

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

```

1ju5a_1/1-109 1  -----SWYWGRLSRQEAVALLQGQ--RHGVFLVVDST-SPGDVVLVSV-----SENS---RVSHYIINSSGPRPPVPPSPAQPPPGVSPSRLRIGD----80
1x6ca_1/1-118 1  -----GSSGSSGWYHGHMSGGQAETLLQAKG-EPWTFVLRRLS-QPGDFVLVSLSDQ-PKAGPGSPLRVTHIKVMCEGG-----RTTVGG----78
2eu0a_1/1-108 1  -----NNLETYEWYNKSIIRDKAELLLDLTG-KEGAFMVRDSTR--PGTYTVSVFTTKAIIISENP-----CIKHYHIKETND-----SPKRYVVAE----77
2iuhb_1/1-115 1  -----NMSLQNAEWYWGDIISREEVNEKLRDT--ADGTFVLVDASTKMHGDIYTLTL-----RKGG---NNKLIKIFHRDG-----KYGFS-----70
2iuib_1/1-110 1  -----NNNMSLQNAEWYWGDIISREEVNEKLRDT--ADGTFVLVDASTKMHGDIYTLTL-----RKGG---NNKLIKIFHRDG-----KYGFS-----72
2iuia_1/1-107 1  -----MSLQNAEWYWGDIISREEVNEKLRDT--ADGTFVLVDASTKMHGDIYTLTL-----RKGG---NNKLIKIFHRDG-----KYGFS-----69
1aouf_1/1-106 1  -----SIQAEWYFGKLRKDAERQLLSFGNPRGTFLVRESEET-TKGAYSLVSLIRDWD-DMKGD---HVKHVKIRKLDN-----G-GYVIT-----75
2ciaa_1/1-98 1  -----SEWYVGNVTRHQAECALNERG-VEGDFLRDSES-SPSDFSVSL-----KASG---KNKHFQVQLVDN-----VVCIG-----63
1ka7a_1/1-107 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1ka6a_1/1-107 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1d1zd_1/1-101 1  -----VAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV69
1d1zc_1/1-101 1  -----VAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV69
1d1za_1/1-101 1  -----VAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV69
1m27a_1/1-104 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1d4wb_1/1-102 1  -----AVAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV70
1d4ta_1/1-104 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDES-VPGVYCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1blk_1/1-114 1  GSVAPVETLEVEKWFRTISRKDAERQLLAPMNAKGSFLIRESES-NKGAFSLVSKDI--TTQGE---VVKHVKIRSLDN-----G-GYVISP-----81
1rjaa_1/1-100 1  -----SEPWFFGCISREAVRRLQAEGNATGAFLLIRESEK-PSADVLSV-----RDTQ---AVRHYKIWRRAAG-----RLHLNE-----67
1cwdl_1/1-98 1  -----GSWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE---VVKHVKIRNLDN-----G-GFYISP-----71
1p13b_1/1-102 1  -----AEWYFGKITRRESERLLNPNPNRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----72
1n2lb_1/1-103 1  -----AEWYFGKITRRESERLLNPNPNRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----72
1n2va_1/1-101 1  -----AEWYFGKITRRESERLLNPNPNRGTFLVRESEET--GAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----70
1n2la_1/1-101 1  -----AEWYFGKITRRESERLLNPNPNRGTFLVRESEET--GAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----70
1lcja_1/1-104 1  -----EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE---VVKHVKIRNLDN-----G-GFYISP-----73
1lkl_1/1-104 1  -----EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE---VVKHVKIRNLDN-----G-GFYISP-----73
1a07a_1/1-105 1  -----SIQAEWYFGKITRRESERLLNNAENPRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----75
1hcsb_1/1-107 1  -----MDSIQAEWYFGKITRRESERLLNNAENPRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----77

```

Conservation

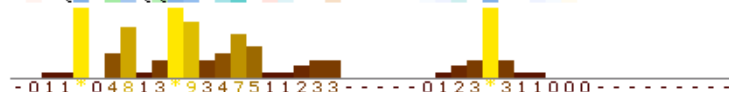


```

1ju5a_1/1-109 81 --QEFDLSPALLEFVKIHYLD-----TTTLIEPVSR-----109
1x6ca_1/1-118 79 -LETFDLTDLVEHFKKKTGIE-EASGAFVYL RQPYVSGPSSG-----118
2eu0a_1/1-108 78 -KYVFDLIPLLIQYHQYNGGG-----LVTRLRYPCVG-----108
2iuhb_1/1-115 71 -PLTFSSVVELINHYRNESLAQYNPKLDVKKLLYPVSKYQQDQVVKKE115
2iuib_1/1-110 73 -PLTFSSVVELINHYRNESLAQYNPKLDVKKLLYPVSKYQ-----110
2iuia_1/1-107 70 -PLTFSSVVELINHYRNESLAQYNPKLDVKKLLYPVSKYQ-----107
1aouf_1/1-106 76 -RAQFETLQQLVQHYSERAA-----LSSRLVVP SHK-----106
2ciaa_1/1-98 64 -QRRFHTMDELVEHYKKAPIFTSEHGEKLYLVRLAQ-----98
1ka7a_1/1-107 73 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEKKSS-----107
1ka6a_1/1-107 73 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEKKSS-----107
1d1zd_1/1-101 70 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEK-----101
1d1zc_1/1-101 70 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEK-----101
1d1za_1/1-101 70 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEK-----101
1m27a_1/1-104 73 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEK-----104
1d4wb_1/1-102 71 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEK-----102
1d4ta_1/1-104 73 HKRYFRKIKNLSIAFQKPDQG-----IVIP LQYPVEK-----104
1blk_1/1-114 82 -RITFPTLQALVQHYSKKGDG-----LCQKLTLPVCNLA-----114
1rjaa_1/1-100 68 -AVSFLSLPELVNHYHRAQSLS-----HGLRLAAPCRKHE-----100
1cwdl_1/1-98 72 -RITFPGHLHLVRHYTNASDG-----LCTRLSR-----98
1p13b_1/1-102 73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----102
1n2lb_1/1-103 73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----103
1n2va_1/1-101 71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----101
1n2la_1/1-101 71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----101
1lcja_1/1-104 74 -RITFPGHLHLVRHYTNASDG-----LCTRLSRPCQT-----104
1lkl_1/1-104 74 -RITFPGHLHLVRHYTNASDG-----LCTRLSRPCQT-----104
1a07a_1/1-105 76 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP-----105
1hcsb_1/1-107 78 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP-----107

