EVOLUTIONARY TRACE

Drew Bryant

Evolutionary Trace (ET)

- Predictive protein sequence-based method
- Analyzes families of homologous proteins
- Extracts highly conserved residues (traces)
- 3-d clustering of traces can identify functionally significant protein features

ET timeline

- 1996: Original ET method paper
- 2002: Applications paper (assigned)
- 2004: Real-valued ET
- 2006: ET report-maker web service

ET input data

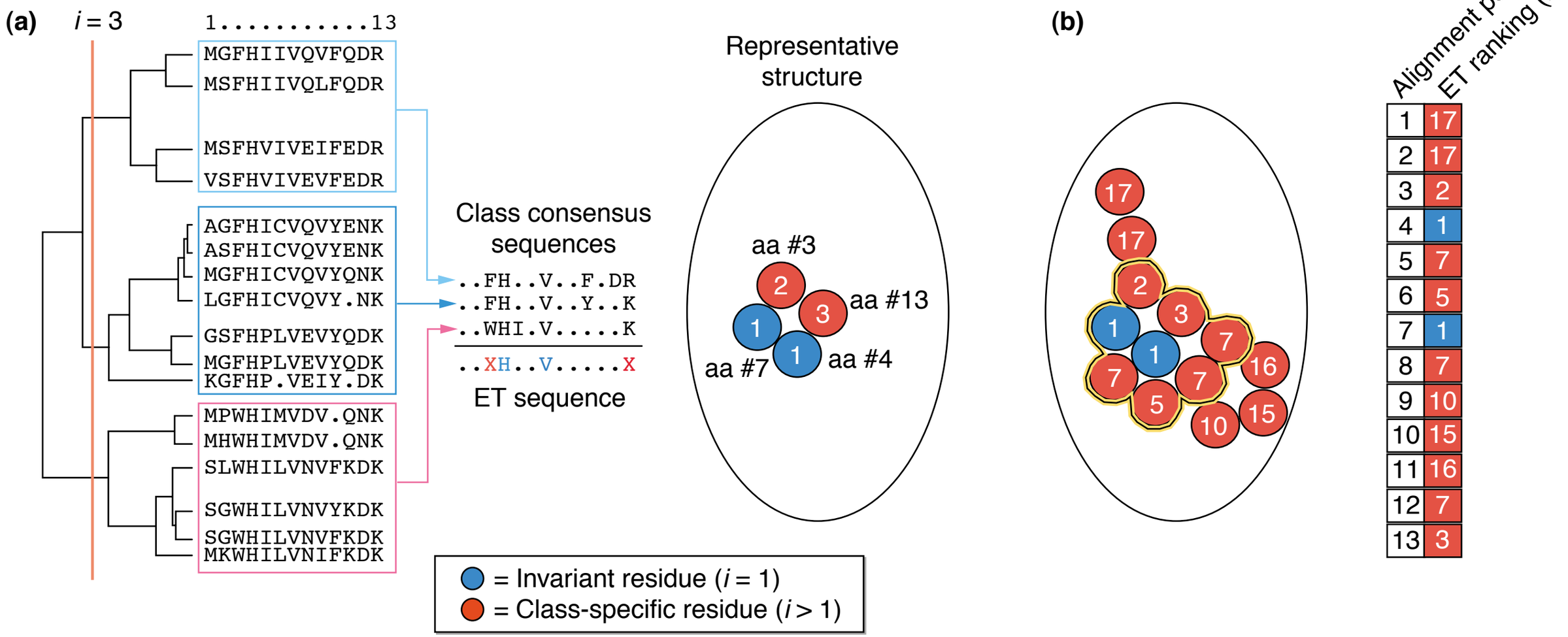
- Protein family(sequences) with *divergently* related sequences in MSA
  - Family is a set of functional homologs
- Tree for sequence family
  - ClustalW, UPGMA, NJ, etc.

Assumptions

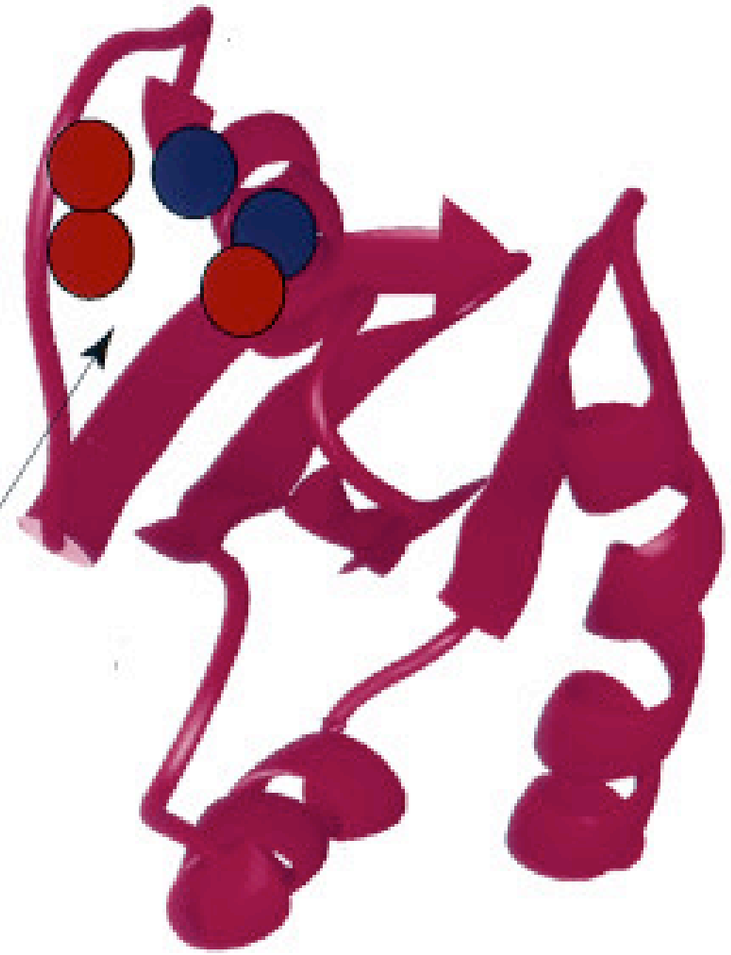
- Functional sites evolve through variations on conserved sequence positions
- Sequence positions are independent of one another
- Sequence identity trees (input data) approximate functional classifications

