MASTER IN TECNOLOGIE BIOINFORMATICHE APPLICATE ALLA MEDICINA PERSONALIZZATA

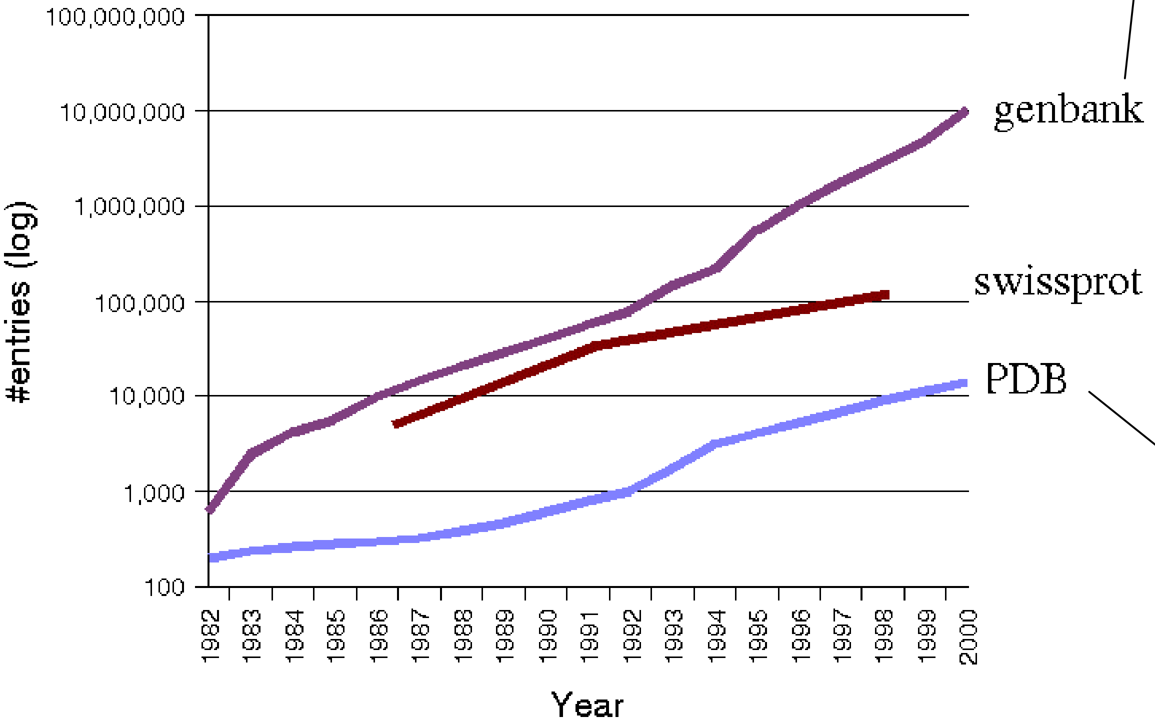
Protein Sequence Analysis

Extraction of Functional Features from Sequence Alignments

Florencio Pazos (CNB-CSIC)

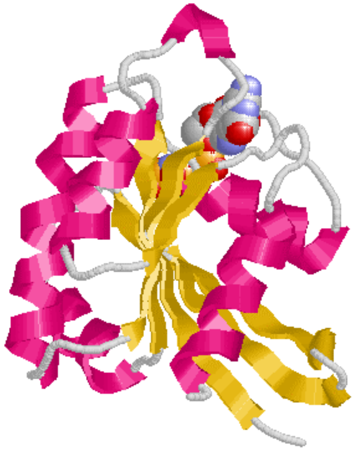
Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es

Interpreting high-throughput data in functional terms

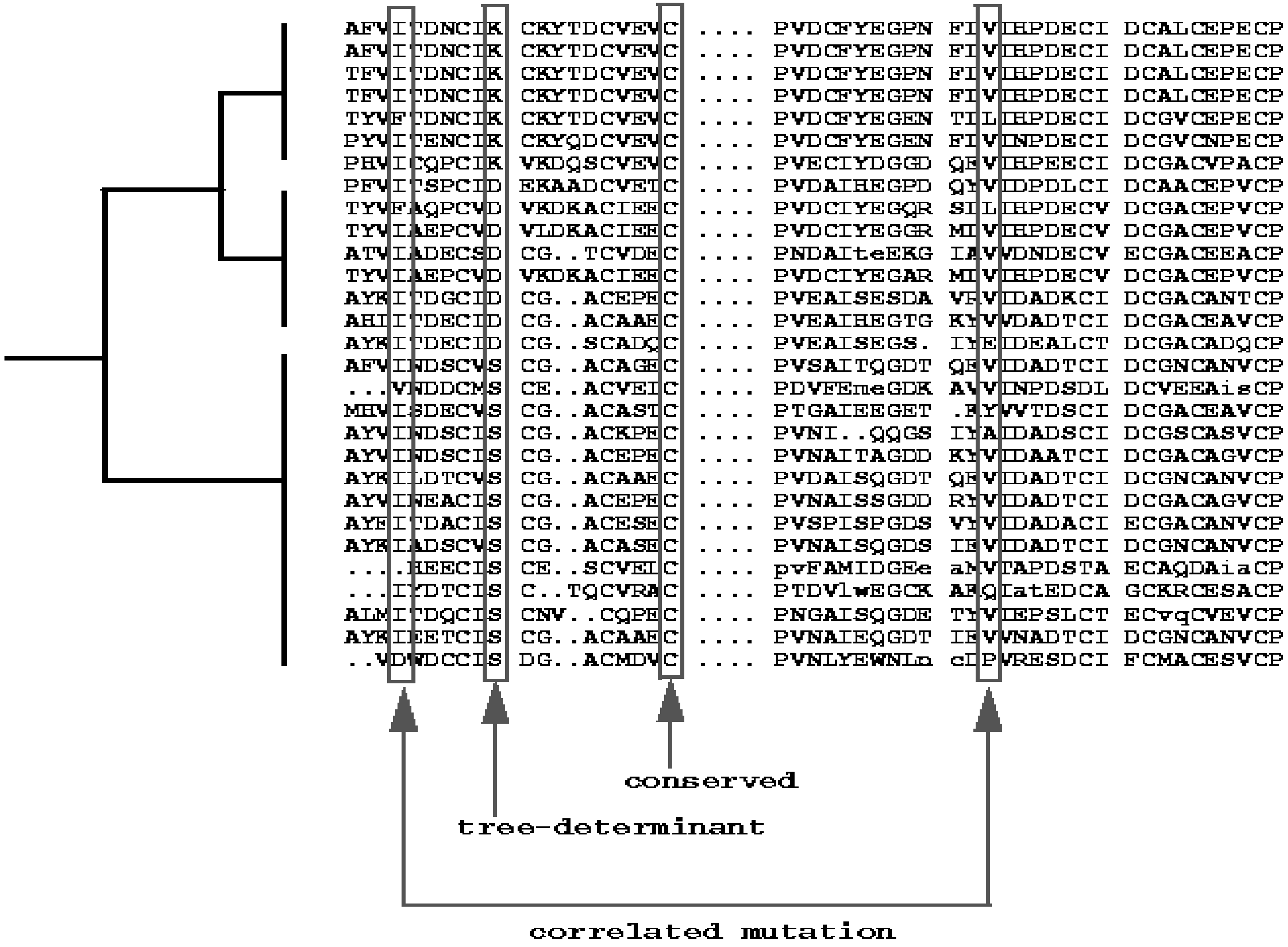


MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

MLEILDTA**GTEQFT**AMRDLYMKNGQGFAL
VYSITAQSTFND**LQDLREQIL**RVKDTEDVPMIL
VGNKCDLEDERV



Prediction of functional regions
Sequence-based methods



Devos, D., Merino, E., Pazos, F. and Valencia, A. (2002) Multiple sequence alignments information in structure and function prediction. In IOS Press, *Artificial Intelligence and Heuristic Methods for Bioinformatics*, pp. 83-94.

Pazos, F. and Bang, J.-W. (2006) Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics*, **1**, 15-23.

