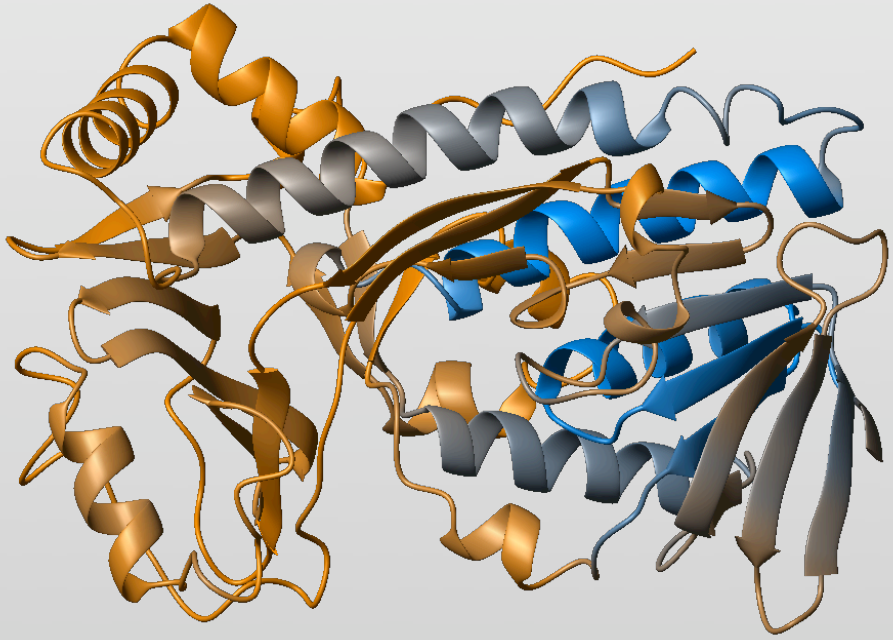
Identification of Structural Domains in Proteins

Marc A. Marti-Renom

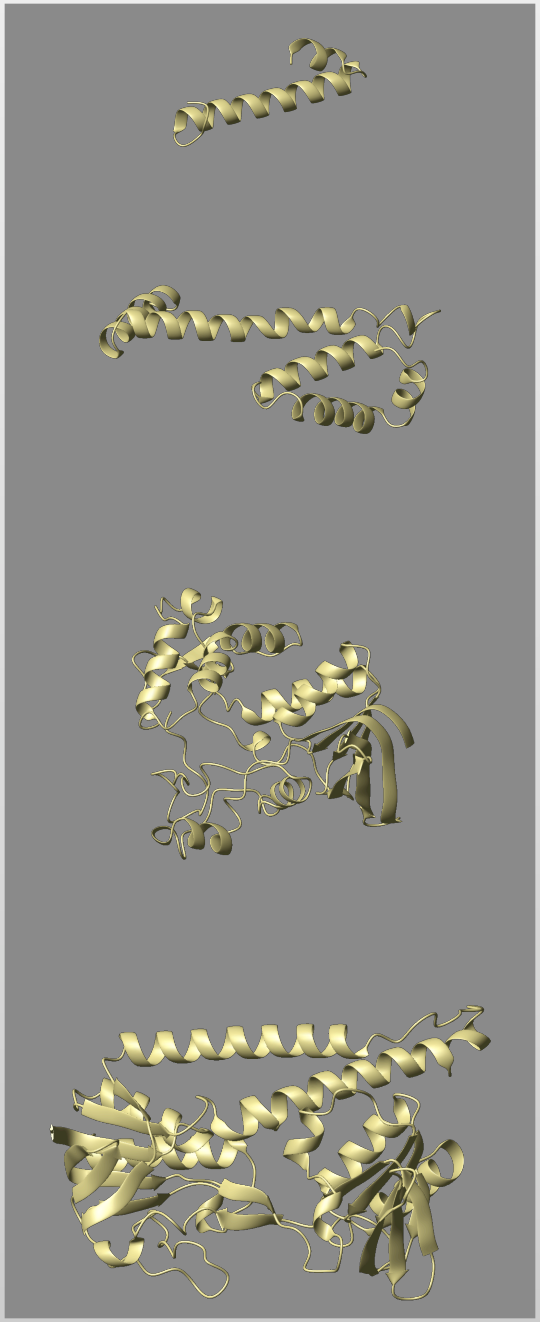
Laboratories of Molecular Biophysics
The Rockefeller University. New York City

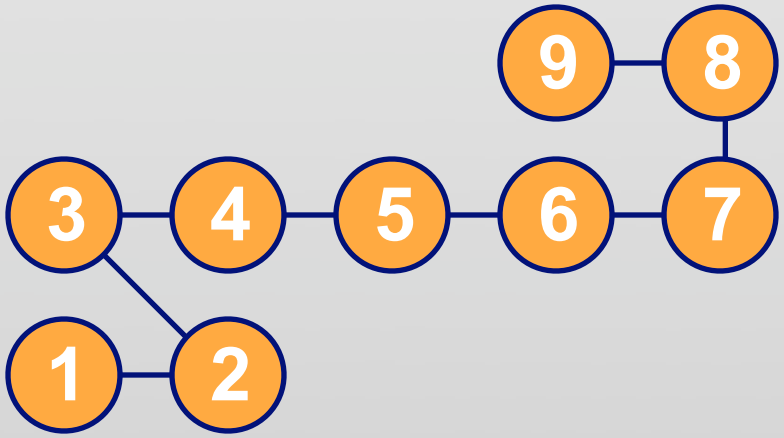
Department of Biopharmaceutical Sciences
University of California, San Francisco

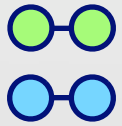
PAR-DOM



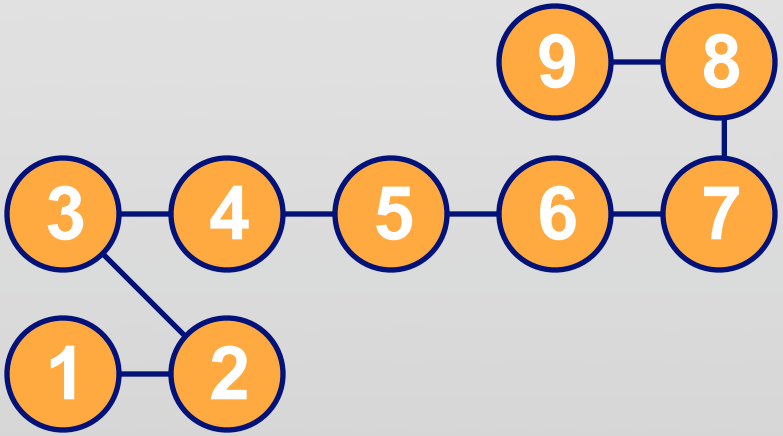
1phh (Oxydoreductase from *Pseudomonas fluorescens*)