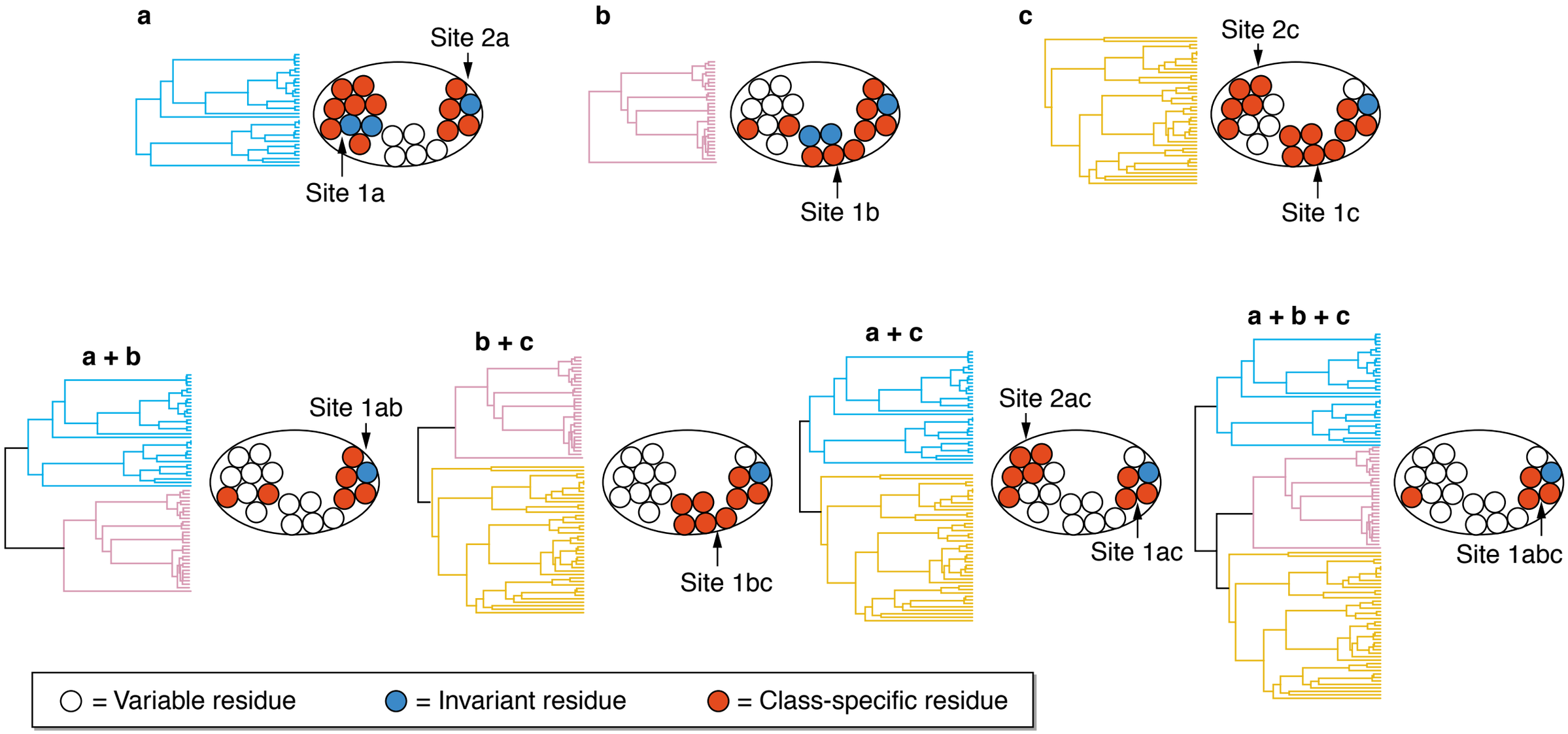
ET overview

- Iteratively partition the tree
  - Increasing number of subgroups delineated by branch points in the tree
    - # of partitions = trace rank
- Trace residue: invariant within branch but variable between branches (class specific)
- Evolutionary rank is the minimum number of partitions to become a trace residue