```

Conservation



AMAS method of calculating conservation

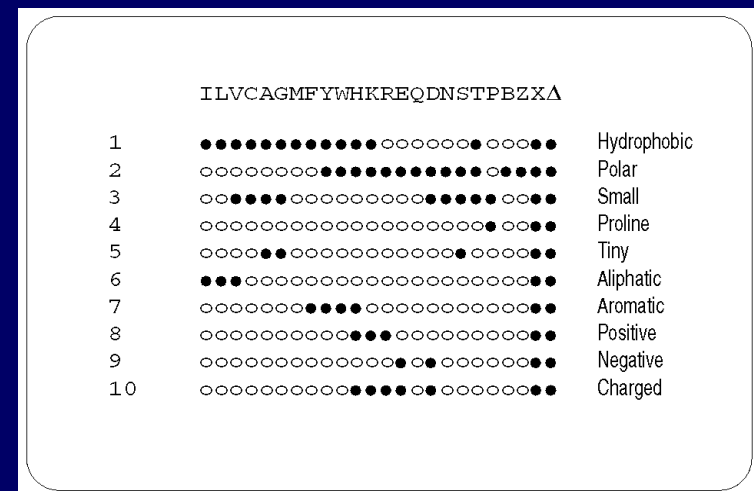
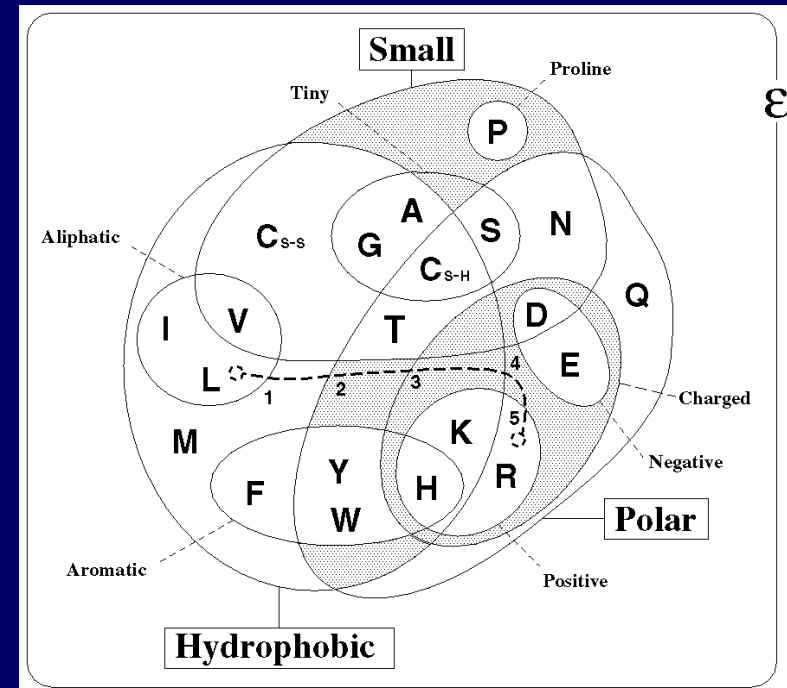
Taylor Venn diagram of amino acid properties

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



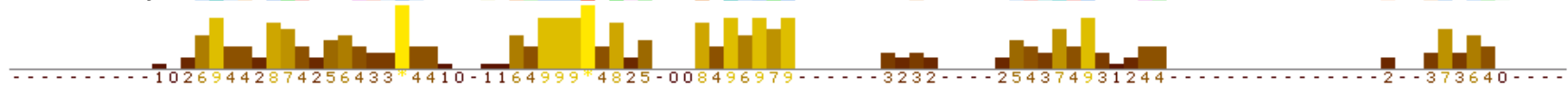
Example Multiple Sequence alignment of 27 SH2 domains

```

1ju5a_1/1-109 1  -----SWYWGRLSRQEAVALLQGQ--RHGVFLVVDST-SPGDVVLVSV-----SENS---RVSHYIINSSGPRPPVPPSPAQPPPGVSPSRLRIGD----80
1x6ca_1/1-118 1  -----GSSGSSGWYHGHMSGGQAETLLQAKG-EPWTFVLRRLS-QPGDFVLVSLSDQ-PKAGPGSPLRVTHIKVMCEGG-----RTTVGG----78
2eu0a_1/1-108 1  -----NNLETYEWYKNSISRDKAELLLDLTG-KEGAFMVRDSTR--PGTYTVSVFTKAIISENP-----CIKHYHIKETND-----SPKRYVVAE----77
2iuhb_1/1-115 1  -----NMSLQNAEWYWGDISREEVNEKLRDT--ADGTFVLRDASTKMHGDTTLTL-----RKGG---NNKLIKIFHRDG-----KYGFS-----70
2iuib_1/1-110 1  -----NNNMSLQNAEWYWGDISREEVNEKLRDT--ADGTFVLRDASTKMHGDTTLTL-----RKGG---NNKLIKIFHRDG-----KYGFS-----72
2iuia_1/1-107 1  -----MSLQNAEWYWGDISREEVNEKLRDT--ADGTFVLRDASTKMHGDTTLTL-----RKGG---NNKLIKIFHRDG-----KYGFS-----69
1aouf_1/1-106 1  -----SIQAEWYFGKLRKDAERQLLSFGNPRGTFLVRESEET-TKGAYSLSDIRWD-DMKGD---HVKHVKIRKLDN-----G-GYVIT-----75
2ciaa_1/1-98 1  -----SEWYVGNVTRHQAECALNERG-VEGDFLRDSES-SPSDFSVSL-----KASG---KNKHFQVQLVDN-----VVCIG-----63
1ka7a_1/1-107 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1ka6a_1/1-107 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1d1zd_1/1-101 1  -----VAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV69
1d1zc_1/1-101 1  -----VAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV69
1d1za_1/1-101 1  -----VAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV69
1m27a_1/1-104 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1d4wb_1/1-102 1  -----AVAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV70
1d4ta_1/1-104 1  -----MDAVAVYHGKISRETGEKLLLATG-LDGSVLLRDS-VPGVVCLCV-----LYHG---YIYTVRVSQTET-----G-SWSAETAPGV72
1blk_1/1-114 1  GSVAPVETLEVEKWFRTISRKDAERQLLAPMNAKGSFLIRES-NGKAFSLSVKDI--TTQGE---VVKHVKIRSLDN-----G-GYVISP-----81
1rjaa_1/1-100 1  -----SEPWFPGCISREAVRRLQAEGNATGAFLLVSEK-PSADVVLVSV-----RDTQ---AVRHVKIWRRAAG-----RLHLNE-----67
1cwdl_1/1-98 1  -----GSWFFKNLSRKDAERQLLAPGNTHGSFLIRES-NTAGSFSLSVRDFD-QNQGE---VVKHVKIRNLND-----G-GFYISP-----71
1p13b_1/1-102 1  -----AEWYFGKITRRESERLLNPNPRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----72
1n2lb_1/1-103 1  -----AEWYFGKITRRESERLLNPNPRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----72
1n2va_1/1-101 1  -----AEWYFGKITRRESERLLNPNPRGTFLVRESEET--GAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----70
1n2la_1/1-101 1  -----AEWYFGKITRRESERLLNPNPRGTFLVRESEET--GAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----70
1lcja_1/1-104 1  -----EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRES-NTAGSFSLSVRDFD-QNQGE---VVKHVKIRNLND-----G-GFYISP-----73
1lkl_1/1-104 1  -----EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRES-NTAGSFSLSVRDFD-QNQGE---VVKHVKIRNLND-----G-GFYISP-----73
1a07a_1/1-105 1  -----SIQAEWYFGKITRRESERLLNNAENPRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----75
1hcsb_1/1-107 1  -----MDSIQAEWYFGKITRRESERLLNNAENPRGTFLVRESEET-TKGAYCLSVSDFD-NAKGL---NVKHVKIRKLDG-----G-GFYITS-----77

