

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

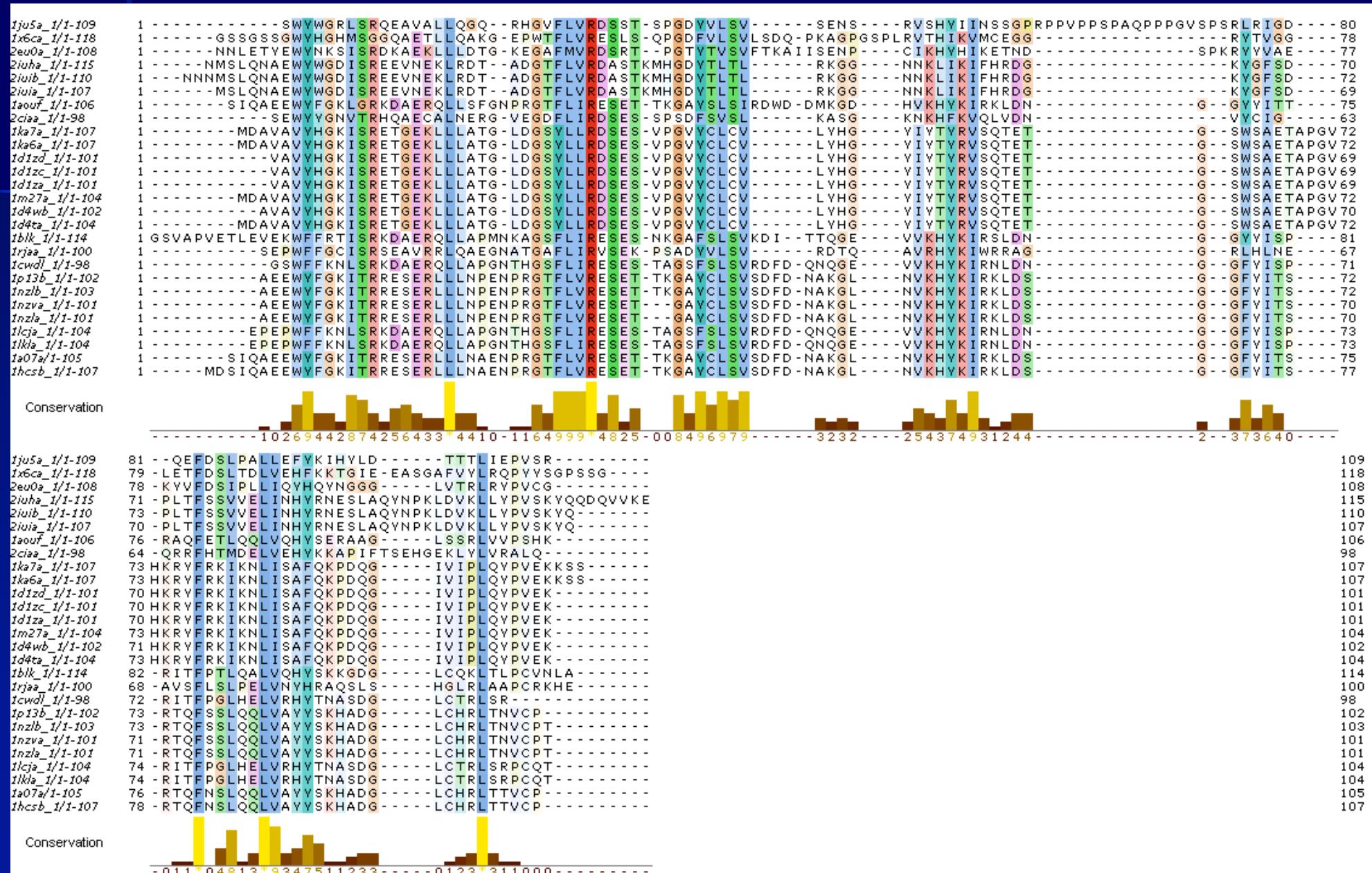
www.jalview.org

Identification of functional sites

- Whole alignment methods
 - Simple visualisation
 - Calculation of “conservation values”

- Sub-family analysis
 - AMAS analysis
 - Tree determinant positions
 - “Evolutionary trace”

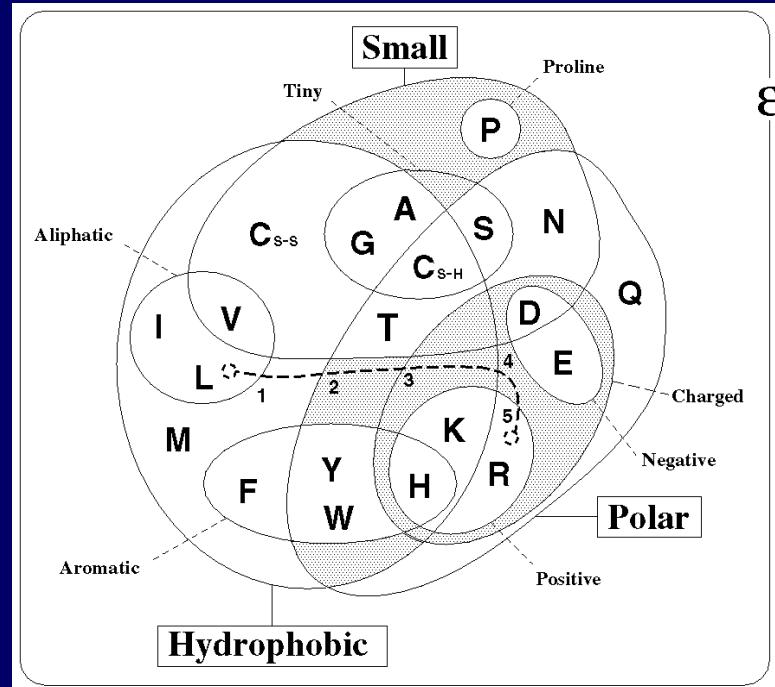
Example Multiple Sequence alignment of 27 SH2 domains



AMAS method of calculating conservation

Taylor Venn diagram of amino acid properties

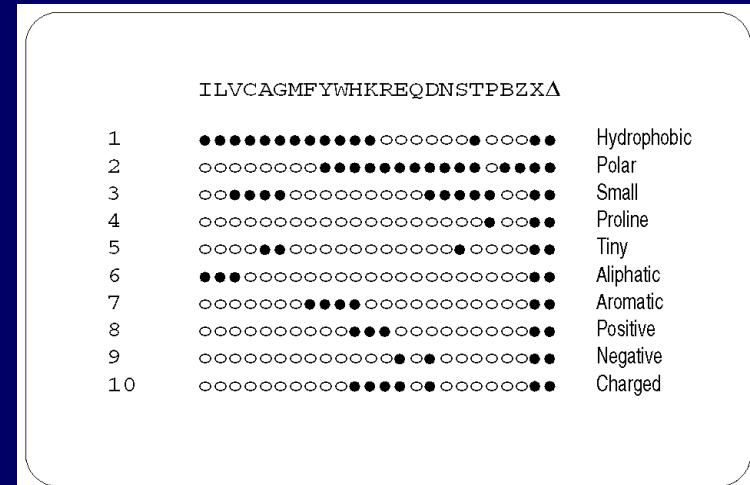
Count maximum number of set boundaries that must be crossed to include all amino acids at an alignment position



