Protein sub-family analysis

Geoff Barton
Division of Computational Biology
School of Life Sciences
University of Dundee, UK

twitter:@gjbarton
blog: geoffbarton.wordpress.com

www.compbio.dundee.ac.uk
www.jalview.org

Identification of functional sites

- Whole alignment methods
  - Simple visualisation
  - Calculation of “conservation values”
- Sub-family analysis
  - AMAS analysis
  - Tree determinant positions
  - “Evolutionary trace”

Example Multiple Sequence alignment of 27 SH2 domains

AMAS method of calculating conservation

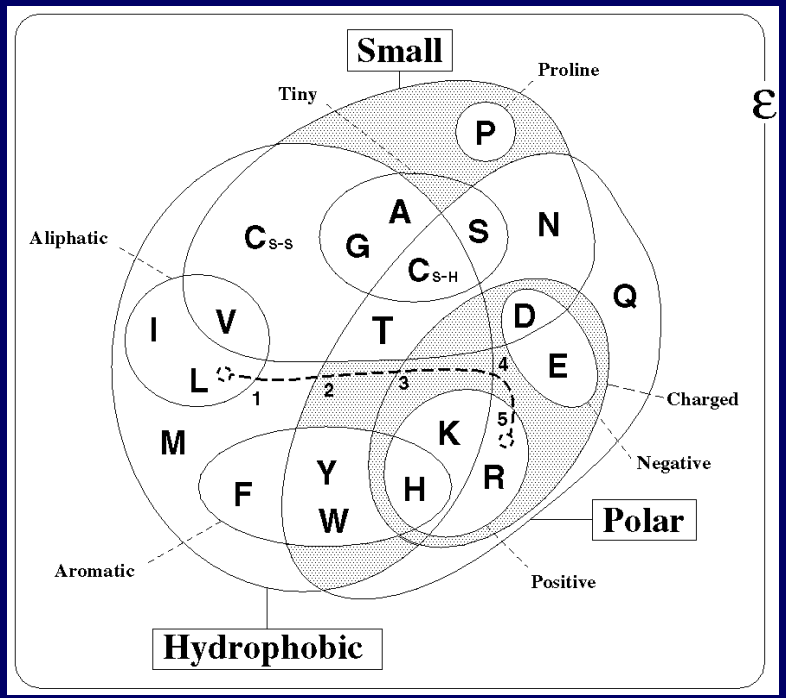
Taylor Venn diagram of amino acid properties

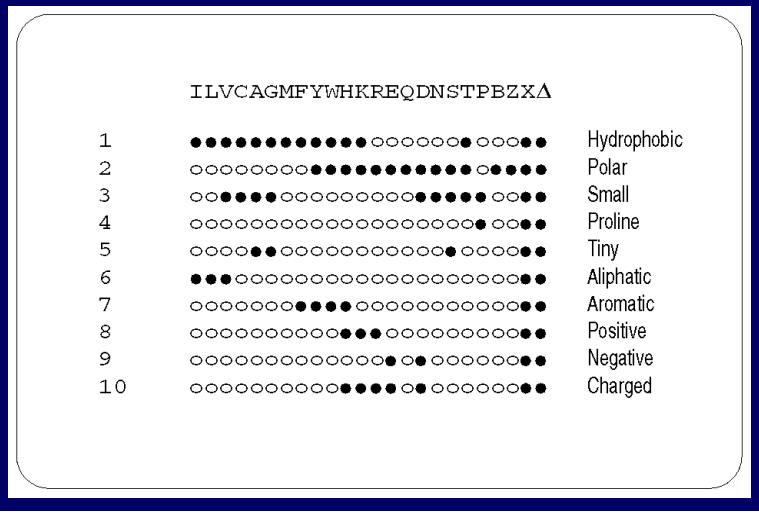
Count maximum number of set boundaries that must be crossed to include all amino acids at an alignment position

This gives a measure of the physico-chemical property variability at the alignment position

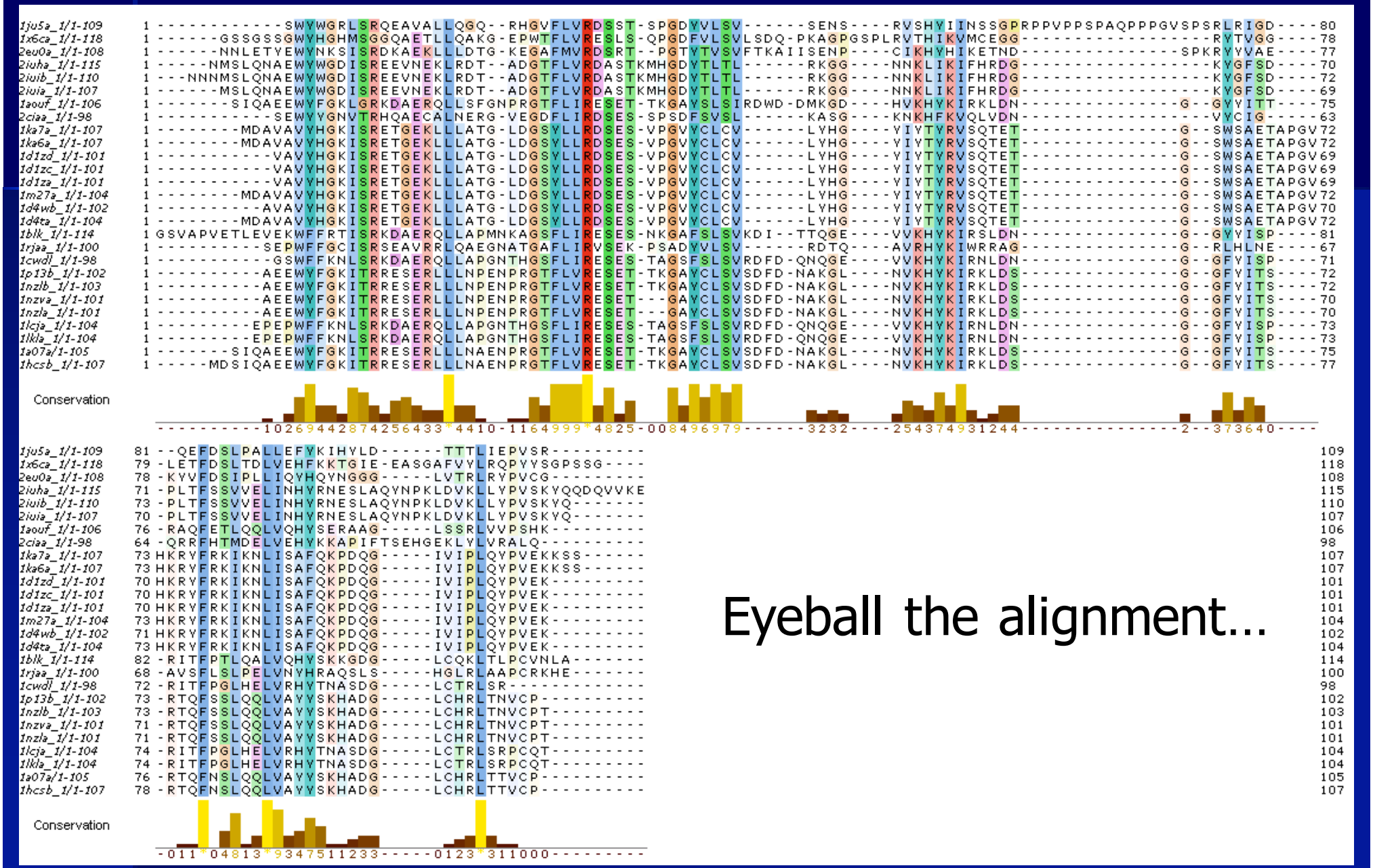
For comprehensive review of methods see:

Valdar, WS (2002) Proteins, 48, 227-41





Example Multiple Sequence alignment of 27 SH2 domains



Eyeball the alignment…

Identification of functional sites

- Whole alignment methods
  - Simple visualisation
  - Calculation of “conservation values”
- Sub-family analysis
  - AMAS analysis
  - “Tree determinant” positions
  - “Evolutionary trace”

Sequence analysis of the Annexins: An example of sub-family analysis

- "Large" number of sequences (for 1990)
- Possess multiple domains
- Unknown tertiary structure at the time of analysis

- Barton, G. J., Freemont, P. F., Newman, R. & Crumpton, M. (1991), "Sequence Analysis of the Annexin Super Gene Family of Proteins" *Eur. J. Biochem*, **198,** 749-760.