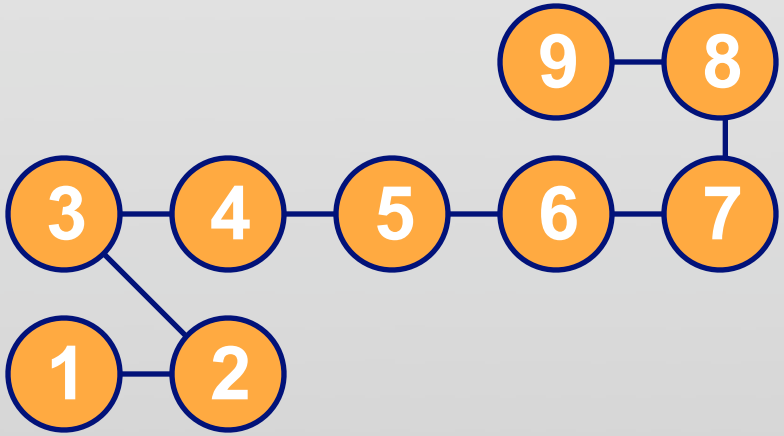


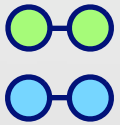

9
8
3
4
5
6
7
1
2



{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

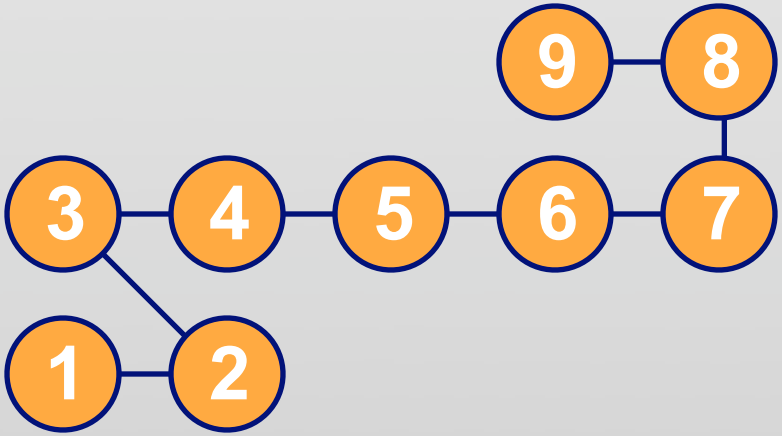


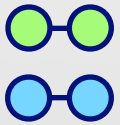




{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

{1,2,3,4}

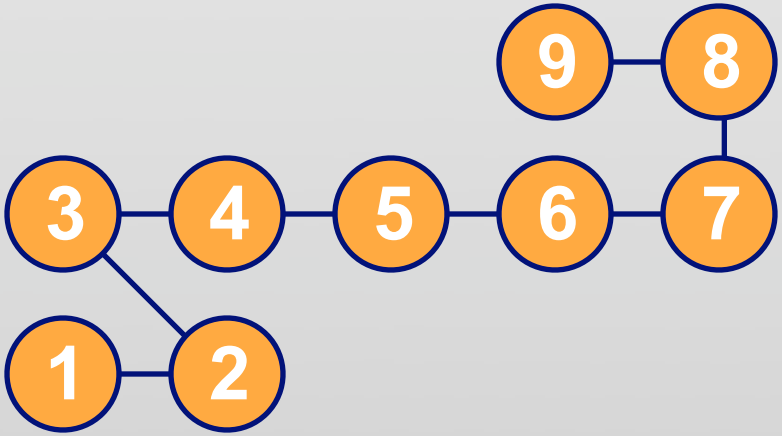


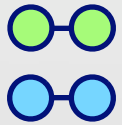


{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

{1,2,3,4}

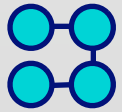
{6,7,8,9}



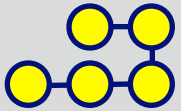


{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

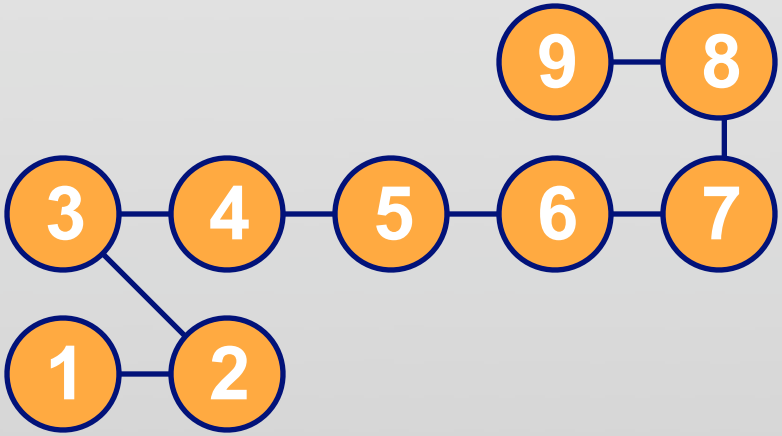
{1,2,3,4}

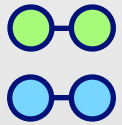


{6,7,8,9}



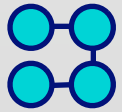
{5,6,7,8,9}



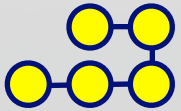


{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

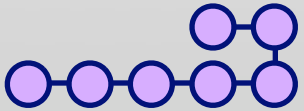
{1,2,3,4}



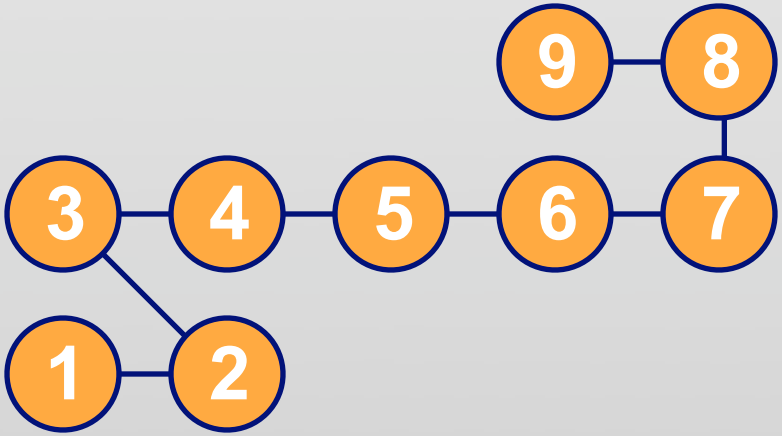
{6,7,8,9}

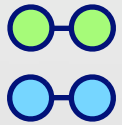


{5,6,7,8,9}



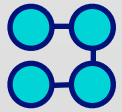
{3,4,5,6,7,8,9}



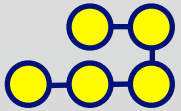


{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

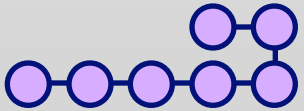
{1,2,3,4}



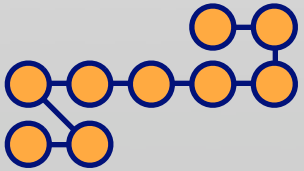
{6,7,8,9}



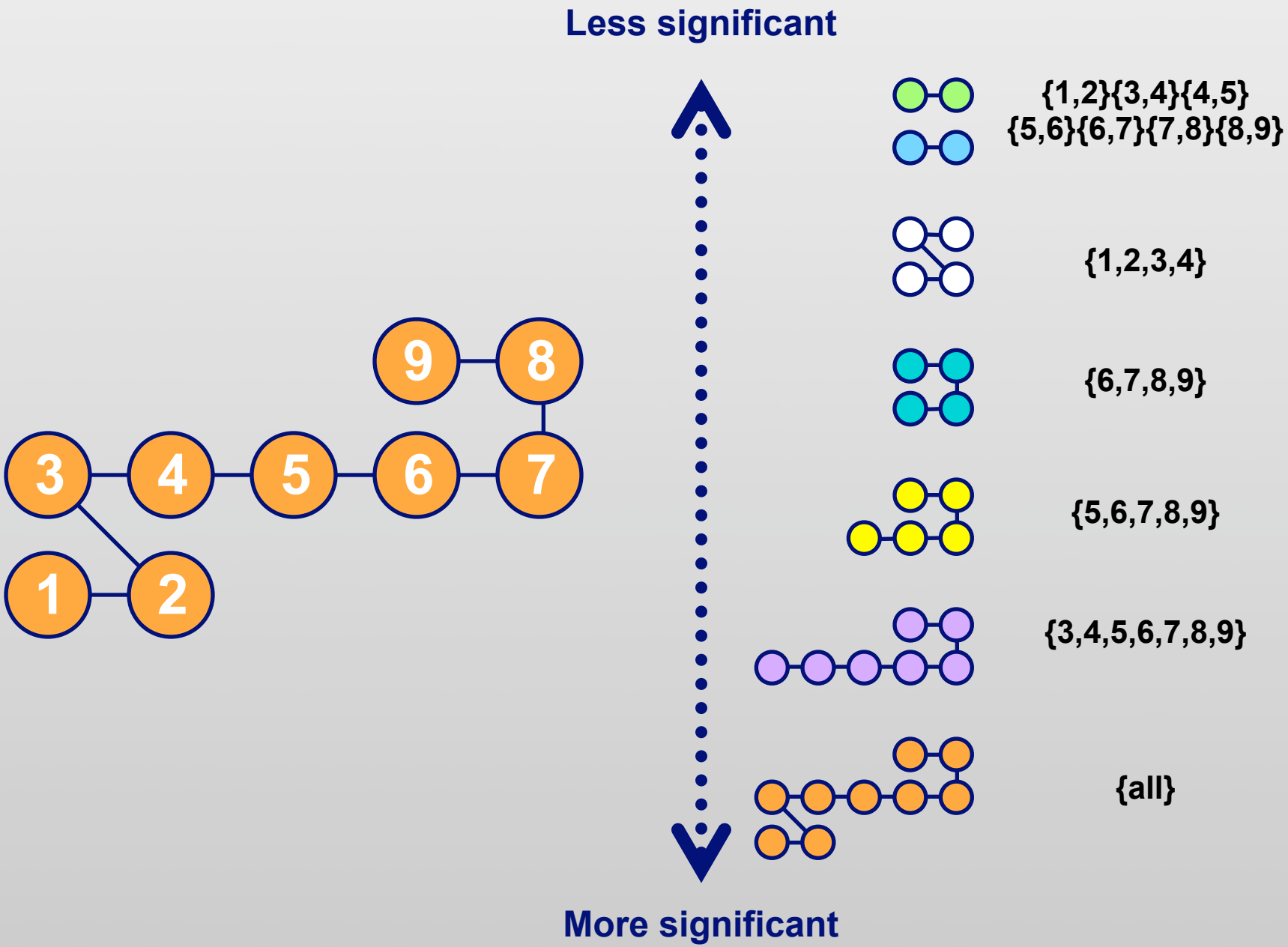
{5,6,7,8,9}

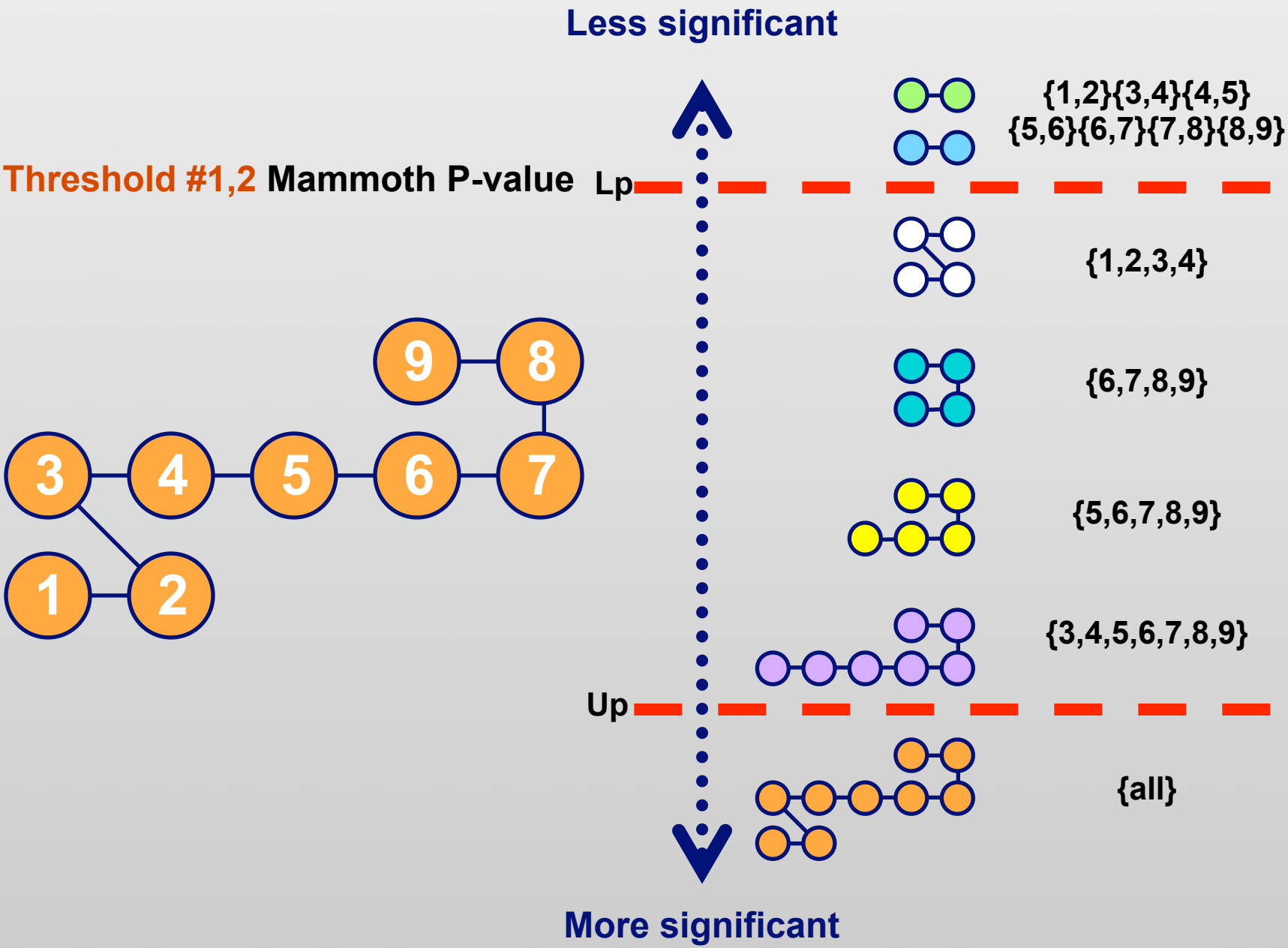


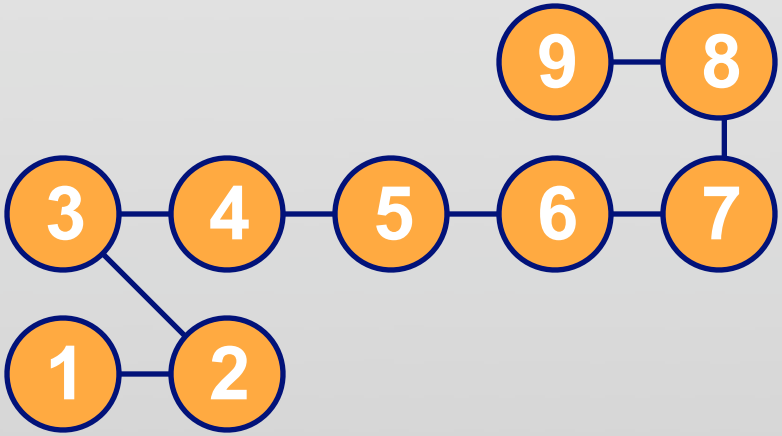
{3,4,5,6,7,8,9}



{all}

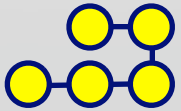

Less significant
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}
{1,2,3,4}
{6,7,8,9}
{5,6,7,8,9}
{3,4,5,6,7,8,9}
{all}
9
8
3
4
5
6
7
1
2
More significant


Less significant
Threshold #1,2 Mammoth P-value
Lp
Up
More significant
1
2
3
4
5
6
7
8
9
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}
{1,2,3,4}
{6,7,8,9}
{5,6,7,8,9}
{3,4,5,6,7,8,9}
{all}