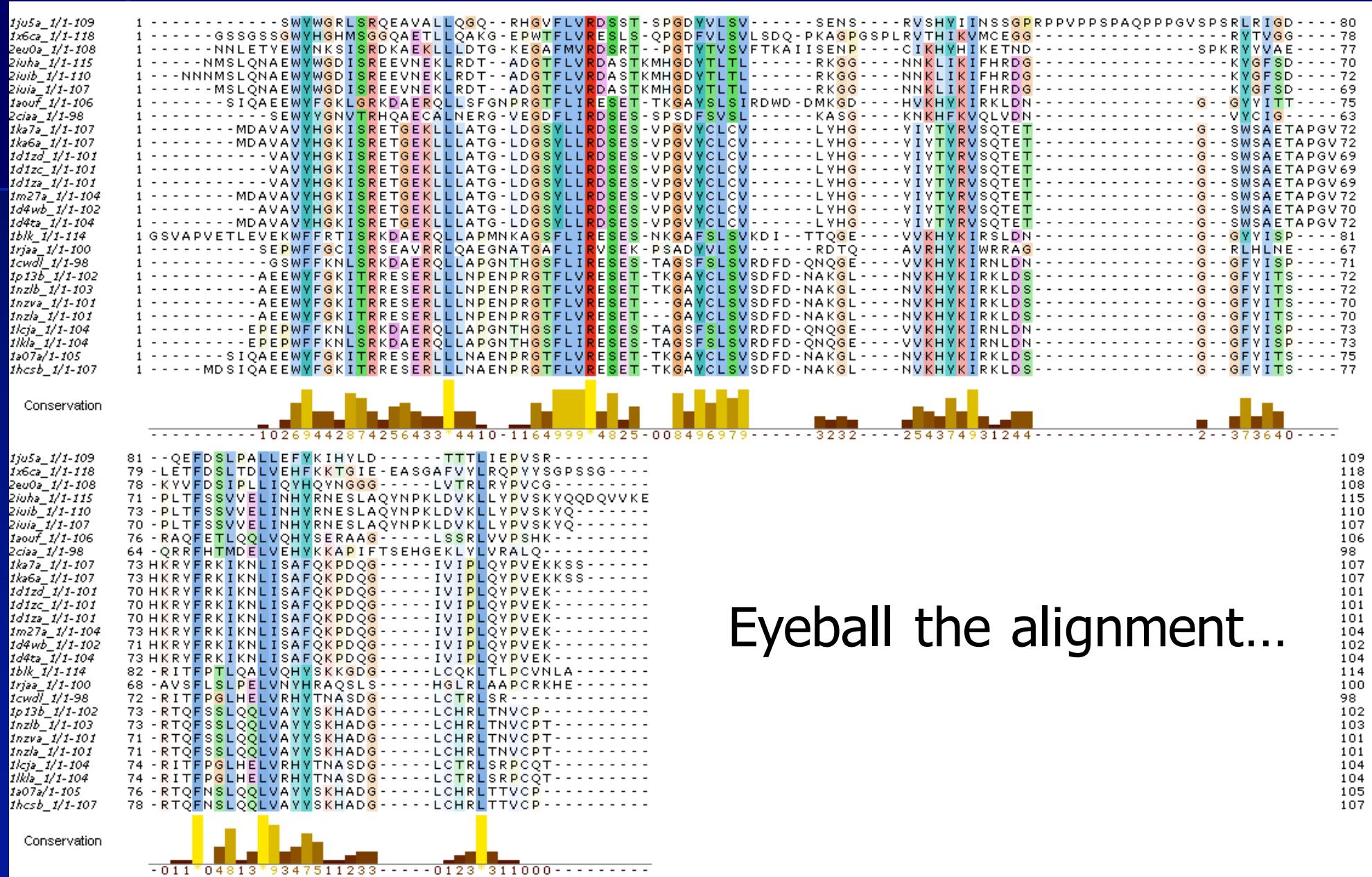
This gives a measure of the physico-chemical property variability at the alignment position

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

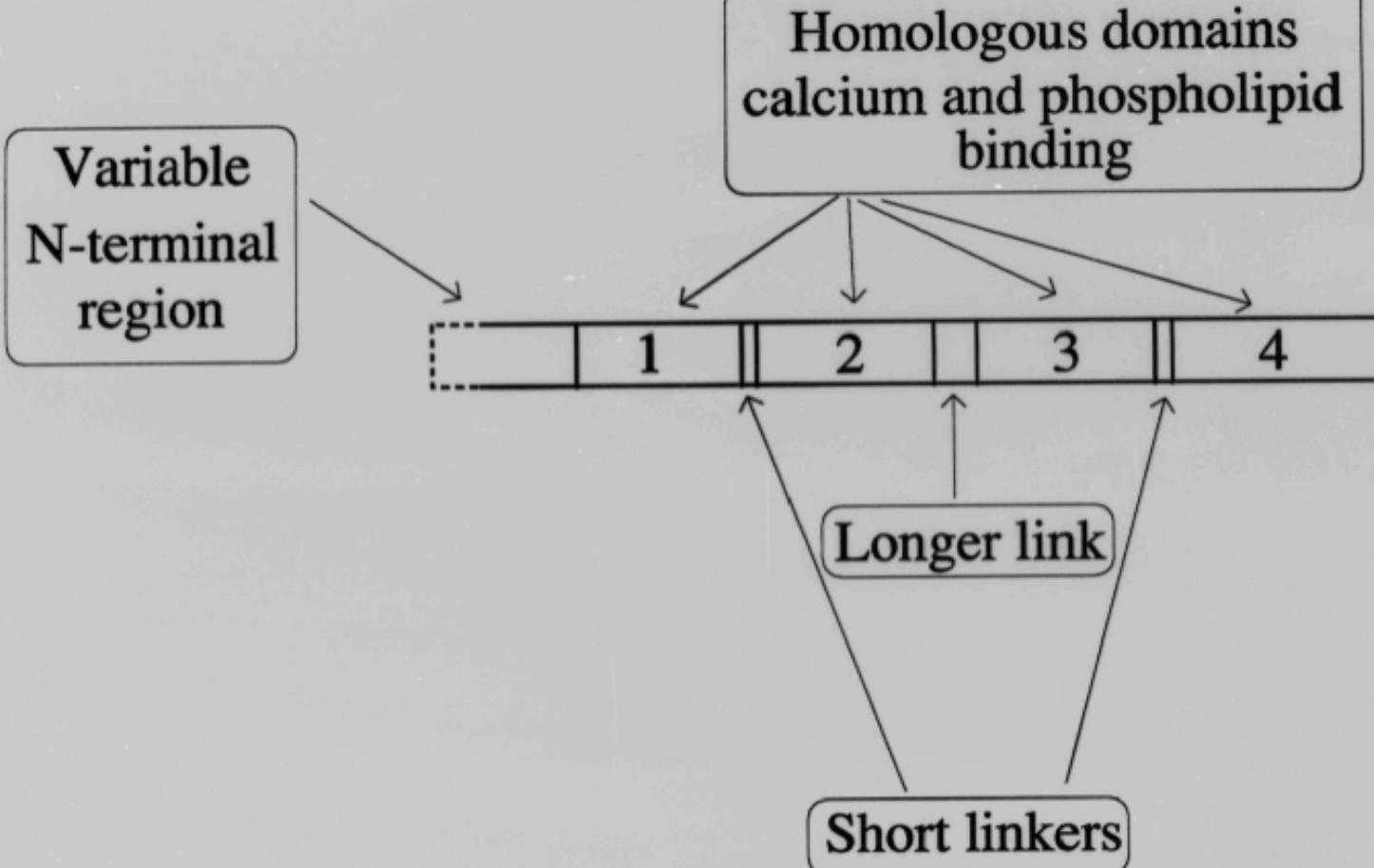
Calcium and phospholipid binding

Wide family - 22 known sequences
(Insect - Human)

Found in many cell types

Implicated in
membrane fusion
exocytosis
cell signalling
anti-inflammatory properties