Annexins

Calcium and phospholipid binding

Wide family - 22 known sequences

(Insect - Human)

Found in many cell types

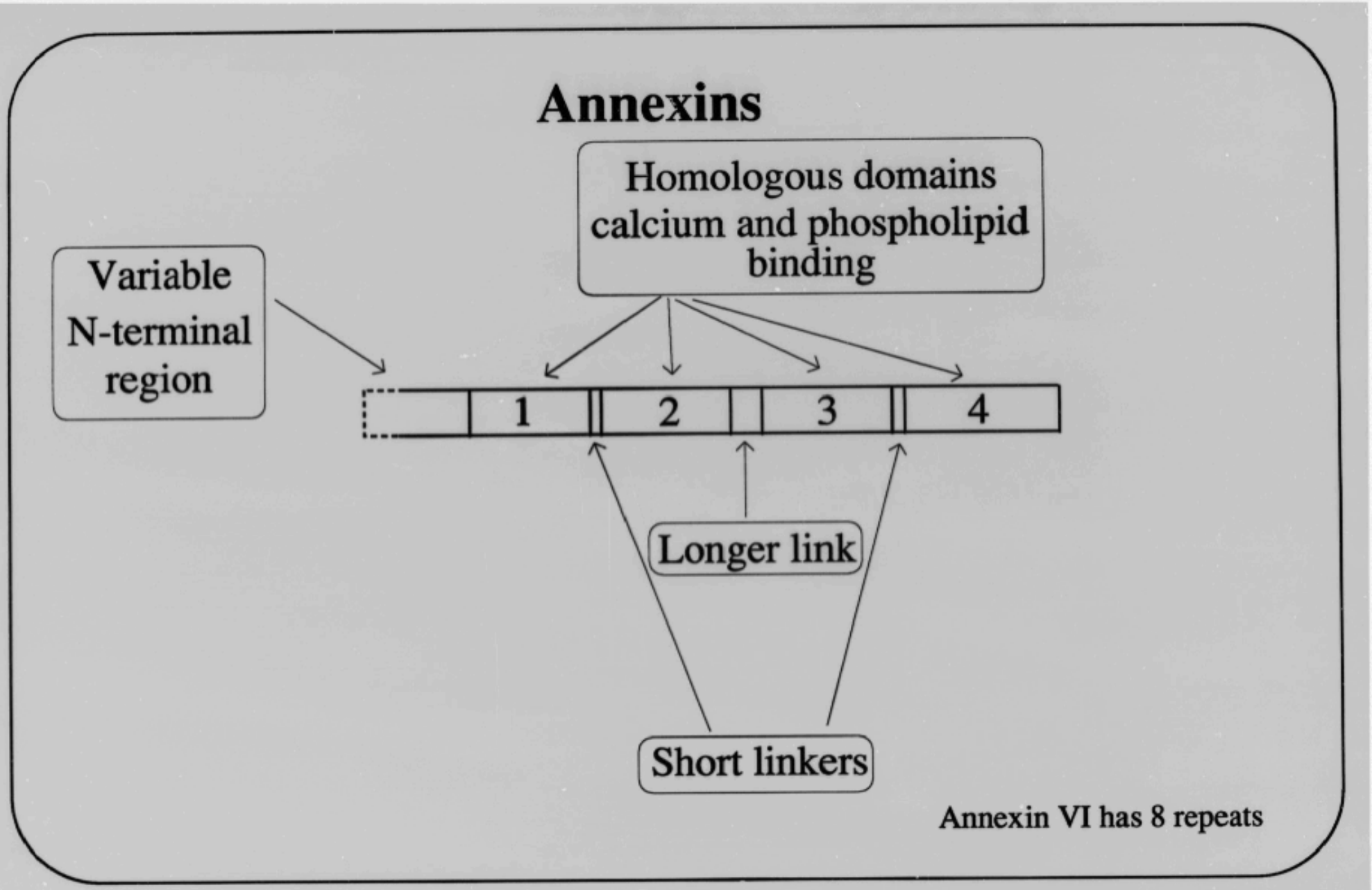
Implicated in

membrane fusion

exocytosis

cell signalling

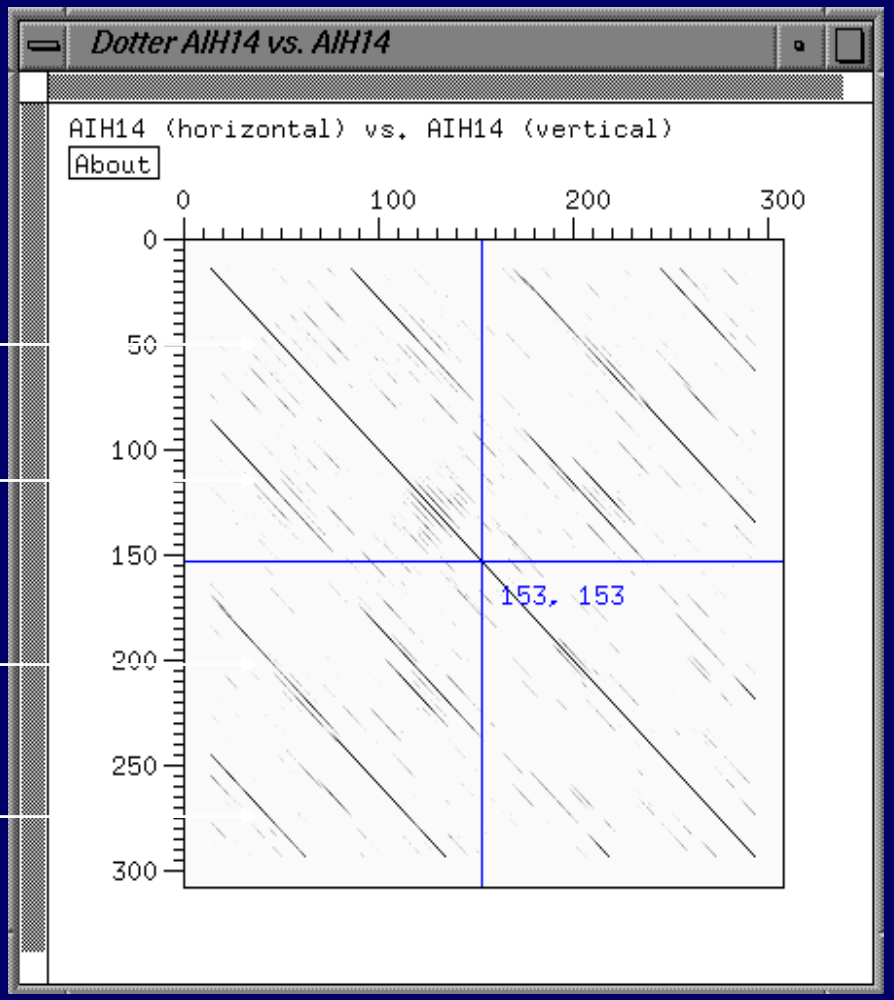
anti-inflammatory properties

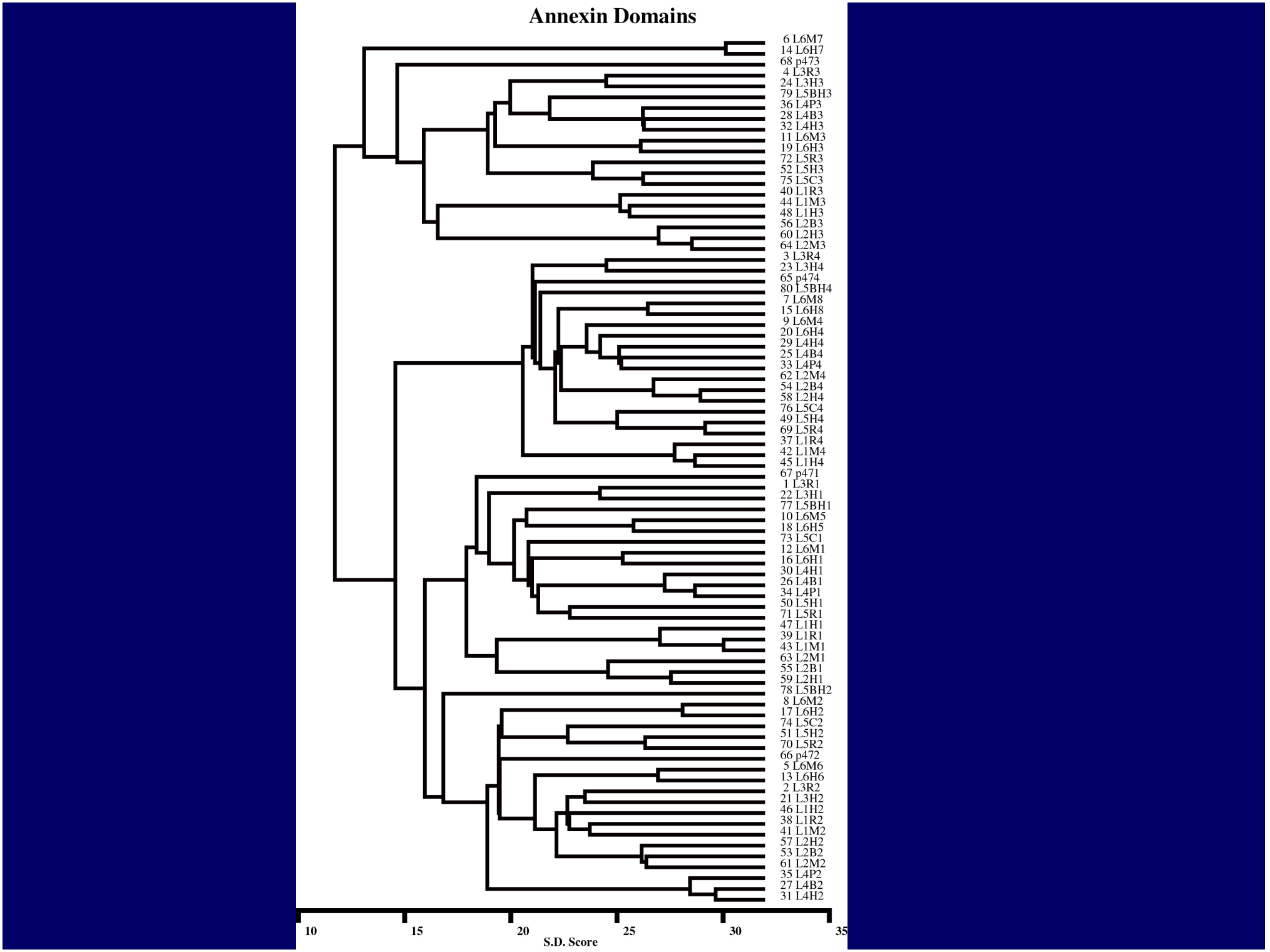

Annexins
Homologous domains calcium and phospholipid binding
Variable N-terminal region
1
2
3
4
Longer link
Short linkers
Annexin VI has 8 repeats