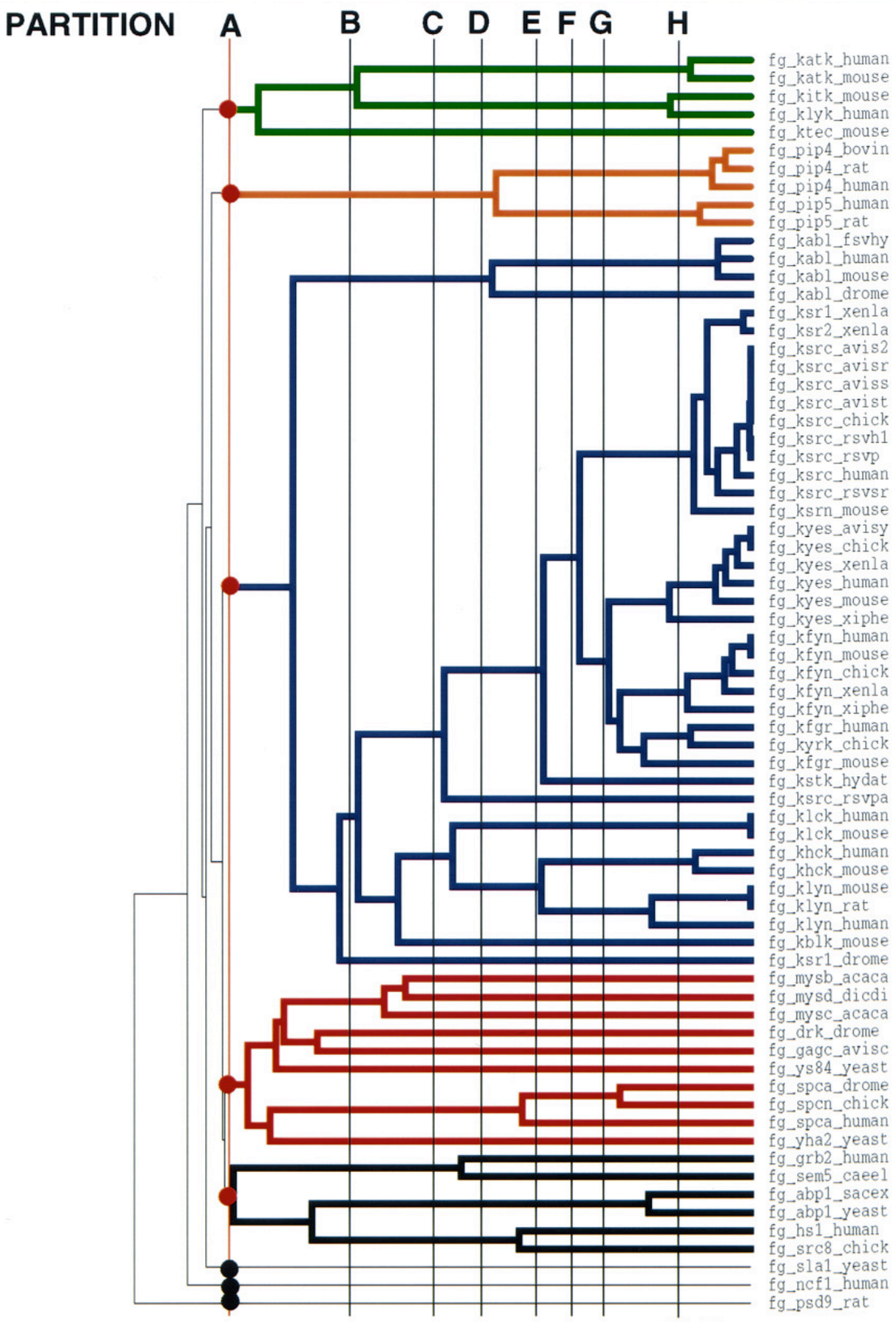
- Partitioning continues until each sequence is a singleton cluster
- Rank determined for all of *N* sequences:
  - Min rank = 1 (i.e. invariant among all sequences)
  - Max rank = *N* (i.e. only invariant within singleton)
- Structural clustering of highly conserved residues is useful
  - Can identify functional sites, binding surfaces: filters noise

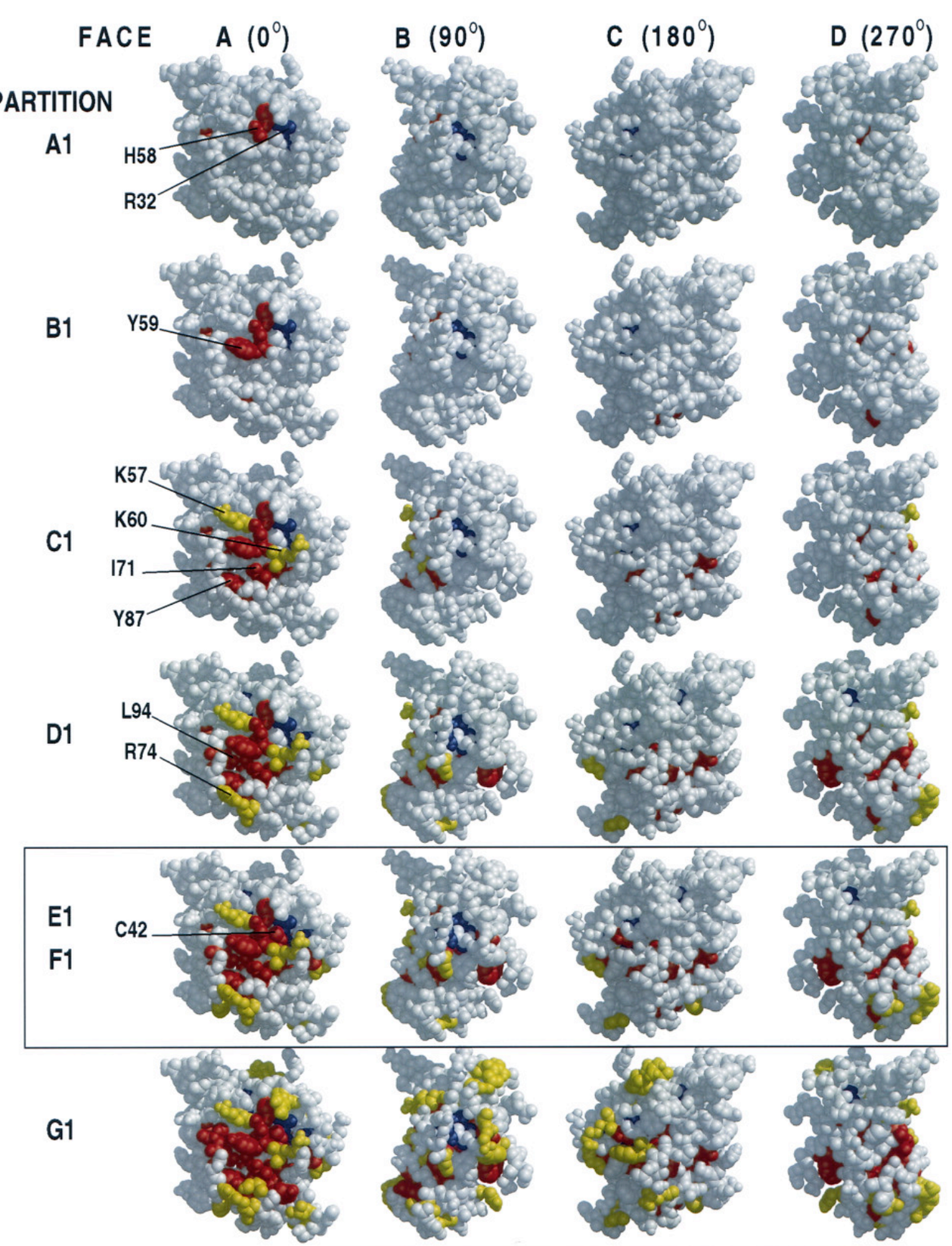
FUNCTIONAL GROUPS	CONSENSUS SEQUENCES	TRACE	MAPPING
1	AERTFTGHKKNM AERTFTGHKRNM AERTFTVHKRNM AEKTFTGHKKNM AE_TFT_HK_NM	AE_TFT_HK_NM	
2	VERTFTGHKEQM VERTYTGKKRQM VERTFTGMKRQM VERT_TG_K_QM	VERT_TG_K_QM	
3	ADR.YTGVKKNM ADR.YTGHKKNM ADR.YTGHKKNN ADR.YTGHKKNM ADR.YTG_KKN_	ADR.YTG_KKN_ XX___T__K_X_	