Annexins

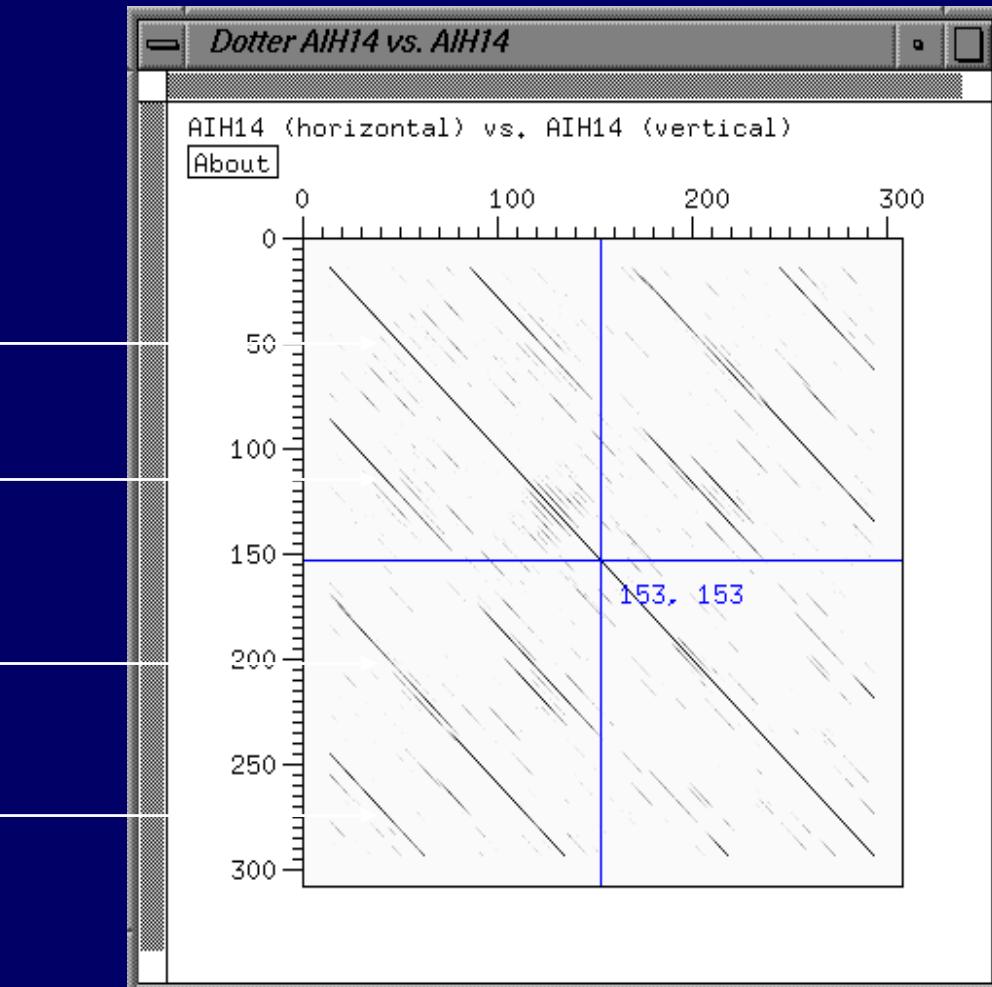


Annexin VI has 8 repeats

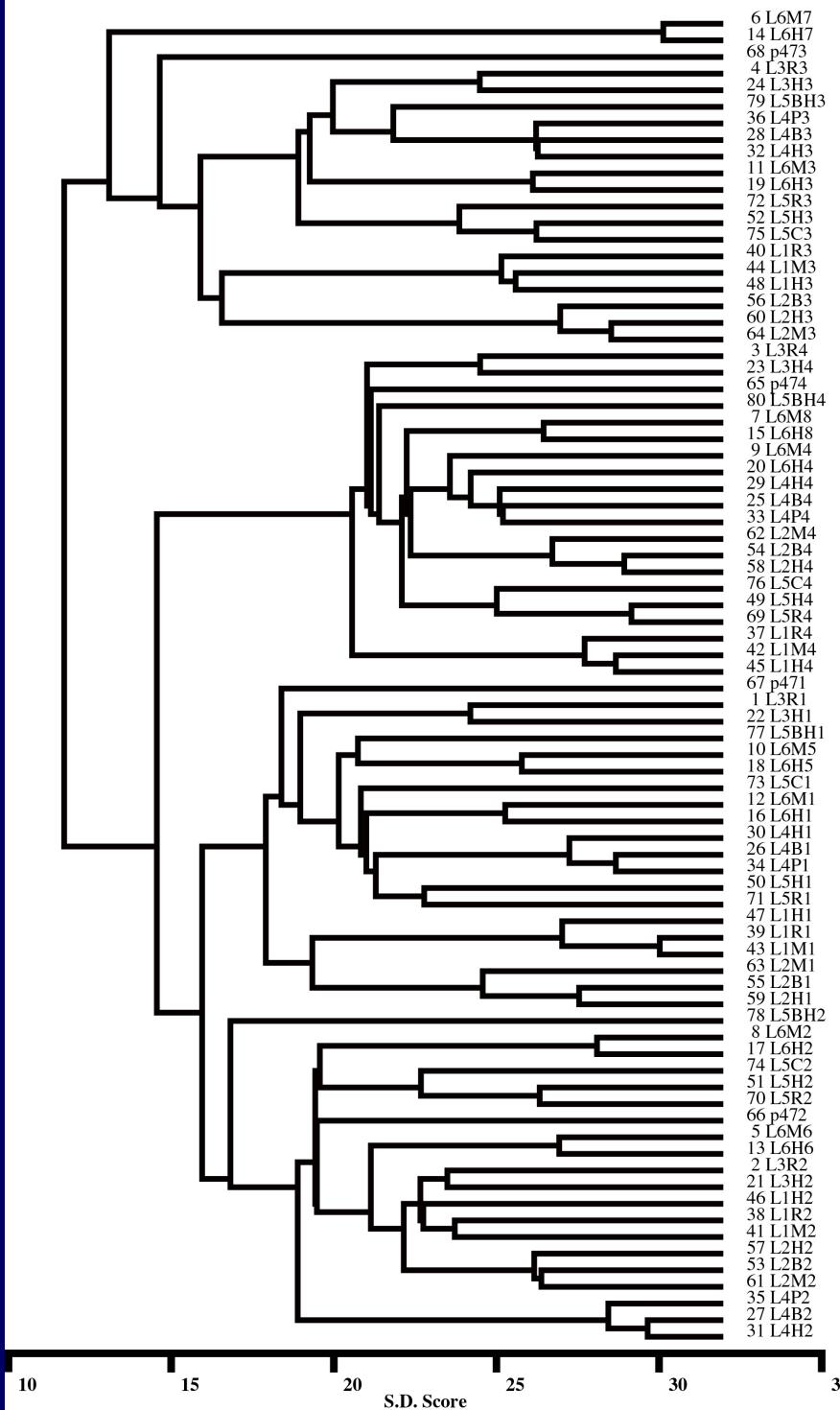
Sequence Analysis of Annexin Domains

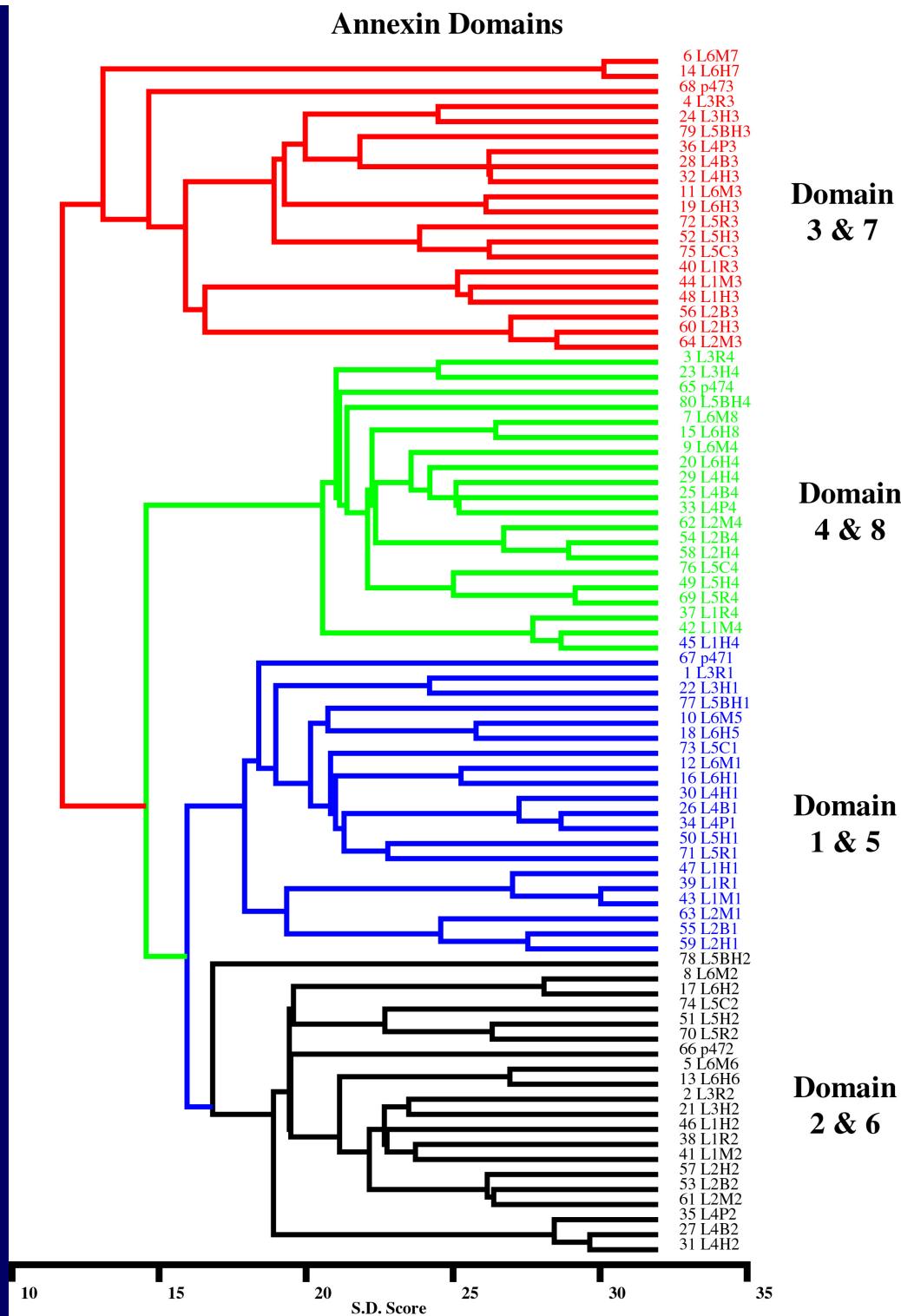
Dot-Plot comparison of Human Annexin I with itself.

Four repeats (domains ?) are visible.