```

Conservation

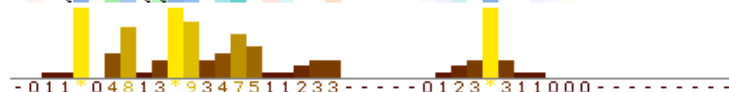


```

1ju5a_1/1-109 81 --QEFDLSPALLEFVKIHYLD-----TTTLIEPVSR-----109
1x6ca_1/1-118 79 -LETFDLTDLVEHFKKKTGIE-EASGAFVYLQPYVSGPSSG-----118
2eu0a_1/1-108 78 -KYVFDLIPLLIQYHQYNGGG-----LVTRLRYPCVG-----108
2iuhb_1/1-115 71 -PLTFSSVVELINHYRNESLAQYNPKLDVKKLYPVSKYQQDQVVKKE115
2iuib_1/1-110 73 -PLTFSSVVELINHYRNESLAQYNPKLDVKKLYPVSKYQ-----110
2iuia_1/1-107 70 -PLTFSSVVELINHYRNESLAQYNPKLDVKKLYPVSKYQ-----107
1aouf_1/1-106 76 -RAQFETLQQLVQHYSERAA-----LSSRLVVP SHK-----106
2ciaa_1/1-98 64 -QRRFHTMDLVEHYKKAPIFTSEHGEKLYLVRLAQ-----98
1ka7a_1/1-107 73 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEKKSS-----107
1ka6a_1/1-107 73 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEKKSS-----107
1d1zd_1/1-101 70 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEK-----101
1d1zc_1/1-101 70 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEK-----101
1d1za_1/1-101 70 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEK-----101
1m27a_1/1-104 73 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEK-----104
1d4wb_1/1-102 71 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEK-----102
1d4ta_1/1-104 73 HKRYFRKIKNLSAFQKPDQG-----IVIPLQYPVEK-----104
1blk_1/1-114 82 -RITFPTLQALVQHYSKKGDG-----LCQKLTLPVCNLA-----114
1rjaa_1/1-100 68 -AVSFLSLPELVNHYHRAQSLS-----HGLRLAAPCRKHE-----100
1cwdl_1/1-98 72 -RITFPGHLHLVRYHTNASDG-----LCTRLSR-----98
1p13b_1/1-102 73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----102
1n2lb_1/1-103 73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----103
1n2va_1/1-101 71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----101
1n2la_1/1-101 71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT-----101
1lcja_1/1-104 74 -RITFPGHLHLVRYHTNASDG-----LCTRLSRPCQT-----104
1lkl_1/1-104 74 -RITFPGHLHLVRYHTNASDG-----LCTRLSRPCQT-----104
1a07a_1/1-105 76 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP-----105
1hcsb_1/1-107 78 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP-----107