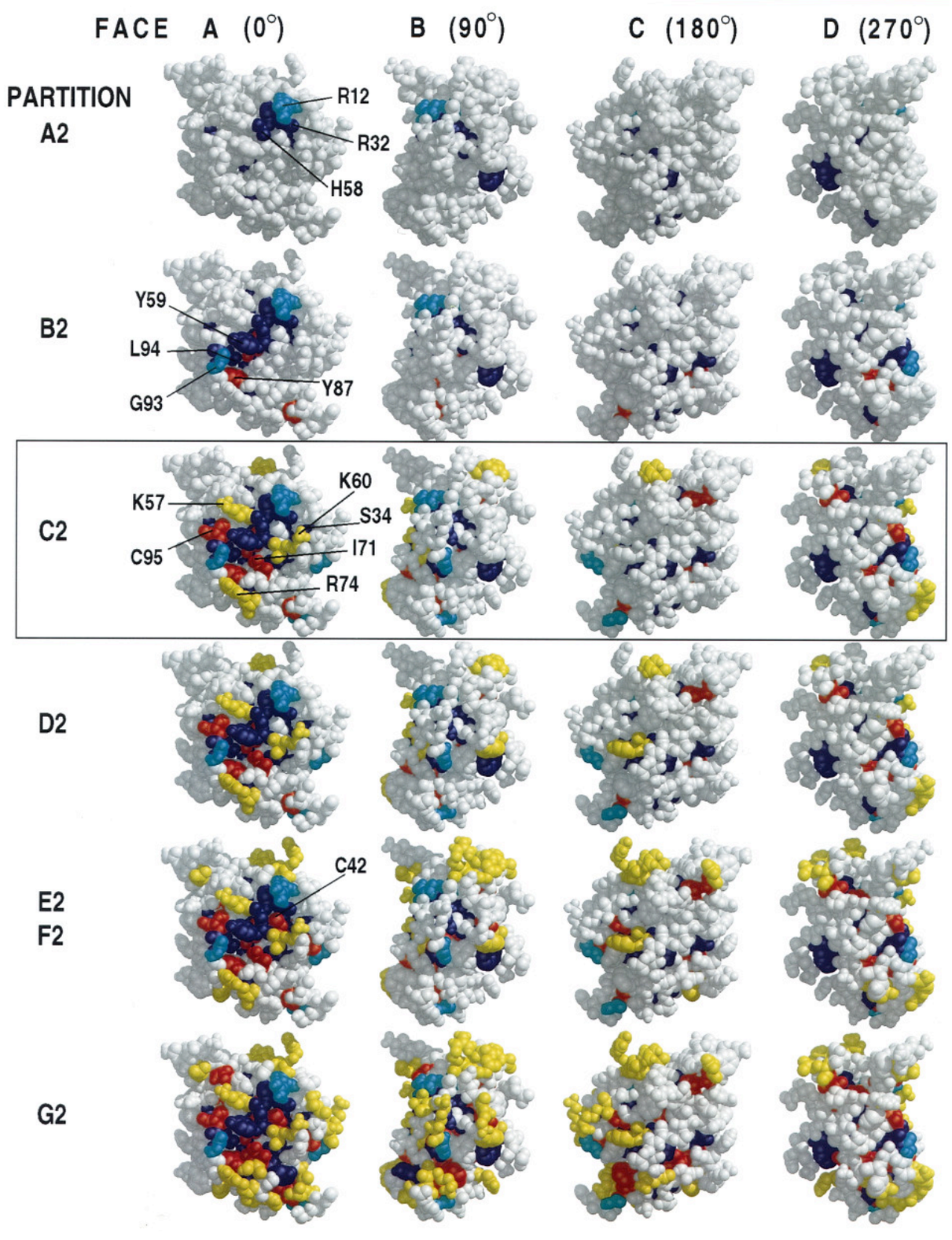


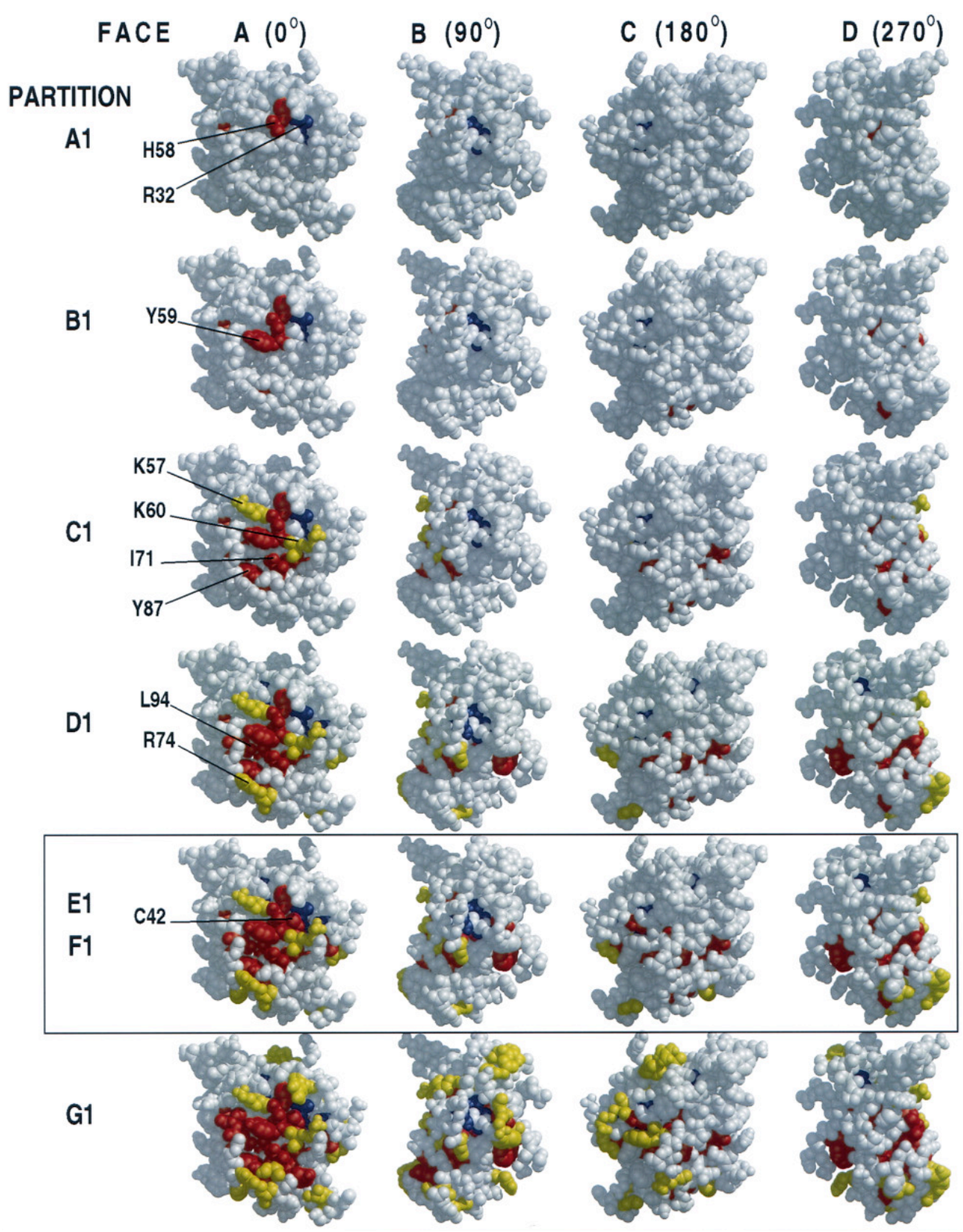
- Larger families can be subdivided
  - Run separate traces for each sub-family
  - Can reveal sub-family specific conservation
- Recombine sub-families to more complex gain/loss of functional sites