Annexin Domains





**Domain
3 & 7**

**Domain
4 & 8**

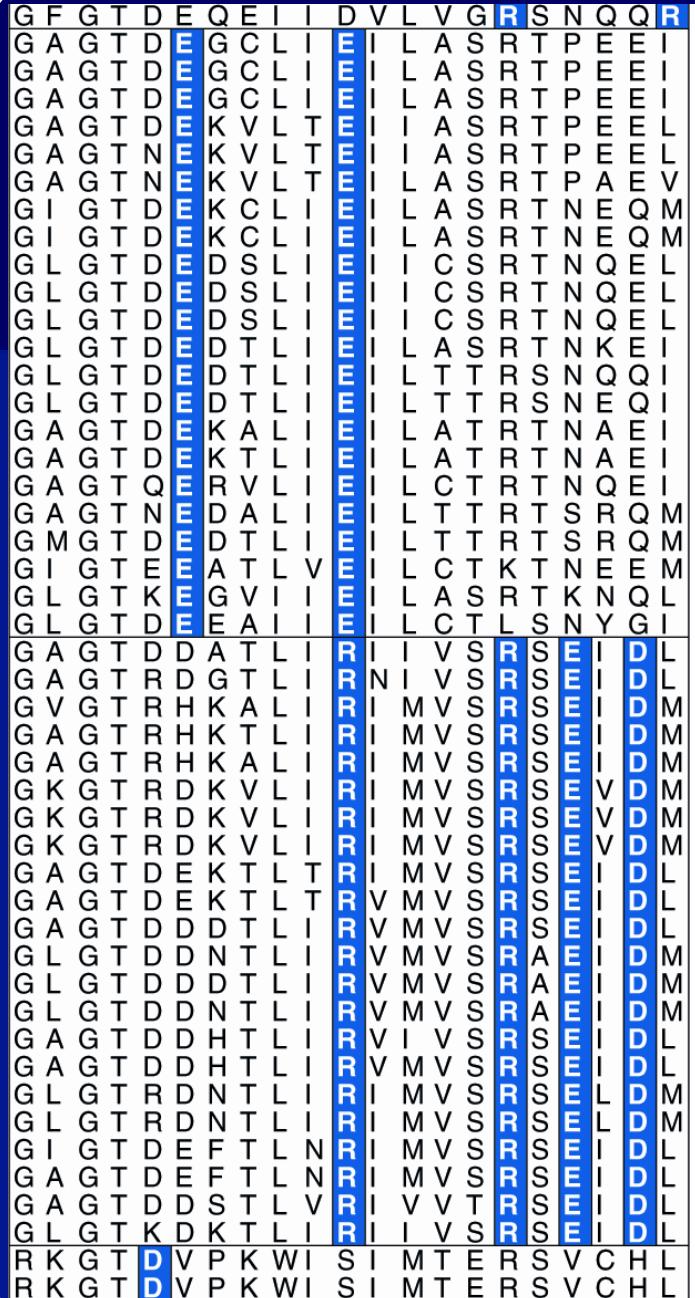
**Domain
1 & 5**

**Domain
2 & 6**

	Repeat 1	Repeat 2	Repeat 3	Repeat 4
1	P A	LL	DDDDNNNDLDDDD N	CCCCCCCS SCCCCS SCCCCC
2	AYYYYYYFFFSSAA YY	TTTKKRRKKKKKKMMMTTMYT	VTAALI I IVVIMMMI I IIVTAC	
3	ATTPPPSPSSSSNNN PP	FFFFFFFFFFPPTTTTPTPFPAPPF	QQTTTQQQKRRRRRRRRRLK	
4	NNNNTS SGGPHGGGDGEDG	TTTSAPLPFFFFPFFFFPFL	H SSSSNNNNNNSNNNS SSSNNNS NT	
5	FFFFF FFFFFFF FFFFFFFF	VVVRRAA AAAAAAATAAVYP T	PLKTTKKKKVKKK I ITTITIN T	
6	DDDDNNNDNNNNNNNDSND	LLLLLQLLLLLLMAQ	AHPPPPP PPPPS PPPP PPPP	
7	AAAAPPVGAEAAPPPPPPA T	IIIAAV VVVMVA AAVVAAA	ASAALLLLA AAAAEEAAAD T	
8	IIEEESS SRRDAMADNNNSSS	DDDDDDEERRLRKQQQTR	F YFFFFY FFFF YYYYYYFFY	
9	RRRSSS SAAAEEAAQVVQ T	QQQSTKQLQQQQQKQEAE	F FFFFFFF FFFFFFL LFFF LFFF	
10	DDDDDDDDDDDDDDDDDDDD	DDDDDDDDDDDDDDDDDDQ	H HAAA AAAAAA AAAAAAAS	
11	AAA VVY AAAAAA AAAAAA	RRRRRQRQQ9999999999999999	NEEEEEDDDDEEEEEE EEEEG	
12	ELLLA AEEEQGKKEEEEQ H	E DDDAAVAADDDTIDDRAQ	H RKKKRKRRK KRRTRRHHH	
13	IINNAAAVTA TTTTA AAAAAA	LL LLLL LLLL LLLL LLLL	L LLLL LLLL LLLL LLLL	
14	ELLLL LLLL LLLL LLLL LLLL	HQK KKKKKKKKKRKR K RKHHHH	Y YHYY Y FFFF YYYYYYFFHHYH	
15	KREEEHHHRRHHHHHHYQRR H	H KKKHHHDAAAAG KKKKAAD H	D DDEER QEELEAKD E QD5	
16	KTTT KKKKKKKKKKTTKKA J	H AAAAAAASSAAA AASAAAH	H KYQEEDD DKRYKKY KKKRQYD	
17	AAA AAAAAA AAAAAA AAAAAA	G GGGGGGGGGGGGGGGGGGG	H AAAAAS SSSSS SAAAAAAAS	
18	MIV MIV I IMMMMMMMMMMM I MIV	V VVEEEELLEEEEEEEL	H MMMMMMMMMMMMMMMMMML LMMH	
19	KKKKRKS SKKKKKEE QKRAKJ	KKKRRRL LKKKKNLGG	N KKKKKKKKKKKKKKKKA	
20	GGTTTVVGGGGGGGGGGGGGG	RRRRRKKKKKBRWWLWL	T	
21	FFKKKKKKLMLL L L L F	G GGGGGGGGGGGGGGGGGGG	GGGGGGGGGGGGGGGGGGGG	
22	GGGGGGGGGGGGGGGGGGGG	GGGGGGGGGGGGGGGGGGGG	AAA VAKKKA AALLLA ALLIAAL	
23	TTVVVVVVT TTTTTT S T T E T	TTTTTTTTTTTTTTTTTTT	GGGGGGGGGGGGGGGGGGGG	
24	DDDDDDDDDDDDDDDDDDDD	TTTTTTTTTTTTTTTTTT	T TTTTTTTTTTTTTTTTTT	
25	E EEEEEEEEEE EEEKEE E	TTTTTTTTTTTTTTTTTT	T DRRRRRRR DDDDDDDDR DDDDKET	
26	QKV VVAADEEQDDDAEEKQ	V VVVVVEEEEEE EEEEEE	H DHHHHH DDEEE DDDDDDDDEEDD T	
27	AAATTTTSTS A AATTA S MITE	GGGGGGGGGGGGGGGGGGGG	AGKKKKKKKDNDNH HNNFFSK	
28	IIIIIIIIII IIIIIIIIII	TTTTTTTTTTTTTTTTTT	T TATAVVI I I I I I I I I	
29	EDNNNDDDNTKDS S NDNNDESND	V VVVVVEEEEEE EEEEEE	H LLLL LLLL LLLL LLLL LLLL	
30	EDNNNDDDNTKDS S NDNNDESND	PPPNNEEEV VVMDAASSET	PT	
31	EDNNNDDDNTKDS S NDNNDESND	KKVVKVTKKKKKRQCTVRI	H DRRRRRRR DDDDDDDDR DDDDKET	
32	VIIIIIIII VVVI I I I V	WWWWWWWWWWWWWWWWWWWW	H DHHHHH DDEEE DDDDDDDDEEDD T	
33	VLLL LLLL LLLL LLLL V I	I INTNII I LLLITI I INNNMHE	AGKKKKKKKDNDNH HNNFFSK	
34	AAATTTTTTTTAAATTTTTT	S S S TTTTTTTTTTTTTTT	T TATAVVI I I I I I I I I	
35	NNNSKKKSKKSSYYHIISSTEG	TTTTTTTTTTTTTTTTTT	H LLLL LLLL LLLL LLLL LLLL	
36	RRRRRRRRRRRRRRRRRRRRR	EEETSTTTT S S T LNNNHTT	PT	
37	G SSSNTTS SNS SNS SSS SSS	RRRRRRRRRRRRRRRRR R R R R	H DRRRRRRR RRRRRRRRRRRRRR	
38	NNNNNNNNNNNTTNNNNNNN	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
39	VDVAEAAAAT A A A A V R K A Q	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
40	QQQQQQQQQQQQQQQQQQQQ	EEEEEQQQKQAEAQRREN	H DRRRRRRR RRRRRRRRRRRR	
41	RRRRRRRRRRRRRRRRRRR	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
42	L QQQQQQQQQQQQQQQQQQQ	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
43	ERDDDDQQQEEQEEQEEQ	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
44	IIIIIIIIII II II II II	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
45	AKAAAKKASA A A R R R R C V V	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
46	E E F F F A A E A S K T T Q Q Q K K A	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
47	AAA A A A A E A S A A A T T S S E Q V	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
48	F F F F Y Y Y F F F F Y Y Y Y H	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
49	KKQQLLKKKKKKKKKKKKQ	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
50	TTRRRQQQT T T A S S S S A H	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
51	S S RR E E E L L L Q T T T H L L A A E T	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
52	YY TTTT TNFFF I I F F F F Y Y	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
53	GGKKKGKGKGKGKGKGKGGE	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
54	KKKKKKKKKKKKKKKKKKQ	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
55	D D E E F T P D D D D D D D D D E A D	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
56	L L L L L L L L L L L L L L L L	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
57	I I P A A D D V L T L I M M M I T K V	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
58	S K S S E E N D D D D T A A E D A D	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
59	D D A A T T V D D D D D D D D D D D	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
60	L L L L L L L M L L L L L L L L L L	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
61	K K K K K K K K K K K K K K K K K Q	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
62	S S S K K S S S S S S S S S S S S S S	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
63	E E A A A A E E E E E E E D D E H	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
64	L L L L L L L L L L L L L L L L	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
65	G S S S T T T T T S S S S T T S S G H	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
66	GGGGGGGGGGGGGGGGGGGG	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
67	KNHHHHHHHHKKNNNDKKHHH	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
68	F M L L L L L F F F F F L L F F F F	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
69	D E T T B E B K T R Q Q R R R R R H D	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
70	V V V V V V V V V V V V V V V V	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
71	L L L L L L M L L L A A L I A L I L B	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
72	L L L L L L M M L L L L L L L L C T B	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
73	V V V V V V V V V V V V V V V V	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
74	L L L L L L V V V V V V V V V V	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
75	A A G G G A A A A S A G G G G G A A G H	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	
76	M F L L L M M L L L M M M L L L L L	TTTTTTTTTTTTTTTTTT	H DRRRRRRR RRRRRRRRRRRR	
77	M F L L L M M M M M M M M M M M M M	EEEEEQQQEEQEEQEQ	H DRRRRRRR RRRRRRRRRRRR	