Prediction of interaction regions

Residue propensities

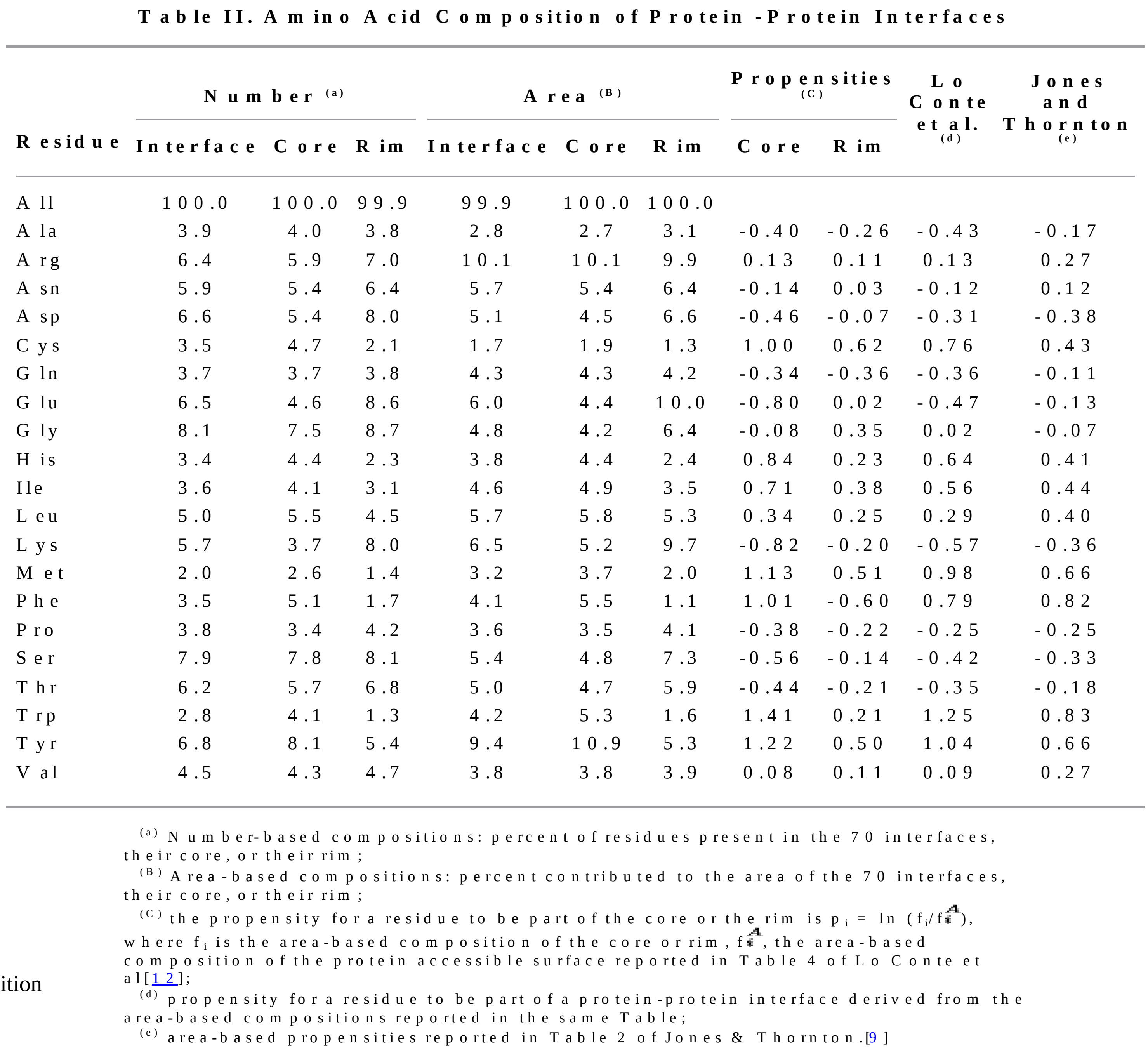


Table II. Amino Acid Composition of Protein-Protein Interfaces

Residue	Number [a]			Area [B]			Propensities [C]		Lo Conte et al. [d]	Jones and Thornton [e]
	Interface	Core	Rim	Interface	Core	Rim	Core	Rim		
All	100.0	100.0	99.9	99.9	100.0	100.0				
Ala	3.9	4.0	3.8	2.8	2.7	3.1	-0.40	-0.26	-0.43	-0.17
Arg	6.4	5.9	7.0	10.1	10.1	9.9	0.13	0.11	0.13	0.27
Asn	5.9	5.4	6.4	5.7	5.4	6.4	-0.14	0.03	-0.12	0.12
Asp	6.6	5.4	8.0	5.1	4.5	6.6	-0.46	-0.07	-0.31	-0.38
Cys	3.5	4.7	2.1	1.7	1.9	1.3	1.00	0.62	0.76	0.43
Gln	3.7	3.7	3.8	4.3	4.3	4.2	-0.34	-0.36	-0.36	-0.11
Glu	6.5	4.6	8.6	6.0	4.4	10.0	-0.80	0.02	-0.47	-0.13
Gly	8.1	7.5	8.7	4.8	4.2	6.4	-0.08	0.35	0.02	-0.07
His	3.4	4.4	2.3	3.8	4.4	2.4	0.84	0.23	0.64	0.41
Ile	3.6	4.1	3.1	4.6	4.9	3.5	0.71	0.38	0.56	0.44
Leu	5.0	5.5	4.5	5.7	5.8	5.3	0.34	0.25	0.29	0.40
Lys	5.7	3.7	8.0	6.5	5.2	9.7	-0.82	-0.20	-0.57	-0.36
Met	2.0	2.6	1.4	3.2	3.7	2.0	1.13	0.51	0.98	0.66
Phe	3.5	5.1	1.7	4.1	5.5	1.1	1.01	-0.60	0.79	0.82
Pro	3.8	3.4	4.2	3.6	3.5	4.1	-0.38	-0.22	-0.25	-0.25
Ser	7.9	7.8	8.1	5.4	4.8	7.3	-0.56	-0.14	-0.42	-0.33
Thr	6.2	5.7	6.8	5.0	4.7	5.9	-0.44	-0.21	-0.35	-0.18
Trp	2.8	4.1	1.3	4.2	5.3	1.6	1.41	0.21	1.25	0.83
Tyr	6.8	8.1	5.4	9.4	10.9	5.3	1.22	0.50	1.04	0.66
Val	4.5	4.3	4.7	3.8	3.8	3.9	0.08	0.11	0.09	0.27

[a] Number-based compositions: percent of residues present in the 70 interfaces, their core, or their rim;

[B] Area-based compositions: percent contributed to the area of the 70 interfaces, their core, or their rim;

[C] the propensity for a residue to be part of the core or the rim is $p_i = \ln(f_i/f_i^{A})$, where f_i is the area-based composition of the core or rim, f_i^{A}, the area-based composition of the protein accessible surface reported in Table 4 of Lo Conte et al[12];

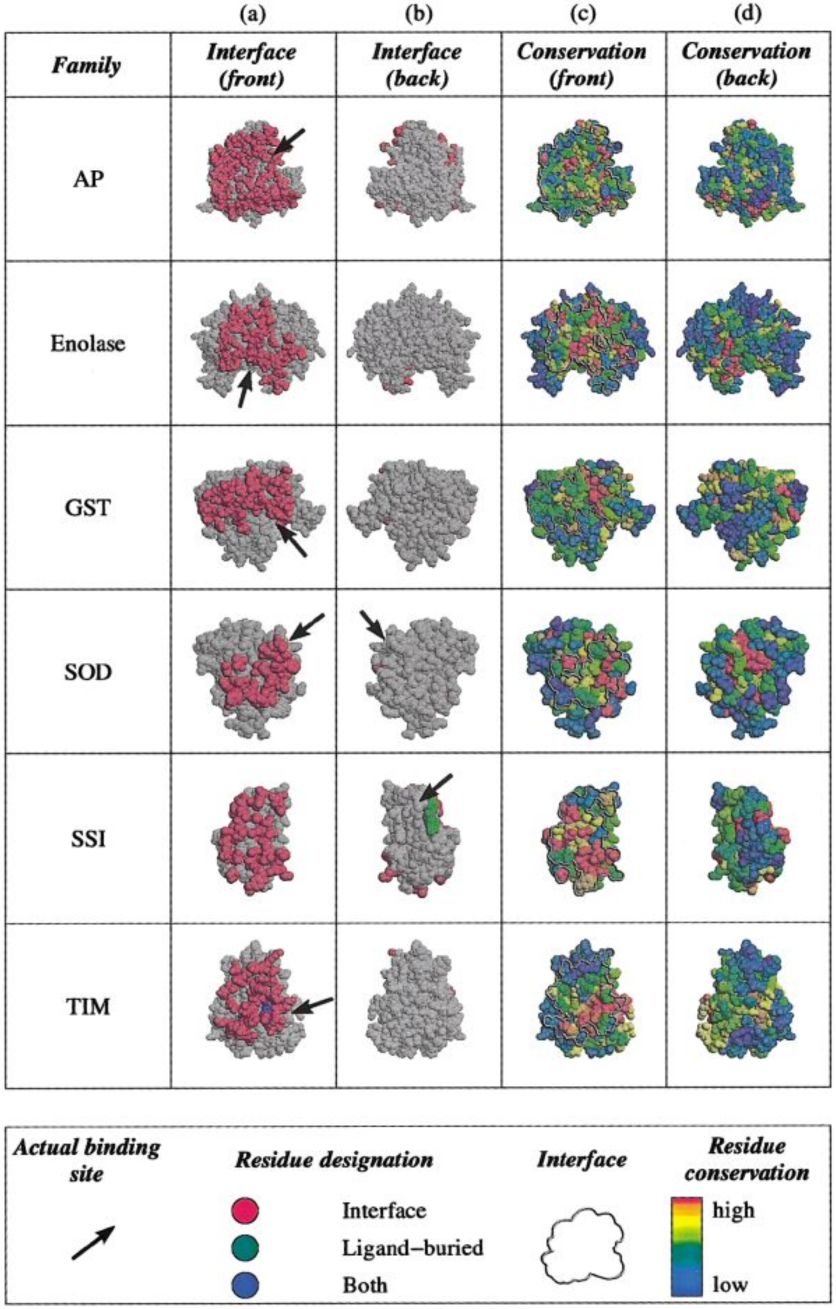
[d] propensity for a residue to be part of a protein-protein interface derived from the area-based compositions reported in the same Table;

[e] area-based propensities reported in Table 2 of Jones & Thornton.[9]

Chakrabarti, P., Janin, J.
Dissecting protein-protein recognition sites.
Proteins 2002 **47:**334-343

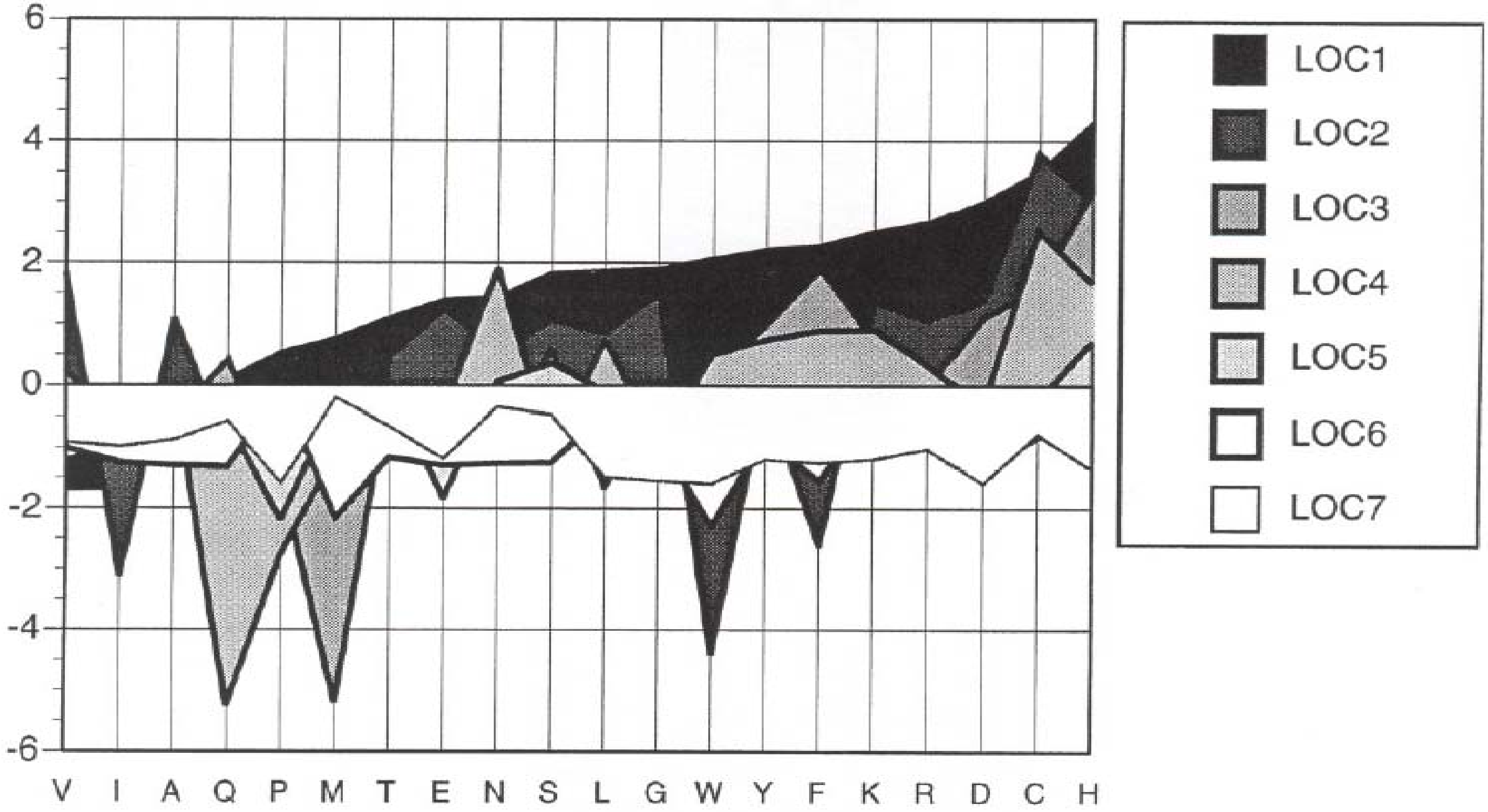
Prediction of interaction regions
Sequence-based methods

Conserved positions



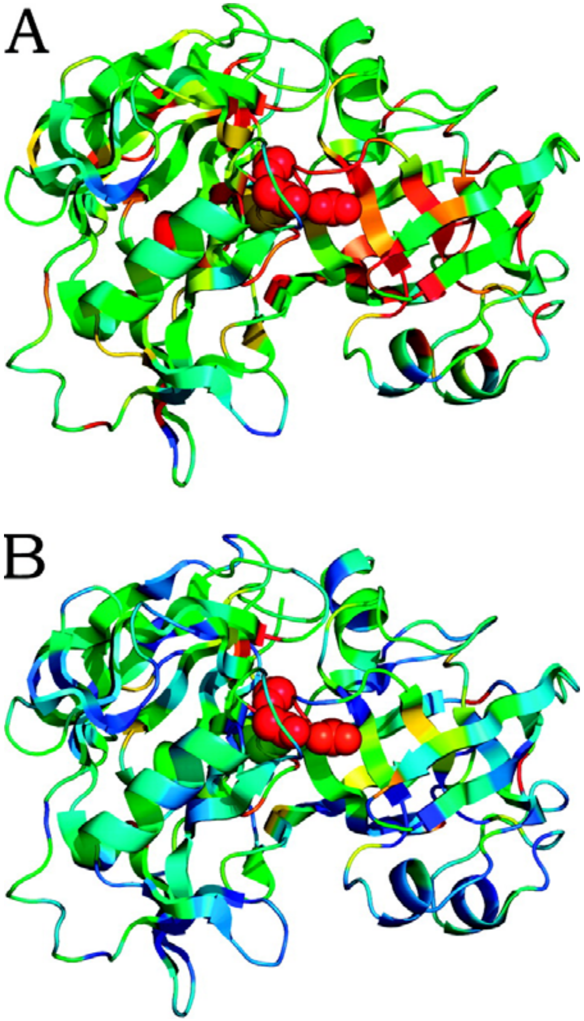
Valdar, W.S., Thornton, J.M.
Protein-protein interfaces: analysis of amino acid conservation in homodimers.
Proteins 2001 **42:**108-124

Are conserved residues always functional?
Are functional residues always conserved?



Ouzounis, C., Perez-Irratxeta, C., Sander, C. and Valencia, A. (1998) Are binding residues conserved? *Pac Symp Biocomput.*, 401-412.