```

Conservation



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

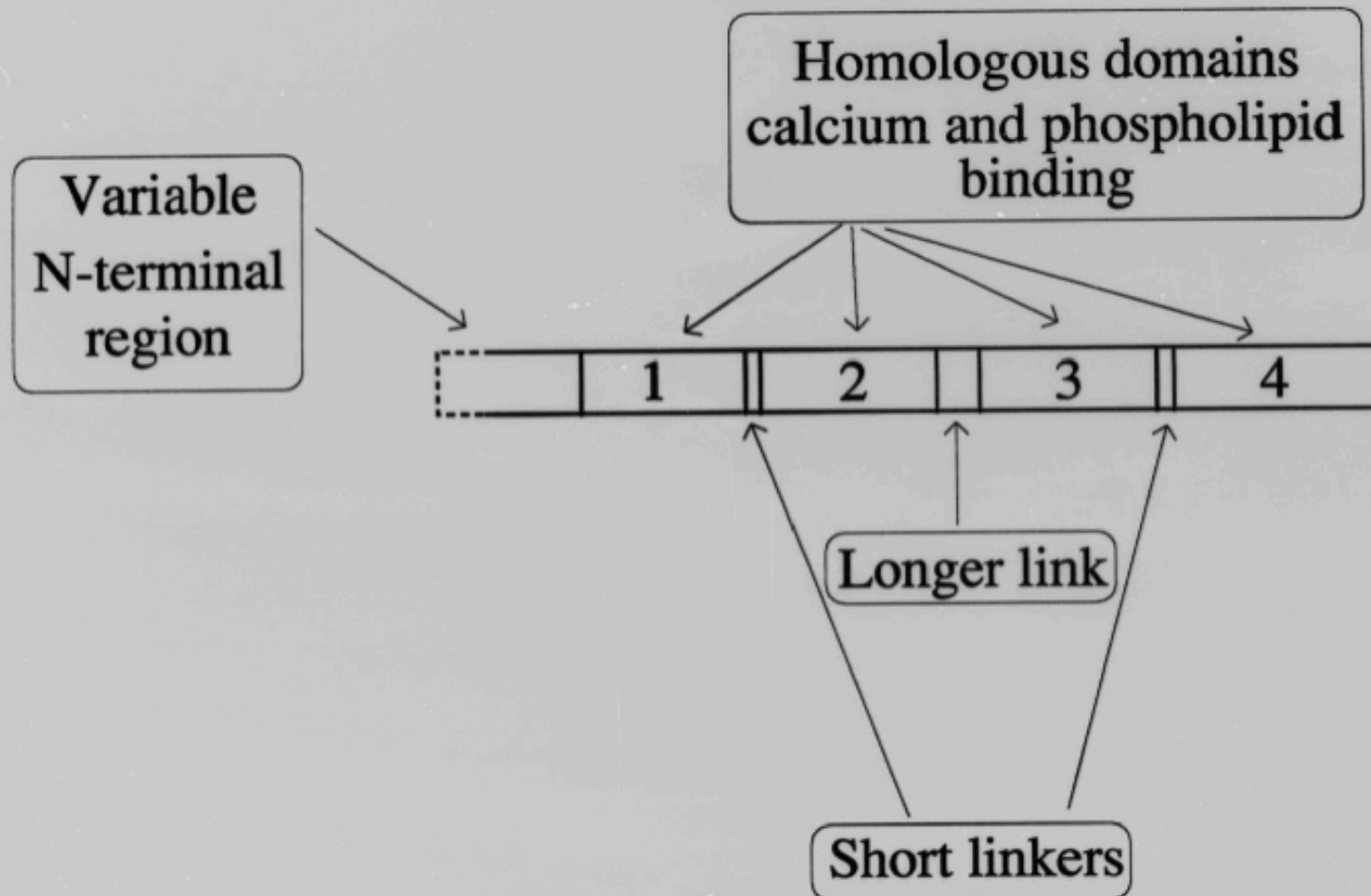
membrane fusion

exocytosis

cell signalling

anti-inflammatory properties

Annexins

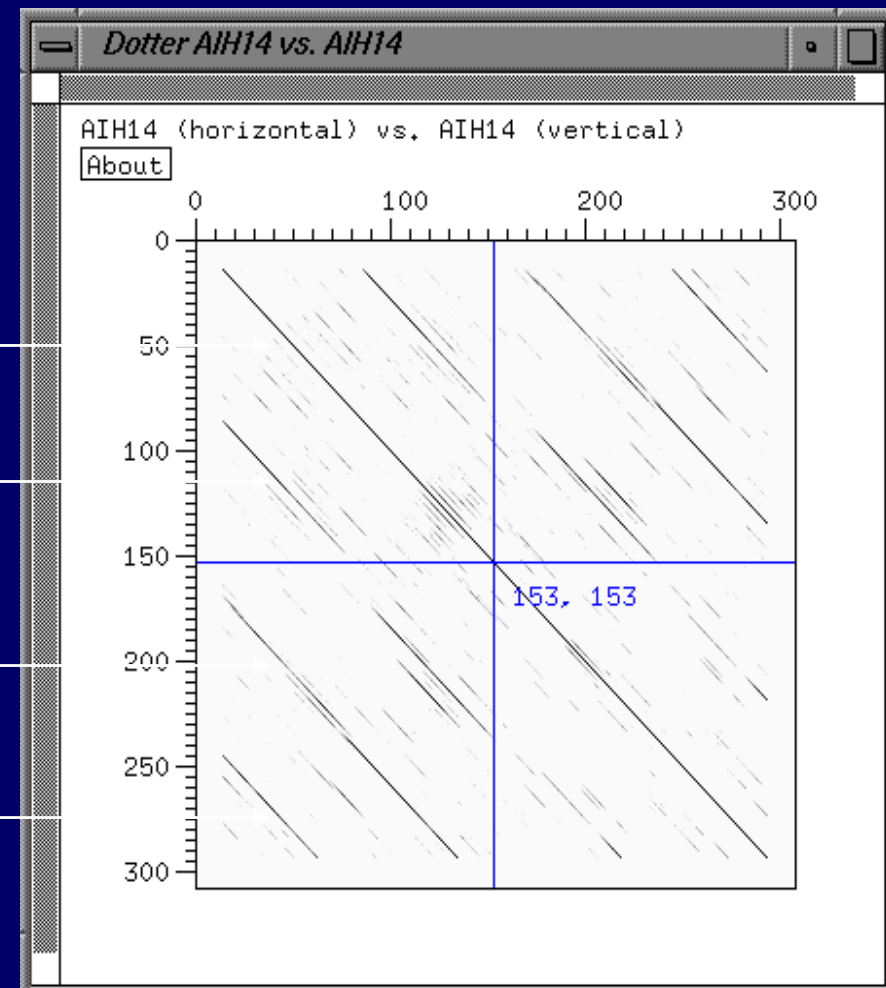


Annexin VI has 8 repeats

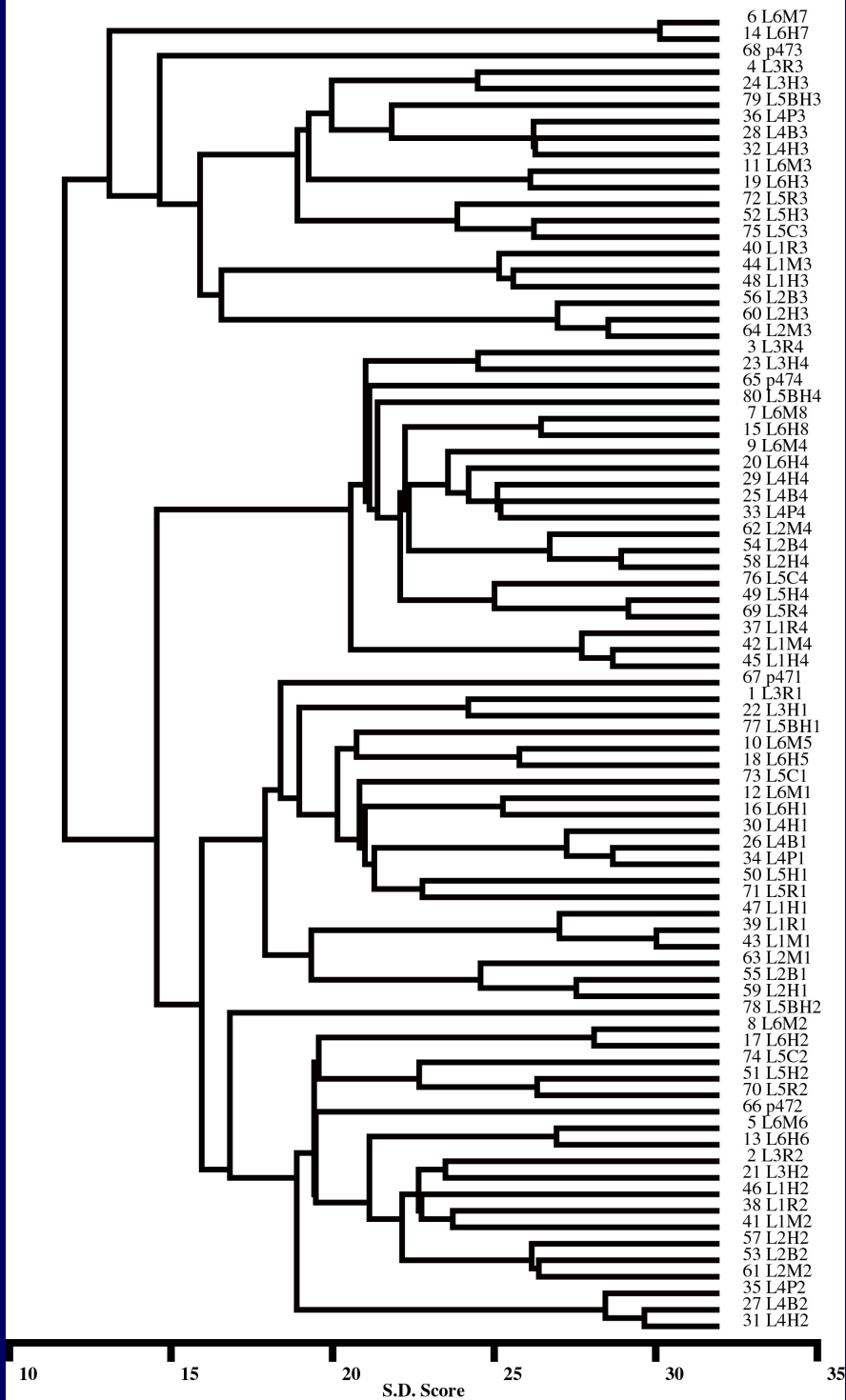
Sequence Analysis of Annexin Domains

Dot-Plot comparison of
Human Annexin I with itself.