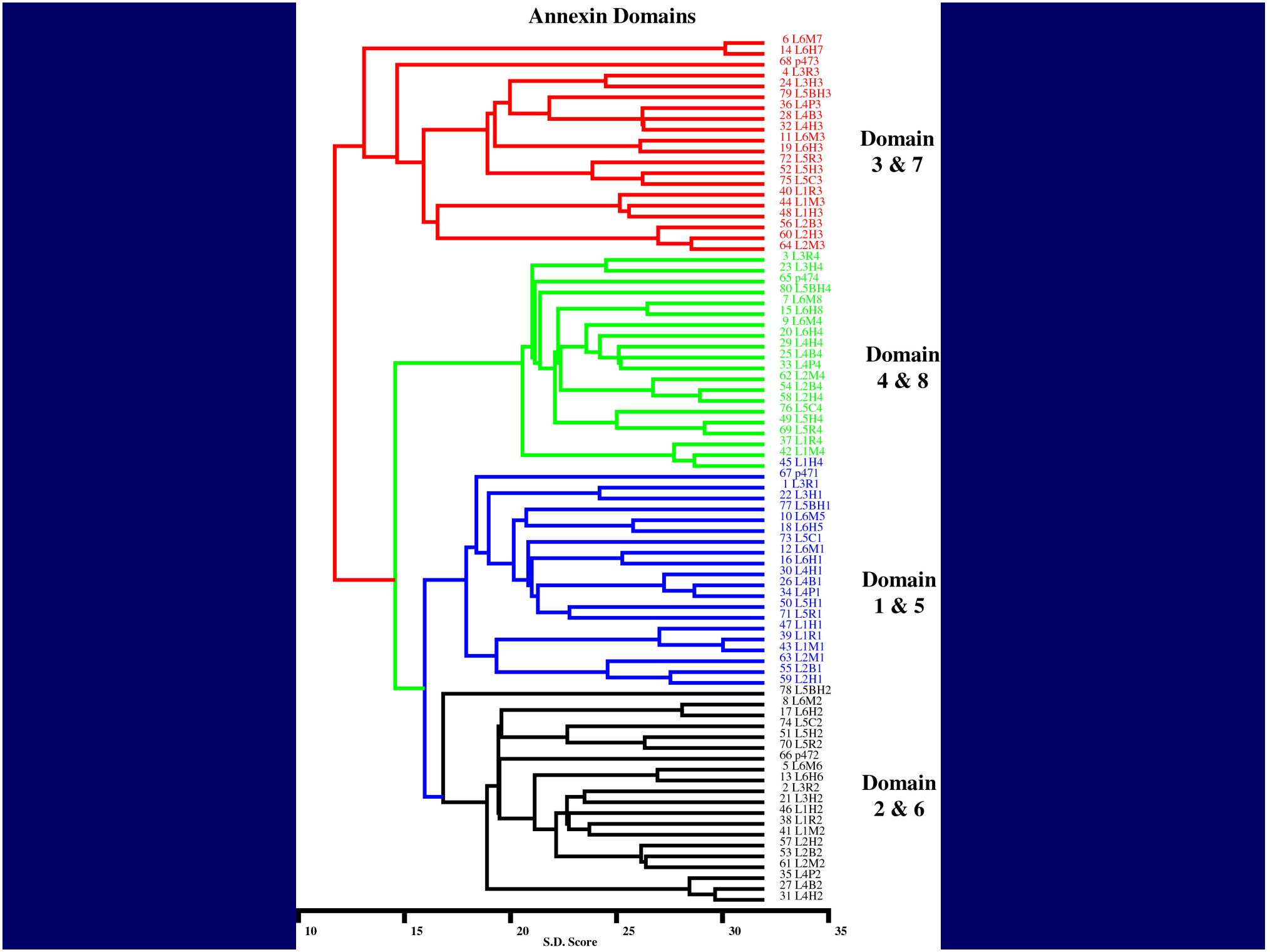
Sequence Analysis of Annexin Domains

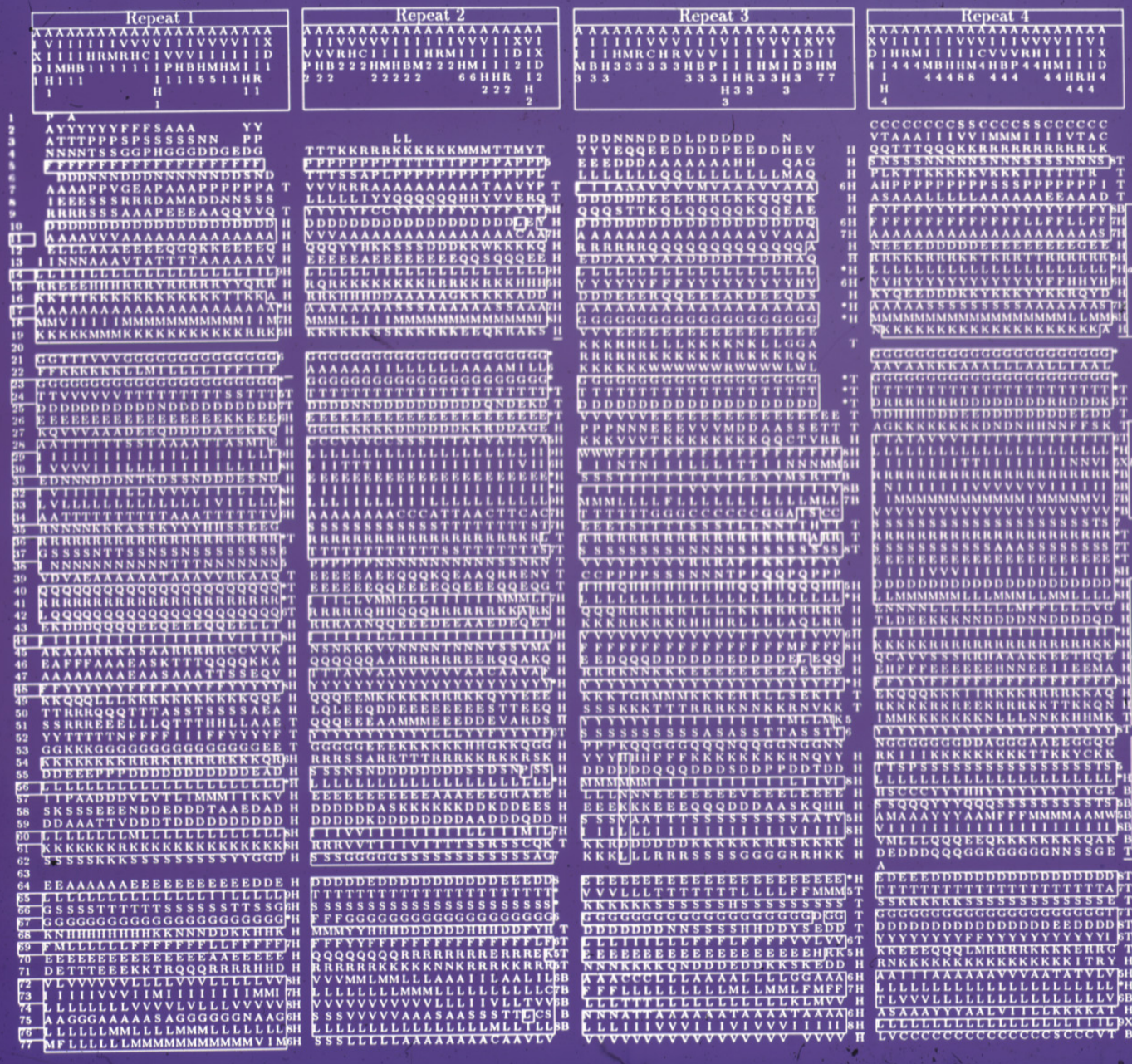
Dot-Plot comparison of Human Annexin I with itself.

Four repeats (domains ?) are visible.