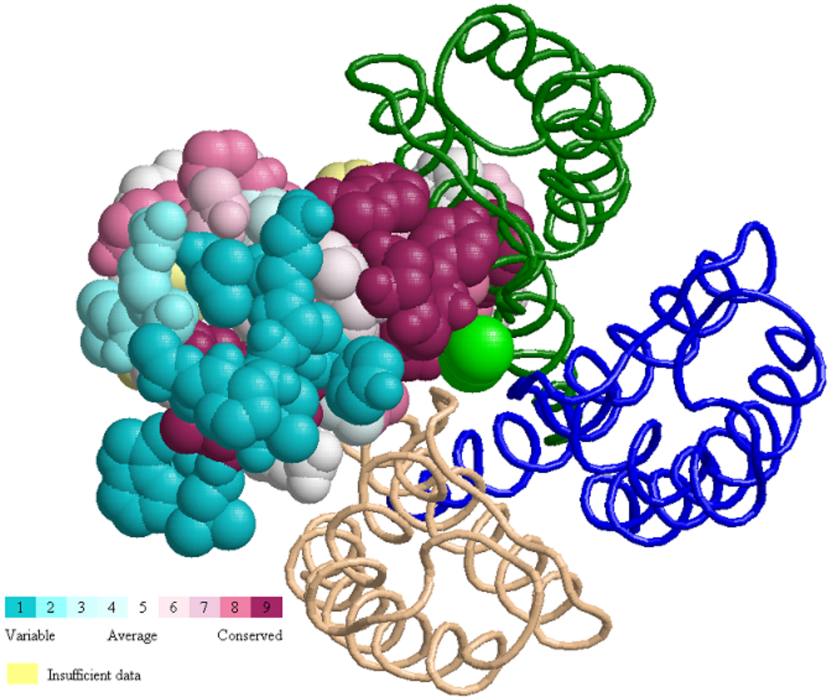
Structural vs. Functional Conservation



Cheng, G., Qian, B., Samudrala, R. and Baker, D. (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, **33**, 5861-5867.

Conserved positions

(not trivial)

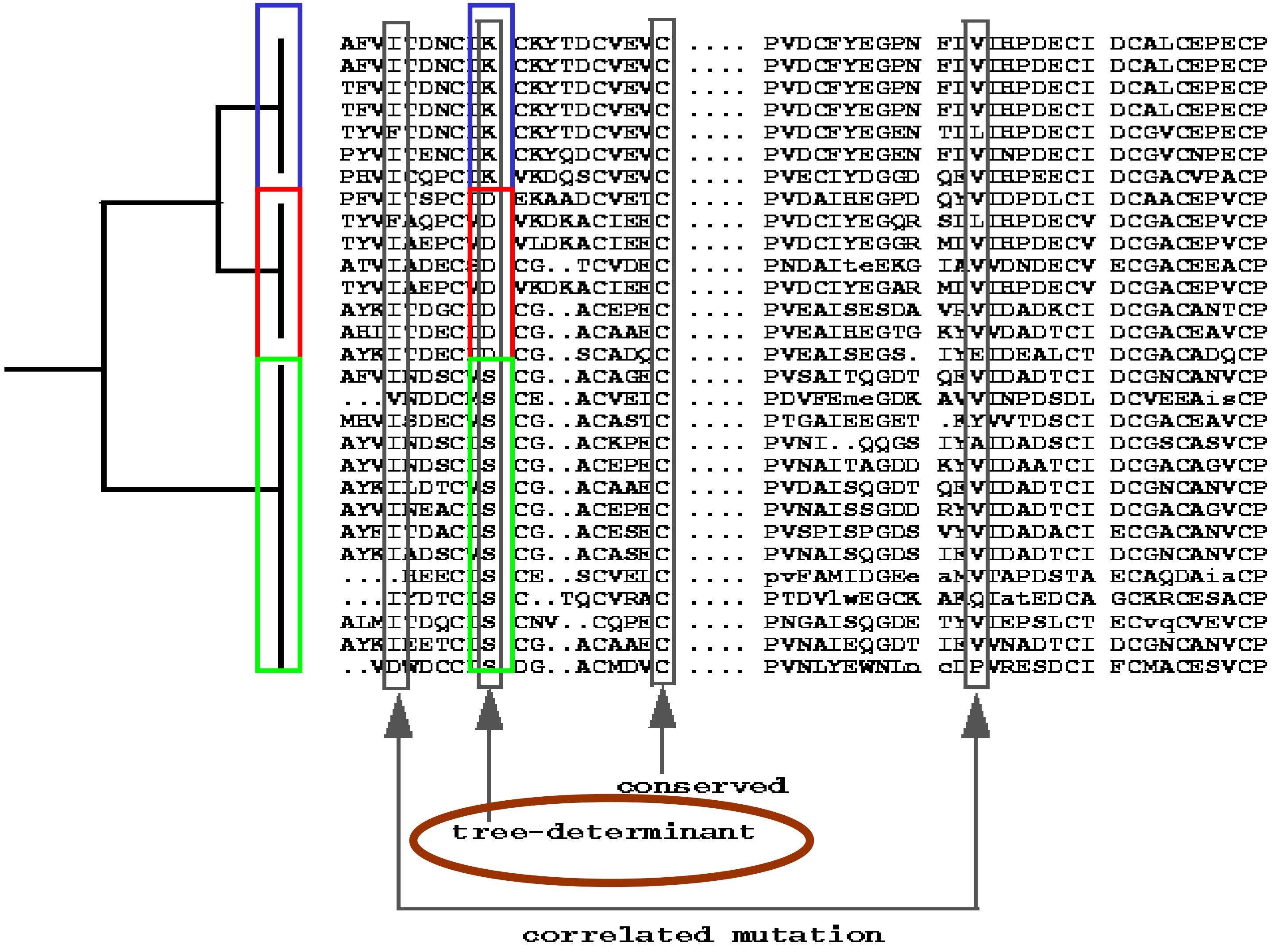


Valdar, W.S. (2002) Scoring residue conservation. *Proteins*, **48,** 227-241.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18,** S71-S77.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R. and Ben-Tal, N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322-1324.

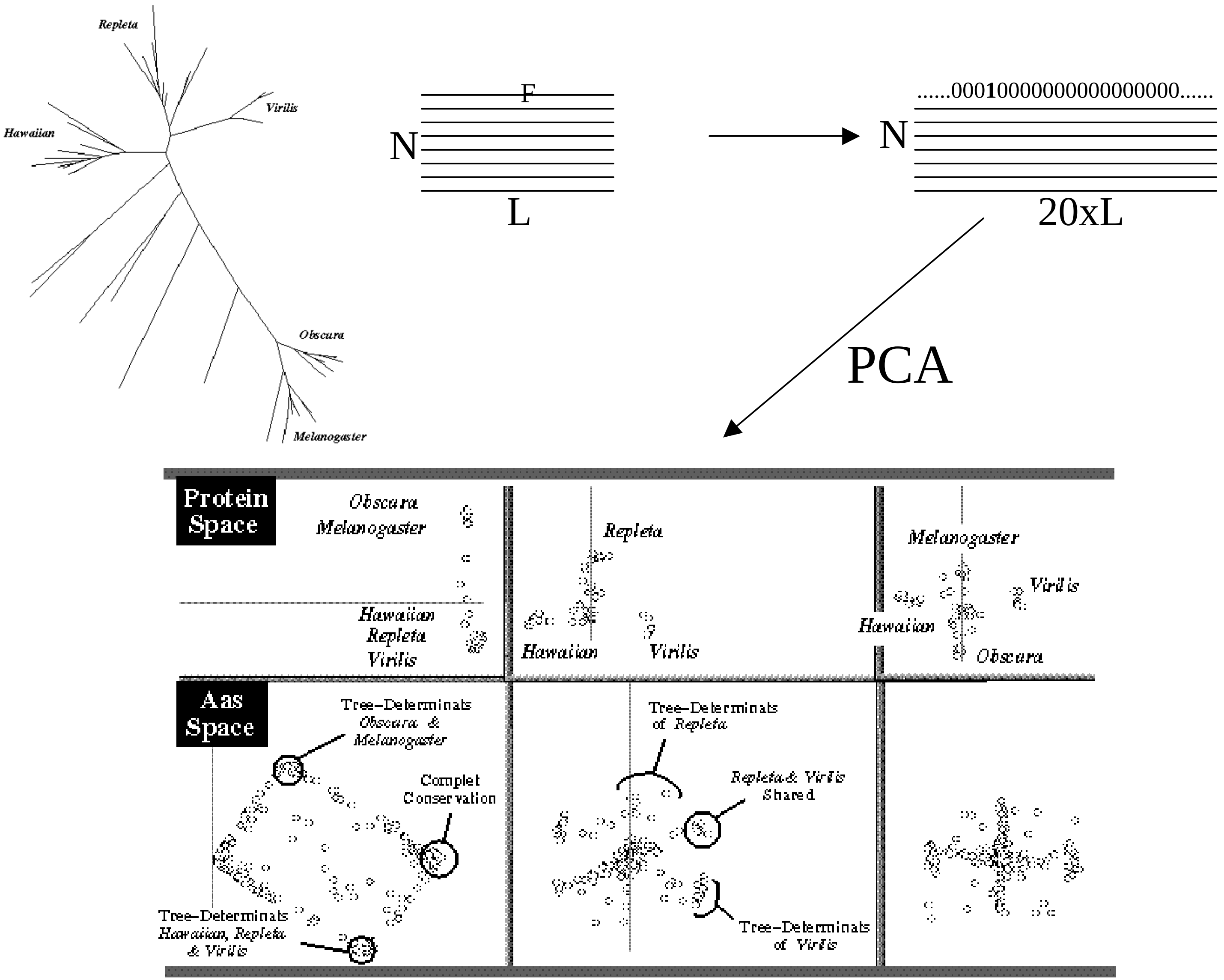
Family-dependent conservation (“tree-determinants”)



```
AFVITDNCIK CKYTDCVEVC .... PVDCFYEGPN FIVIHPDECI DCALCEPECP
AFVITDNCIK CKYTDCVEVC .... PVDCFYEGPN FIVIHPDECI DCALCEPECP
TFVITDNCIK CKYTDCVEVC .... PVDCFYEGPN FIVIHPDECI DCALCEPECP
TFVITDNCIK CKYTDCVEVC .... PVDCFYEGPN FIVIHPDECI DCALCEPECP
TYVFTDNCIK CKYTDCVEVC .... PVDCFYEGEN TILIHPDECI DCGVCEPECP
PYVITENCIK CKYQDCVEVC .... PVDCFYEGEN FIVINPDECI DCGVCNPECP
PHVICQPCIK VKDQSCVEVC .... PVECIYDGGD QEVIHPEECI DCGACVPACP
PFVITSPCID EKAADCVETC .... PVDAIHEGPD QYVIDPDLCI DCAACEPVCP
TYVFAQPCVD VKDKACIEEC .... PVDCIYEGQR SILIHPDECV DCGACEPVCP
TYVIAEPCVD VLDKACIEEC .... PVDCIYEGGR MIVIHPDECV DCGACEPVCP
ATVIADECSD CG..TCVDEC .... PNDAIteEKG IAVVDNDECV ECGACEEACP
TYVIAEPCVD VKDKACIEEC .... PVDCIYEGAR MIVIHPDECV DCGACEPVCP
AYKITDGCID CG..ACEPEC .... PVEAISESDA VKVIDADKCI DCGACANTCP
AHIITDECID CG..ACAAEC .... PVEAIHEGTG KYVVDADTCI DCGACEAVCP
AYKITDECID CG..SCADQC .... PVEAISEGS. IYEIDEALCT DCGACADQCP
AFVINDSCVS CG..ACAGEC .... PVSAITQGDT QEVIDADTCI DCGNCANVCP
...VNDDCVS CE..ACVEIC .... PDVFEmeGDK AVVINPDSDL DCVEEAisCP
MHVISDECVS CG..ACASTC .... PTGAIEEGET .KYVVTDSCI DCGACEAVCP
AYVINDSCIS CG..ACKPEC .... PVNI..QQGS IYAIDADSCI DCGSCASVCP
AYVINDSCIS CG..ACEPEC .... PVNAITAGDD KYVIDAATCI DCGACAGVCP
AYKILDTCVS CG..ACAAEC .... PVDAISQGDT QEVIDADTCI DCGNCANVCP
AYVINEACIS CG..ACEPEC .... PVNAISSGDD RYVIDADTCI DCGACAGVCP
AYFITDACIS CG..ACESEC .... PVSPISPGDS VYVIDADACI ECGACANVCP
AYKIADSCVS CG..ACASEC .... PVNAISQGDS IEVIDADTCI DCGNCANVCP
....HEECIS CE..SCVEIC .... pvFAMIDGEe aMVTAPDSTA ECAQDAiaCP
...IYDTCIS C..TQCVRAC .... PTDVlwEGCK AKQIatEDCA GCKRCESACP
ALMITDQCIS CNV..CQPEC .... PNGAISQGDE TYVIEPSLCT ECvqCVEVCP
AYKIEETCIS CG..ACAAEC .... PVNAIEQGDT IEVVNADTCI DCGNCANVCP
..VDWDCCIS DG..ACMDVC .... PVNLYEWNLn cIPVRESDCI FCMACESVCP
```