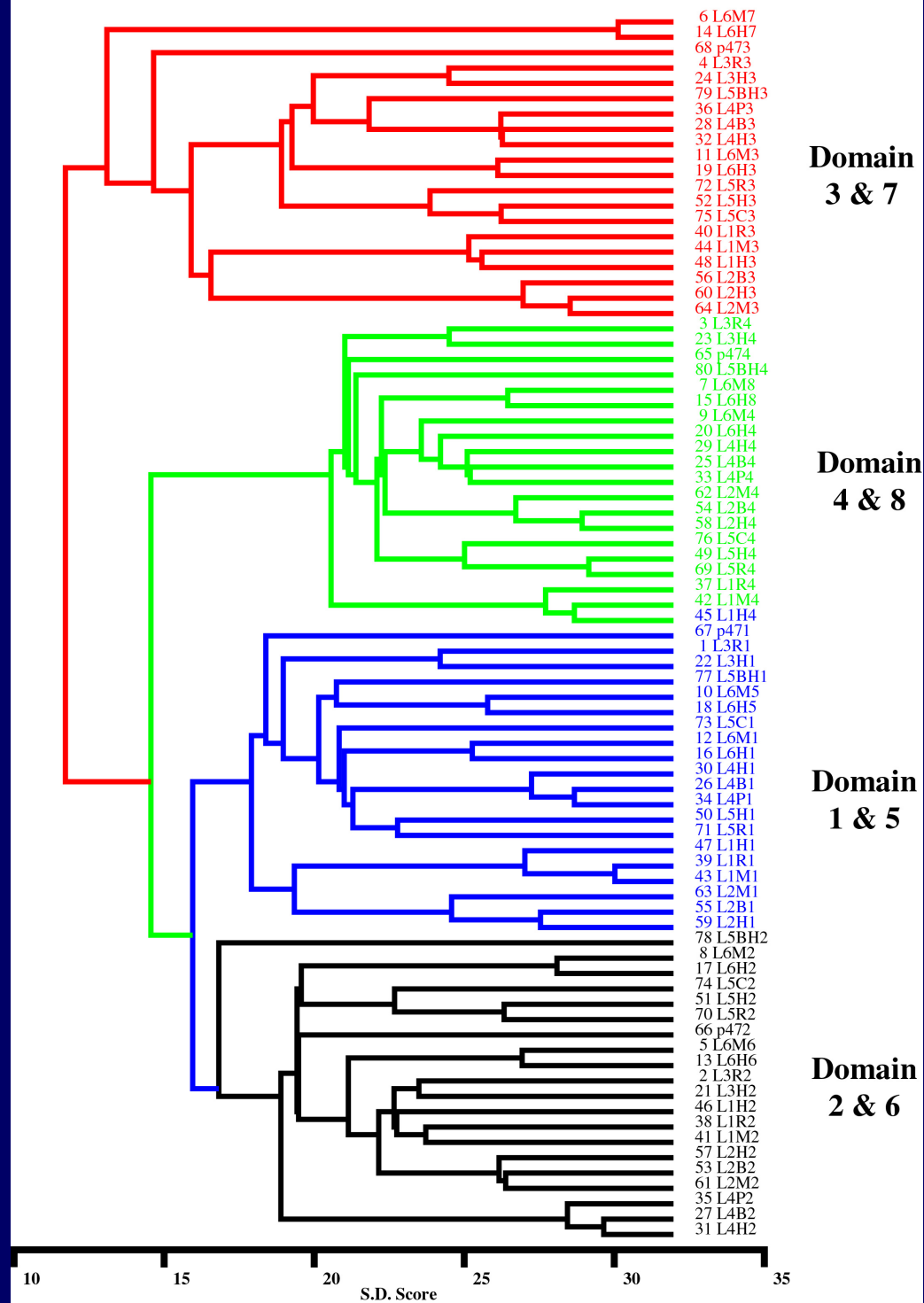
Four repeats (domains ?)
are visible.



Annexin Domains



Annexin Domains



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

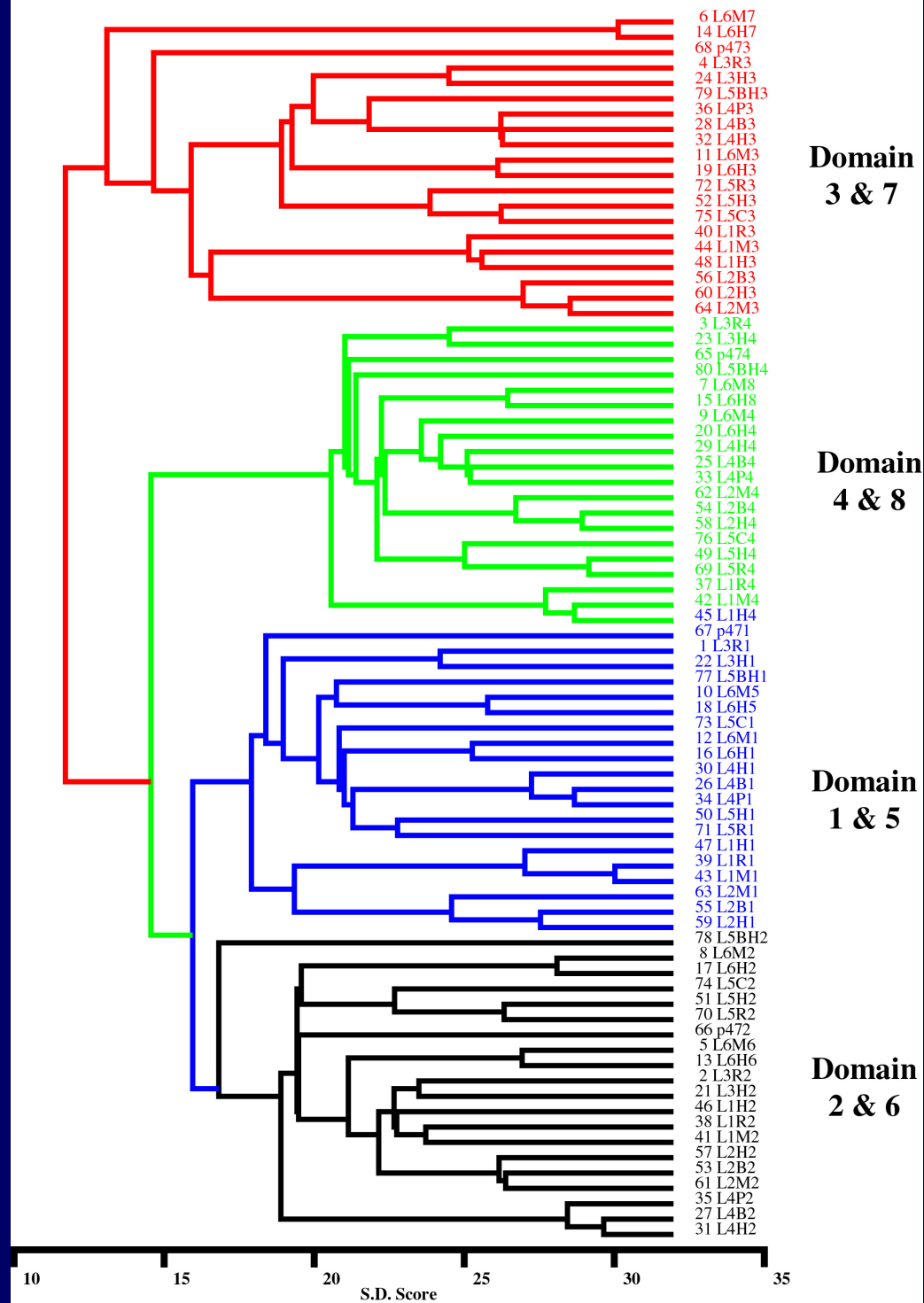
Annexin V showing Glu-Arg salt bridge between
helix 2 of domain II and helix 2 of domain IV



