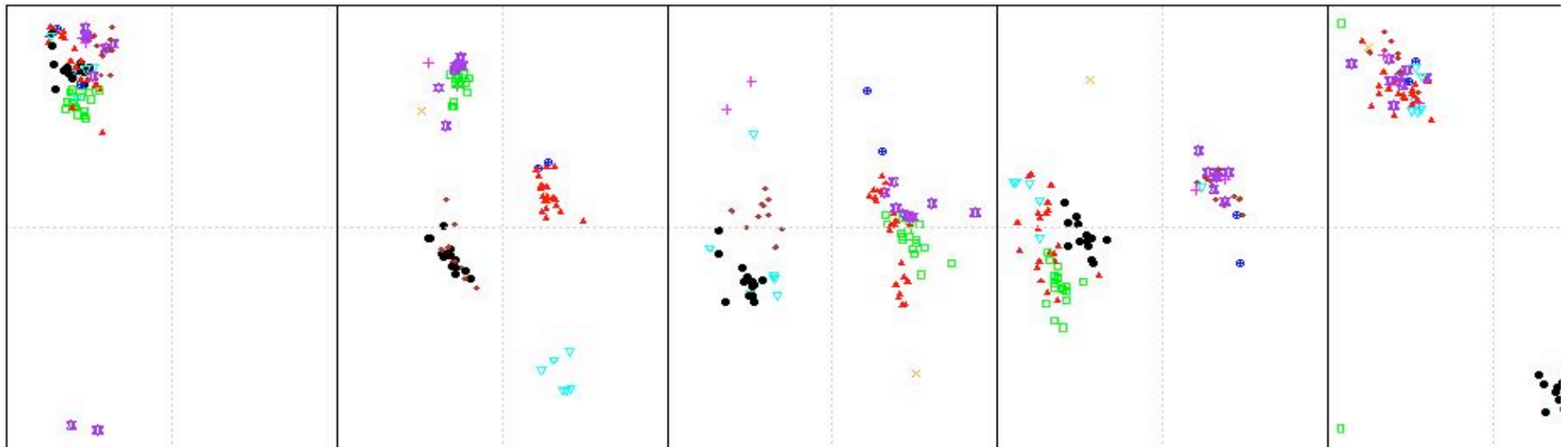
- Hierarchical clustering – different results wrt classification of outliers

# RESULTS

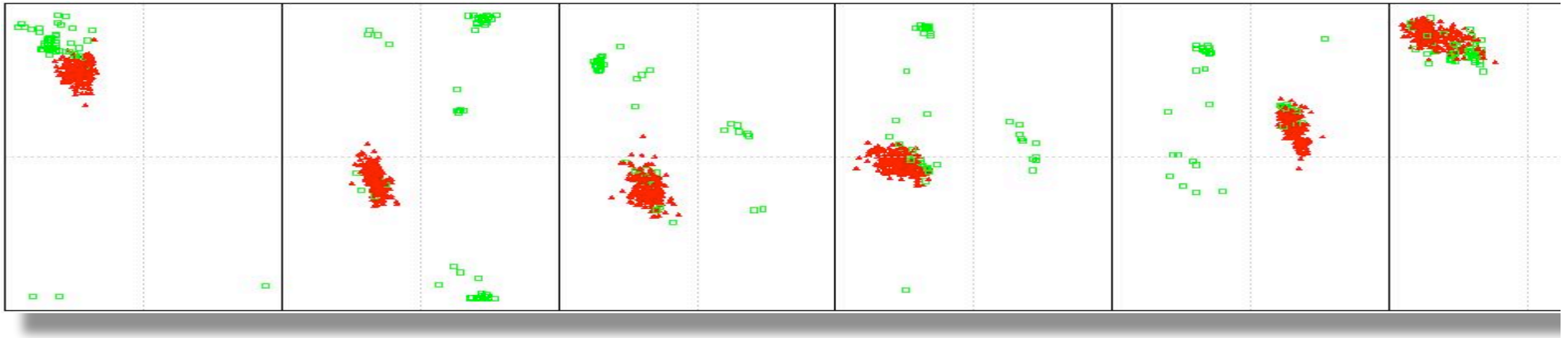


Model based (VVI, 11 clusters)

Hierarchical (11 clusters)

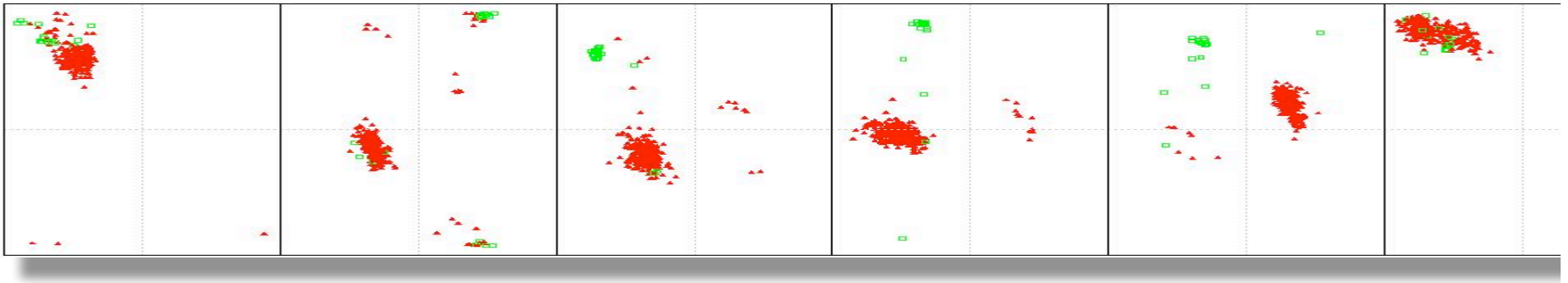


# RESULTS



Model based (VAV, 2 clusters)

Hierarchical (2 clusters)



# RELATIONSHIP TO SEQUENCE

- Clustered sequence in an *ad hoc* way:
  - Translate sequence 20 letter amino acid code to 2 letter amino acid code (Glycine or Other)
  - Each of  $n-1$  most frequent sequences is its own cluster
  - All other sequences are a single cluster

# MUTUAL INFORMATION

- Quantifies how much information about the value of one random variable is revealed by knowing the value of another random variable
- More appropriate than chi-square to use in this context, since each pair (x,y) is likely not to be observed  $\geq 5$  times
- Account for sampling bias by permuting the data and calculating independent and excess information

$$I(x, y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

$$\hat{I}_E(x, y | D) = I(x, y | D) - \frac{1}{N} \sum_{i=1}^N I(x, y | \sigma_i(D))$$

# RESULTS

- Mutual Information
  - More “self-information” in model based clusters
    - this just indicates that distribution of hierarchical clusters are not as balanced as model based clusters

L e n g t h	S v s S	H v s H	M v s M	S v s H	S v s M	H v s M
4	0 . 9	0 . 5	0 . 5	0 . 7	0 . 7	1 . 1
5	2 . 2	2 . 3	2 . 4	0 . 9	0 . 8	2
6	1	0 . 3	0 . 6	0 . 0 3	0 . 2	0 . 1 5

S = Sequence, H = Hierarchical, M = Model-based

# REFERENCES

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>.

Fiser A, Do RK, Sali A. Modeling of loops in protein structures. *Protein Sci.* 2000 Sep; 9(9):1753-73

Creighton T. *Proteins: Structure and Molecular Properties*. Freeman (1993)

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein with atomic level accuracy. *Science*. 2003. Nov 21(5659): 1364-8.

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.