Family-dependent conservation ("tree-determinants")

SequenceSpace

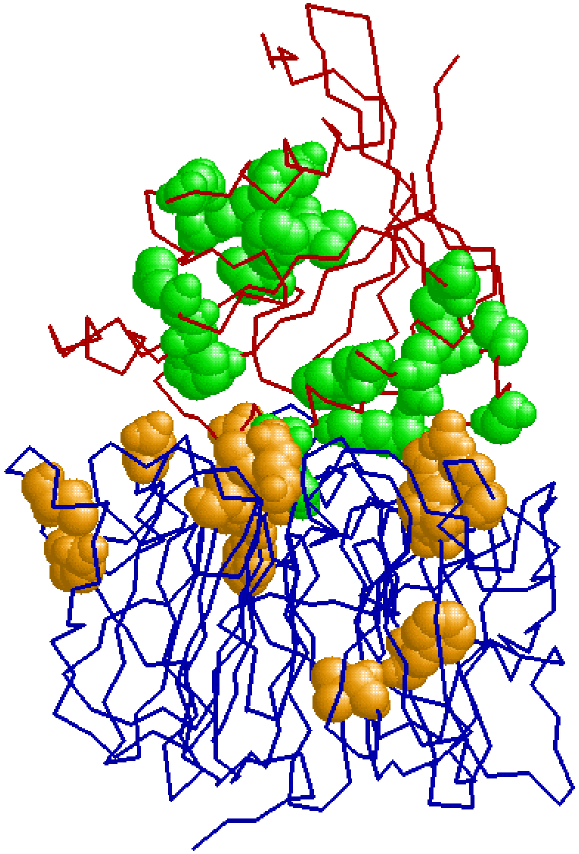


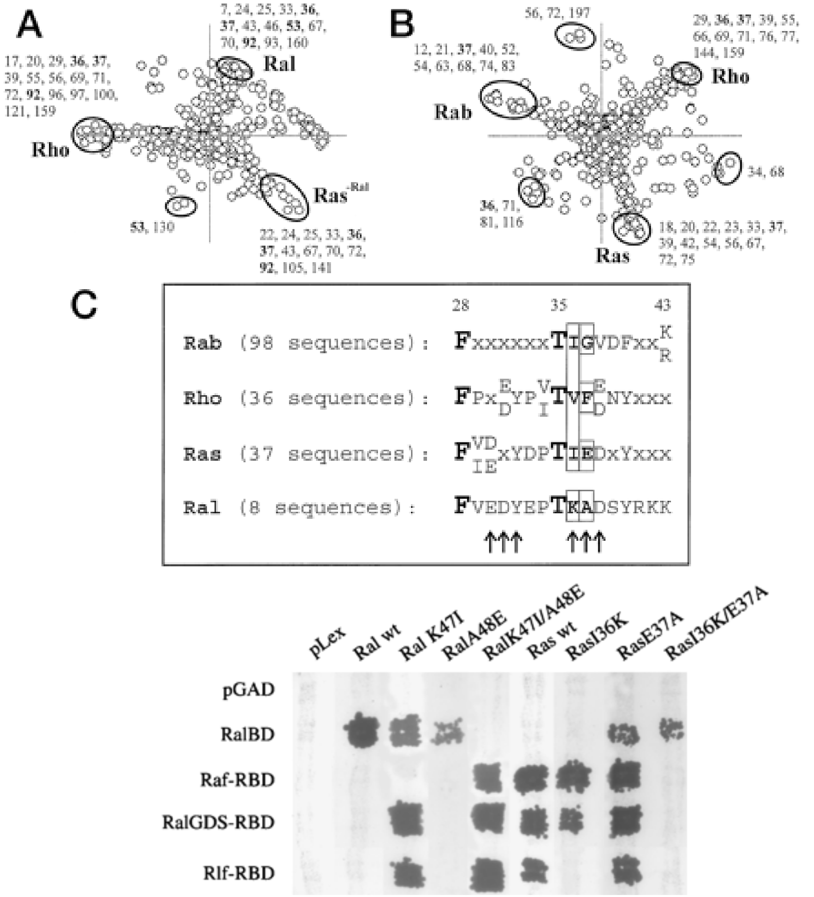
Casari, G., Sander, C., Valencia, A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol*, **2**, 171-178.

Prediction of interaction regions. Sequence-based methods

Tree-determinants

SequenceSpace



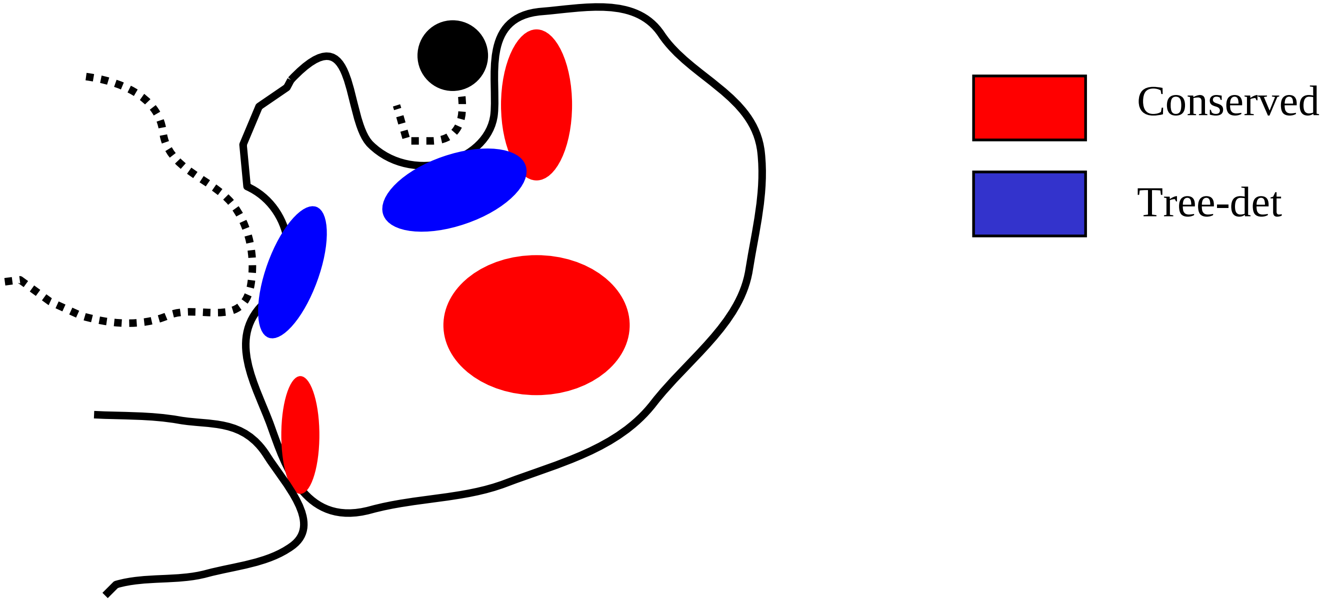


Azuma, Y., Renault, L., Garcia-Ranea, J. A., Valencia, A., Nishimoto, T. & Wittinghofer, A. (1999). Model of the ran RCC1 interaction using biochemical and docking experiments. *J Mol Biol* **289**(4), 1119-1130.

Bauer, B., Mirey, G., Vetter, I.R., Garcia-Ranea, J.A., Valencia, A., Wittinghofer, A., Camonis, J.H. and Cool, R.H. (1999) Effector recognition by the small GTP-binding proteins Ras and Ral. *J Biol Chem*, **274**, 17763-17770.

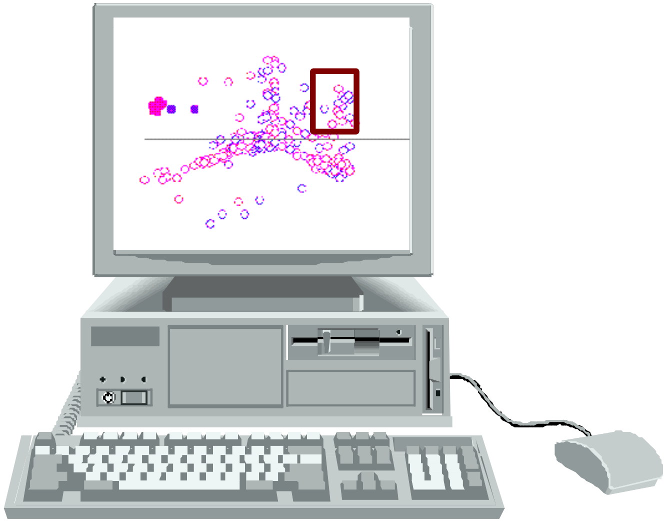
Prediction of functional regions
Sequence-based methods

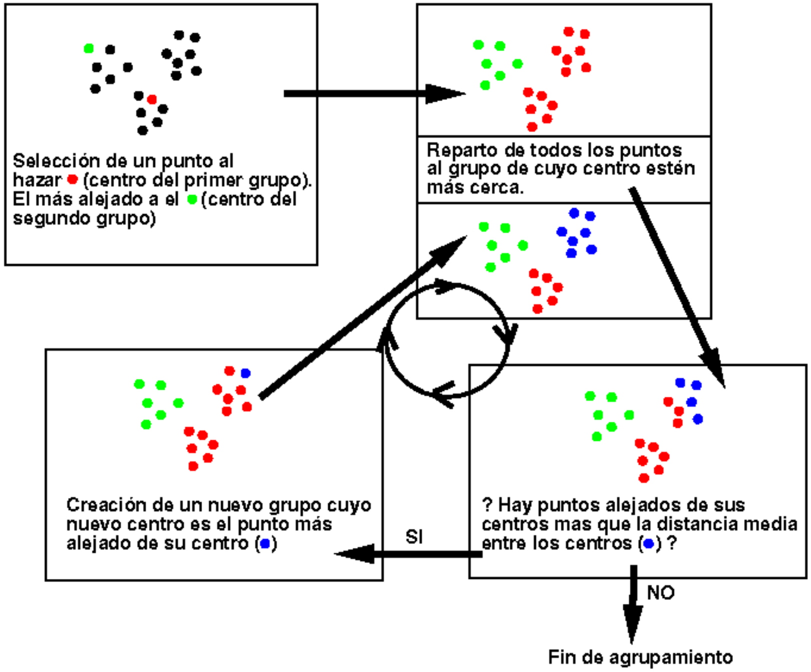
Tree-determinants vs conserved residues



Pazos, F. and Bang, J.-W. (2006) Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics*, **1**, 15-23.

SequenceSpace clustering





Selección de un punto al hazar ● (centro del primer grupo). El más alejado a el ● (centro del segundo grupo)

Reparto de todos los puntos al grupo de cuyo centro estén más cerca.

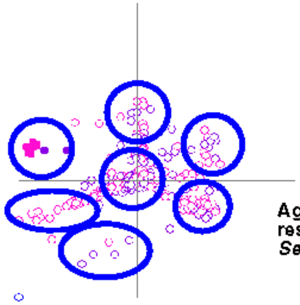
? Hay puntos alejados de sus centros mas que la distancia media entre los centros (●) ?