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

Generalise lessons learned in Annexin study

Annexin Domains



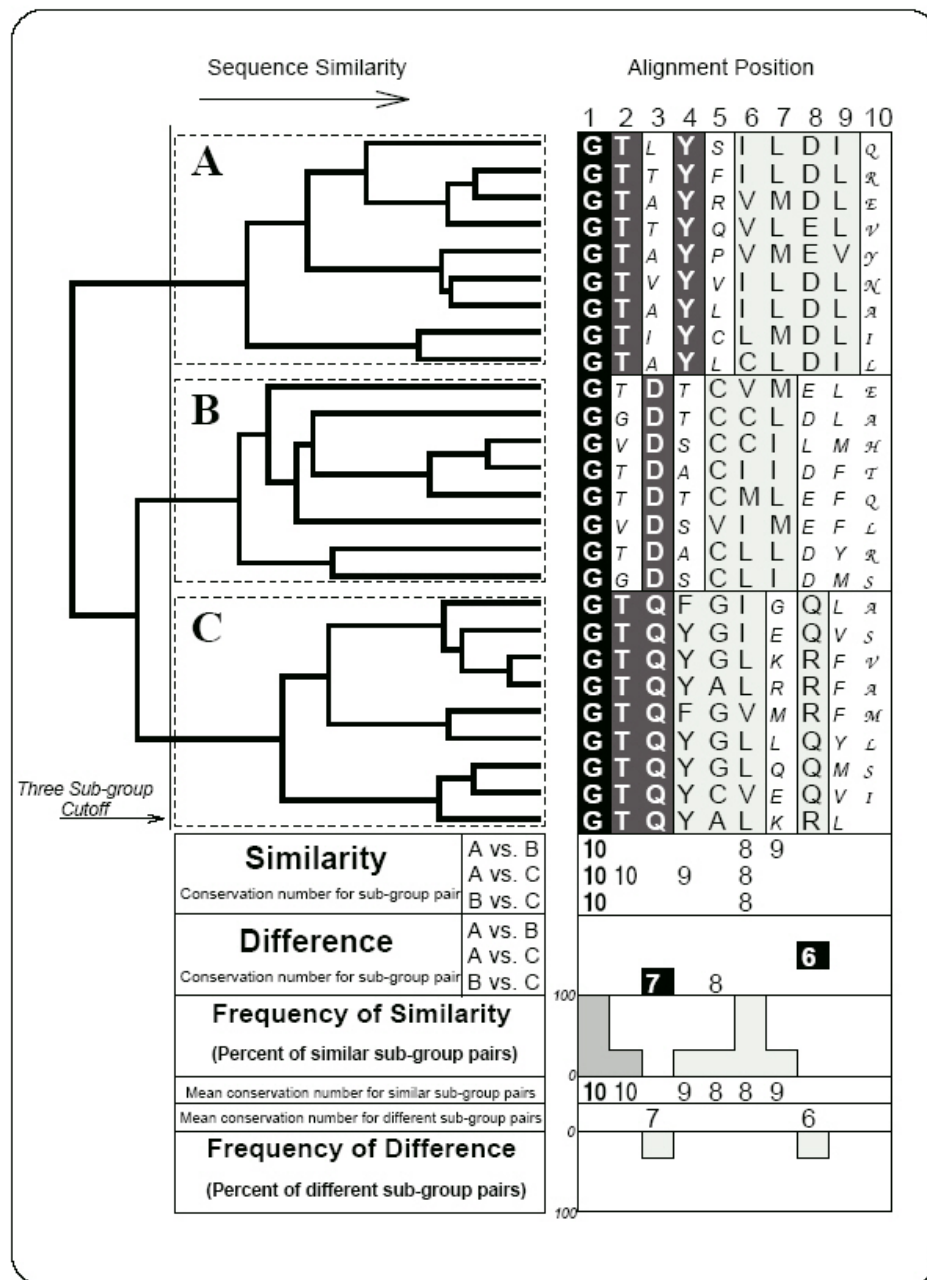


Figure 3

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

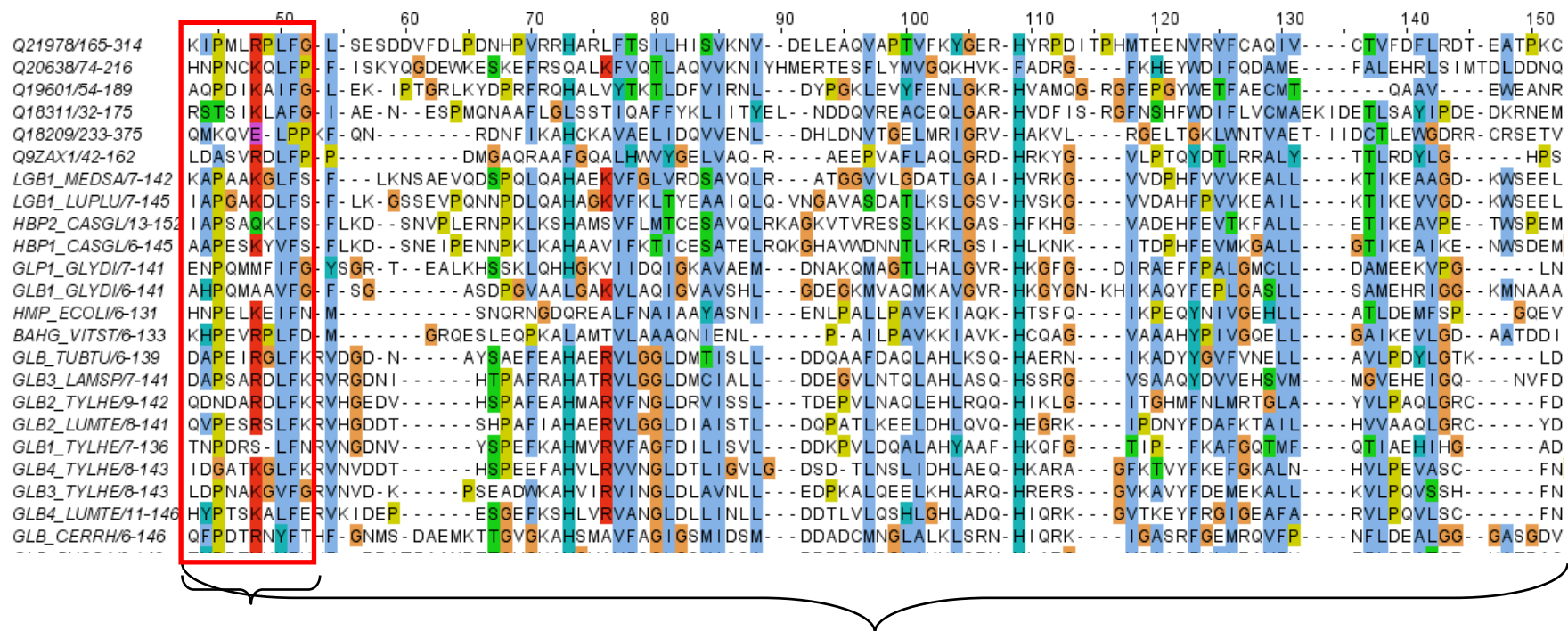
- Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci. 1993 Dec;9(6):745-56.
- Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol. 1996;266:497-512.
- Methods Enzymol paper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

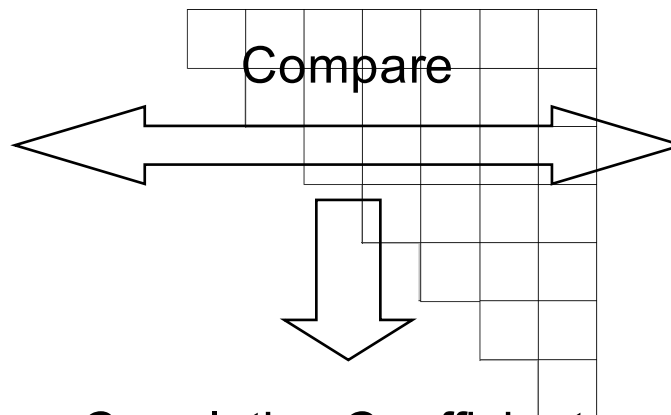
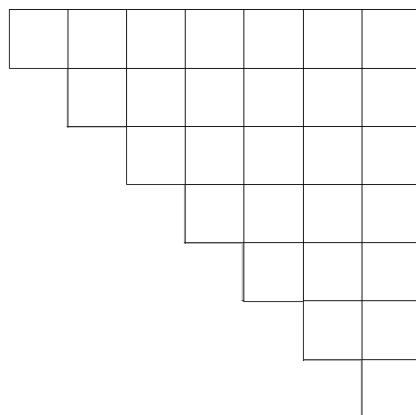
- Positions in the alignment that are most responsible for the topology of the phylogenetic tree derived from the complete alignment
- These positions may be functionally important.