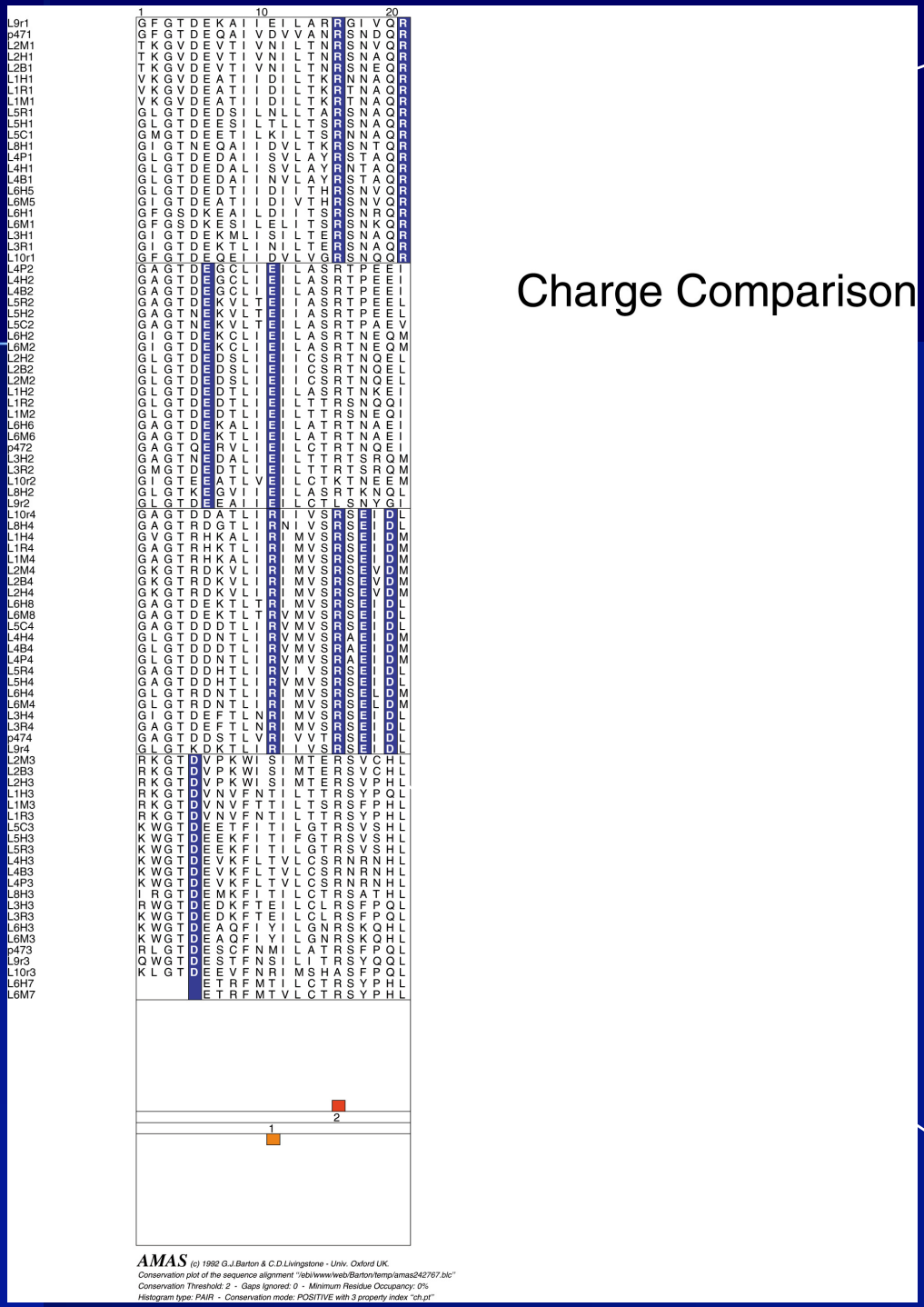


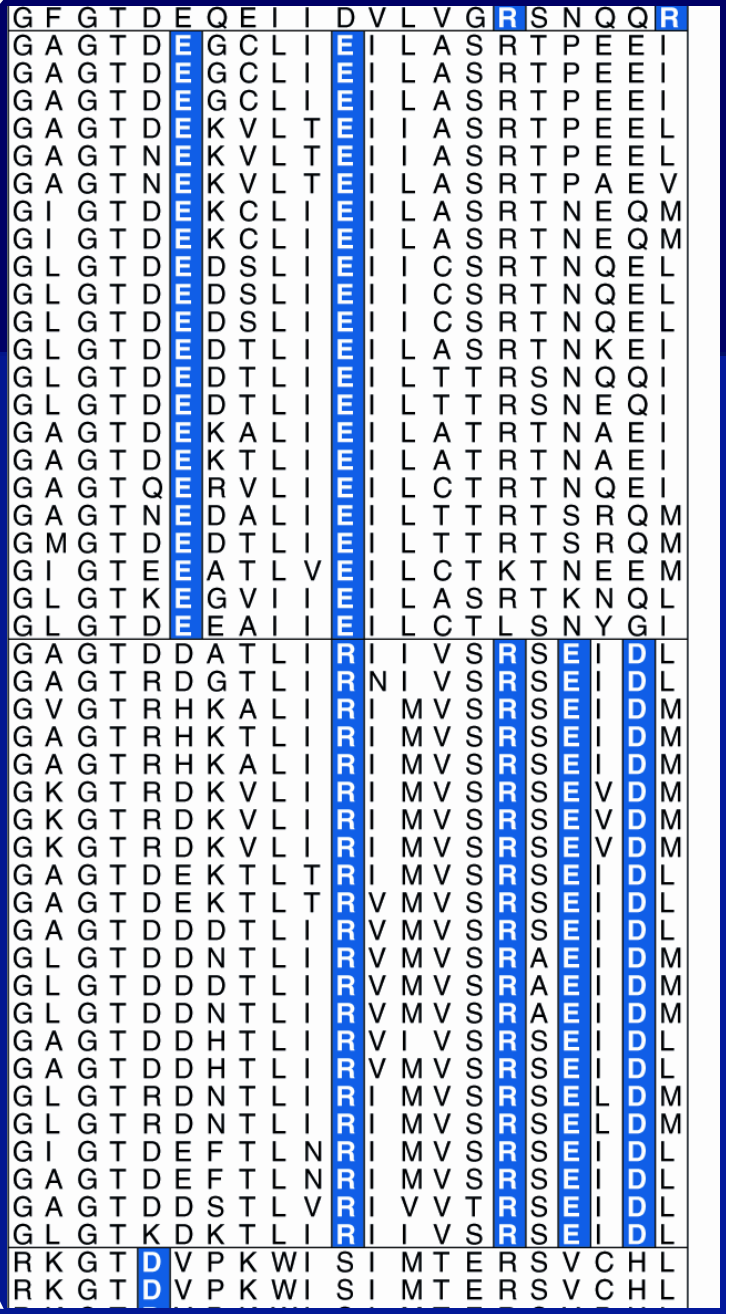

Annexin Domains
6 L6M7
14 L6H7
68 p473
4 L3R3
24 L3H3
79 L5BH3
36 L4P3
28 L4B3
32 L4H3
11 L6M3
19 L6H3
72 L5R3
52 L5H3
75 L5C3
40 L1R3
44 L1M3
48 L1H3
56 L2B3
60 L2H3
64 L2M3
3 L3R4
23 L3H4
65 p474
80 L5BH4
7 L6M8
15 L6H8
9 L6M4
20 L6H4
29 L4H4
25 L4B4
33 L4P4
62 L2M4
54 L2B4
58 L2H4
76 L5C4
49 L5H4
69 L5R4
37 L1R4
42 L1M4
45 L1H4
67 p471
1 L3R1
22 L3H1
77 L5BH1
10 L6M5
18 L6H5
73 L5C1
12 L6M1
16 L6H1
30 L4H1
26 L4B1
34 L4P1
50 L5H1
71 L5R1
47 L1H1
39 L1R1
43 L1M1
63 L2M1
55 L2B1
59 L2H1
78 L5BH2
8 L6M2
17 L6H2
74 L5C2
51 L5H2
70 L5R2
66 p472
5 L6M6
13 L6H6
2 L3R2
21 L3H2
46 L1H2
38 L1R2
41 L1M2
57 L2H2
53 L2B2
61 L2M2
35 L4P2
27 L4B2
31 L4H2
10
15
20
25
30
35
S.D. Score


Annexin Domains
6 L6M7
14 L6H7
68 p473
4 L3R3
24 L3H3
79 L5BH3
36 L4P3
28 L4B3
32 L4H3
11 L6M3
19 L6H3
72 L5R3