SI

Creación de un nuevo grupo cuyo nuevo centro es el punto más alejado de su centro (●)

NO

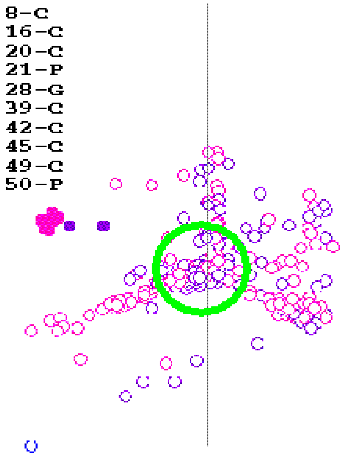
Fin de agrupamiento

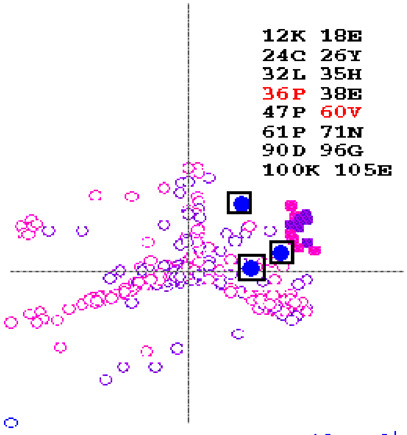


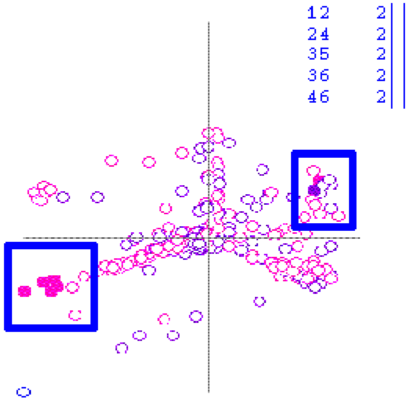
Agrupamiento del espacio de residuos generado por *SequenceSpace*

del Sol Mesa, A., Pazos, F., Valencia, A. Automatic Methods for Predicting Functionally Important Residues. *Journal of Molecular Biology* 2003 **326:**1289-1302

Lopez-Romero, P. *et al.* (2005). *Submited*







>1 (Trash)

> 2 (50-P)
8-C
16-C
20-C
21-P
28-G
39-C
42-C
45-C
49-C
50-P

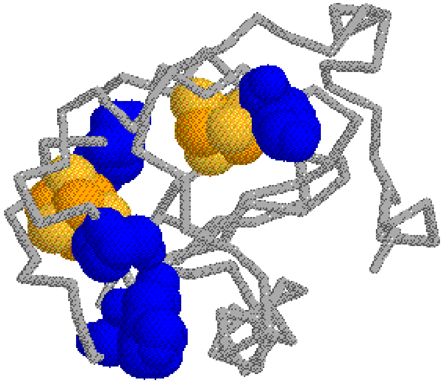
> 3 (36-P)
18-E
36-P
54-I

> 4 (19-E)
2-Y
4-I
15-A
19-E
25-I
43-G
48-V
51-V

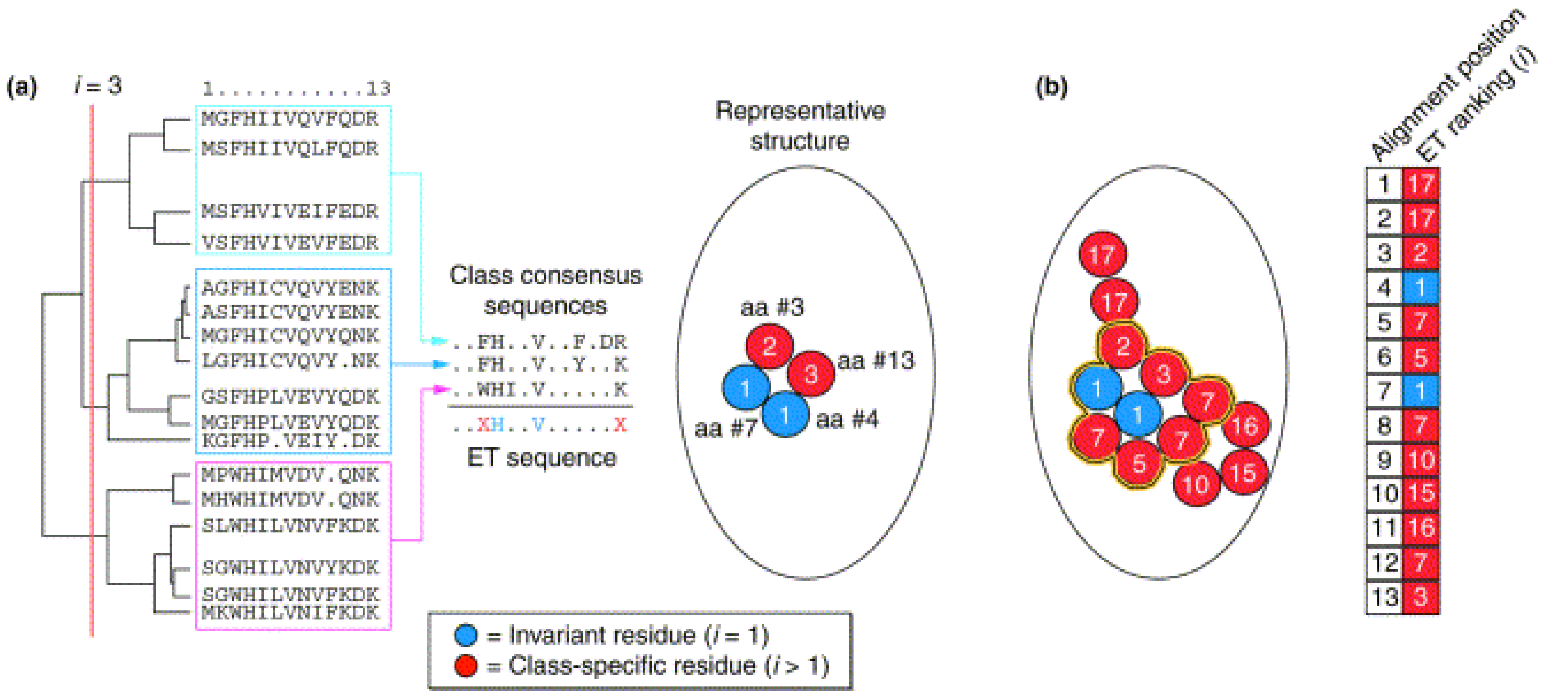
> 5 (33-V)
11-C
33-V

> 14 (58-D)
1-T
12-K
24-C
26-Y
32-L
35-H
38-E
47-P
58-D
61-P
71-N
90-D
96-G
98-K
100-K
105-E
106-R

>15 ...
....



Family-dependent conservation
Evolutionary Trace



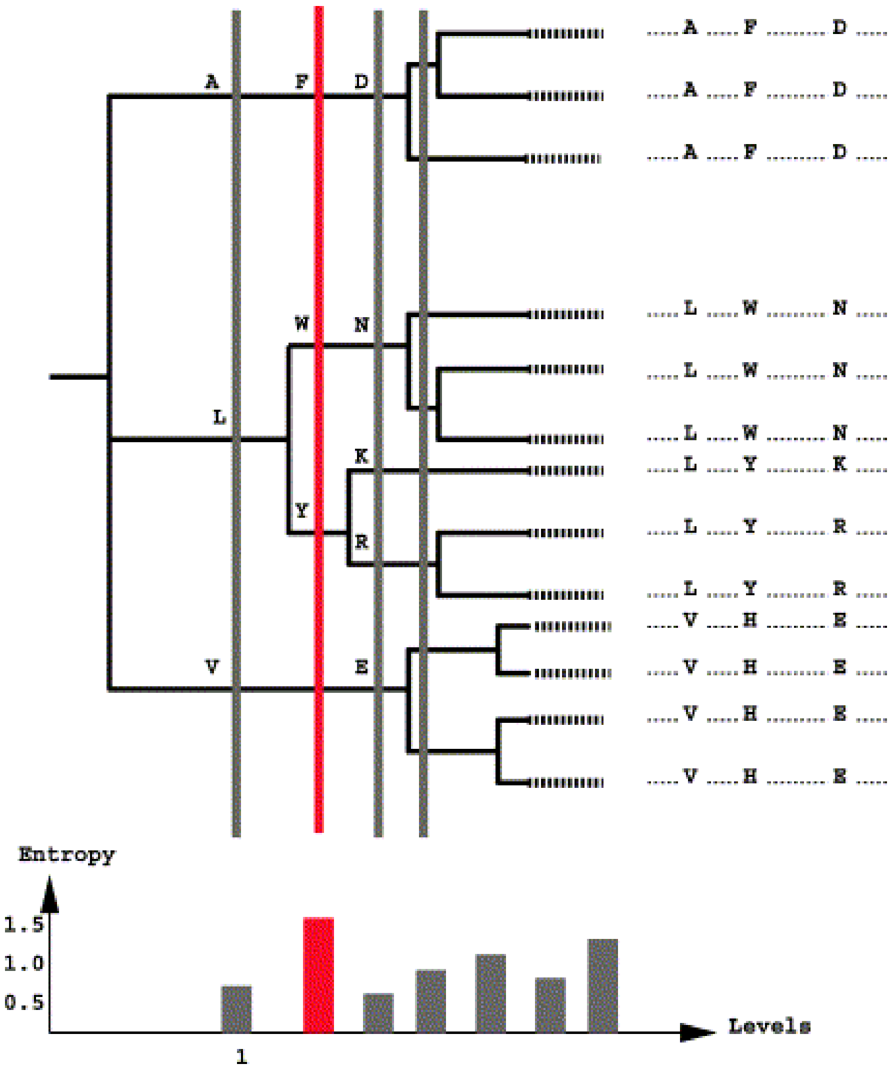
Current Opinion in Structural Biology

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An Evolutionary Trace method defines binding surfaces common to protein families. *J Mol Biol* **257,** 342-358.

Mihalek, I., Res, I. and Lichtarge, O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*, **336,** 1265-1282.

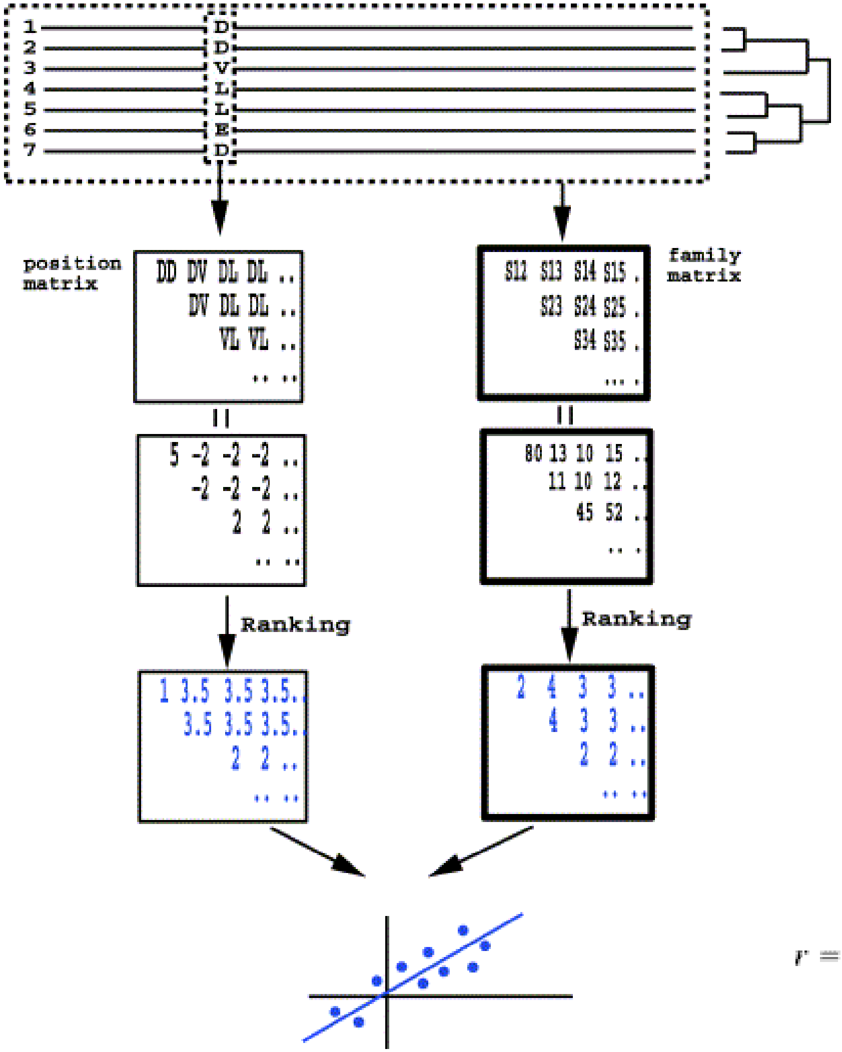
Mihalek, I., Res, I. and Lichtarge, O. (2006) A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins. *Bioinformatics.*, **22**, 149-156.