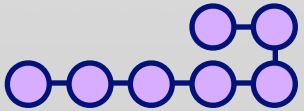


{1,2,3,4}

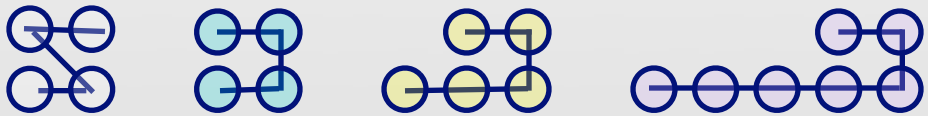
{6,7,8,9}

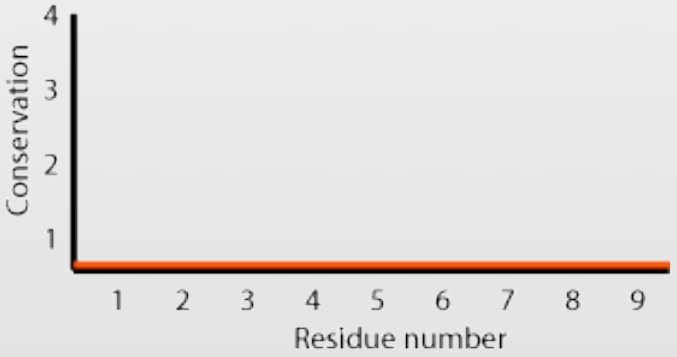


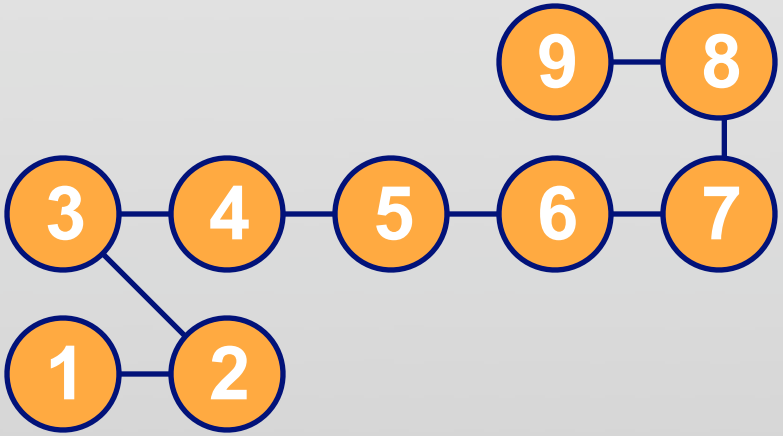
{5,6,7,8,9}



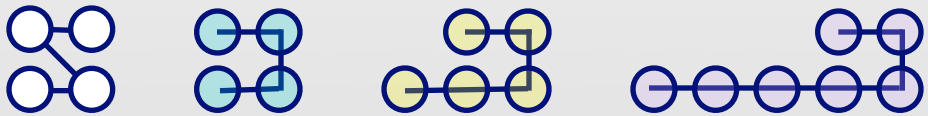
{3,4,5,6,7,8,9}

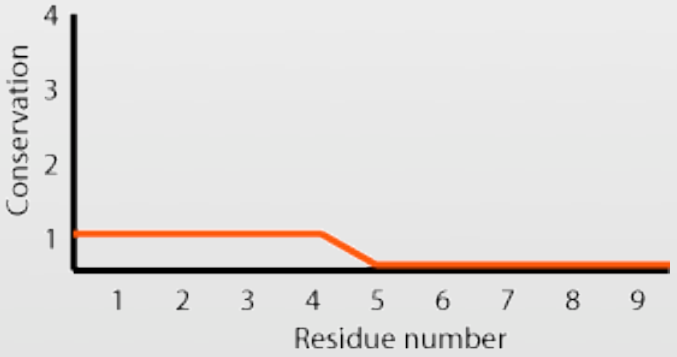


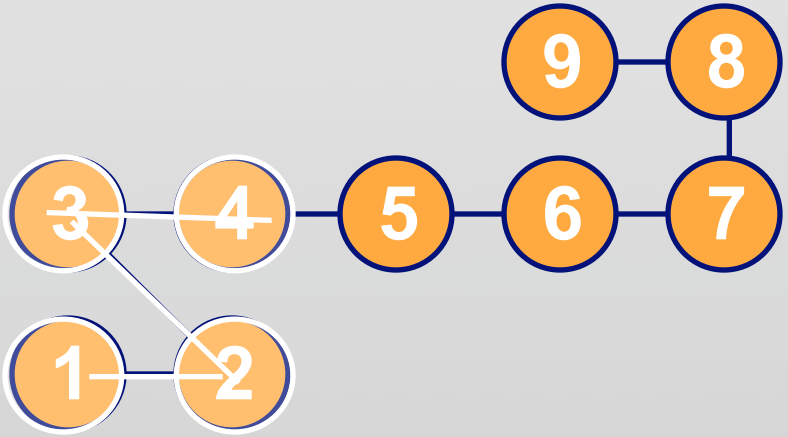




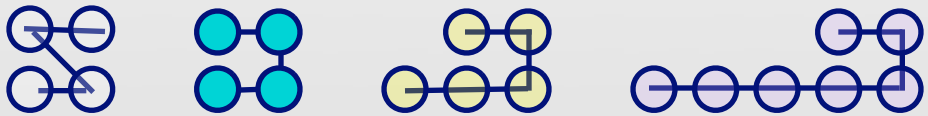
#	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

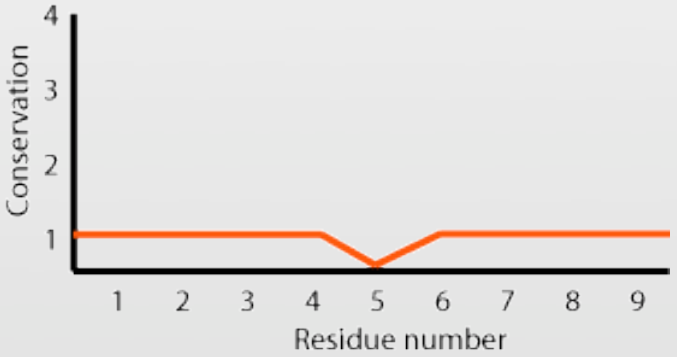


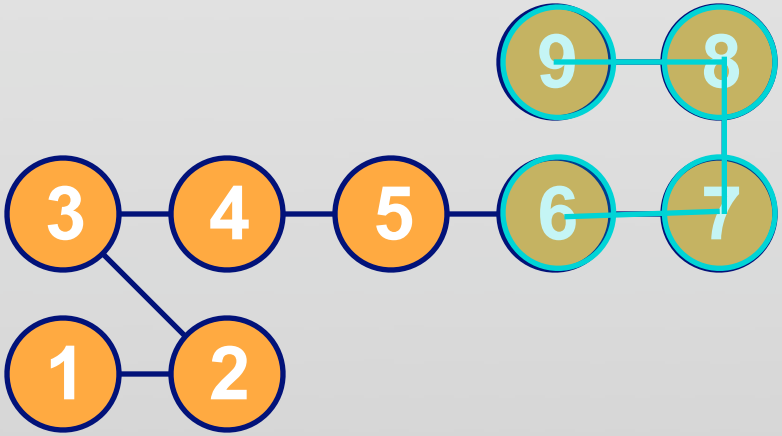




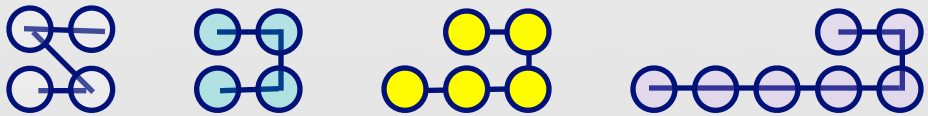
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

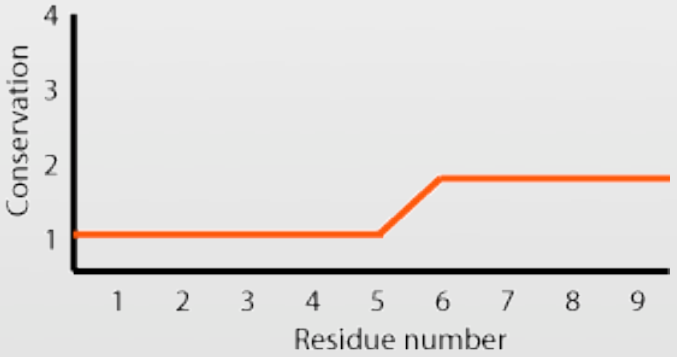


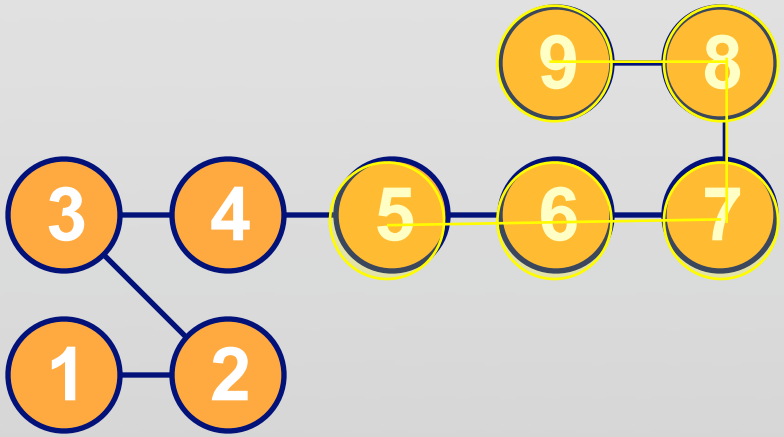




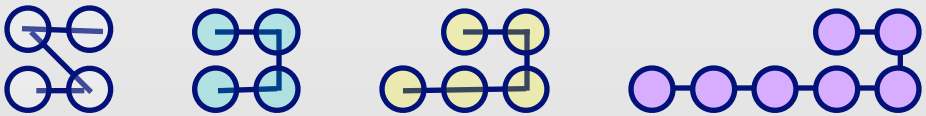
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	1	1	1	1
9	0	0	0	0	0	1	1	1	1

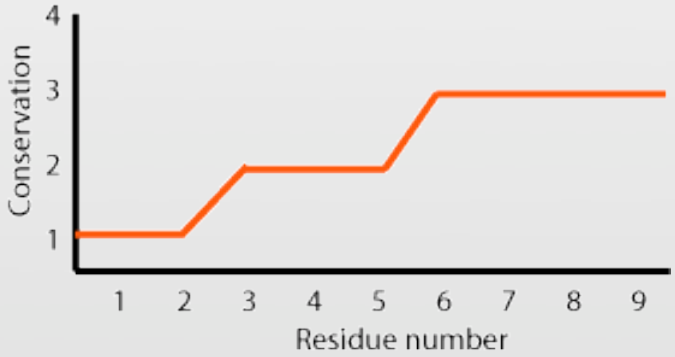


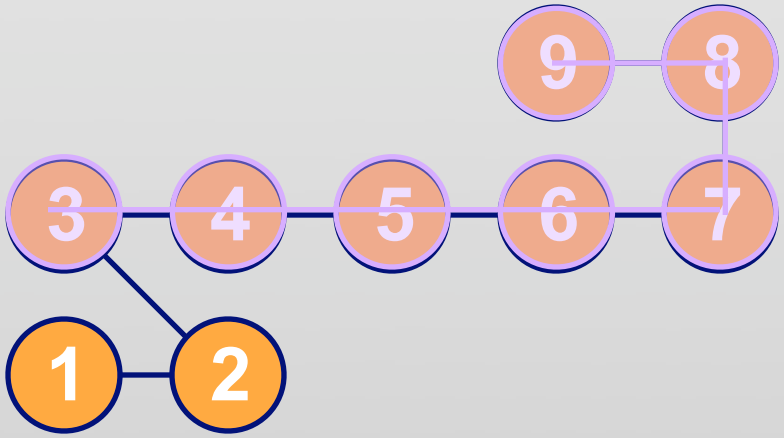




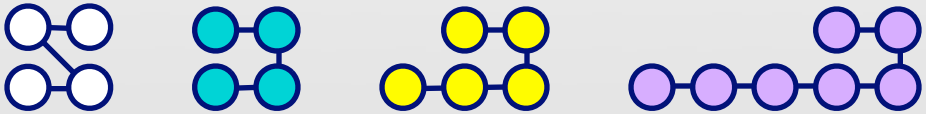
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	1	1	1	1	1
6	0	0	0	0	1	2	2	2	2
7	0	0	0	0	1	2	2	2	2
8	0	0	0	0	1	2	2	2	2
9	0	0	0	0	1	2	2	2	2

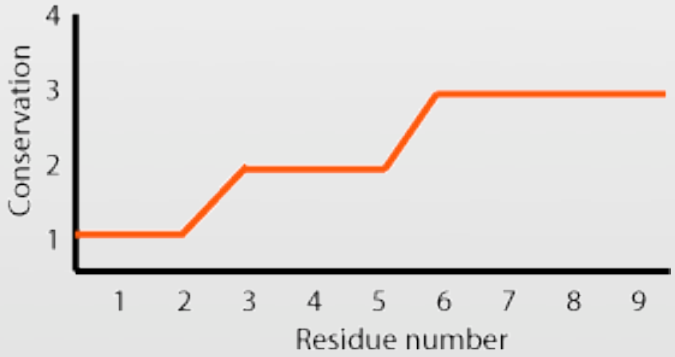


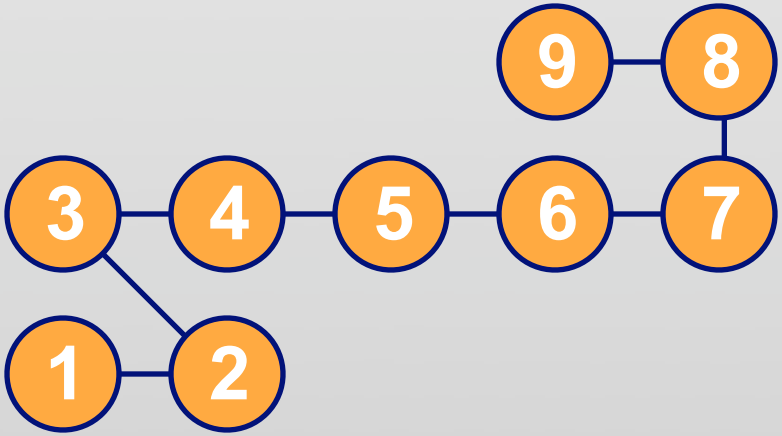




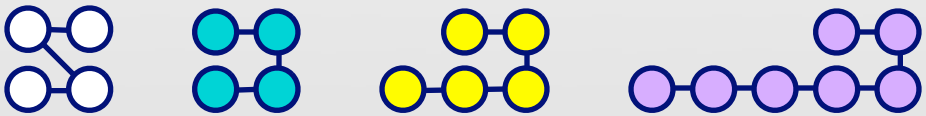
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3