52 L5H3
75 L5C3
40 L1R3
44 L1M3
48 L1H3
56 L2B3
60 L2H3
64 L2M3
3 L3R4
23 L3H4
65 p474
80 L5BH4
7 L6M8
15 L6H8
9 L6M4
20 L6H4
29 L4H4
25 L4B4
33 L4P4
62 L2M4
54 L2B4
58 L2H4
76 L5C4
49 L5H4
69 L5R4
37 L1R4
42 L1M4
45 L1H4
67 p471
1 L3R1
22 L3H1
77 L5BH1
10 L6M5
18 L6H5
73 L5C1
12 L6M1
16 L6H1
30 L4H1
26 L4B1
34 L4P1
50 L5H1
71 L5R1
47 L1H1
39 L1R1
43 L1M1
63 L2M1
55 L2B1
59 L2H1
78 L5BH2
8 L6M2
17 L6H2
74 L5C2
51 L5H2
70 L5R2
66 p472
5 L6M6
13 L6H6
2 L3R2
21 L3H2
46 L1H2
38 L1R2
41 L1M2
57 L2H2
53 L2B2
61 L2M2
35 L4P2
27 L4B2
31 L4H2
Domain 3 & 7
Domain 4 & 8
Domain 1 & 5
Domain 2 & 6
10
15
20
25
30
35
S.D. Score


Repeat 1
Repeat 2
Repeat 3
Repeat 4

Charge Comparison





Annexin
Predictions

1. 5 Helices

2. Core residues (hydrophobic patterns)

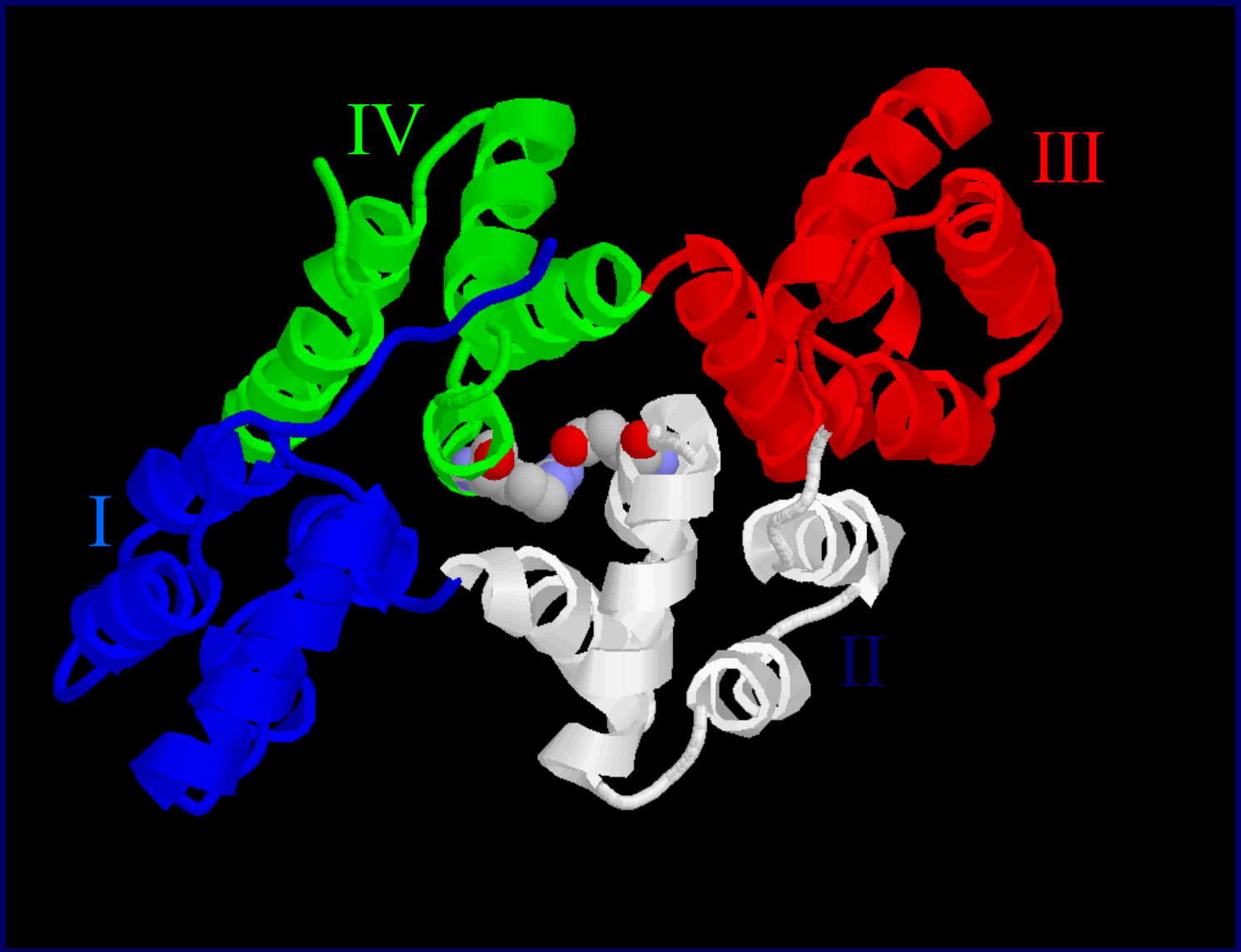
3. Conserved Glu in repeats II
   and Arg in repeats IV
   form a salt bridge