		10	20
L9r1	G F G T D E K A I I E I L A R R G I V O R		
p471	G F G T D E Q A I I V D V V A N R R S N D O R		
L2M1	T K G V D E E V T I V N L T N R S N D O R		
L2H1	T K G V D E E V T I V N L T N R S N D O R		
L3H1	V K G V D E E V T I V N L T N R S N D O R		
L1R1	V K G V D E E A T I I D I I L T K P R T N A O R		
L1M1	V K G V D E E A T I I D I I L T K P R T N A O R		
L5R1	G L G T D E E D S I L N L L T A R S N N A O O R		
L5H1	G L G T D E E D S I L N L L T S R S N N A O O R		
L5C1	G M G T D E E E A T I I L K I L T S R S N N A O O R		
L8H1	G G I G T D E E Q A I I D V L T K R S N T Q R		
L4P1	G G I G T D E E Q A I I S S L A R R S N D O R		
L4H1	G G I G T D E E D A I I S S L A R R S N D O R		
L4B1	G G I G T D E E D A I I N V L A Y P S T A O R		
L6H5	G G I G T D E E D A I I D I I T H R S N V Q O R		
L6M5	G F F G S D K E A I I L D I I T S R S N N V Q O R		
L6H1	G F F G S D K E S I I L E L I T S R S N N V Q O R		
L6M1	G F F G S D K E S I I L E L I T S R S N N V Q O R		
L3H1	G I G T D E E K M L I S I I L E R S N A O O R		
L3P1	G I G T D E E K M L I S I I L E R S N A O O R		
L10I	G F G T D E E E A I I V L V E R S N A O O R		
L4P2	G A G T D E E Q C L I E I L A S R T P E E E I		
L4H2	G A G T D E E Q C L I E I L A S R T P E E E I		
L4B2	G A G T D E E Q C L I E I L A S R T P E E E I		
L5R2	G A G T D E E K V L I T E I I A S R T P E E E I		
L5H2	G A G T D E E K V L I T E I I A S R T P E E E I		
L5C2	G A G T D E E K V L I T E I I A S R T P E E E I		
L6H2	G I G T D E E K C L I E I L A S R T P E E E I		
L2H2	G G I G T D E E K C L I E I L A S R T P E E E I		
L2B2	G G I G T D E E K C L I E I L A S R T P E E E I		
L2M2	G G I G T D E E D S L I E I L A S R T P E E E I		
L1H2	G G I G T D E E D S L I E I L A S R T P E E E I		
L1R2	G G I G T D E E D T L I E I L T T R S N N Q Q E I		
L1M2	G G I G T D E E D T L I E I L T T R S N N Q Q E I		
L6H6	G A G T D E E K A L I I E I L A T R T N A A E I		
L6M6	G A G T D E E K A L I I E I L A T R T N A A E I		
L4P7	G A G T D E E K A L I I E I L A T R T N A A E I		
L3H2	G A G T D E E K A L I I E I L A T R T N A A E I		
L3R2	G M G T D D D T L I E I L T T R T S S R Q M		
L10r	G I G T D E E A T L V E I L C T K T N N E E M		
L8H2	G L G T D E E G V I I E I L A S R T K N Q L		
L9r2	G L G T D E E E A I I E I L C T L S N Y G I		
L10r4	G A G T D D D A T L I R I I V S R S S E I D L		
L8H4	G A G T R D D G T L I R I I V S R S S E I D L		
L1H4	G V G T R H K A L I R I I M V S R S S E I D M		
L1P4	G A G T R H K A L I R I I M V S R S S E I D M		
L1M4	G A G T R H K A L I R I I M V S R S S E I D M		
L2M4	G K G T R D K V L I R I I M V S R S S E I V D M		
L2B4	G K G T R D K V L I R I I M V S R S S E I V D M		
L2H4	G K G T R D K V L I R I I M V S R S S E I V D M		
L6H8	G A G T D E E K T L I T R I I M V S R S S E I D L		
L6M8	G A G T D E E K T L I T R I I M V S R S S E I D L		
L5C4	G A G T D D D D T L I T R I I M V S R S S E I D M		
L4H4	G L G T D D D D T L I T R I I M V S R S S E I D M		
L4B4	G L G T D D D D T L I T R I I M V S R S S E I D M		
L4P4	G L G T D D D D H T L I T R I I M V S R S S E I D M		
L5R4	G A G T D D D D H T L I T R I I M V S R S S E I D L		
L5H4	G A G T D D D D H T L I T R I I M V S R S S E I D L		
L6H4	G L G T D D D D N T L I T R I I M V S R S S E I L D M		
L6M4	G L G T D D D D N T L I T R I I M V S R S S E I L D M		
L3H4	G I G T D E E F T L I N R I I M V S R S S E I D L		
L3P4	G A G T D E E F T L I N R I I M V S R S S E I D L		
p474	G A G T D E E F T L I N R I I M V S R S S E I D L		
L9r4	G L G T D K D K T L I R I I Y V S R S S E I D L		
L2M3	R K G T D V P K W I S I M T E R S V C H L		
L2B3	R K G T D V P K W I S I M T E R S V C H L		
L2H3	R K G T D V P K W I S I M T E R S V C H L		
L1H3	R K G T D V N V V F N T I L T T R S Y P Q L		
L1M3	R K G T D V N V V F N T I L T T R S Y P Q L		
L1R3	R K G T D V N V V F N T I L T T R S Y P Q L		
L5C3	K W G T D E E E T I I I T I Q T R S S V S H L		
L5H3	K W G T D E E E K F I I I T F G T R S S V S H L		
L5R3	K W G T D E E E K F I I I T L G T R S S V S H L		
L4H3	K W G T D E E E K F I I I T L G T R S S V S H L		
L4B3	K W G T D E E E K F I I I T L G T R S S V S H L		
L4P3	K W G T D E E E K F I I I T L G T R S S V S H L		
L8H3	I R G T D E E M K F I I T L C T R S A T H L		
L3H3	R W G T D E E D K F I I E L C L R S F F P Q L		
L3P3	K W G T D E E D K F I I E L C L R S F F P Q L		
L6H3	K W G T D E E A Q F I I Y I L G N P R S K O Q H L		
L6M3	K W G T D E E A Q F I I Y I L G N P R S K O Q H L		
p473	R L G T D D S C F N M I L A T R S V C H L		
L9r3	Q W G T D E E S T F N S I L I T R S Y Q Q L		
L10r3	K L G T D E E T R F M T I L C T R S Y P H L		
L6H7		1	
L6M7		2	