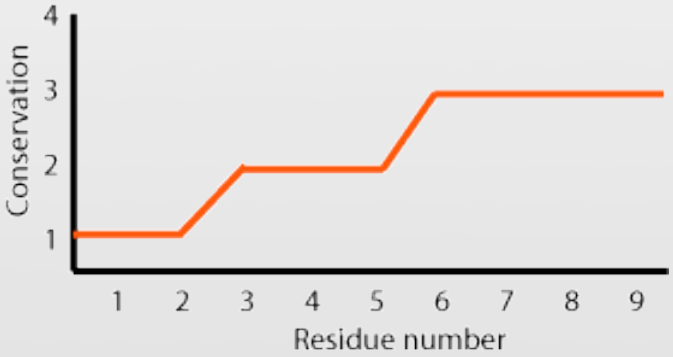


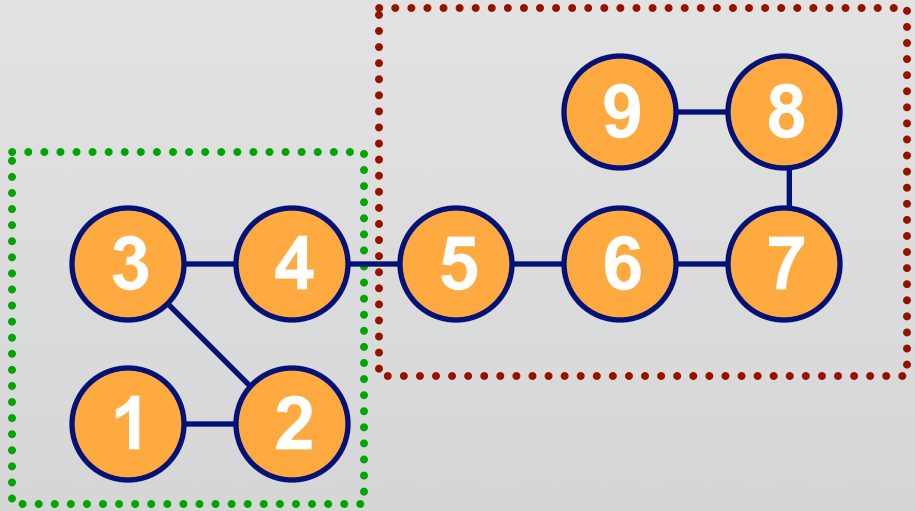




#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3







#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3

Threshold #3 MCL Cluster level (-I)

Stijn van Dongen (http://micans.org/mcl/)

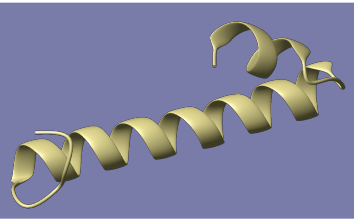
Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)

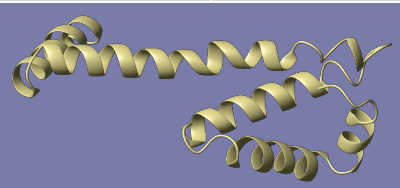
High P-values → fewer partitions

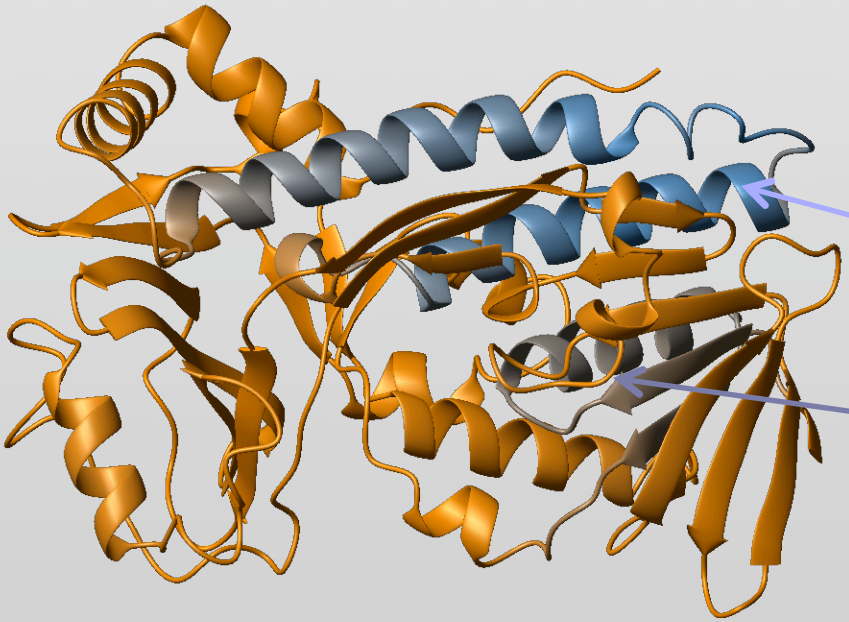
Threshold #3 → Cluster Level (-l)

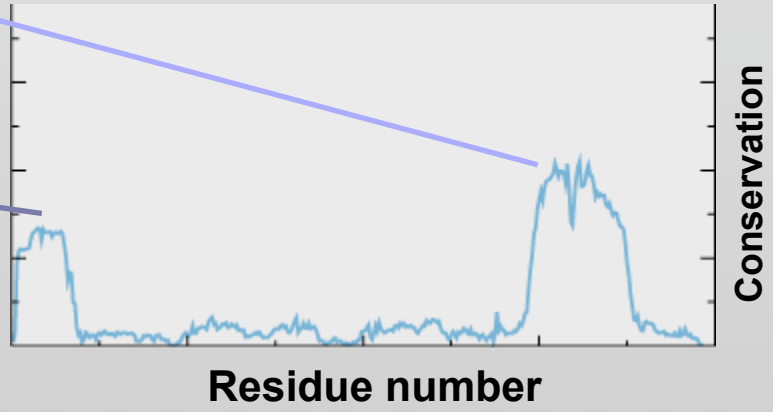
Low –l cluster value → fewer partitions

Applied to the ~40,000 chains in PDB (Dec 2002)

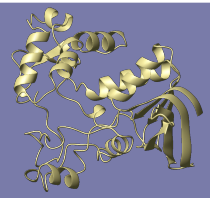
1phh	290-329	**2.7Å**	**3.1**
1hadB	72-111		

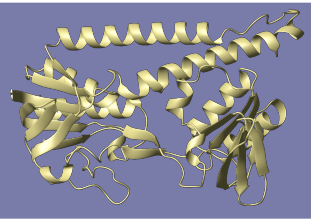
1phh	279-373	**3.9Å**	**4.7**
1bke	310-410		

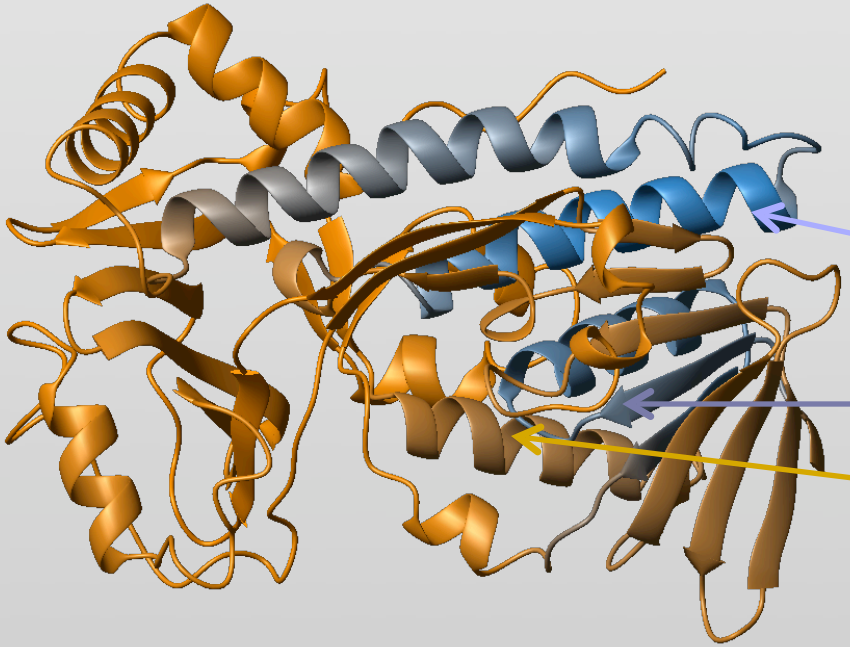


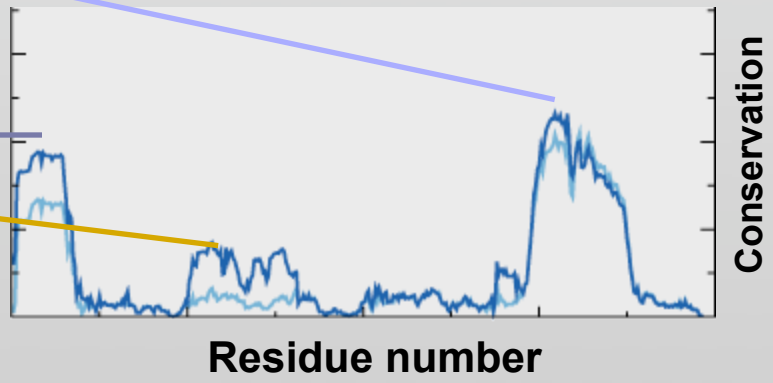


1phh (Oxydoreductase from *Pseudomonas fluorescens*)

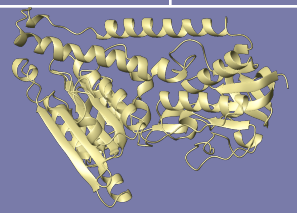
1phh	1-213	**3.0Å**	**8.1**
1qjdA	125-379		

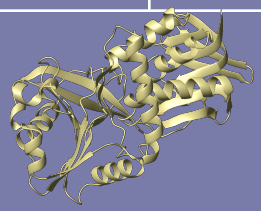
1phh	1-319	**3.6Å**	**9.8**
1gerA	3-327		

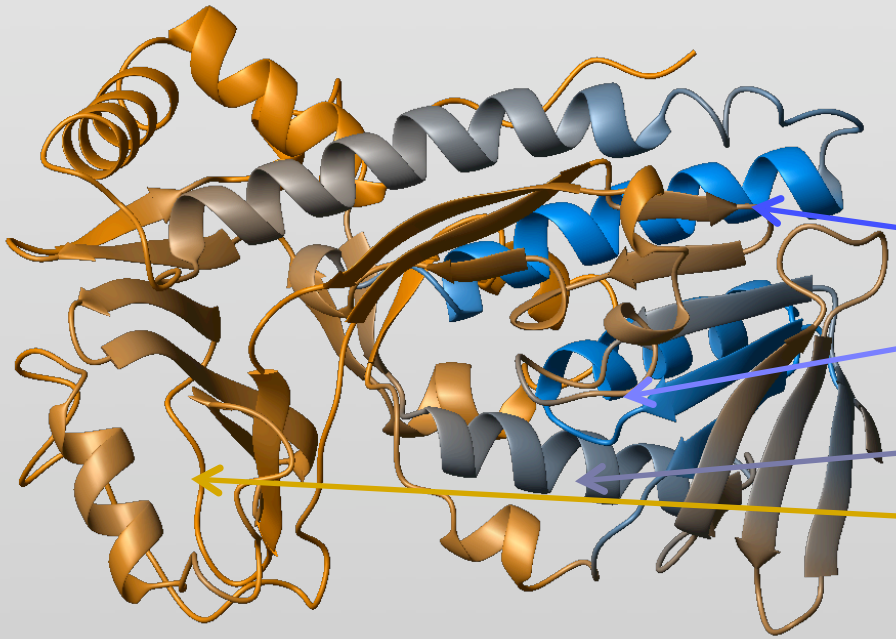


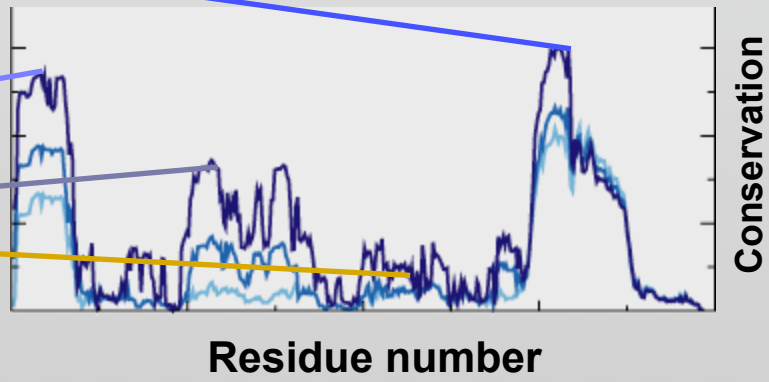


1phh (Oxydoreductase from *Pseudomonas fluorescens*)