SH2 protein-protein interaction domain



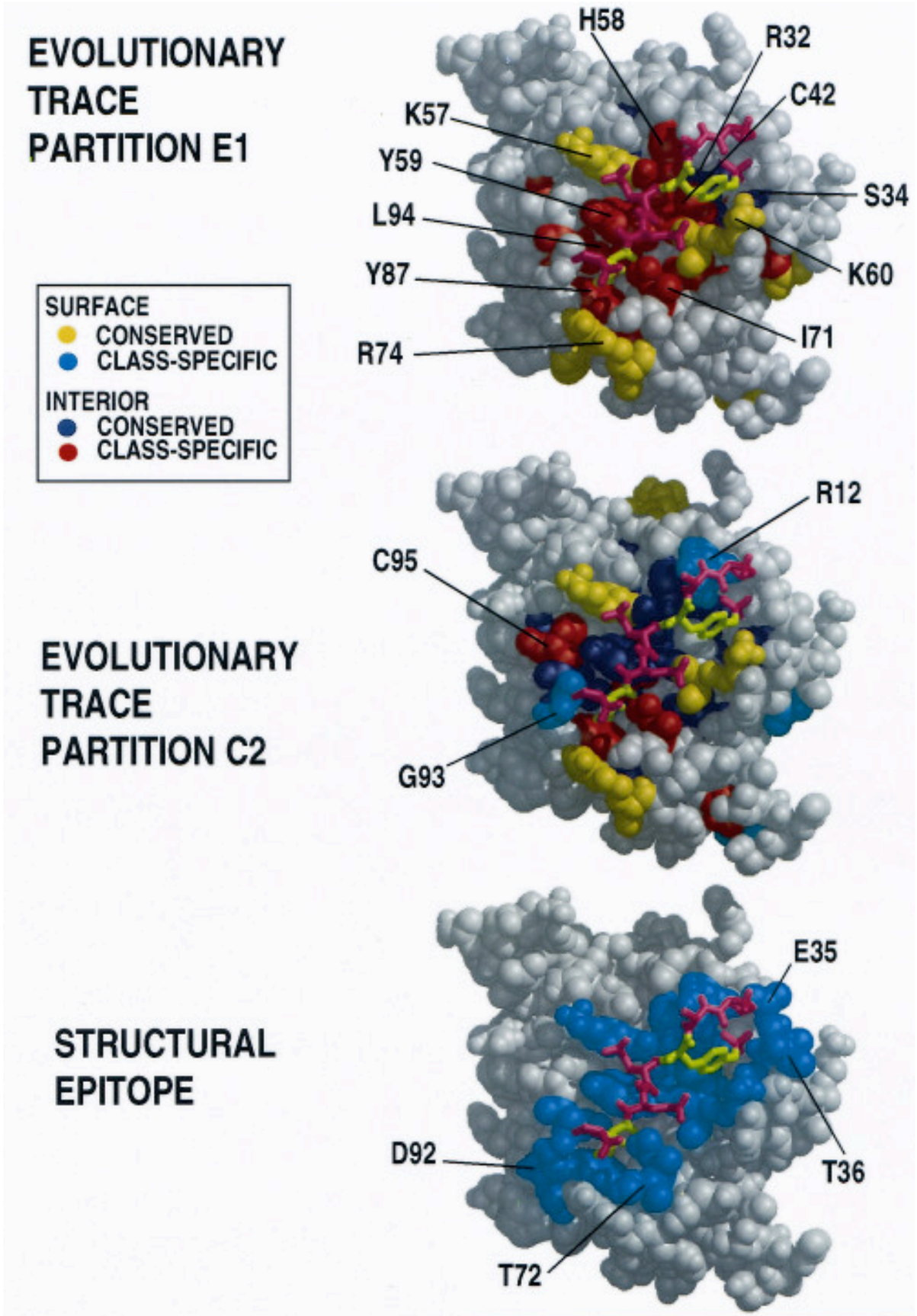






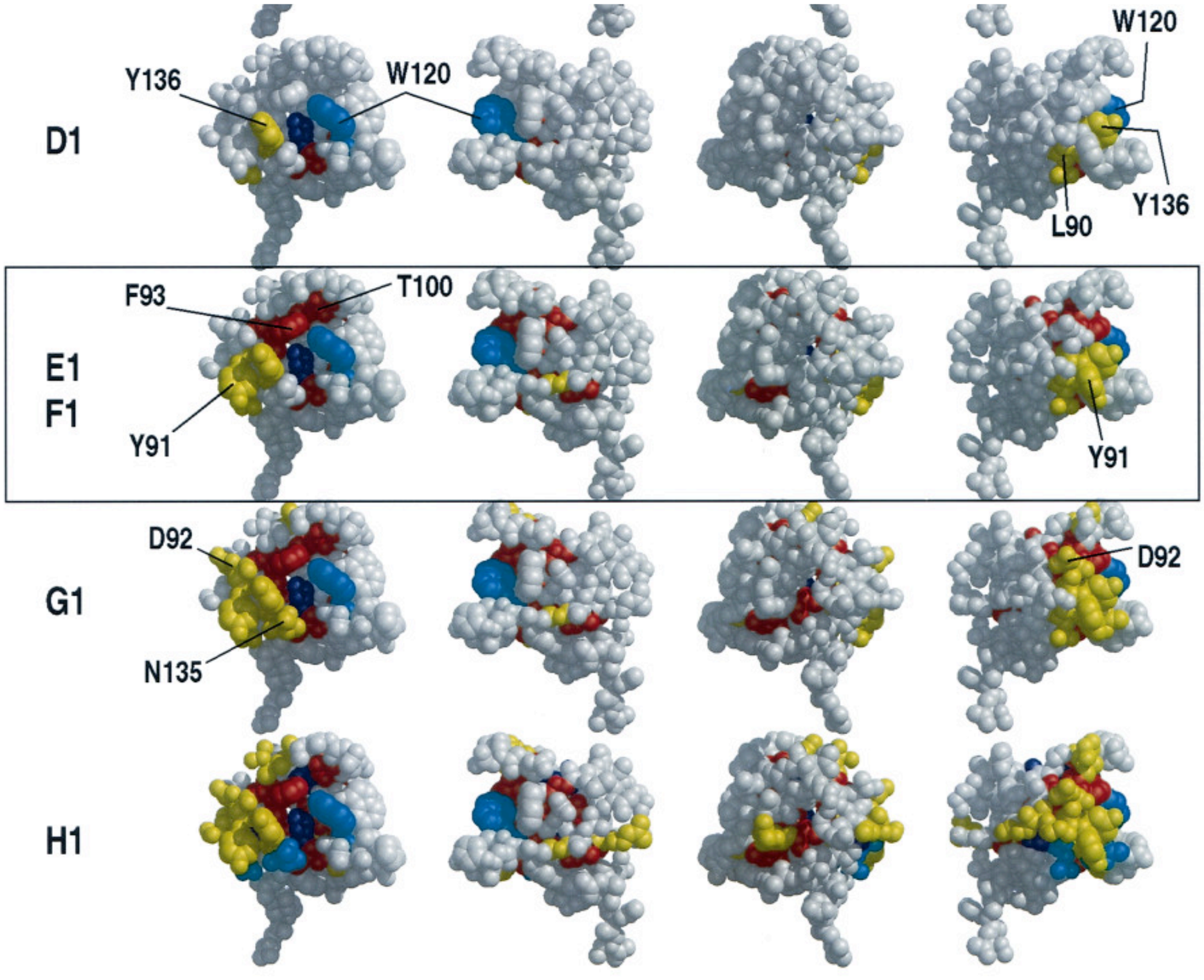