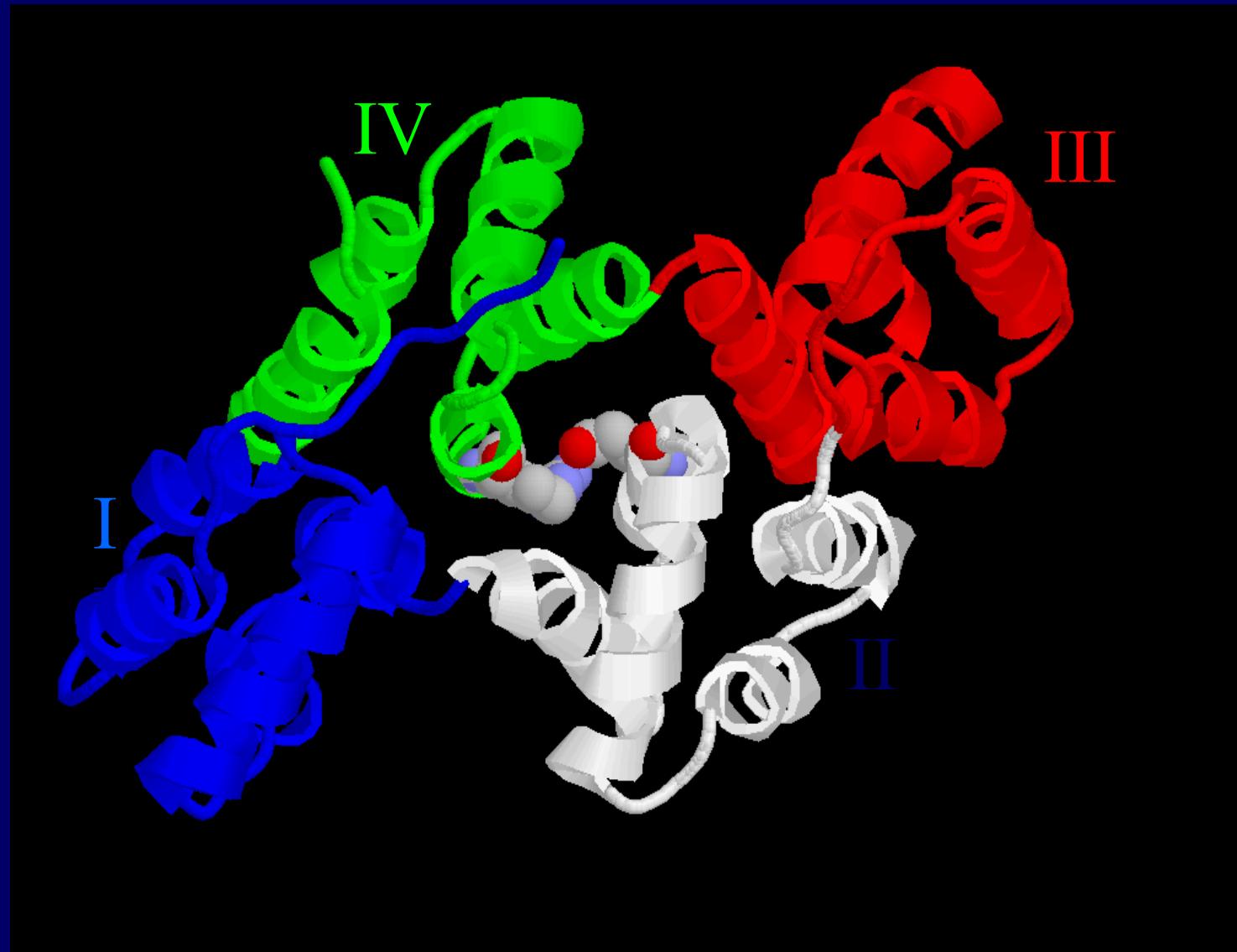
Charge Comparison



Annexin Predictions

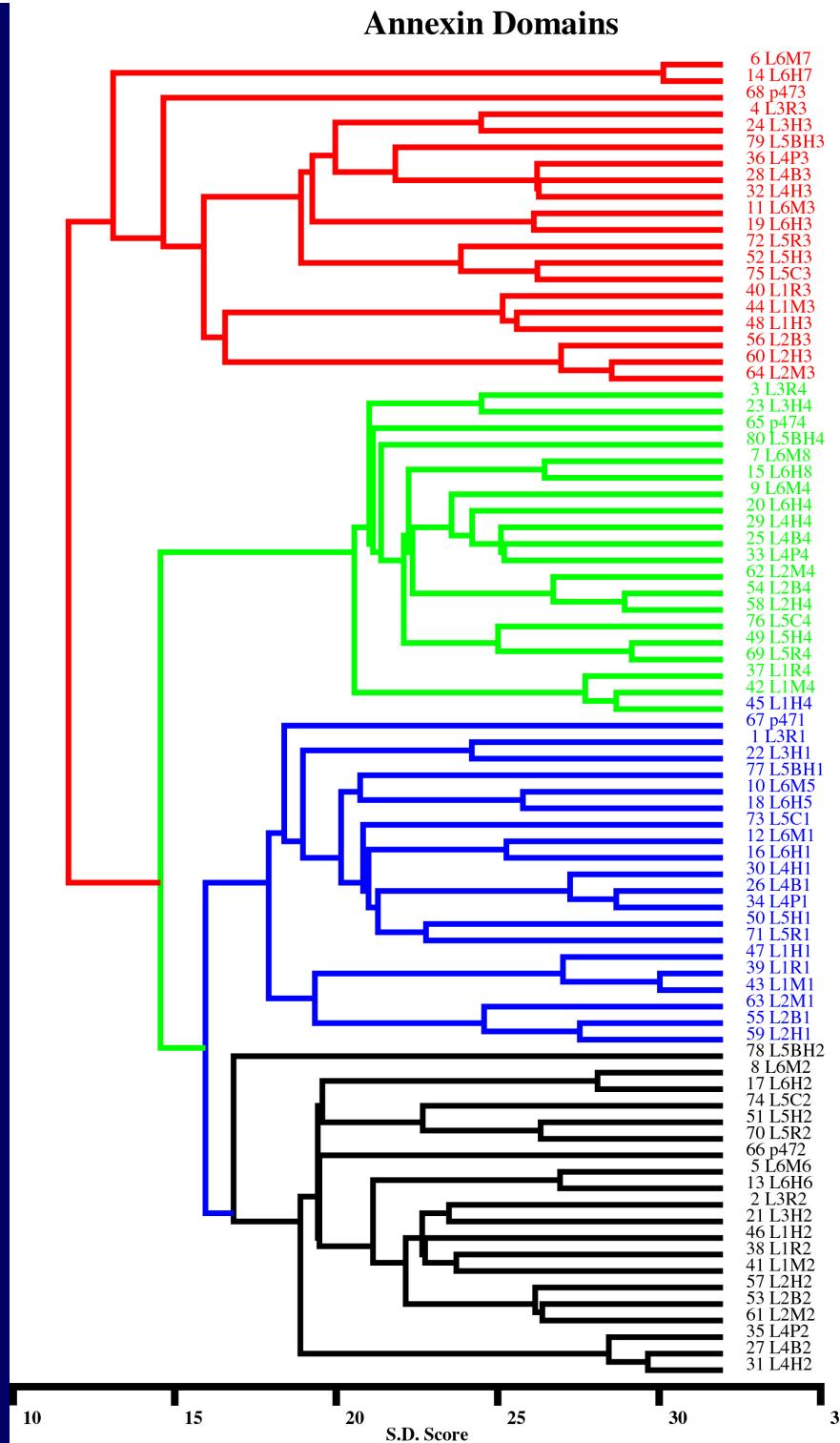
1. 5 Helices
2. Core residues (hydrophobic patterns)
3. Conserved Glu in repeats II
and Arg in repeats IV
form a salt bridge
4. Helix a in repeat III shorter
5. Not like uteroglobin
6. Helix a - helix b loop important in repeat III