Level Entropy Method


A
F
D
W
N
L
K
Y
R
V
E
..... A ..... F .... D .....
..... A ..... F .... D .....
..... A ..... F .... D .....
..... L ..... W .... N .....
..... L ..... W .... N .....
..... L ..... W .... N .....
..... L ..... Y .... K .....
..... L ..... Y .... R .....
..... L ..... Y .... R .....
..... V ..... H .... E .....
..... V ..... H .... E .....
..... V ..... H .... E .....
..... V ..... H .... E .....
Entropy
1.5
1.0
0.5
1
Levels

del Sol Mesa, A., Pazos, F., Valencia, A. (2003). Automatic Methods for Predicting Functionally Important Residues. *Journal of Molecular Biology* **326:**1289-1302

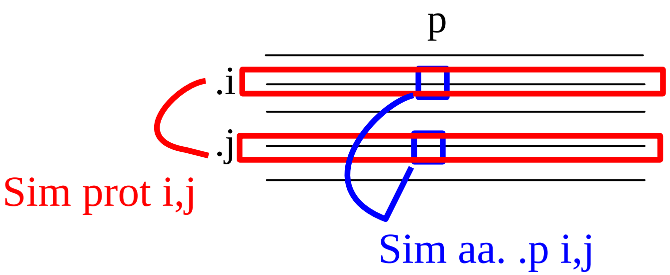
MTreedet

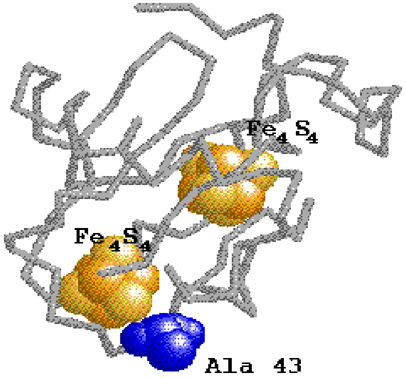


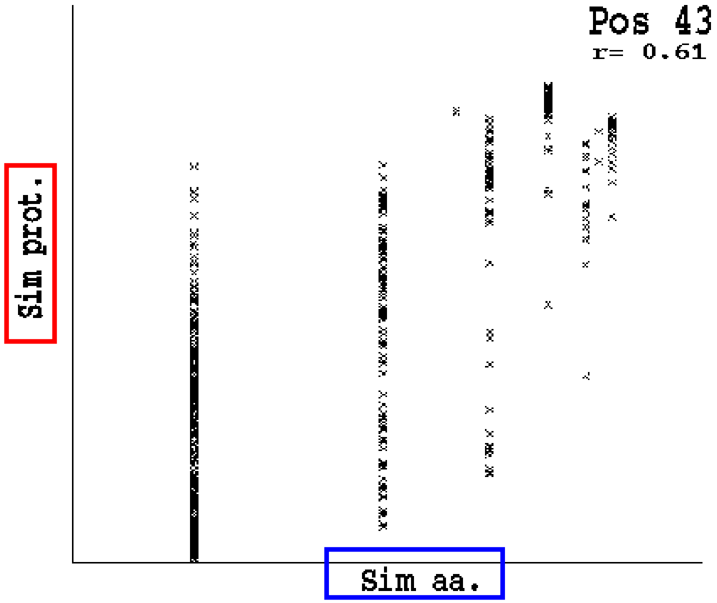
$$r = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}}$$

del Sol Mesa, A., Pazos, F., Valencia, A. (2003). Automatic Methods for Predicting Functionally Important Residues. *Journal of Molecular Biology* 2003 **326:**1289-1302

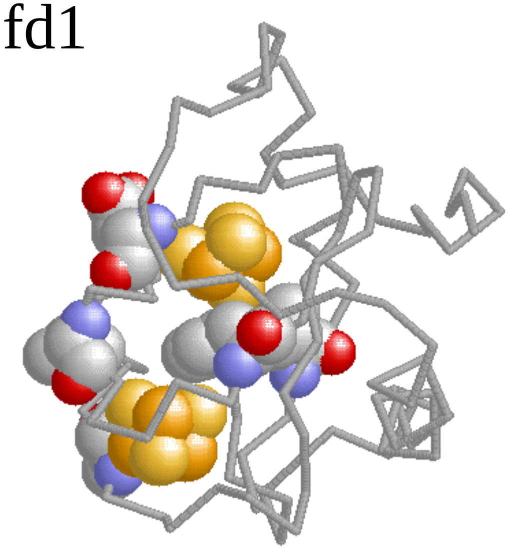
MTreedet





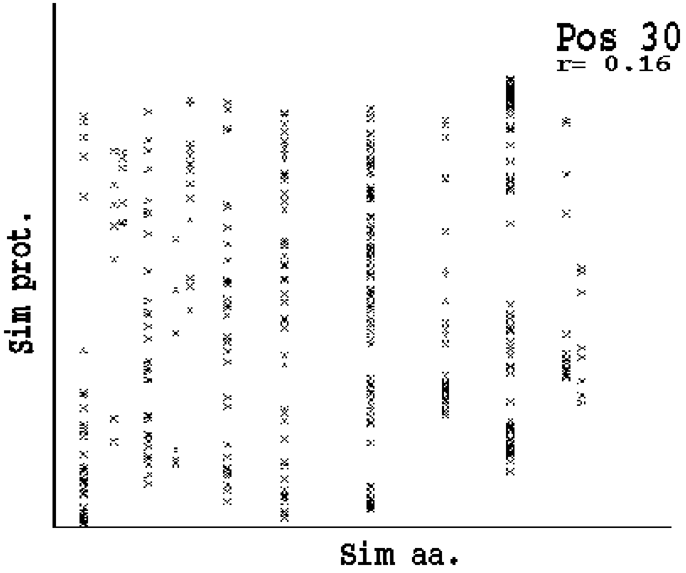


5fd1

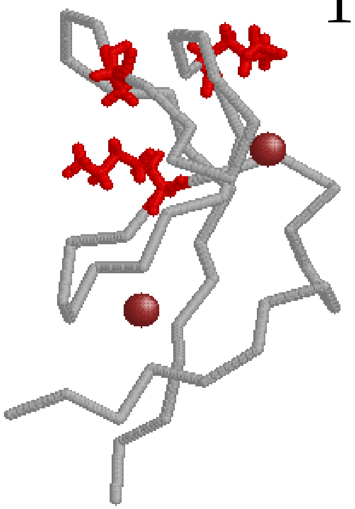


43	0.6067
19	0.5951
25	0.5886
4	0.5864
35	0.5109
26	0.5086
41	0.5038
32	0.4559
29	0.4435
24	0.4256
45	0.3958
39	0.3952
40	0.3823
48	0.3671
12	0.3646
22	0.3627
36	0.3560

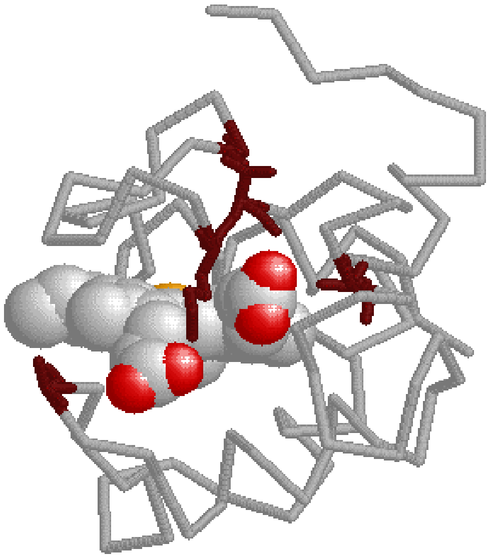
.....
...
.



MTreedet

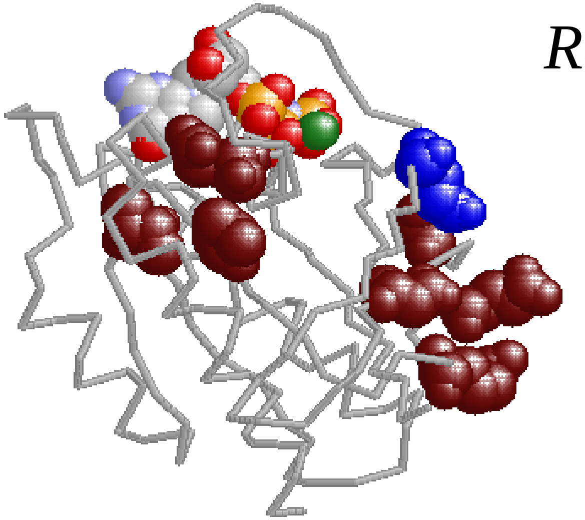


1far



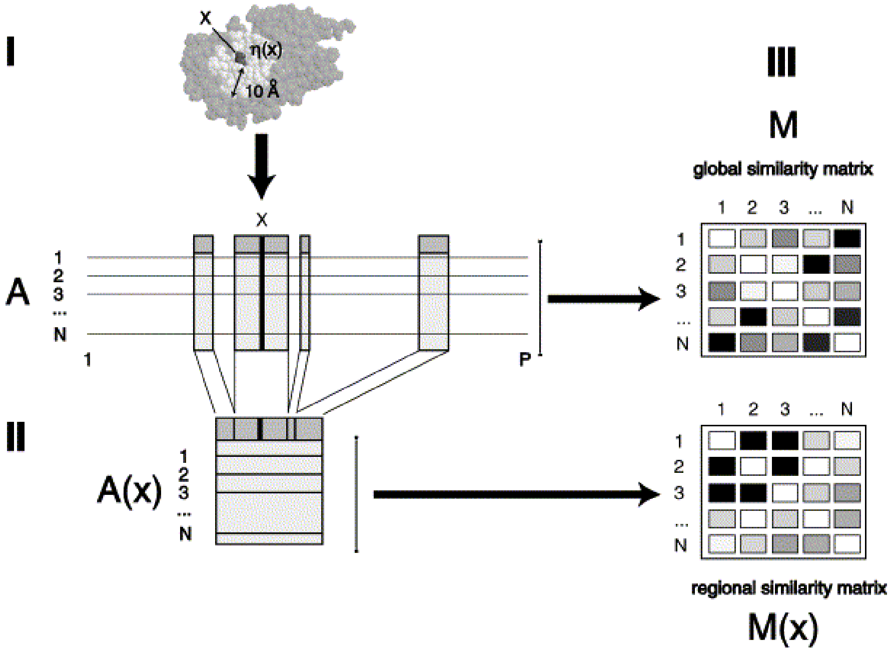
1cyj

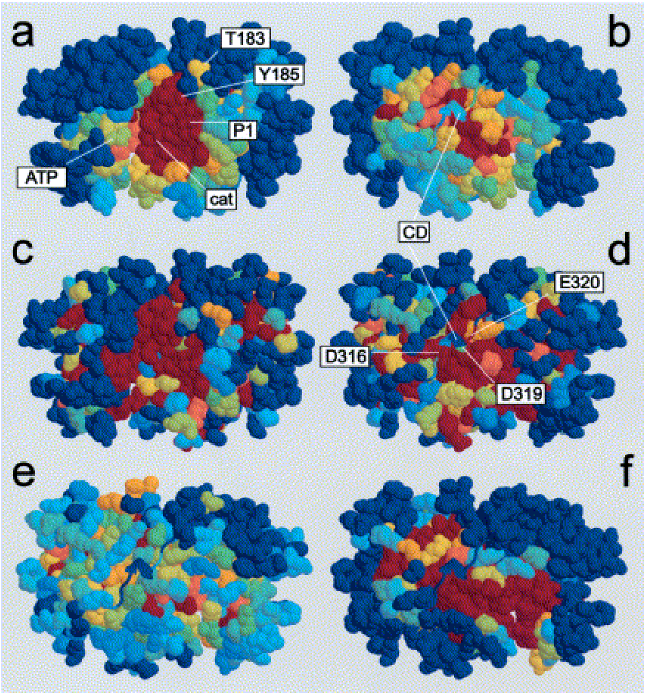
```
 37       0.7650
 54       0.6894
 65       0.6693
 73       0.6413
 22       0.6290
 81       0.6240
 70       0.6123
144       0.6002
 75       0.5797
.....
...
.
```



Ras (5p21)