1phh	1-378	**3.8Å**	**10.3**
1feaC	2-464		

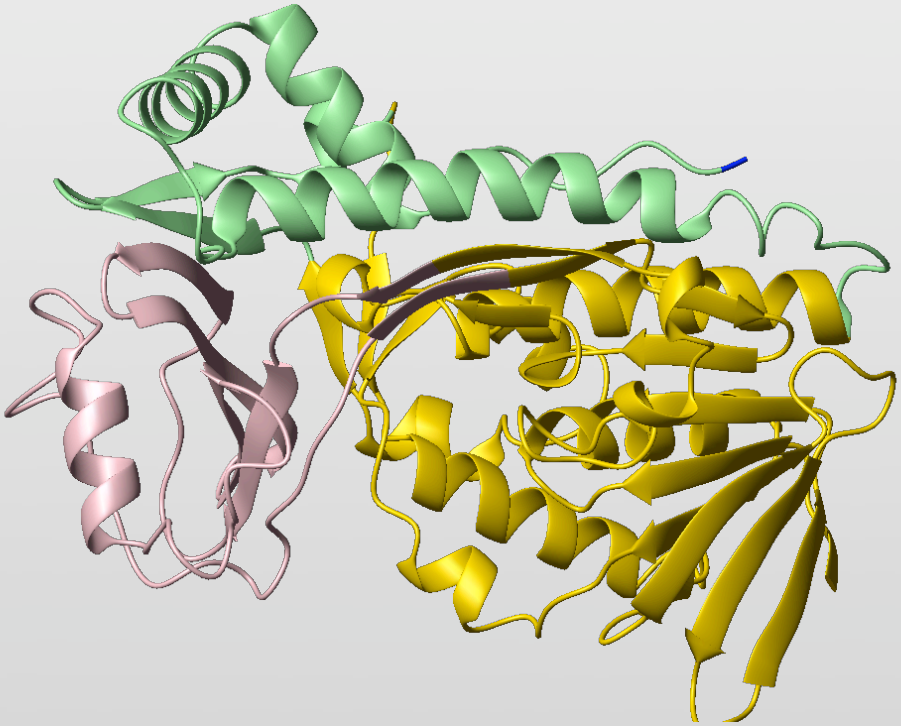
1phh	1-316	**3.8Å**	**17.2**
1l9dB	2-364		

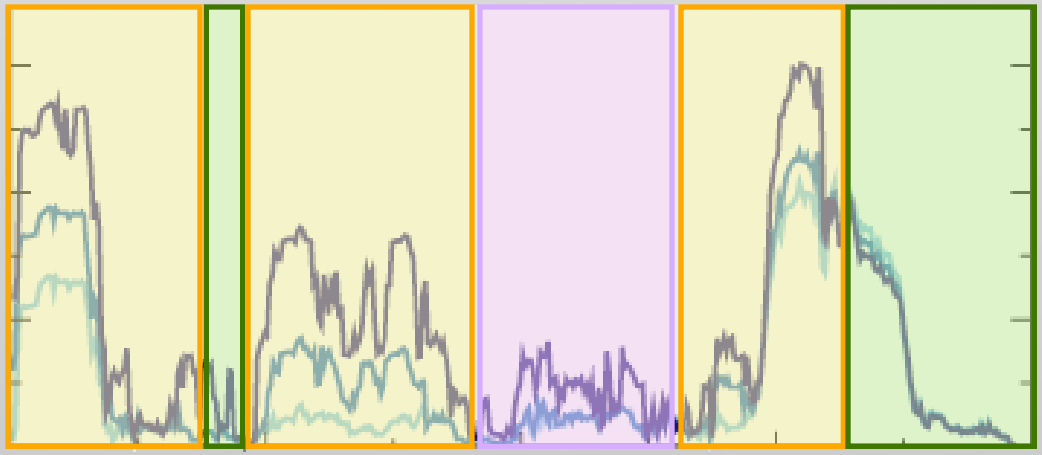




1phh (Oxydoreductase from *Pseudomonas fluorescens*)

1phh (Oxydoreductase from *Pseudomonas fluorescens*)





ONE DOMAIN: (30)

1aak	1tlk
1bbhA	1ula
1bbpA	1wsyA
1brd	2ace
1fxiA	2azaA
1gky	2ccyA
1gmpA	2gmfA
1gox	2rn2
1ofv	2stv
1pyp	2tmvP
1rbp	3chy
1rcb	3cla
1rveA	3drf
1snc	4blmA
1tie	5p21

TWO DOMAINS: (20)

Code	Authors	SCOP
1ezm	1-134 | 135-298	1-153 | 154-298
1fnb	19-161 | 162-314	Not defined
1gpb	19-489 | 490-841	One domain
1lap	1-150 | 171-484	1-159 | 160-484
1pfkA	0-138,251-301 | 139-250,302-319	One domain
1ppn	1-10,112-208 | 21-111,209-212	One domain
1rhd	1-158 | 159-293	1-149 | 150-293
1sgt	22-123,234,245 | 129,233	One domain
1vsgA	1-29,92-251 | 42-75,266-362	One domain
1wsyB	9-52,86-204 | 53-85,205-393	Not defined
2cyp	3-145,266-294 | 164-265	One domain
2had	1-155,230-310 | 156-229	One domain
3cd4	1-98 | 99-178	1-97 | 98-178
3gapA	1-129 | 139-208	Not defined
3pgk	1-185,403-415 | 200-392	One domain
4gcr	1-83 | 84-174	1-85 | 86-174
5fbpA	6-201 | 202-335	One domain
8adh	1-175,319-374 | 176-318	1-174,325-374 | 175-324
8atcA	1-137,288-310 | 144-283	1-150 | 151-310
8atcB	8-97 | 101-152	8-100 | 101-153

THREE DOMAINS: (2)

Code	Authors	SCOP
1phh	1-175 | 176-290 | 291-394	1-173,276-394 | 174-275
3grs	18-157,294-364 | 158-283 | 365-478	18-165,291-363 | 166-290 | 364-478

FOUR DOMAINS: (3)

Code	Authors	SCOP
1atnA	1-32,70-144,338-372 | 33-69 | 145-180,270-337 | 181-269	1-146 | 147-375
3pmgA	1-188 | 192-315 | 325-403 | 408-561	Not defined
8acn	2-200 | 201-317 | 320-513 | 538-754	2-528 | 529-754

Islam *et al.* Prot. Eng. (1995)

Lp: 3-6, Up: 4-30, I: 1.2-5

ONE DOMAIN: (30)
100%

1aak	1tlk
1bbhA	1ula
1bbpA	1wsyA
1brd	2ace
1fxiA	2azaA
1gky	2ccyA
1gmpA	2gmfA
1gox	2rn2
1ofv	2stv
1pyp	2tmvP
1rbp	3chy
1rcb	3cla
1rveA	3drf
1snc	4blmA
1tie	5p21

OVERALL:
(49/55 OK) 89.1%

Definition:
OK if Same # dom.
> 85% correct

TWO DOMAINS: (20) 80%

Code	Authors	Result
1ezm	1-134 | 135-298	76.09%
1fnb	19-161 | 162-314	86.78%
1gpb	19-489 | 490-841	84.31% → 3 domains
1lap	1-150 | 171-484	96.33%
1pfkA	0-138,251-301 | 139-250,302-319	97.80%
1ppn	1-10,112-208 | 21-111,209-212	93.53%
1rhd	1-158 | 159-293	99.32%
1sgt	22-123,234,245 | 129,233	87.34%
1vsgA	1-29,92-251 | 42-75,266-362	57.99% → 3 domains
1wsyB	9-52,86-204 | 53-85,205-393	88.28%
2cyp	3-145,266-294 | 164-265	87.91%
2had	1-155,230-310 | 156-229	93.20%
3cd4	1-98 | 99-178	100.0%
3gapA	1-129 | 139-208	96.97%
3pgk	1-185,403-415 | 200-392	96.92%
4gcr	1-83 | 84-174	100.0%
5fbpA	6-201 | 202-335	94.83%
8adh	1-175,319-374 | 176-318	77.48%
8atcA	1-137,288-310 | 144-283	97.99%
8atcB	8-97 | 101-152	100.0%

THREE DOMAINS: (2) 50%

Code	Authors	Result
1phh	1-175 | 176-290 | 291-394	82.70%
3grs	18-157,294-364 | 158-283 | 365-478	98.22%

FOUR DOMAINS: (3) 67%

Code	Authors	Result
1atnA	1-32,70-144,338-372 | 33-69 | 145-180,270-337 | 181-269	73.85% → 3 domains
3pmgA	1-188 | 192-315 | 325-403 | 408-561	93.75%
8acn	2-200 | 201-317 | 320-513 | 538-754	89.53%

for single values, e.g. Lp=3, Up=8, l=1.5

ONE DOMAIN: (30) 97%

1aak	1tlk
1bbhA	1ula
1bbpA	1wsyA
1brd	2ace
1fxiA	2azaA
1gky	2ccyA
1gmpA	2gmfA
1gox	2rn2
1ofv	2stv
1pyp	2tmvP
1rbp	3chy
1rcb	3cla
1rveA	3drf
1snc	4blmA
1tie	5p21

OVERALL:
(38/55 OK) 69.1%

Definition:
OK if Same # dom.
> 85% correct

TWO DOMAINS: (20) 30%

Code	Authors	Result
1ezm	1-134 | 135-298	45.12% → 1 domain
1fnb	19-161 | 162-314	51.53% → 1 domain
1gpb	19-489 | 490-841	84.31% → 3 domains
1lap	1-150 | 171-484	67.60% → 1 domain
1pfkA	0-138,251-301 | 139-250,302-319	97.17%
1ppn	1-10,112-208 | 21-111,209-212	53.23% → 1 domain
1rhd	1-158 | 159-293	97.95%
1sgt	22-123,234,245 | 129,233	45.41% → 1 domain
1vsgA	1-29,92-251 | 42-75,266-362	39.50%
1wsyB	9-52,86-204 | 53-85,205-393	81.68%
2cyp	3-145,266-294 | 164-265	87.91%
2had	1-155,230-310 | 156-229	76.05% → 1 domain
3cd4	1-98 | 99-178	98.87%
3gapA	1-129 | 139-208	65.15% → 1 domain
3pgk	1-185,403-415 | 200-392	96.92%
4gcr	1-83 | 84-174	52.02% → 1 domain
5fbpA	6-201 | 202-335	94.22%
8adh	1-175,319-374 | 176-318	61.39% → 1 domain
8atcA	1-137,288-310 | 144-283	53.18% → 1 domain
8atcB	8-97 | 101-152	63.83% → 1 domain

THREE DOMAINS: (2) 50%

Code	Authors	Result
1phh	1-175 | 176-290 | 291-394	44.53% → 1 domain
3grs	18-157,294-364 | 158-283 | 365-478	96.67%

FOUR DOMAINS: (3) 67%

Code	Authors	Result
1atnA	1-32,70-144,338-372 | 33-69 | 145-180,270-337 | 181-269	55.80% → 2 domains
3pmgA	1-188 | 192-315 | 325-403 | 408-561	90.62%
8acn	2-200 | 201-317 | 320-513 | 538-754	89.39%

Values Employed…

SINGLE VALUES

pI=3 pM=8 I=1.5

pI=3 pM=16 I=1.5

pI=3 pM=4 I=1.5

pI=3 pM=14 I=1.5

pI=3 pM=12 I=1.5

All these combinations resulted in 69% accuracy (38/55).

VALUES BASED ON LENGTH

Length: <=250

pI=3 pM=8 I=2.0

Length: >250 and <=300

pI=3 pM=8 I=1.5

Length: >300 and <=350

pI=4 pM=8 I=2.0

Length: >350 and <=400

pI=3 pM=4 I=1.5

Length: >400

pI=3 pM=6 I=1.5

This algorithm resulted in 80% accuracy (44/55).