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



**Domain
3 & 7**

**Domain
4 & 8**

**Domain
1 & 5**

**Domain
2 & 6**

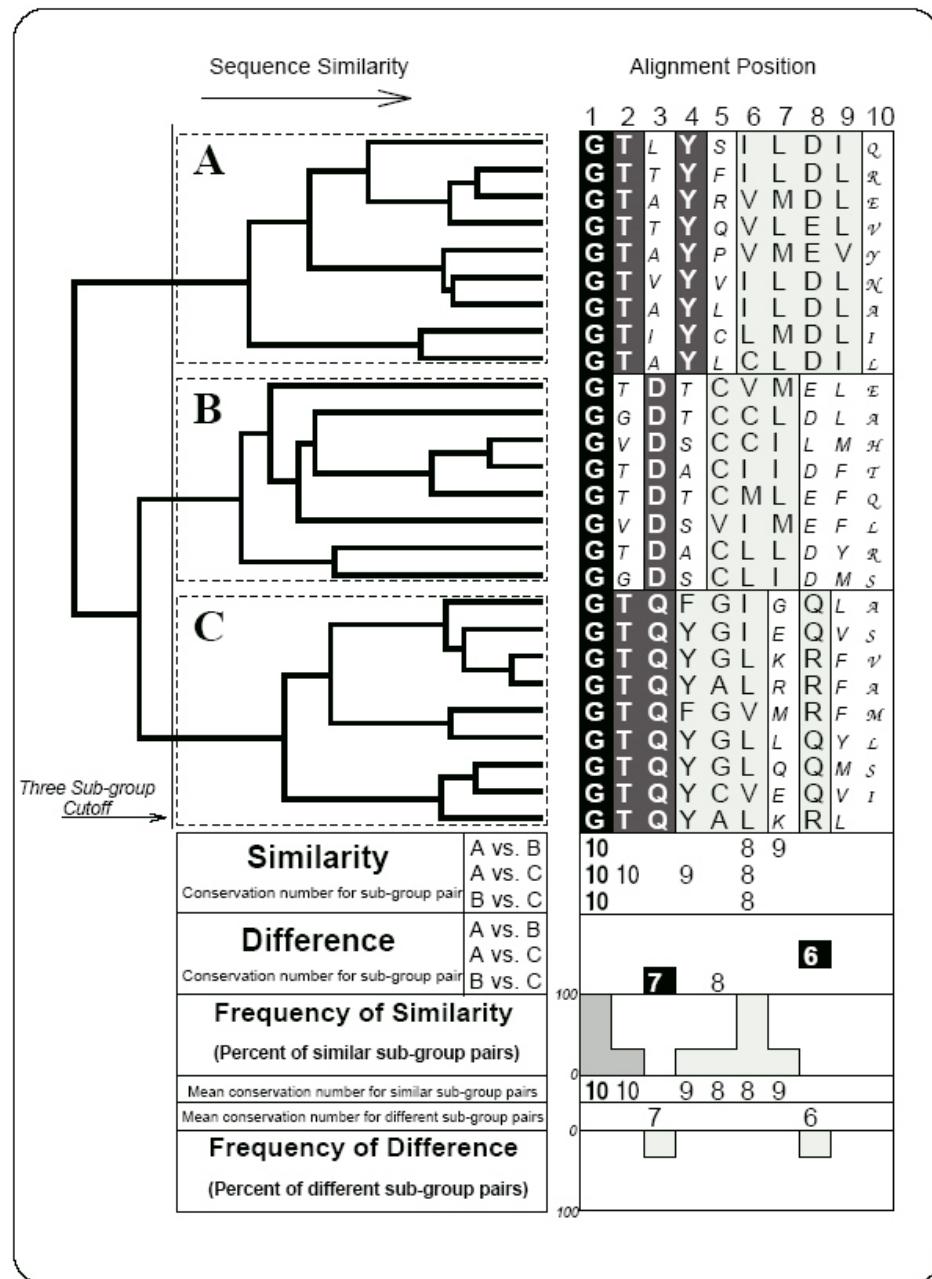


Figure 3

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/>

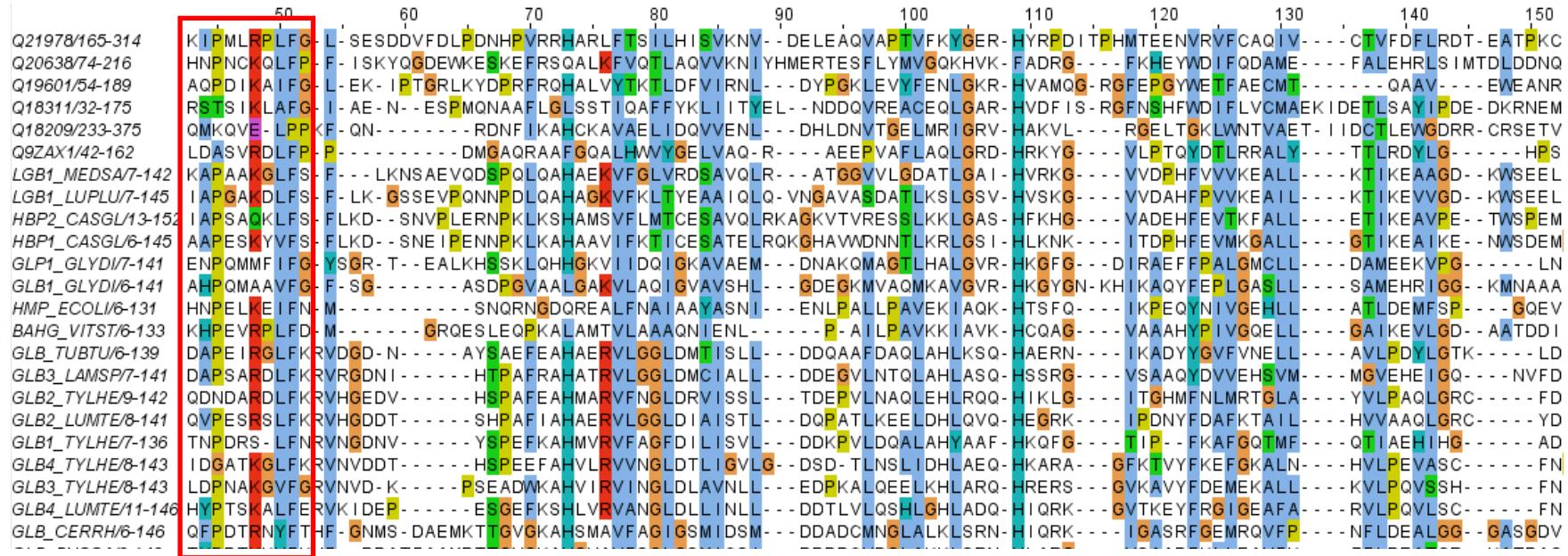
Identification of functional sites

- Whole alignment methods
 - Simple visualisation
 - Calculation of “conservation values”

- Sub-family analysis
 - AMAS analysis
 - “Tree determinant” positions
 - “Evolutionary trace”

Tree determinants

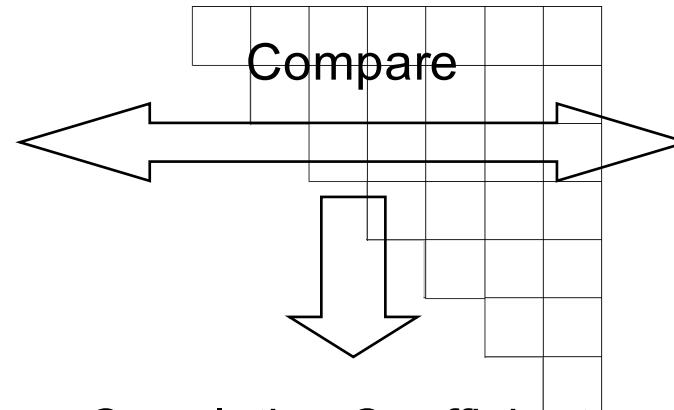
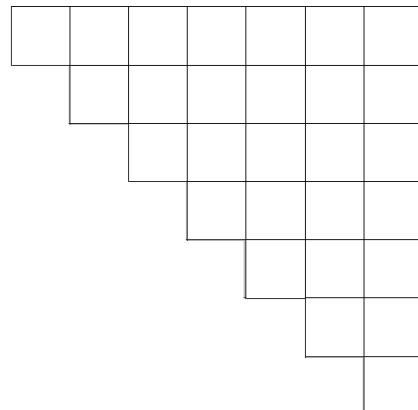
- Positions in the alignment that are most responsible for the topology of the phylogenetic tree derived from the complete alignment
- These positions may be functionally important.