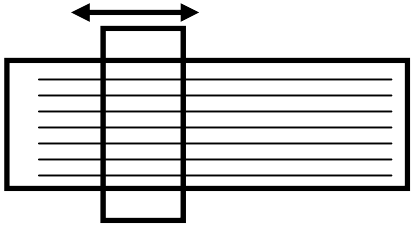
3D cluster

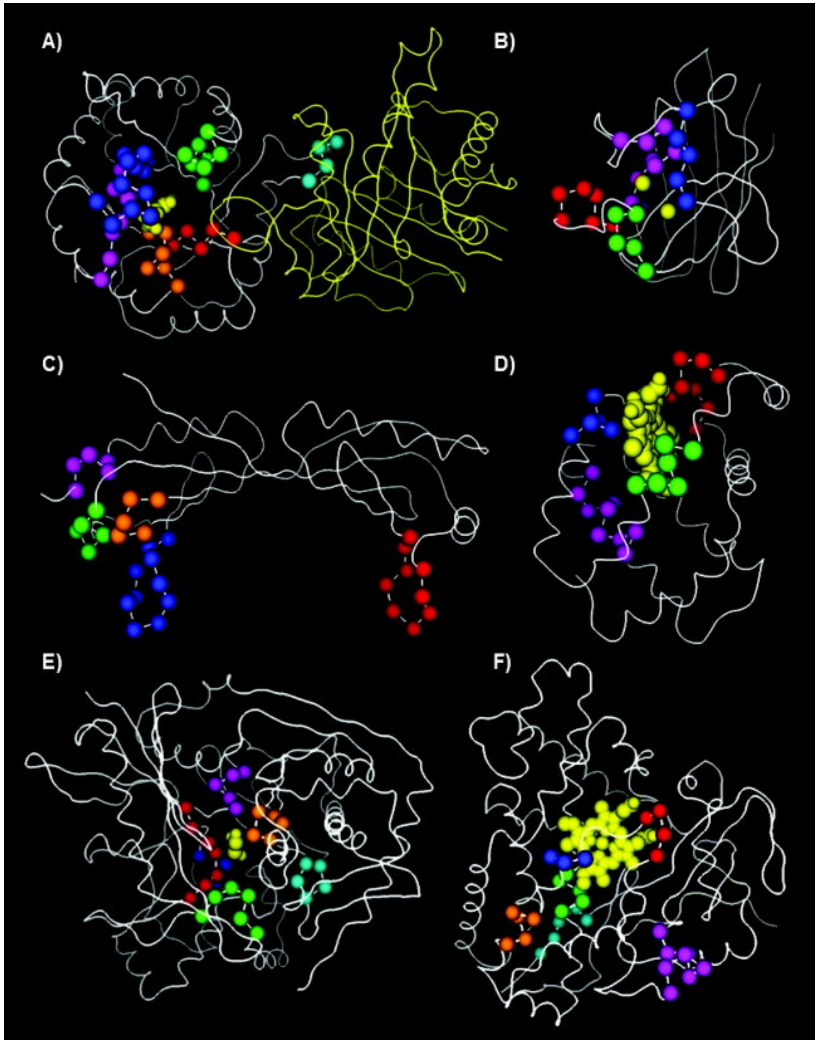




Landgraf, R., Xenarios, I. and Eisenberg, D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol*, **307**, 1487-1502.

Phylogenetic motifs

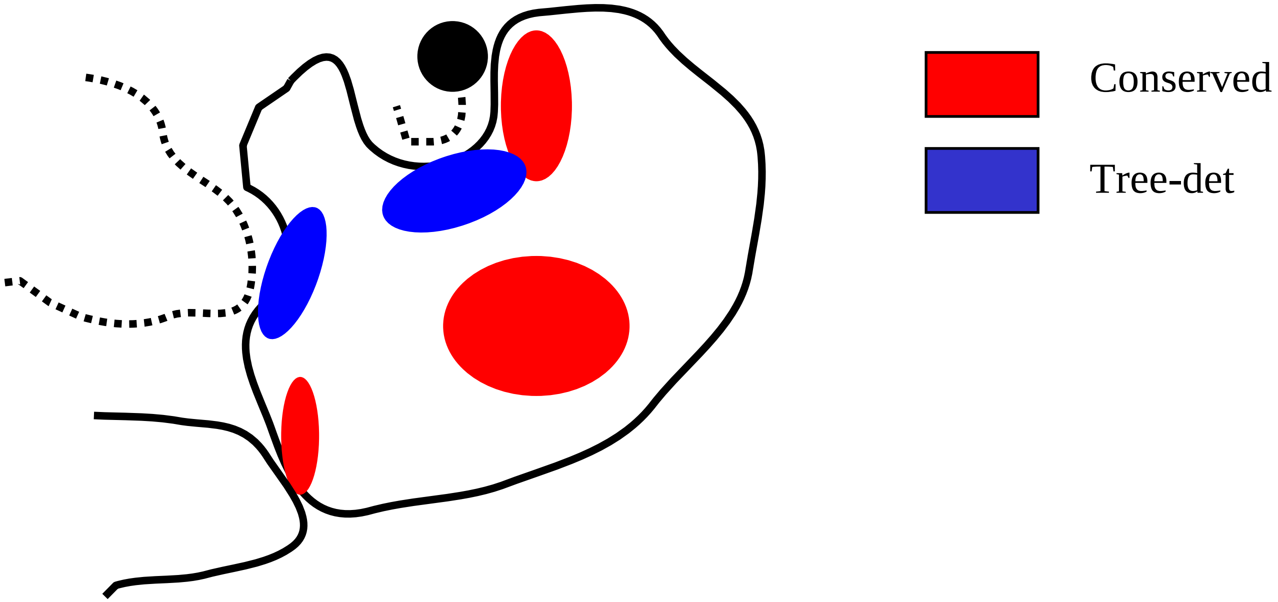




La, D., Sutch, B. and Livesay, D.R. (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins*, **58**, 309-320.

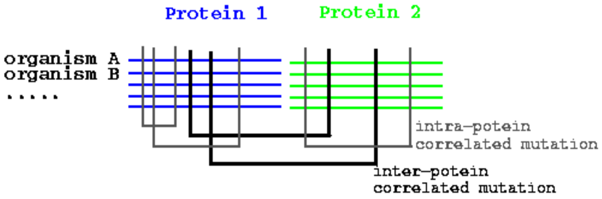
Prediction of functional regions
Sequence-based methods

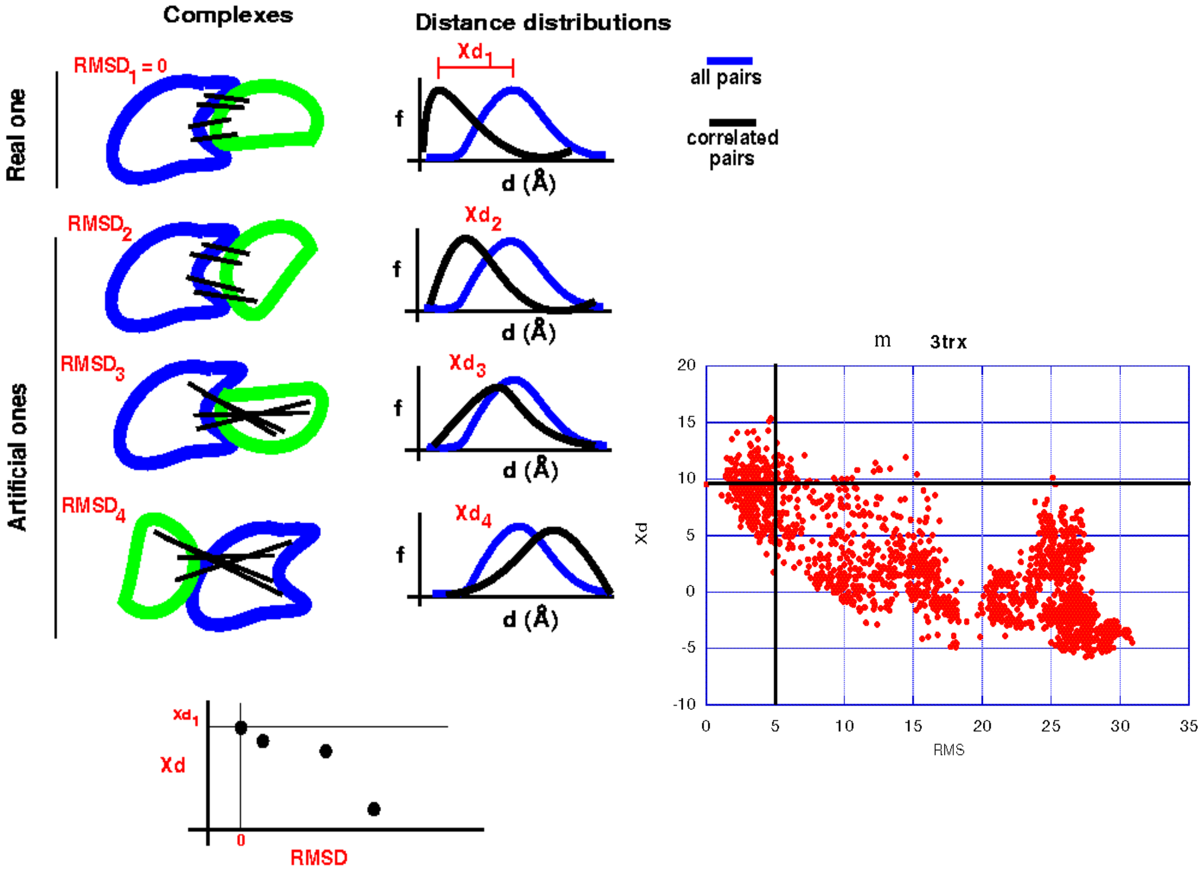
Tree-determinants vs conserved residues

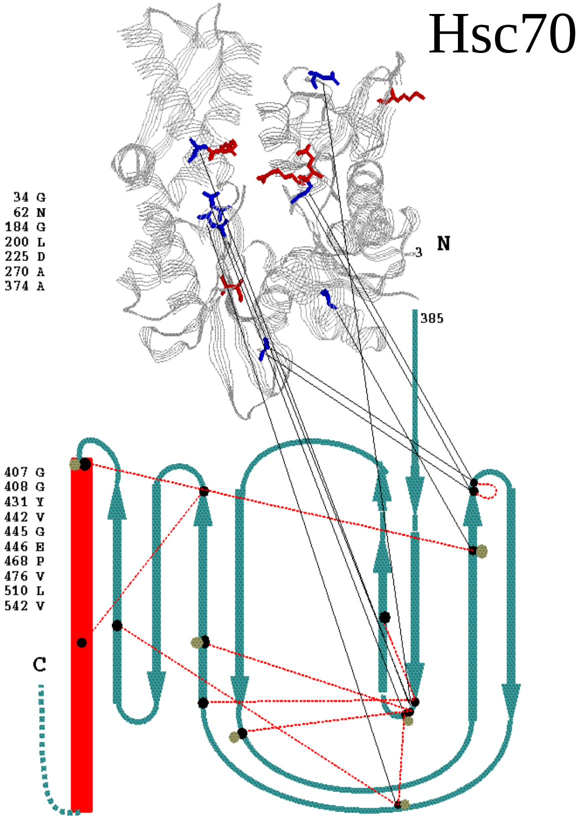


Pazos, F. and Bang, J.-W. (2006) Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics*, **1**, 15-23.