Results (for values based on length)

ONE DOMAIN: (30) 97%

1aak	1tlk
1bbhA	1ula
1bbpA	1wsyA
1brd	2ace
1fxiA	2azaA
1gky	2ccyA
1gmpA	2gmfA
1gox	2rn2
1ofv	2stv
1pyp	2tmvP
1rbp	3chy
1rcb	3cla
1rveA	3drf
1snc	4blmA
1tie	5p21

OVERALL:

(44/55 OK) 80%

Definition:

OK if Same # dom.

> 85% correct

TWO DOMAINS: (20) 60%

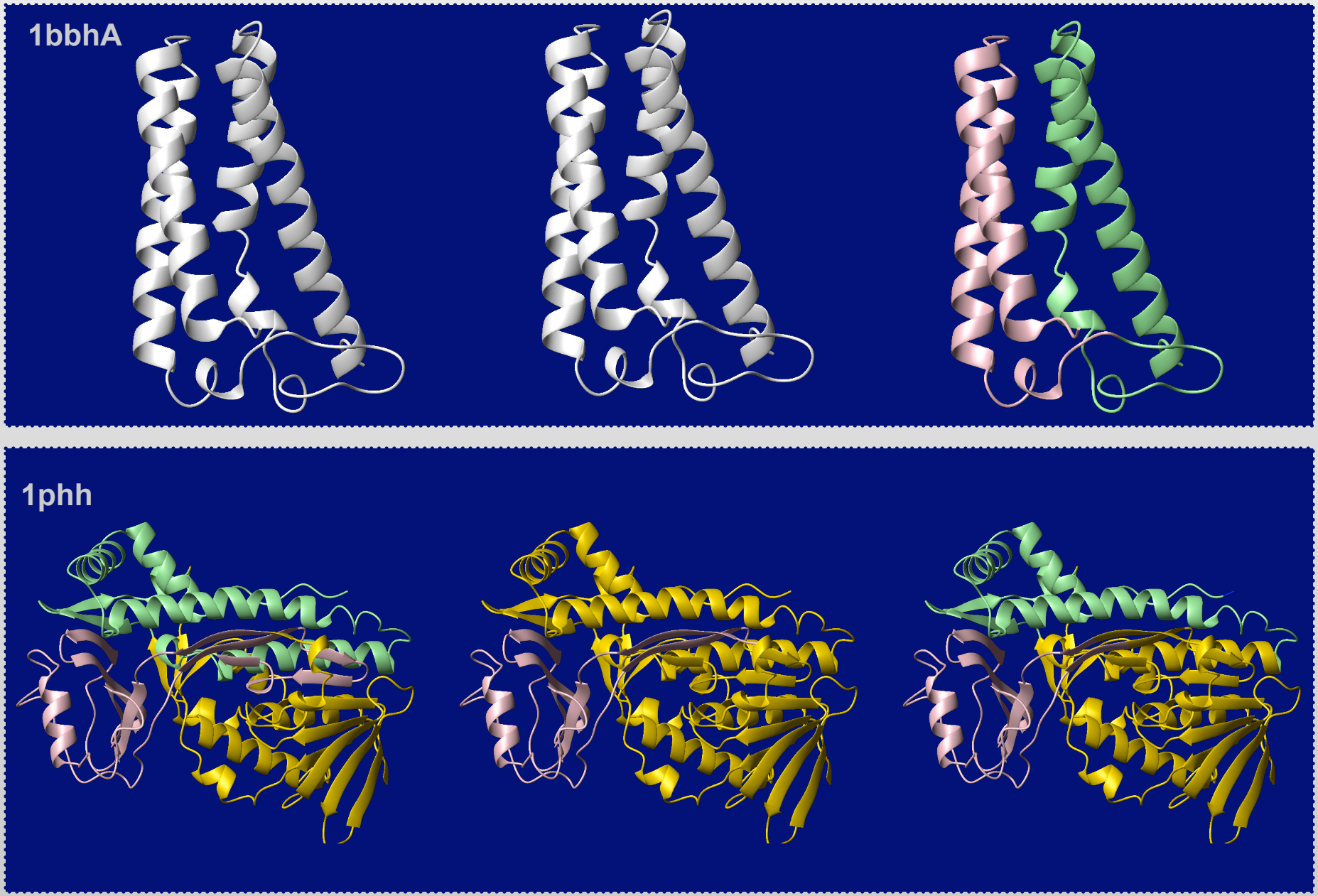
Code	Authors	Result
1ezm	1-134 | 135-298	45.12% → 1 domain
1fnb	19-161 | 162-314	51.53% → 1 domain
1gpb	19-489 | 490-841	82.00% → 3 domains
1lap	1-150 | 171-484	67.60% → 1 domain
1pfkA	0-138,251-301 | 139-250,302-319	96.23%
1ppn	1-10,112-208 | 21-111,209-212	90.55%
1rhd	1-158 | 159-293	97.95%
1sgt	22-123,234,245 | 129,233	86.46%
1vsgA	1-29,92-251 | 42-75,266-362	52.66% → 3 domains
1wsyB	9-52,86-204 | 53-85,205-393	88.28%
2cyp	3-145,266-294 | 164-265	87.91%
2had	1-155,230-310 | 156-229	76.05% → 1 domain
3cd4	1-98 | 99-178	98.87%
3gapA	1-129 | 139-208	74.24%
3pgk	1-185,403-415 | 200-392	96.92%
4gcr	1-83 | 84-174	95.95%
5fbpA	6-201 | 202-335	93.92%
8adh	1-175,319-374 | 176-318	73.19%
8atcA	1-137,288-310 | 144-283	96.99%
8atcB	8-97 | 101-152	97.87%

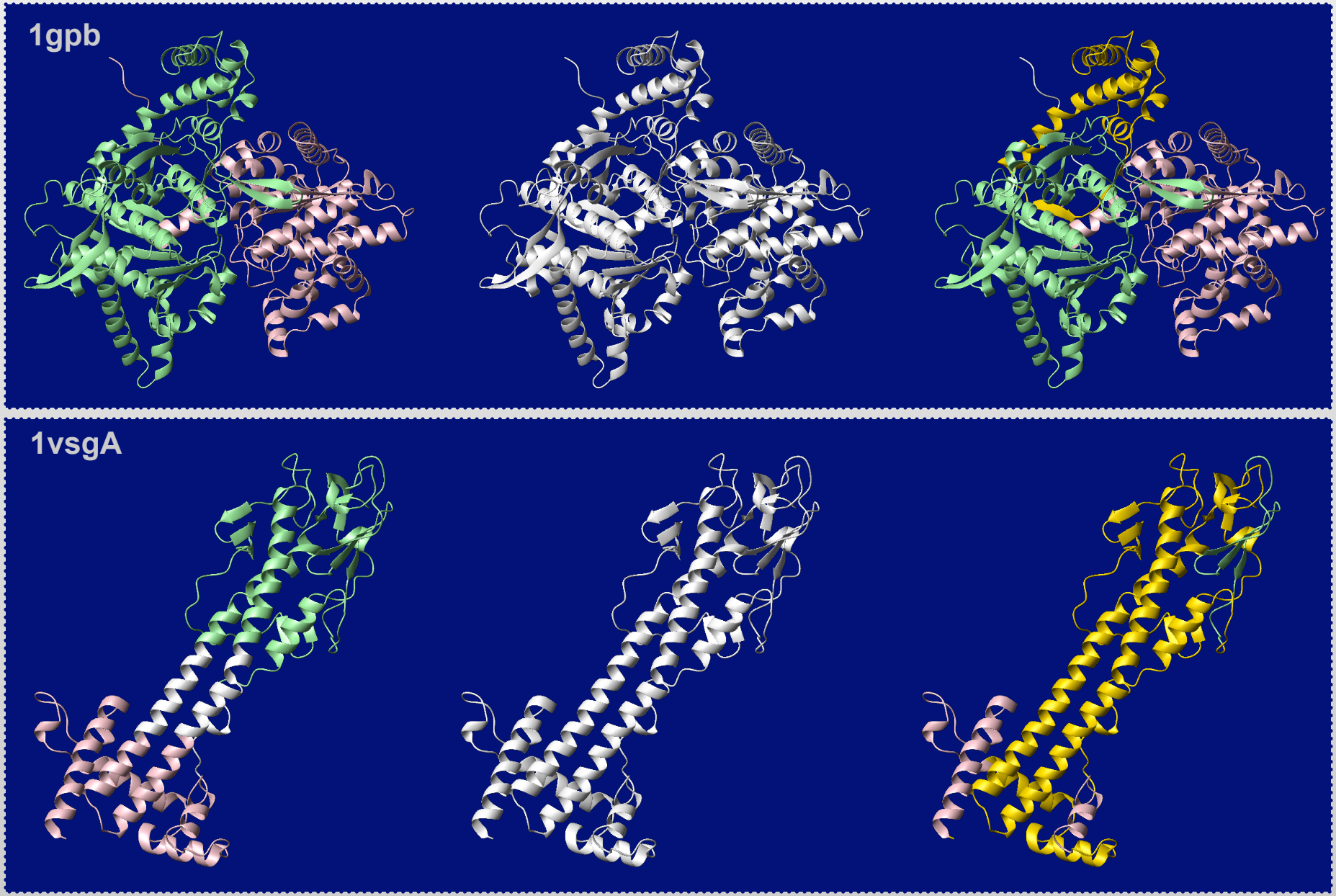
THREE DOMAINS: (2) 50%

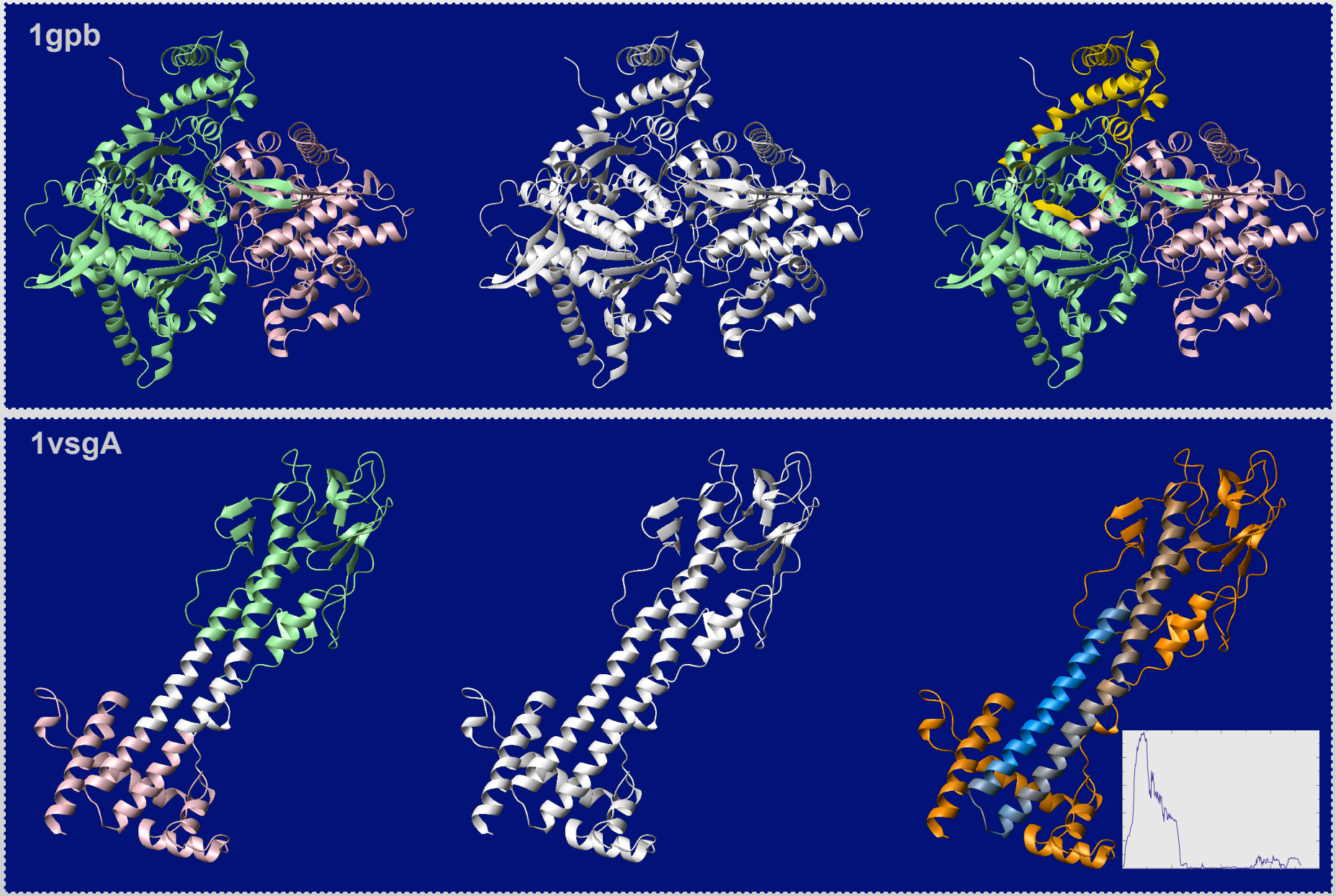
Code	Authors	Result
1phh	1-175 | 176-290 | 291-394	44.53% → 1 domain
3grs	18-157,294-364 | 158-283 | 365-478	93.78%