sub-family trace

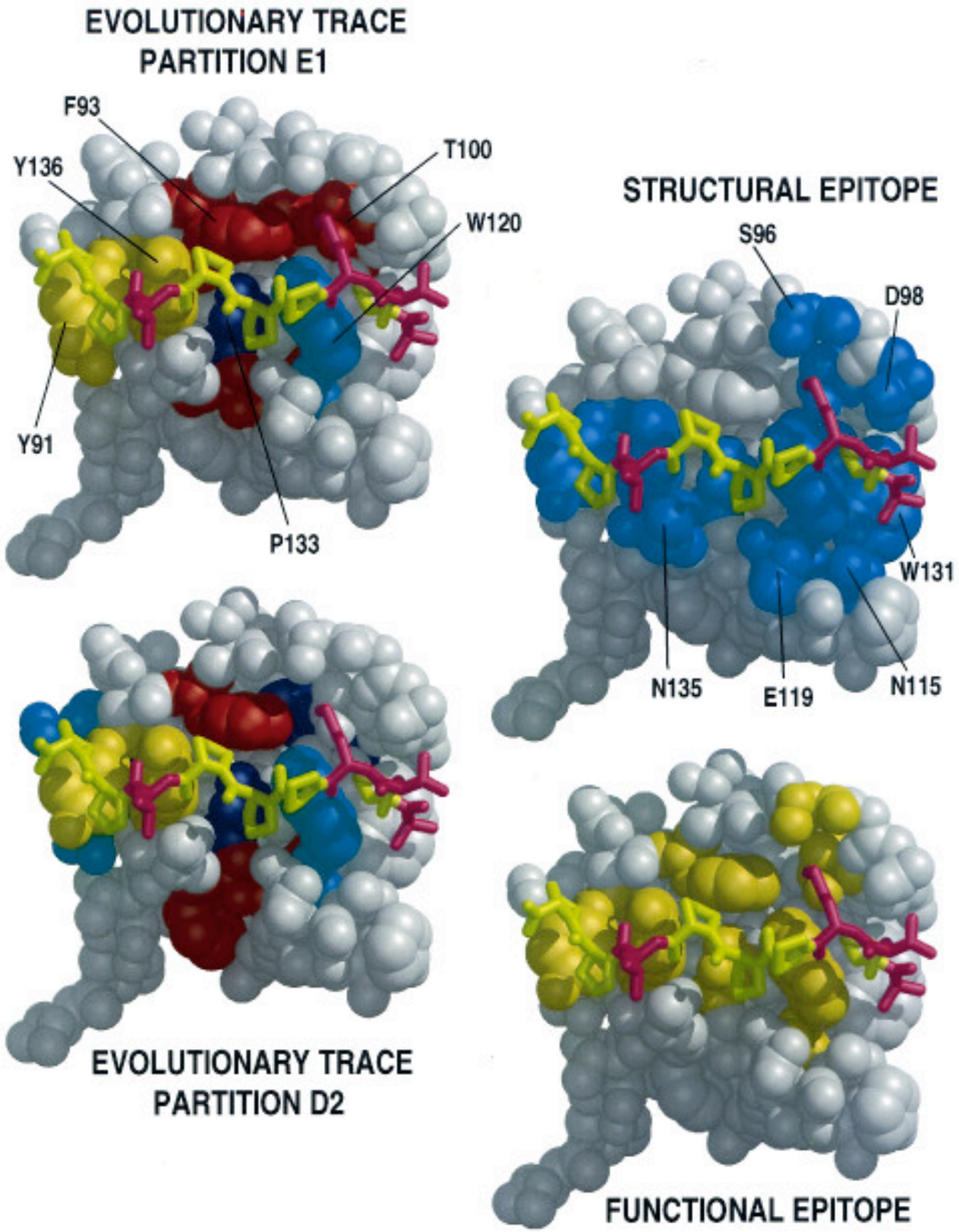
whole-family trace



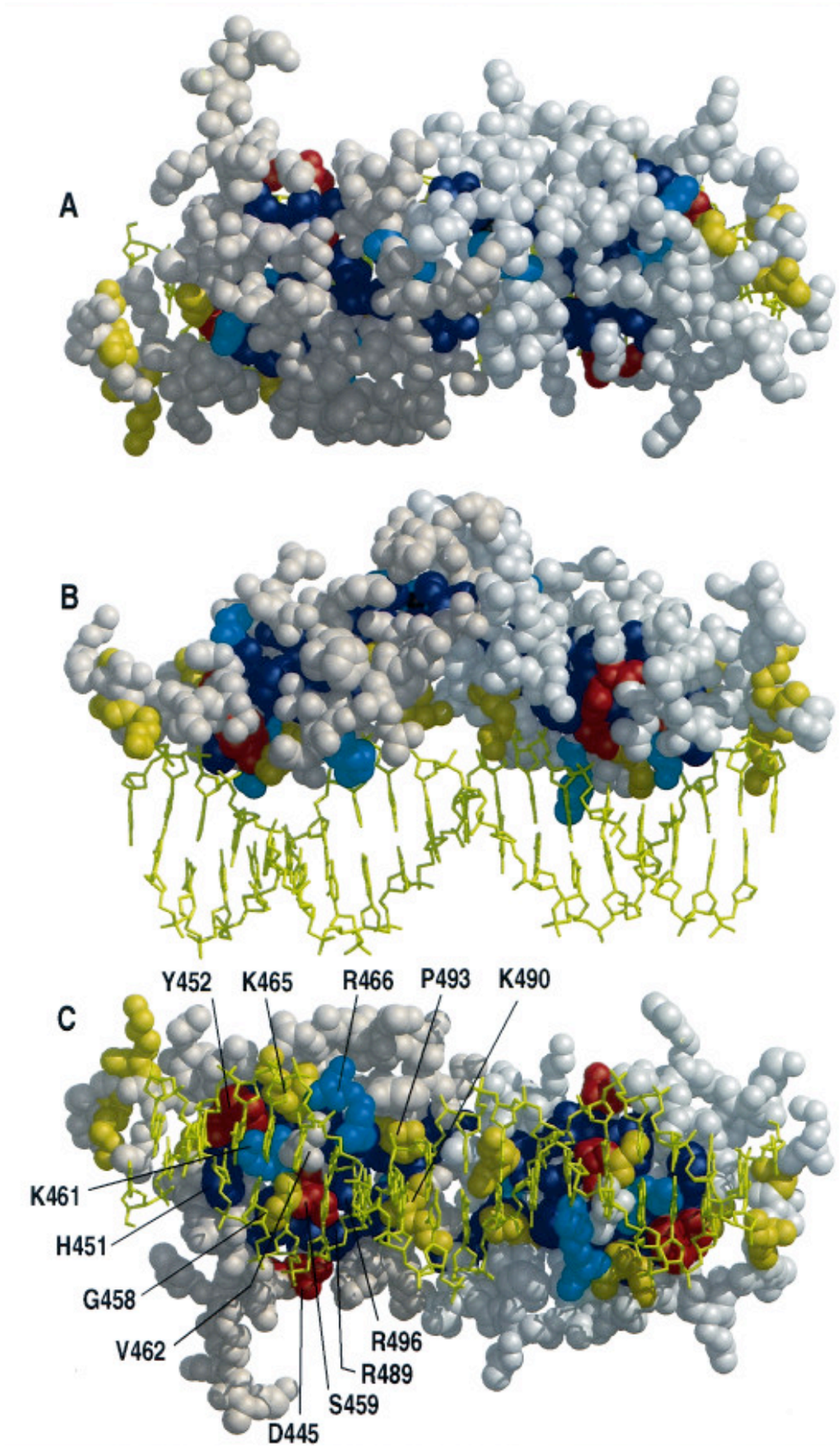
- SH2 protein-protein interaction domain
- Bound to peptide
- High-ranked ET residues cluster along protein-protein interaction surface