4. Helix a in repeat III shorter

5. Not like uteroglobin

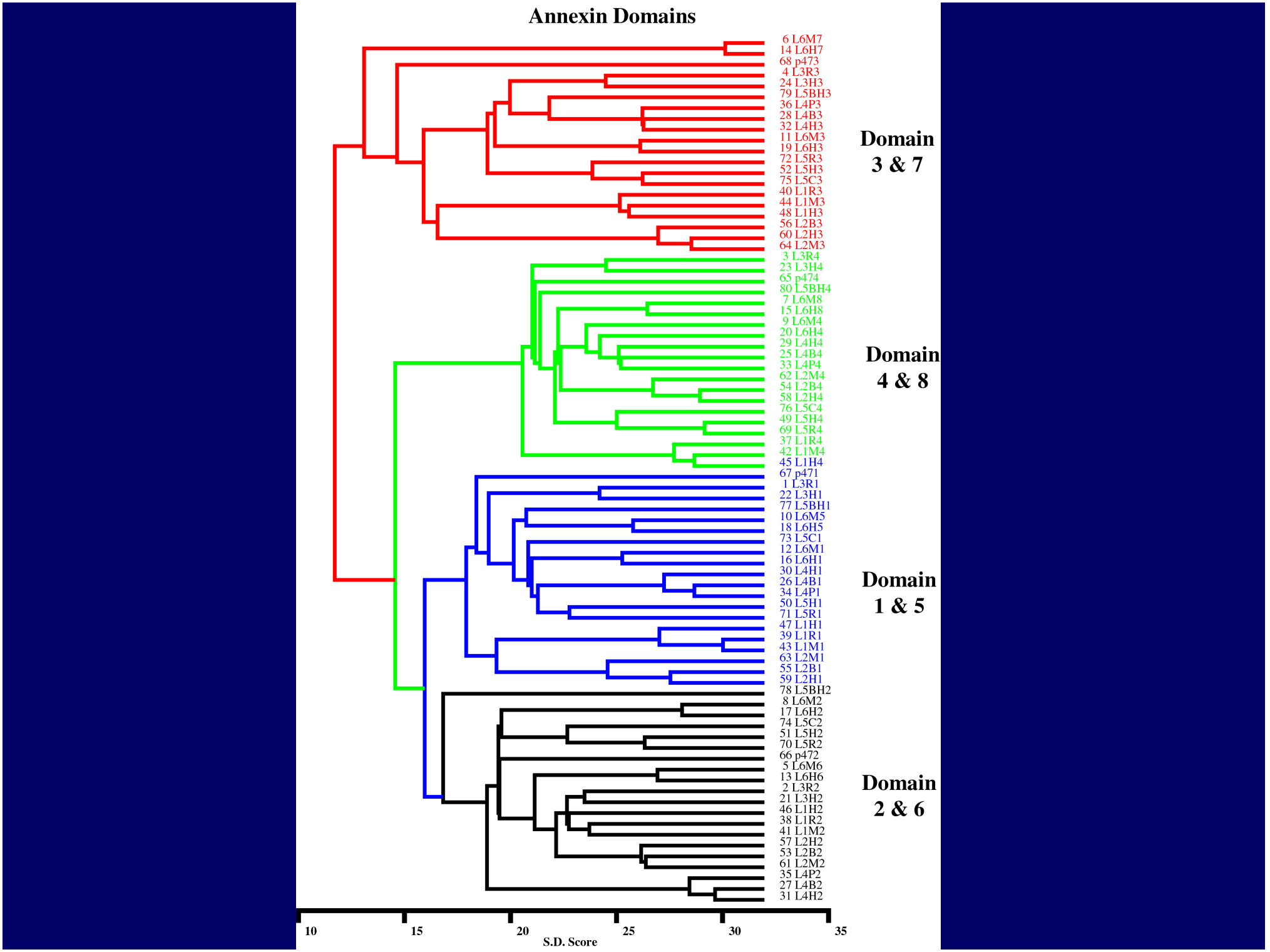
6. Helix a - helix b loop important in repeat III

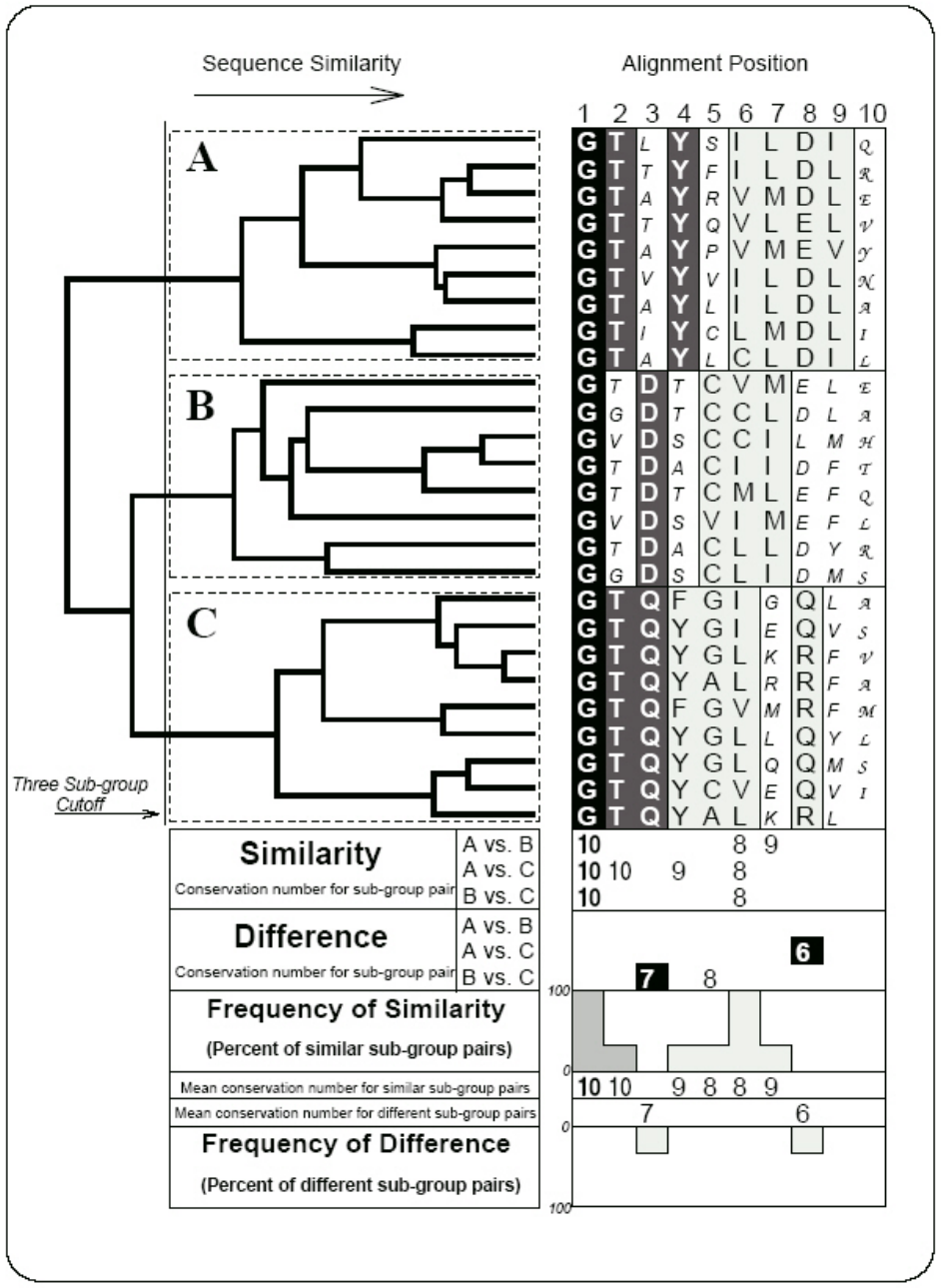
Annexin V showing Glu-Arg salt bridge between helix 2 of domain II and helix 2 of domain IV



Analysis of similarities and differences between sub-families can reveal functionally important residues

Generalise lessons learned in Annexin study


Annexin Domains
6 L6M7
14 L6H7
68 p473
4 L3R3
24 L3H3
79 L5BH3
36 L4P3
28 L4B3
32 L4H3
11 L6M3
19 L6H3
72 L5R3
52 L5H3
75 L5C3
40 L1R3
44 L1M3
48 L1H3
56 L2B3
60 L2H3
64 L2M3
3 L3R4
23 L3H4
65 p474
80 L5BH4
7 L6M8
15 L6H8
9 L6M4
20 L6H4
29 L4H4
25 L4B4
33 L4P4
62 L2M4
54 L2B4
58 L2H4
76 L5C4
49 L5H4
69 L5R4
37 L1R4
42 L1M4
45 L1H4
67 p471
1 L3R1
22 L3H1
77 L5BH1
10 L6M5
18 L6H5
73 L5C1
12 L6M1
16 L6H1
30 L4H1
26 L4B1
34 L4P1
50 L5H1
71 L5R1
47 L1H1
39 L1R1
43 L1M1
63 L2M1
55 L2B1
59 L2H1
78 L5BH2
8 L6M2
17 L6H2
74 L5C2
51 L5H2
70 L5R2
66 p472
5 L6M6
13 L6H6
2 L3R2
21 L3H2
46 L1H2
38 L1R2
41 L1M2
57 L2H2
53 L2B2
61 L2M2
35 L4P2
27 L4B2
31 L4H2
Domain 3 & 7
Domain 4 & 8
Domain 1 & 5
Domain 2 & 6
10
15
20
25
30
35
S.D. Score



KEY	
D	Identical across all sequences
Y	Identical within one sub-group
Q	Conserved within one sub-group
E	Unconserved within one sub-group
S	Unconserved across all sequences

Figure 3

Principles of sub-family analysis

See what happens to conservation when you put two sub-families Together.