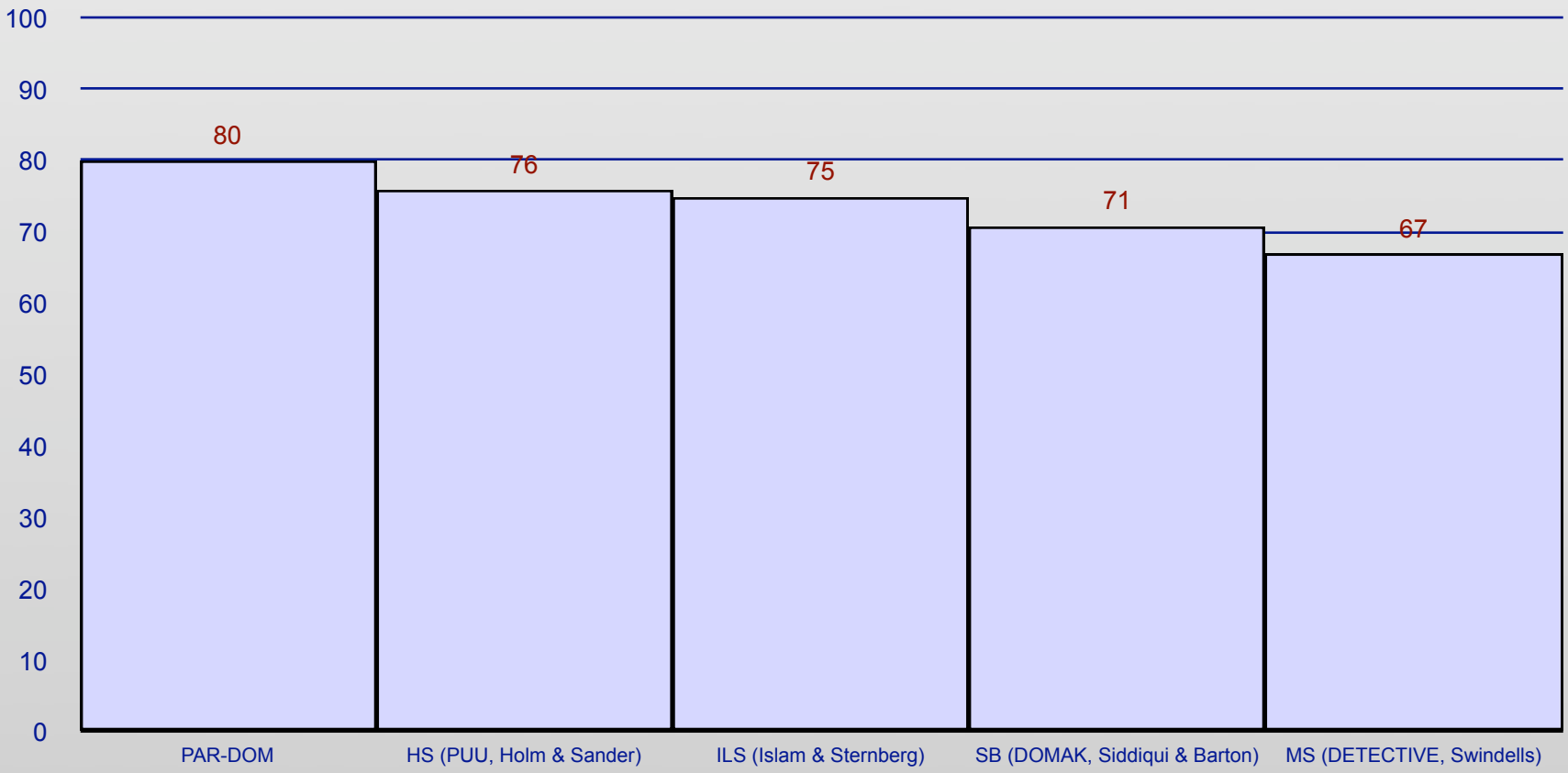
FOUR DOMAINS: (3) 67%

Code	Authors	Result
1atnA	1-32,70-144,338-372 | 33-69 | 145-180,270-337 | 181-269	73.32% → 3 domains
3pmgA	1-188 | 192-315 | 325-403 | 408-561	90.62%
8acn	2-200 | 201-317 | 320-513 | 538-754	89.39%

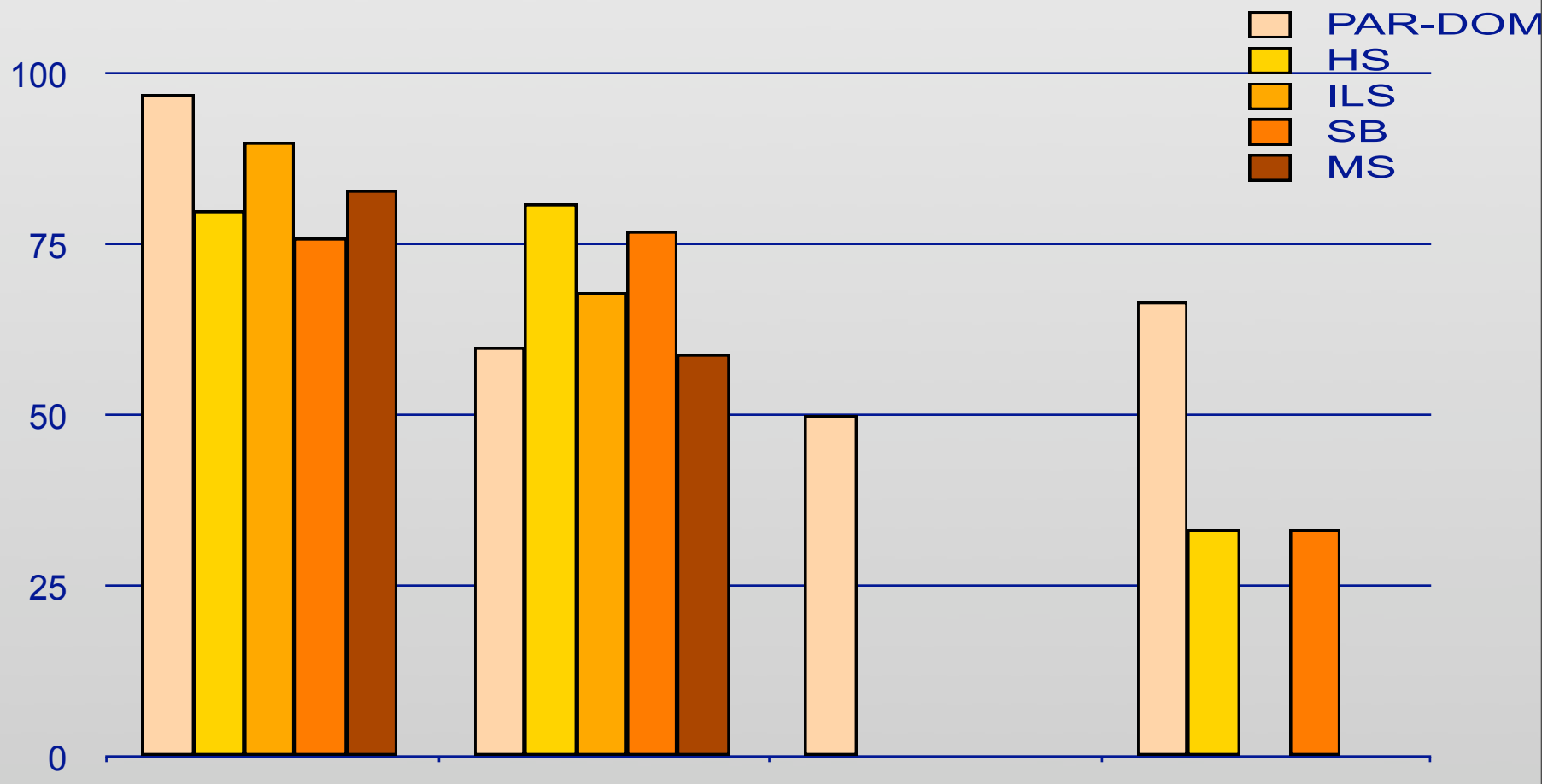

Authors
SCOP
PAR-DOM
1bbhA
1phh


Authors
SCOP
PAR-DOM
1gpb
1vsgA


Authors
SCOP
PAR-DOM
1gpb
1vsgA



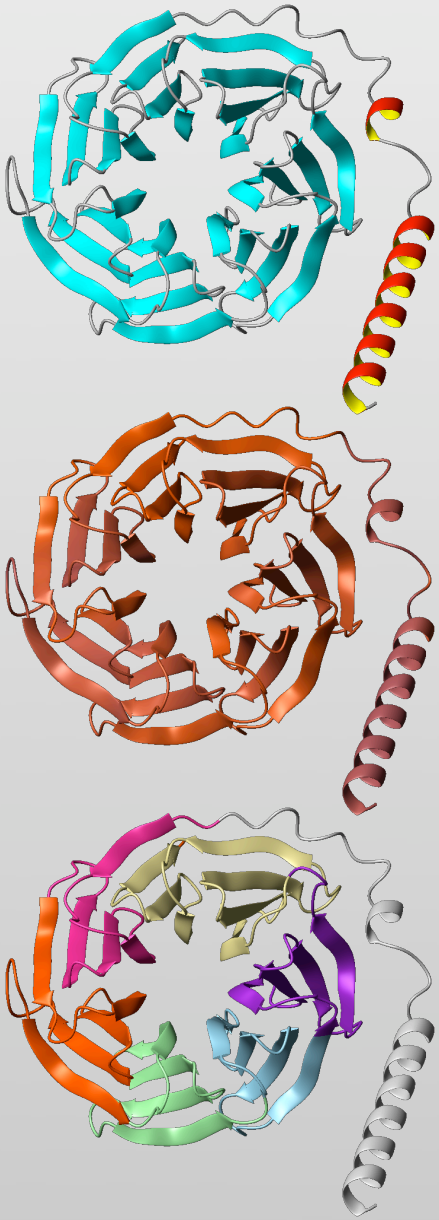
Jones *et al.* Prot. Sci. (1998)

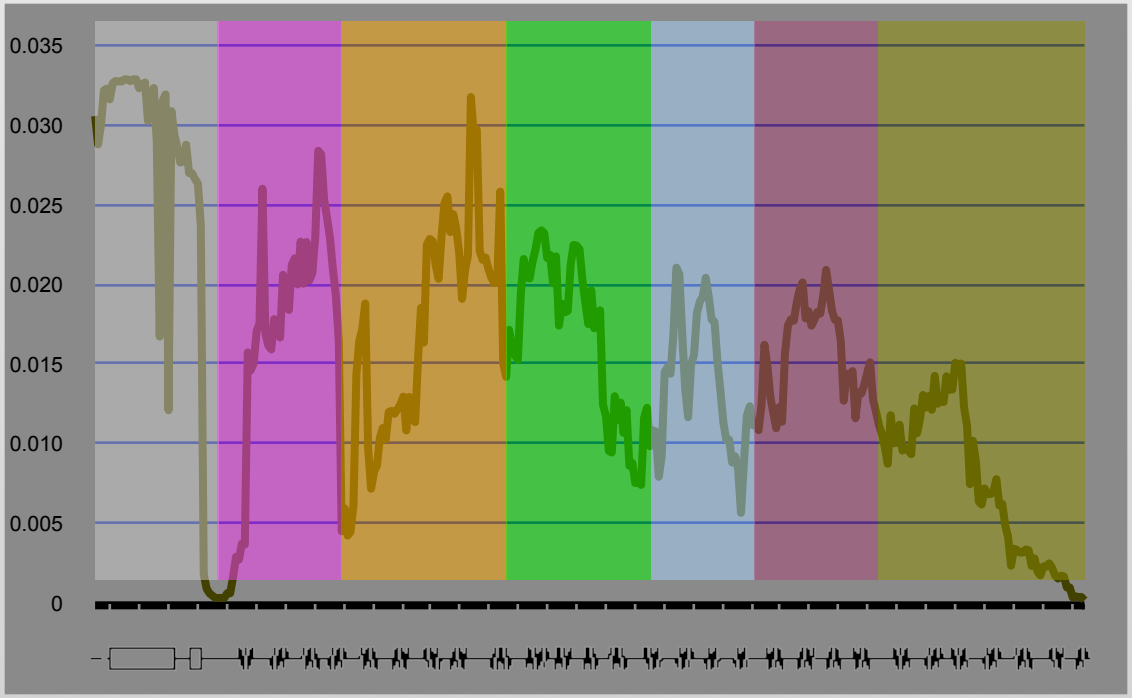


Jones *et al.* Prot. Sci. (1998)