SH3 protein interaction domain





- SH3 protein-protein interaction domain
- Simulated protein partner peptide shown bound to interface
- Highly ranked ET residues cluster at interface



- Nuclear hormone receptor
- DNA-binding protein
- Clusters of highly-ranked residues occur at DNA interface

Real-valued traces

- ET is integer-valued by nature of discrete partitioning of tree
- Real-valued ET ranks combine entropy of a partition with standard ET rank

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left(- \sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \right)$$

- Nice webserver that returns detailed reports

- http://mammoth.bcm.tmc.edu/report_maker

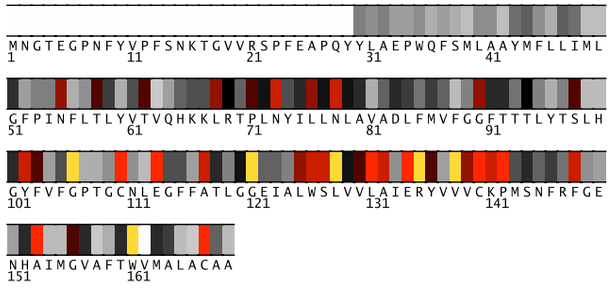


Fig. 1. Residues 1-169 in 1f88A colored by their relative importance. (See Appendix, Fig.11, for the coloring scheme.)

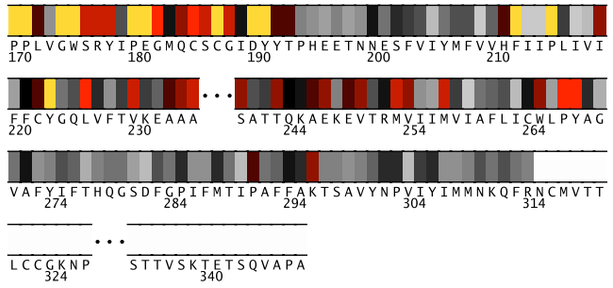


Fig. 2. Residues 170-348 in 1f88A colored by their relative importance. (See Appendix, Fig.11, for the coloring scheme.)

2.2 Multiple sequence alignment for 1f88A

For the chain 1f88A, the alignment 1f88A.msf (attached) with 613 sequences was used. The alignment was downloaded from the HSSP database, and fragments shorter than 75% of the query as well as duplicate sequences were removed. It can be found in the attachment to this report, under the name of 1f88A.msf. Its statistics, from the *alistat* program are the following:

```
Format:              MSF
Number of sequences: 613
Total number of residues:    176261
Smallest:            78
Largest:             338
Average length:      287.5
Alignment length:    338
Average identity:    53%
Most related pair:   99%
Most unrelated pair: 0%
Most distant seq:    35%
```

Furthermore, <1% of residues show as conserved in this alignment.

The alignment consists of 99% eukaryotic (93% vertebrata, 4% arthropoda) sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 1f88A.descr.

2.3 Residue ranking in 1f88A

The 1f88A sequence is shown in Figs. 1–2, with each residue colored according to its estimated importance. The full listing of residues in 1f88A can be found in the file called 1f88A.ranks_sorted in the attachment.

2.4 Top ranking residues in 1f88A and their position on the structure

In the following we consider residues ranking among top 25% of residues in the protein . Figure 3 shows residues in 1f88A colored by their importance: bright red and yellow indicate more conserved/important residues (see Appendix for the coloring scheme). A Pymol script for producing this figure can be found in the attachment.

2.4.1 Clustering of residues at 25% coverage. Fig. 4 shows the top 25% of all residues, this time colored according to clusters they belong to. The clusters in Fig.4 are composed of the residues listed

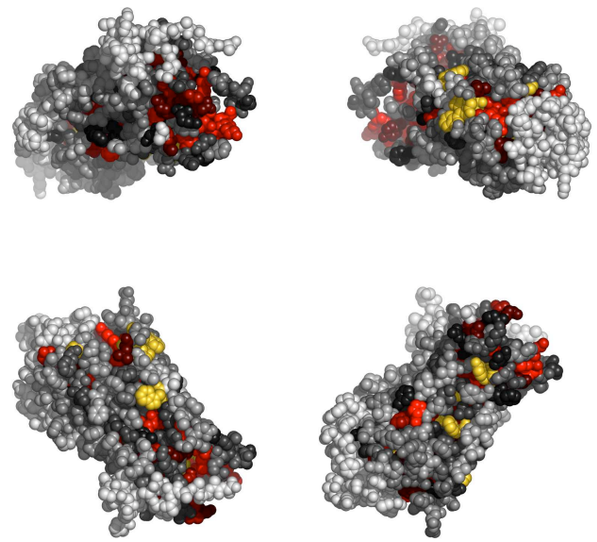


Fig. 3. Residues in 1f88A, colored by their relative importance. Clockwise: front, back, top and bottom views.

in Table 1.

Table 1.

cluster color	size	member residues
red	83	55,58,62,68,69,71,73,76,78 90,98,102,103,106,110,113 117,121,125,126,127,128,130 131,132,134,135,136,138,139 140,141,142,148,153,156,161 167,170,171,172,174,175,176 177,178,180,181,182,184,185 186,187,188,190,191,192,193 211,212,215,219,222,223,226 230,233,234,235,243,244,246 247,249,250,253,254,257,265

continued in next column

Fin