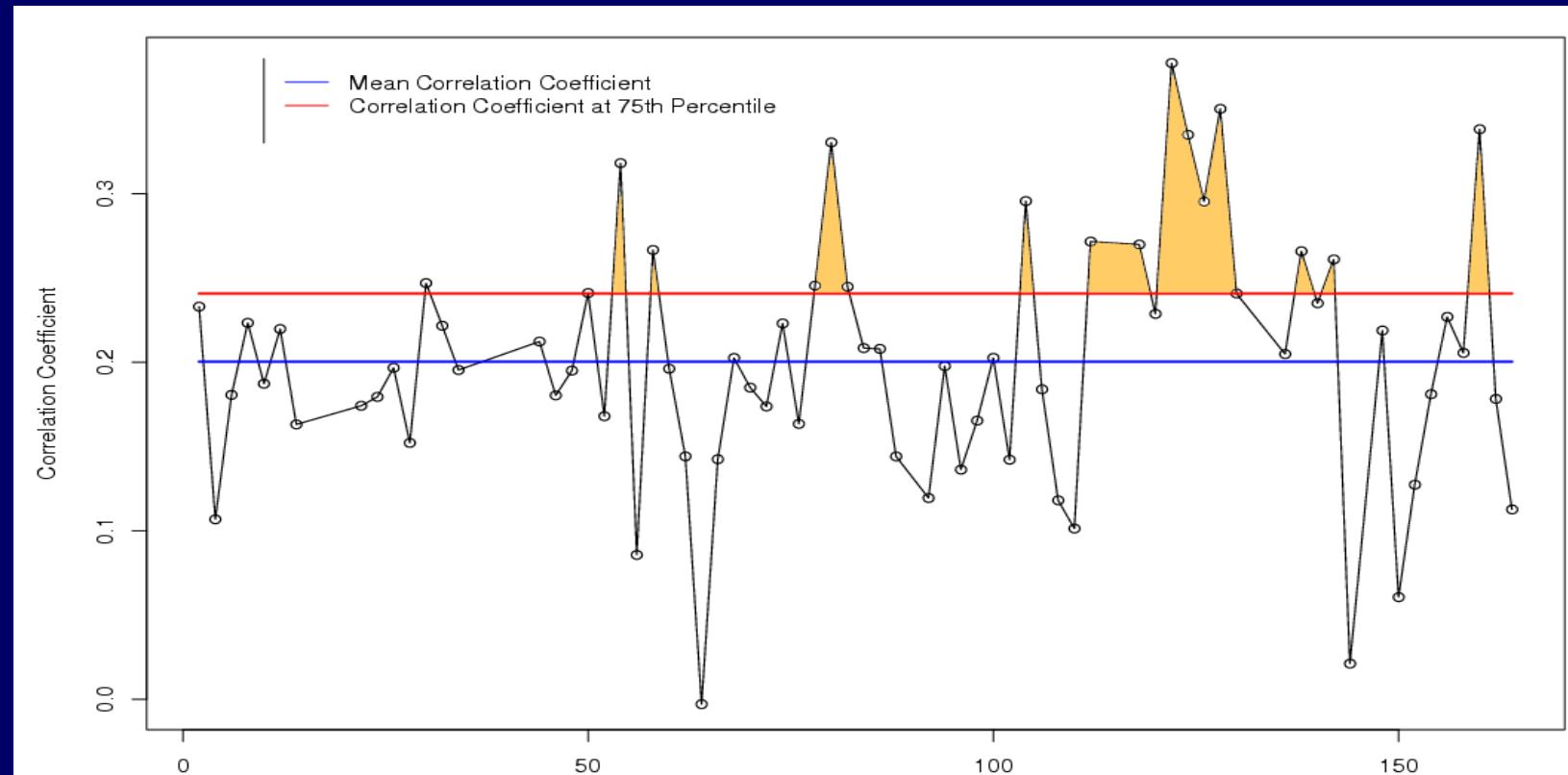
Generate Local Matrix

Generate Global Distance Matrix

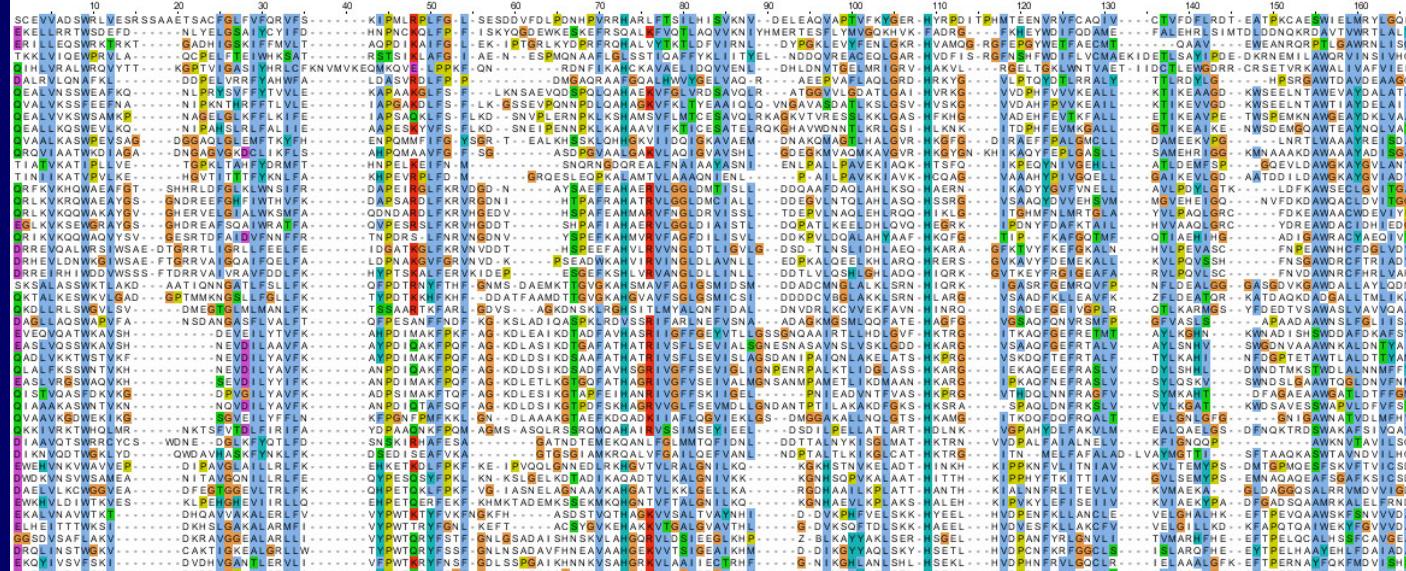
Saved Global Matrix



Correlation Coefficient



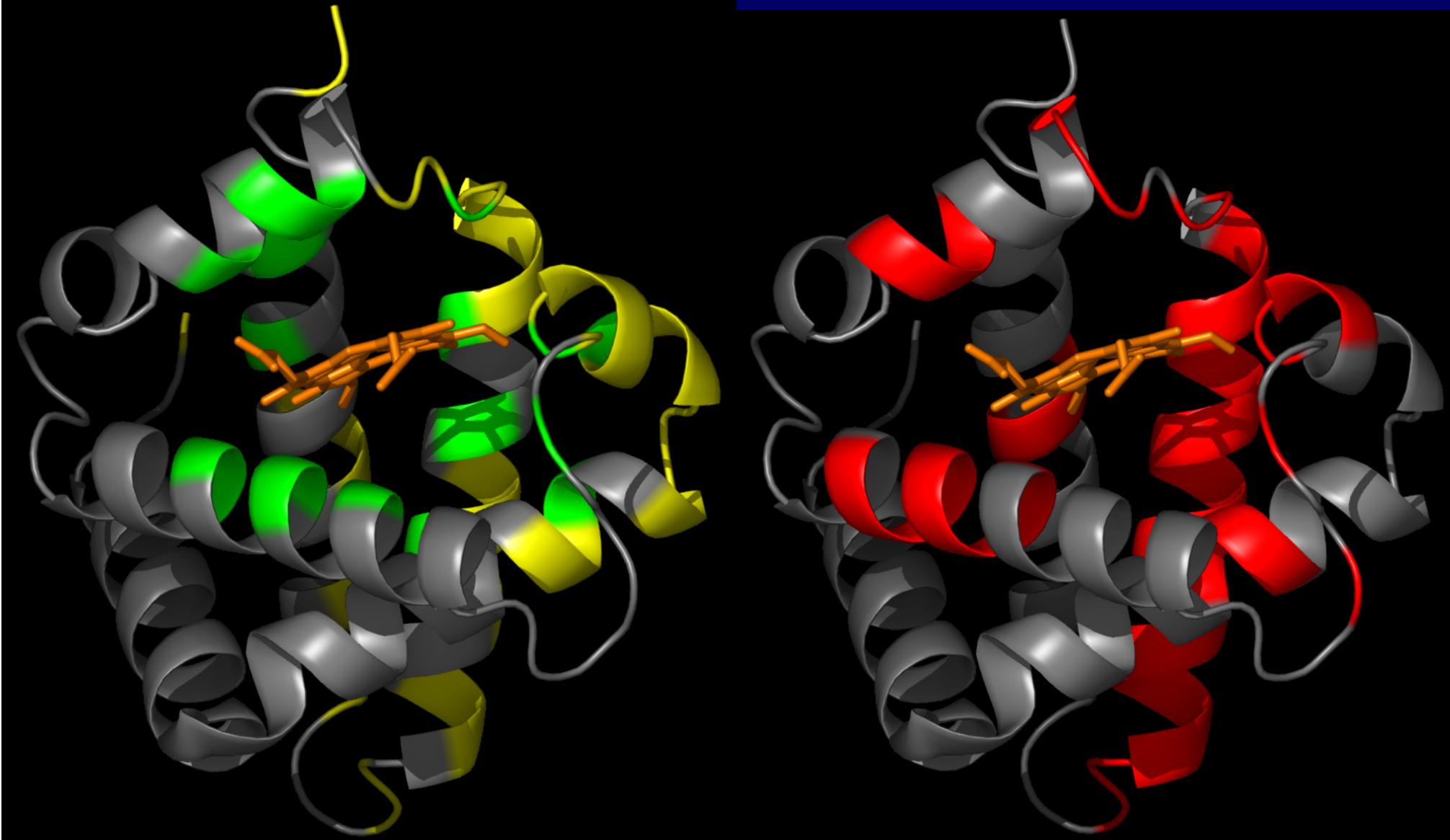
Partial Multiple Sequence Alignment (Globins)



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