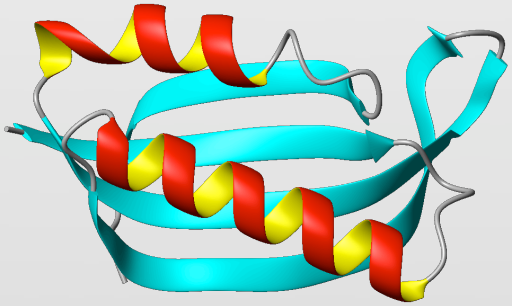
Domains → Fragments

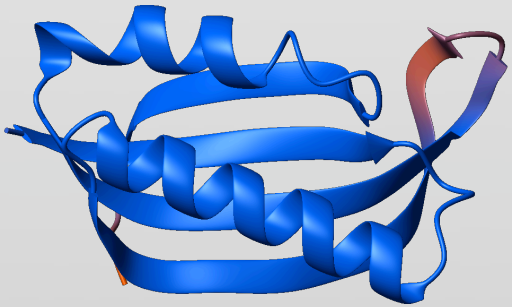
G-protein (1gotB) *all-β* → *7 bladed beta propeller domain*

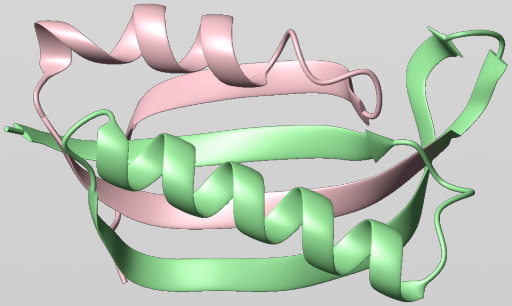


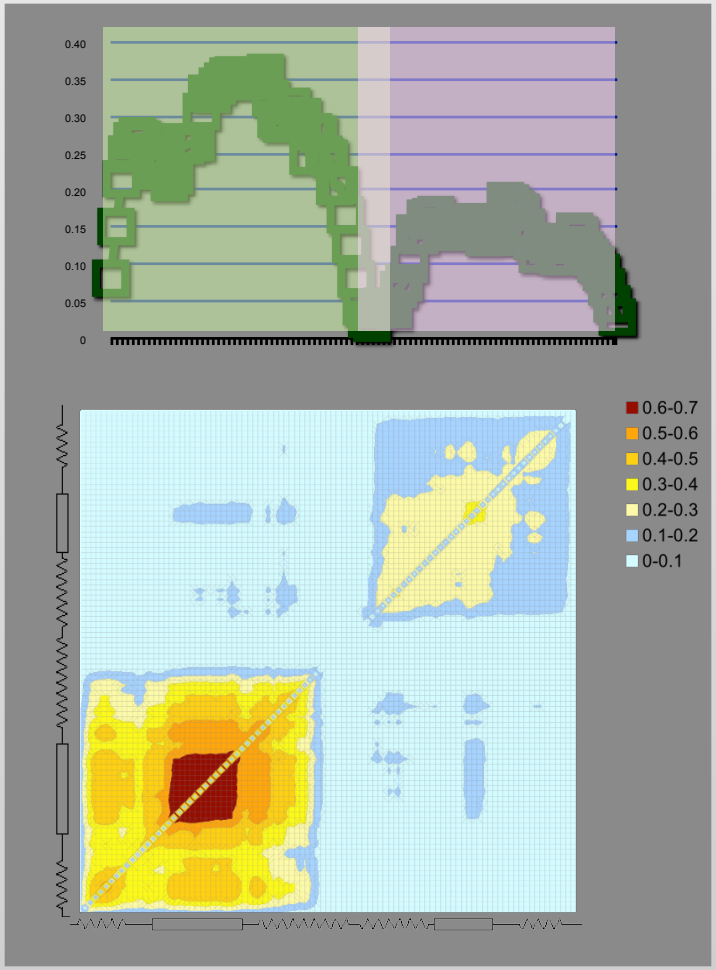


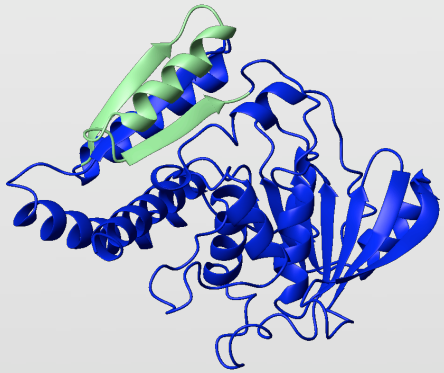
Ribosomal protein S6 (1ris) $\alpha+\beta$ → *Ferrodoxin Like domain*



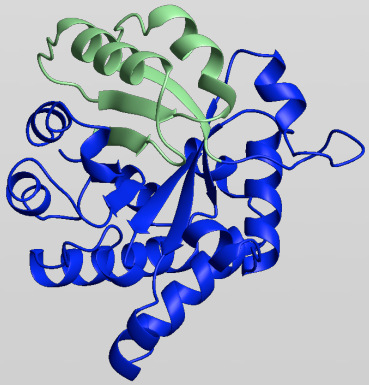






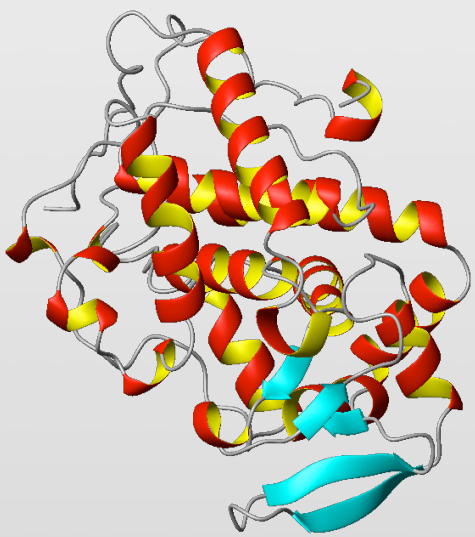


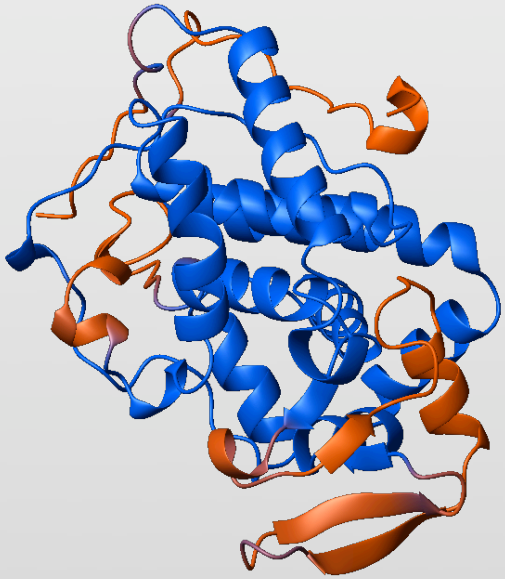
1ee9A 17.9% id. 2.3Å

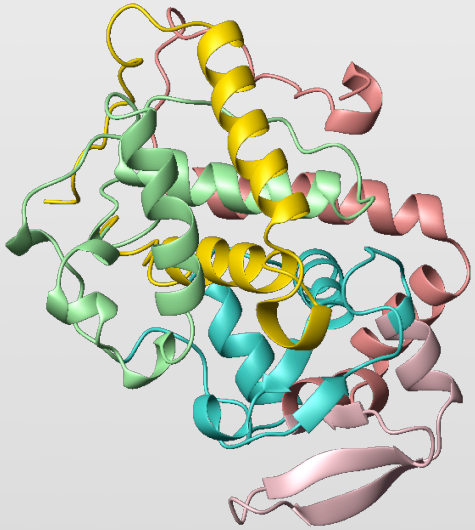


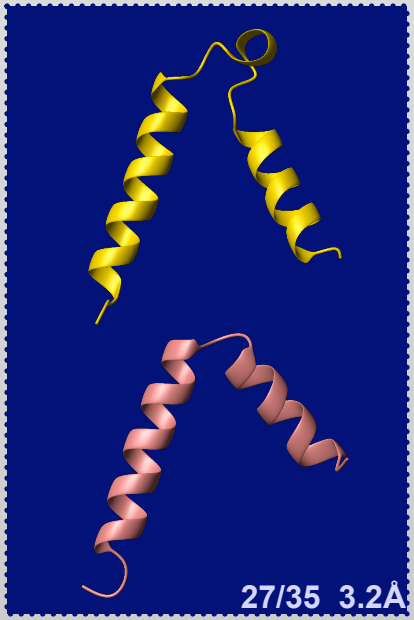
6timB 11.1% id. 2.6Å

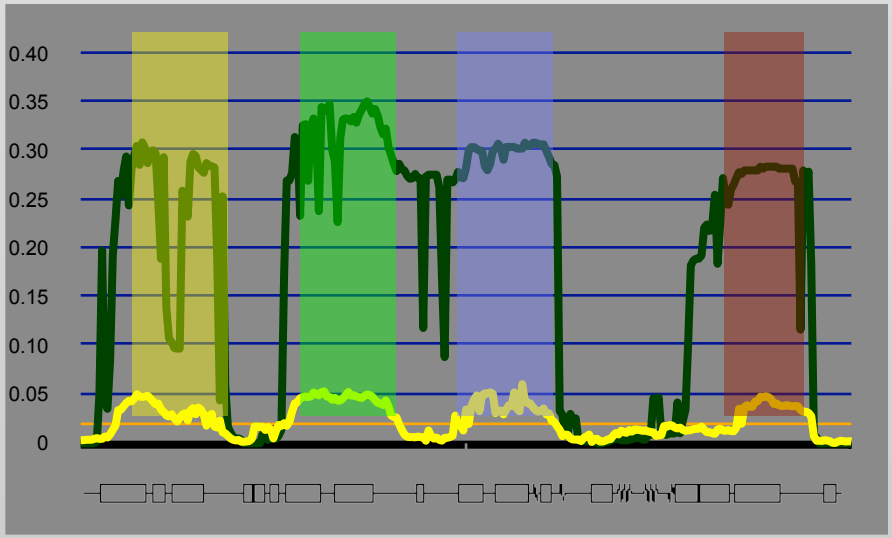
Cytochrome C Peroxidase (2cyp) *all-α → CCP-like domain*

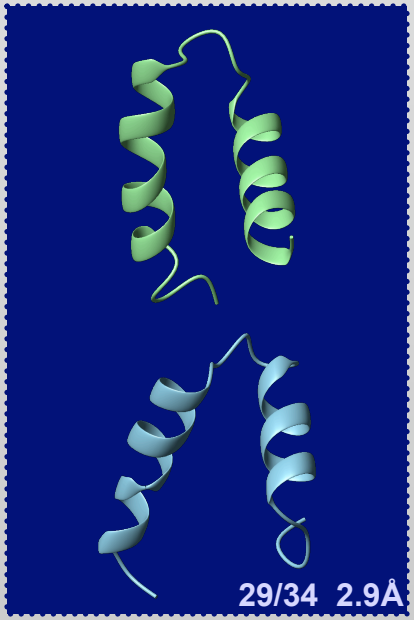












- ✓ **PAR-DOM**
- ✓ **Domains ←→ Fragments**

Acknowledgments

The Rockefeller University Presidential Fellowship

Andrej Sali
David Katz
Angel Ortiz

Frank Alber
Fred Davis
Damien Devos
Narayanan Eswar
Dmitry Korkin
M. S. Madhusudhan
Ursula Pieper
Andrea Rossi
Min-yi Shen
Maya Topf

http://www.salilab.org