Does it stay high?
Implies- position is important to both and doing a similar job.

Does it go from high to low?
Implies- position is important to both but the position is important for novel features of the two sub-families.

References on Sub-family analysis

- Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci. 1993 Dec;9(6):745-56.
- Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol. 1996;266:497-512.

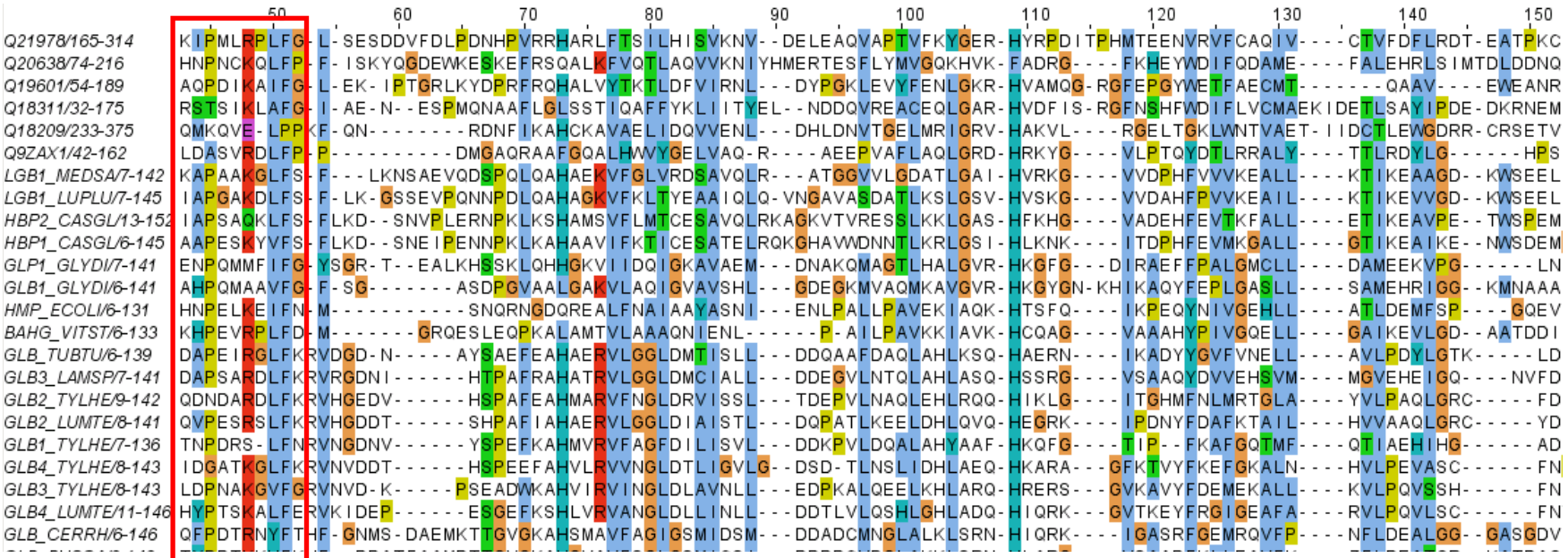
- Methods Enzymol paper includes summary of first paper.
- Copies are available on: http://www.compbio.dundee.ac.uk/ftp/pdf/

Identification of functional sites

- Whole alignment methods
  - Simple visualisation
  - Calculation of “conservation values”
- Sub-family analysis
  - AMAS analysis
  - “Tree determinant” positions
  - “Evolutionary trace”

Tree determinants

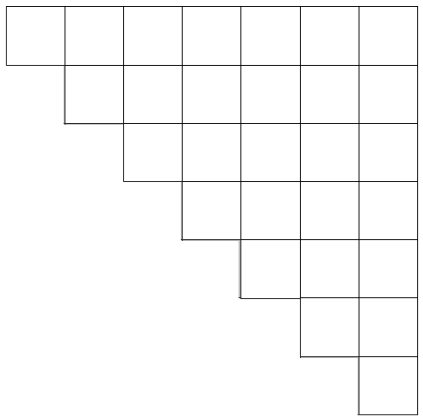
- Positions in the alignment that are most responsible for the topology of the phylogenetic tree derived from the complete alignment
- These positions may be functionally important.