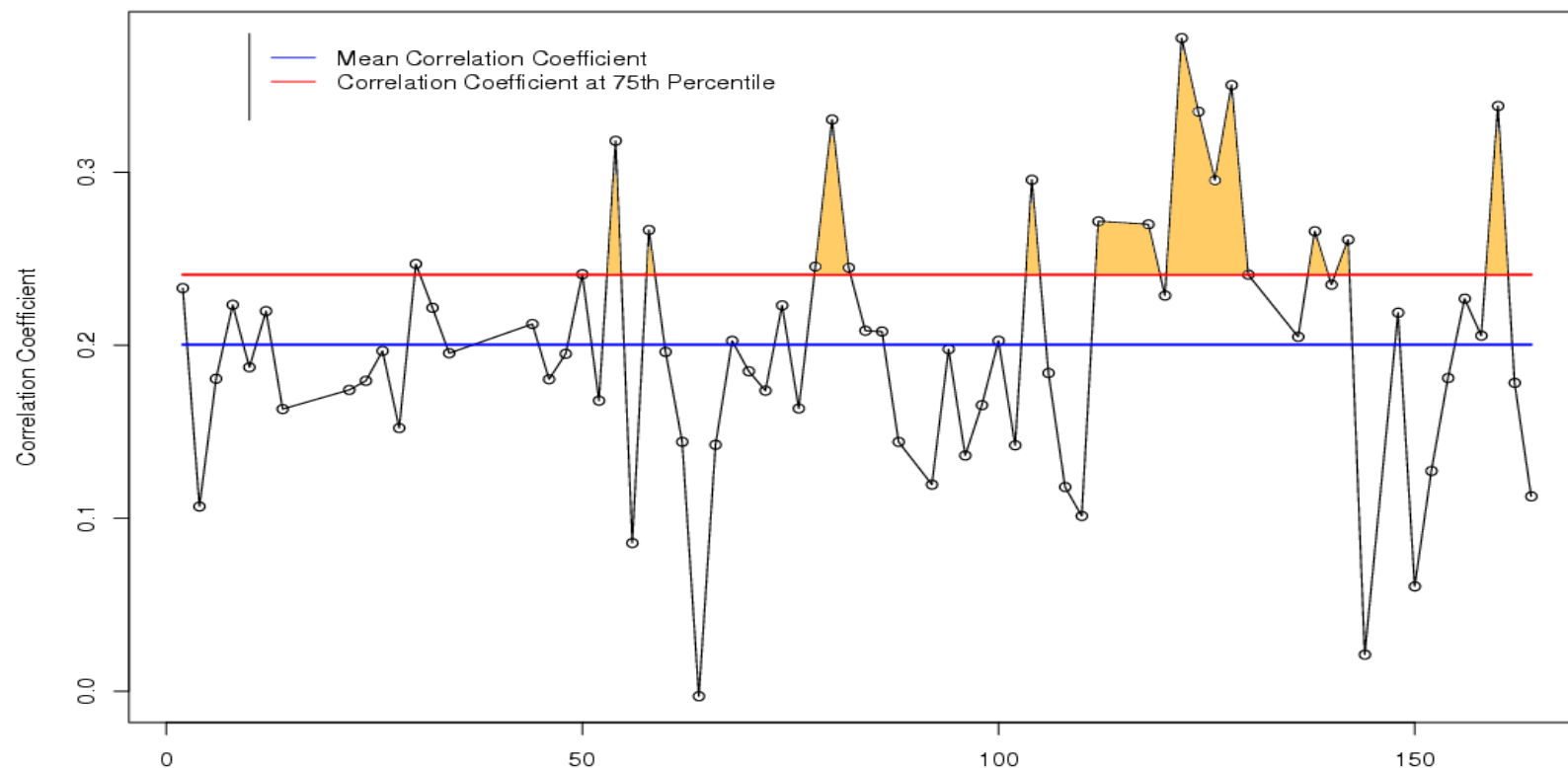
Generate Local Matrix

Generate Global Distance Matrix

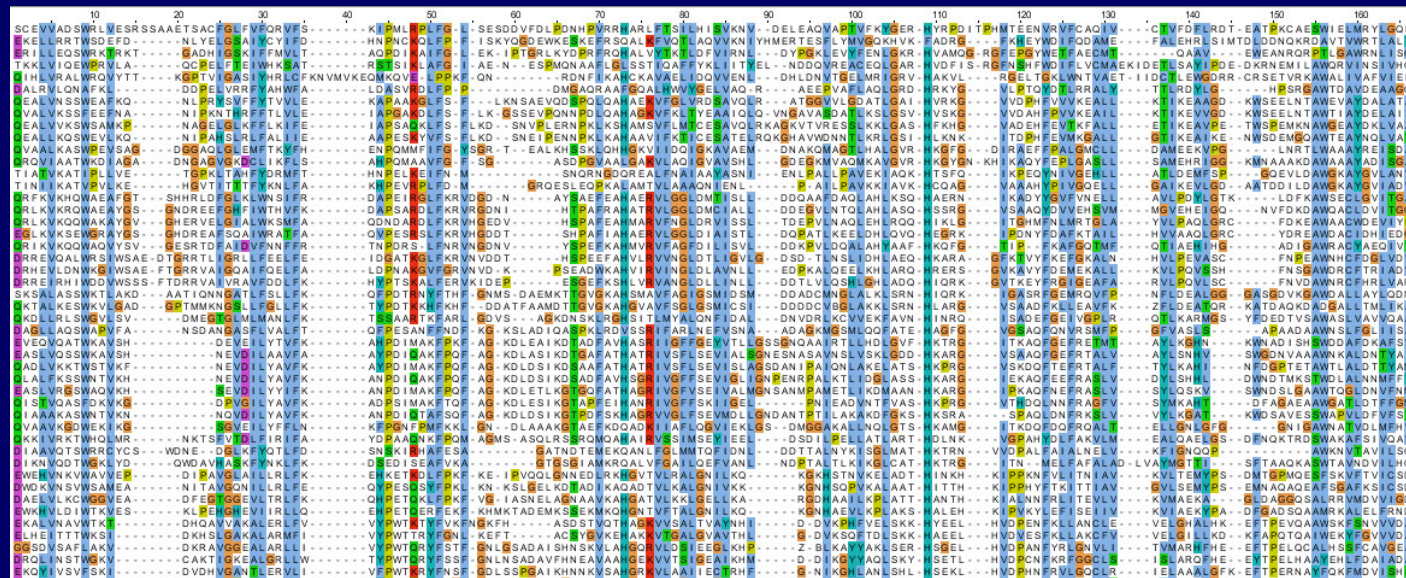
Saved Global Matrix



Correlation Coefficient