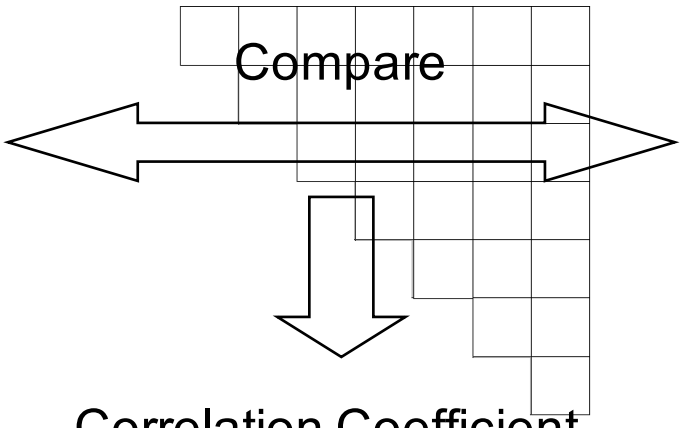


Generate Local Matrix

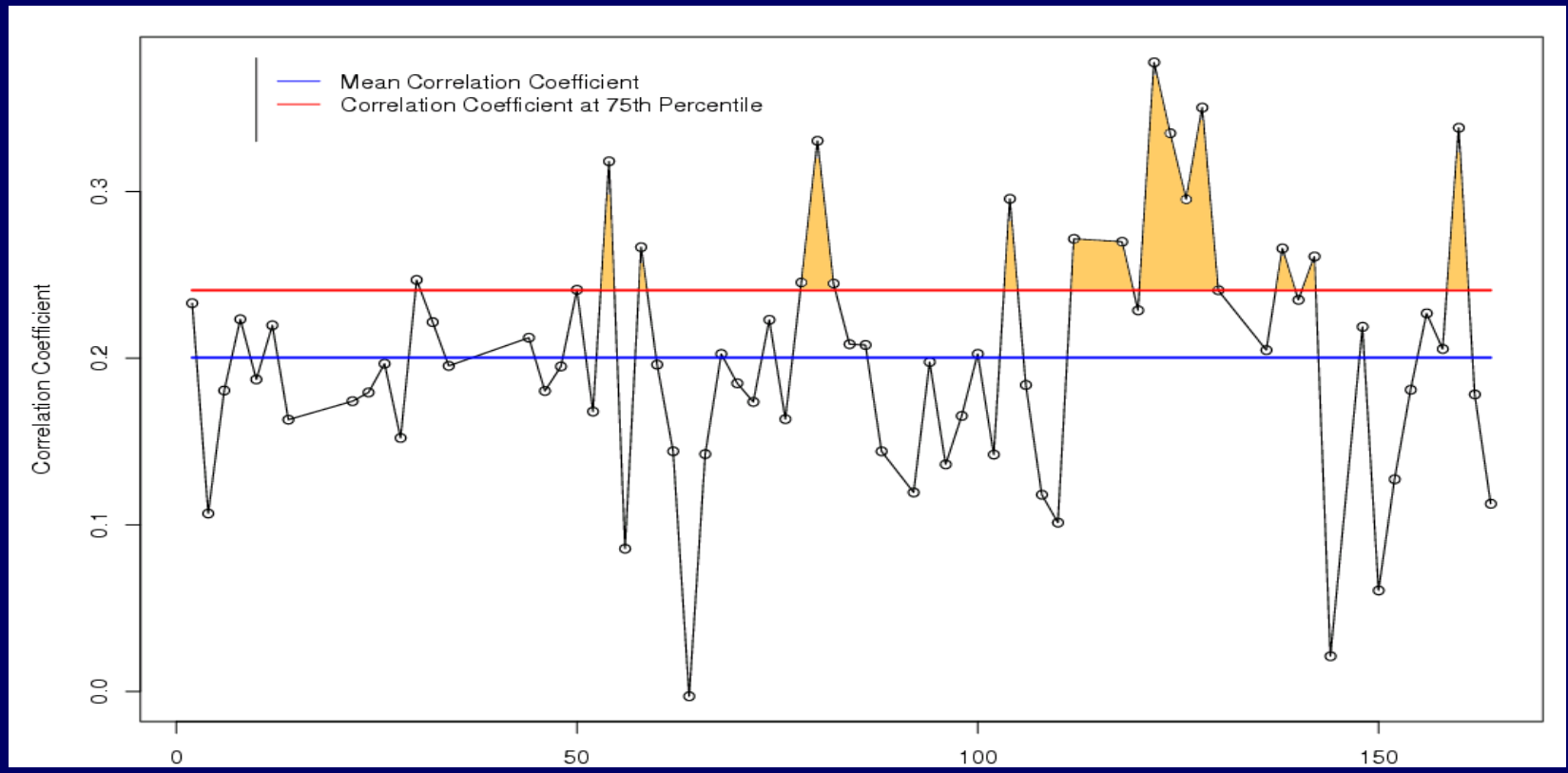
Generate Global Distance Matrix

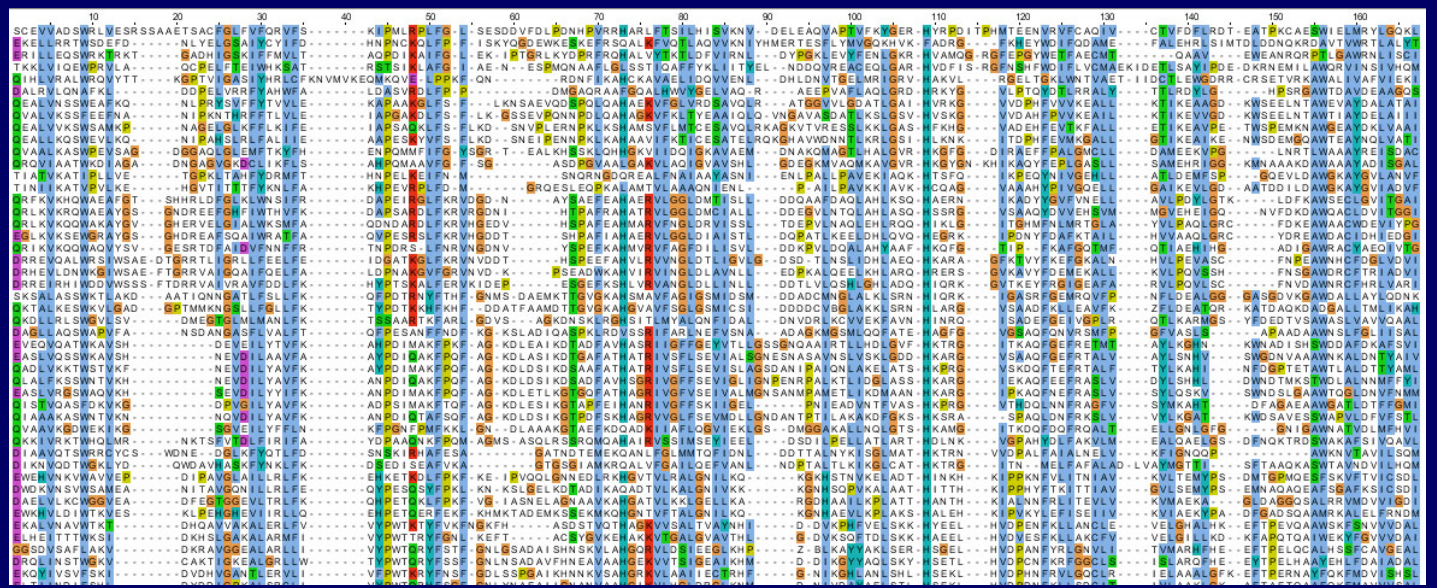
Saved Global Matrix





Correlation Coefficient





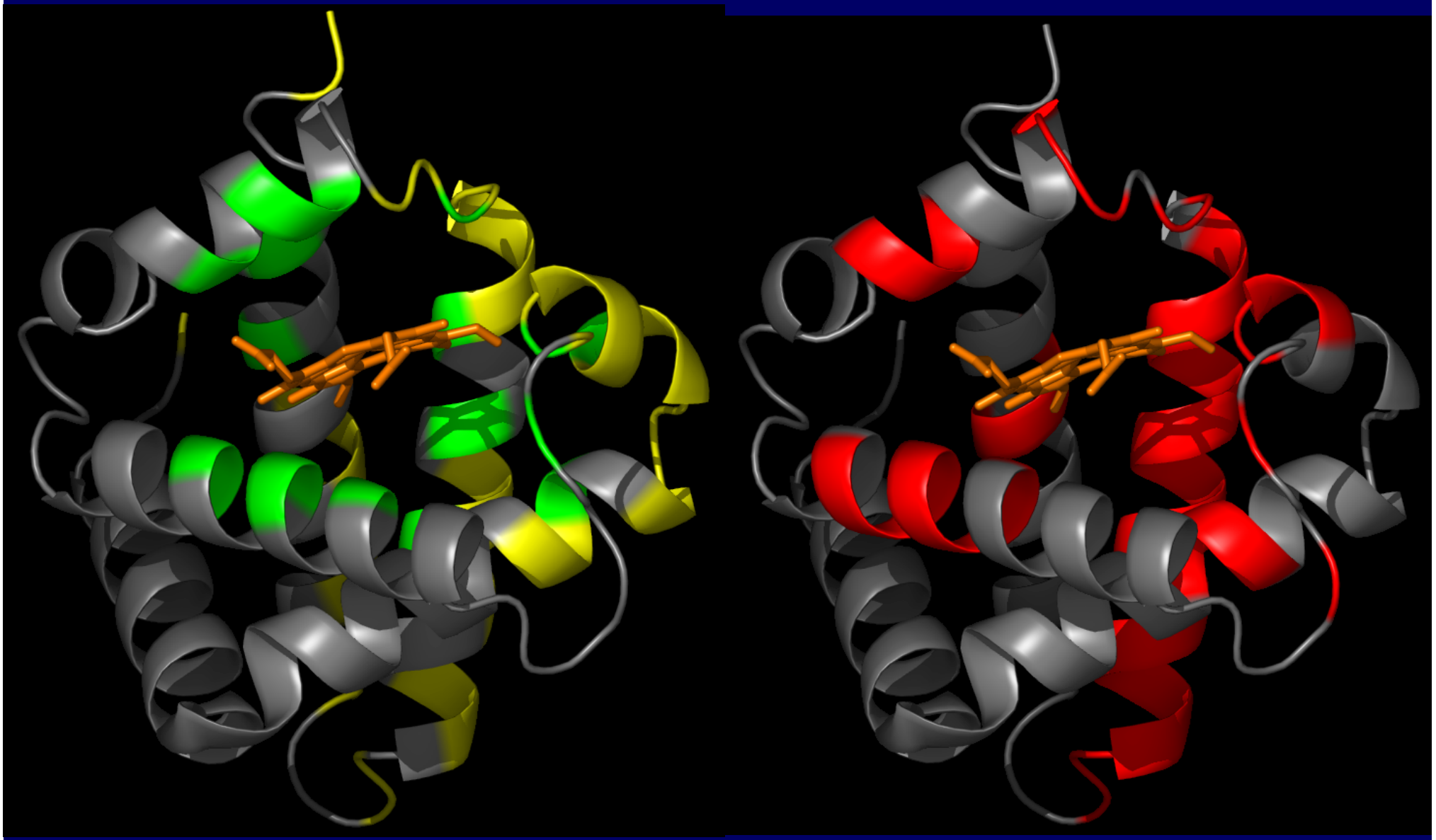
Partial Multiple Sequence Alignment

(Globins)

Observed Haeme Contacts (green)

Observed Protein Contacts (yellow)

Prediction (window size 3, interval 2)



Papers on tree determinants/predicting functional sites

- del Sol Mesa A., Pazos F., Valencia A. Automatic methods for predicting functionally important residues. J. Mol. Biol. 2003;326:1289–1302. [PubMed]

- Mihalek I., Res I., Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. J. Mol. Biol. 2004;336:1265–1282. [PubMed]

Practicals on sub-family analysis