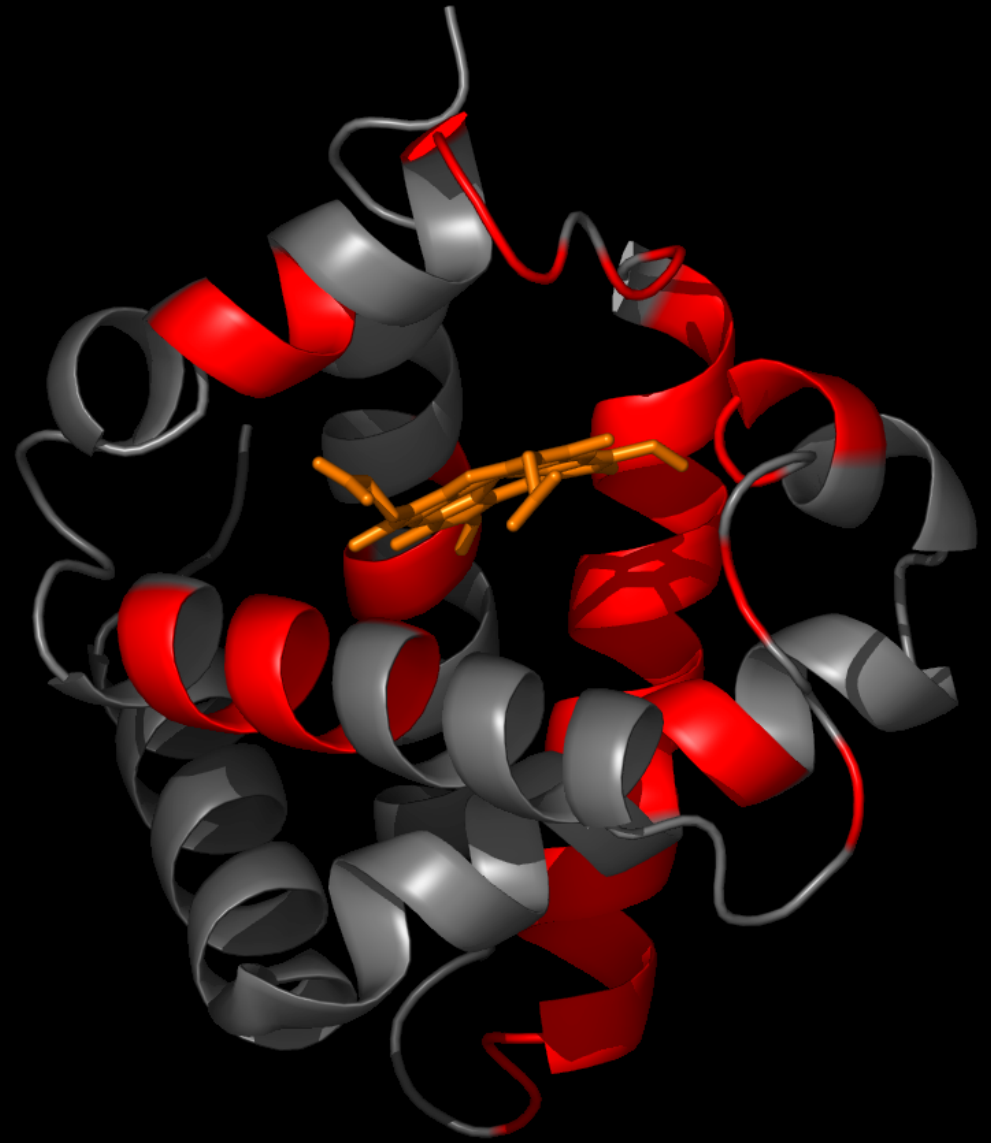
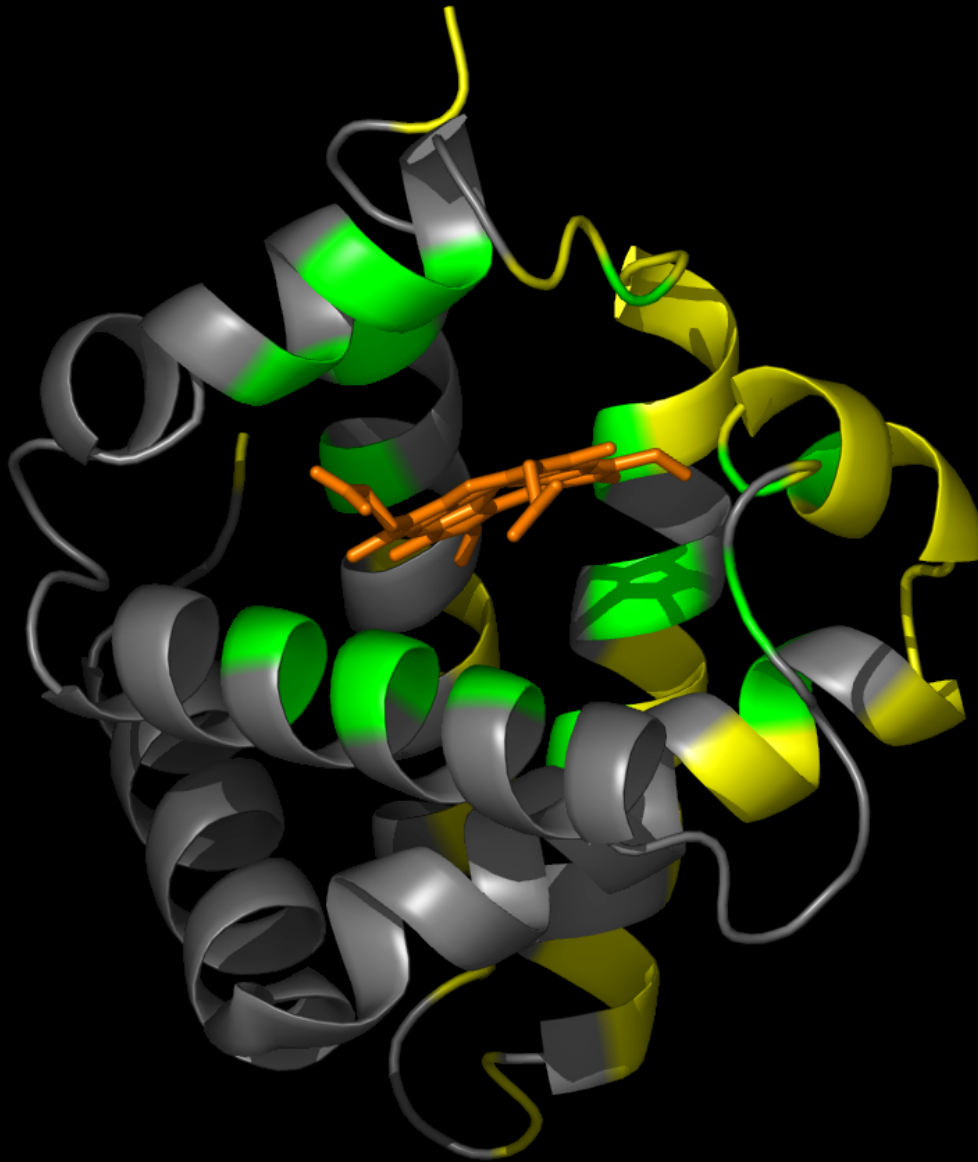


Partial Multiple Sequence Alignment (Globins)



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