Prediction of interaction regions
Sequence-based methods
Correlated mutations

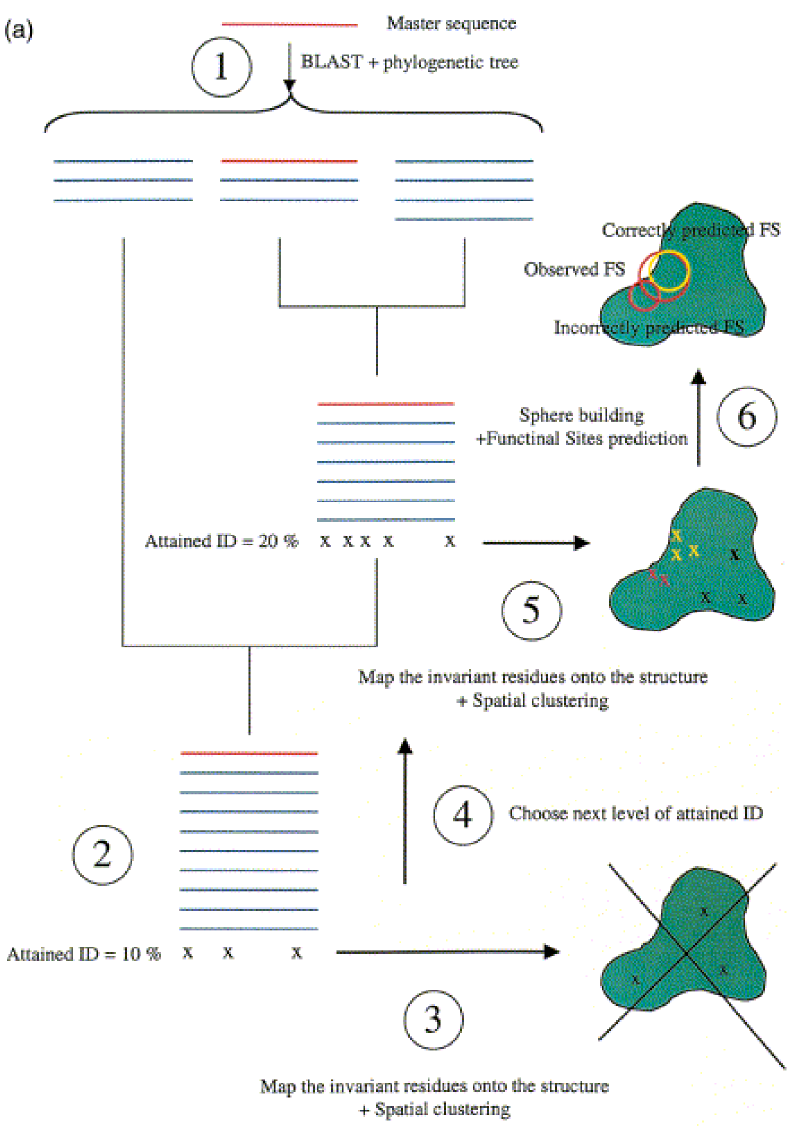


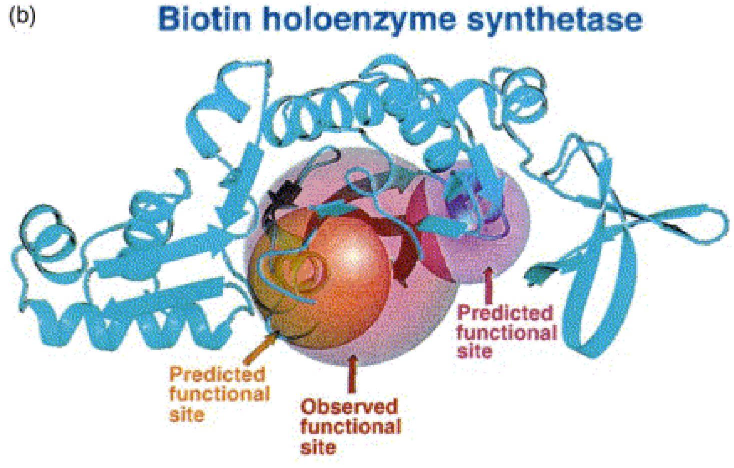




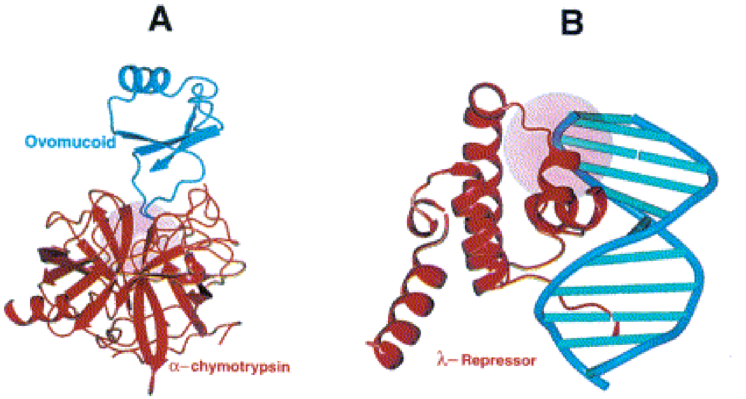
Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J Mol Biol* **271**(4), 511-523.

3D function



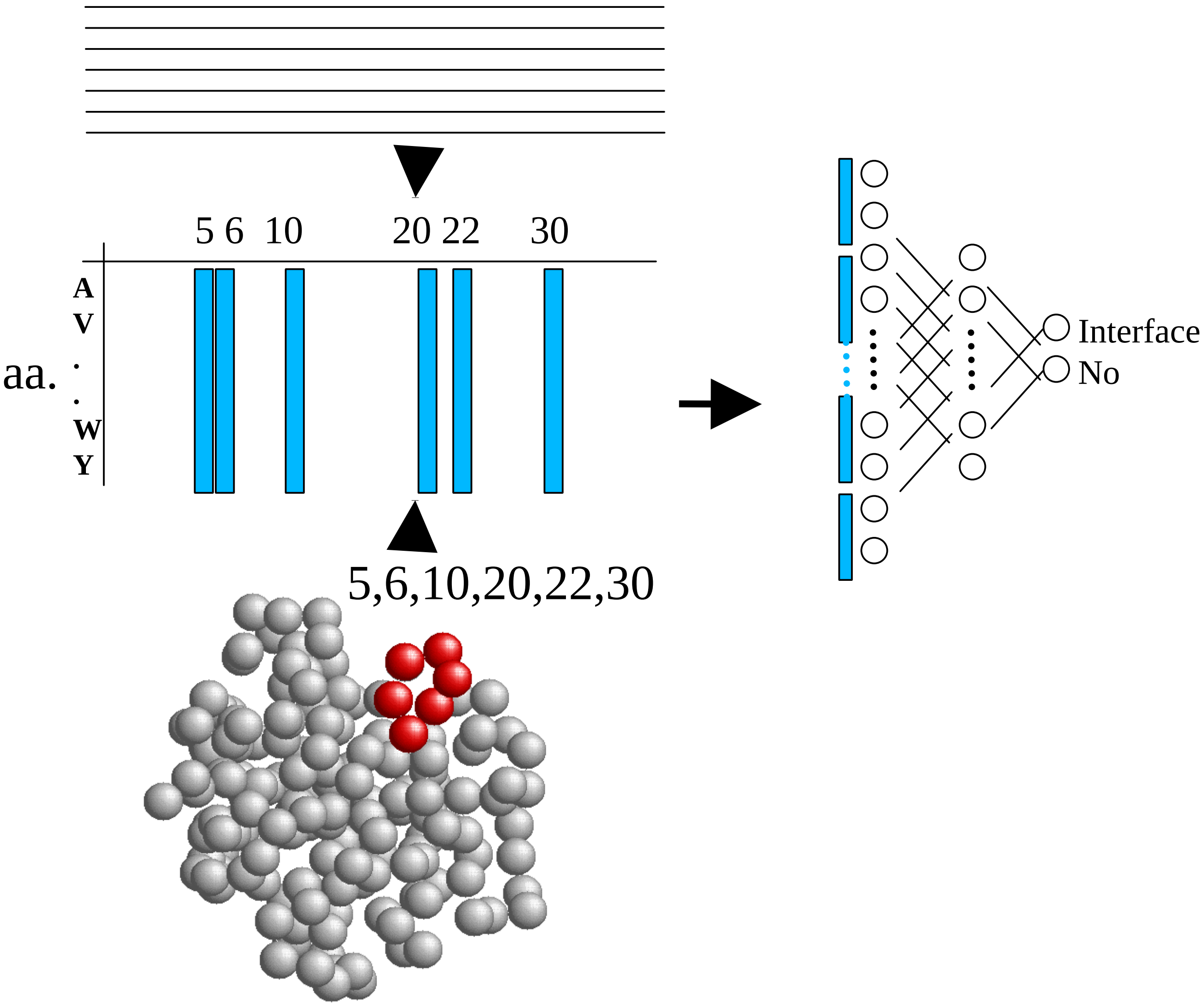


(b)
Biotin holoenzyme synthetase
Predicted functional site
Predicted functional site
Observed functional site



Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J.E. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, **311**, 395-408.

Evolutionary information + 3D structure

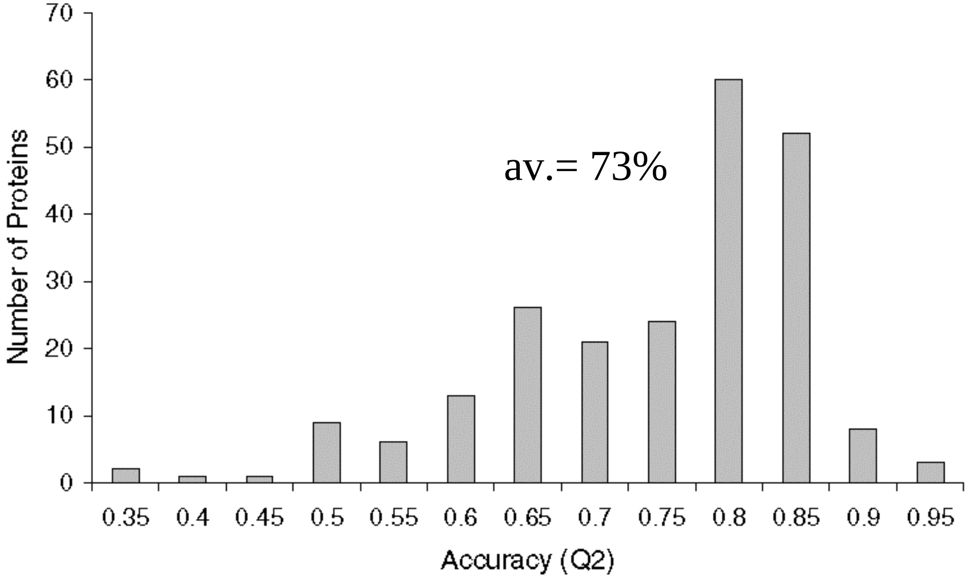


Multiple sequence alignment

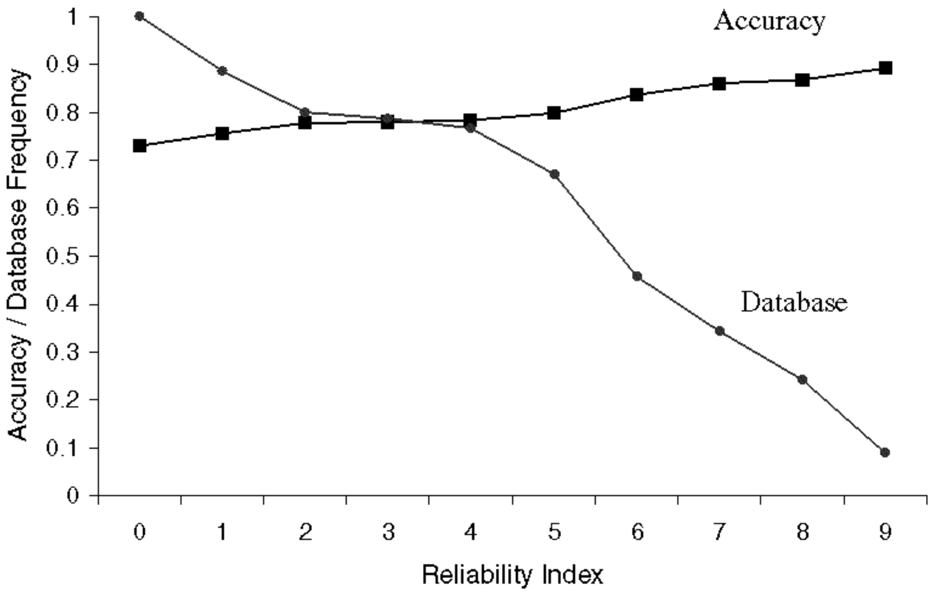
Sequence profiles

3D str.
Surface patch

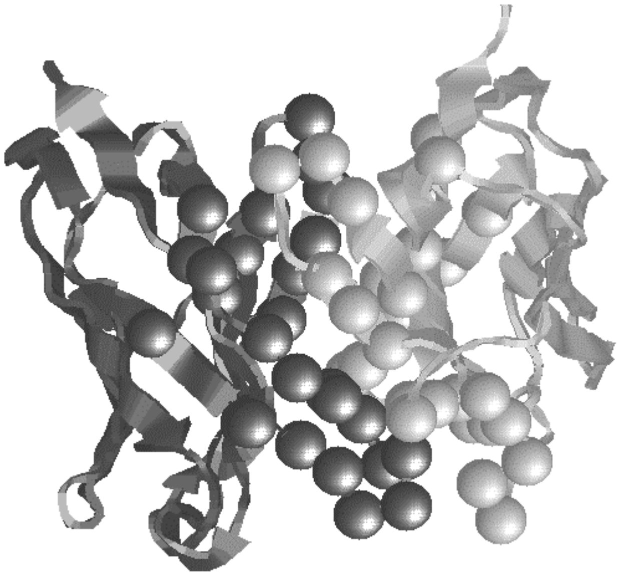
- Piero Fariselli, Florencio Pazos, Alfonso Valencia & Rita Casadio (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem.* **269(5):** 1356-1361.

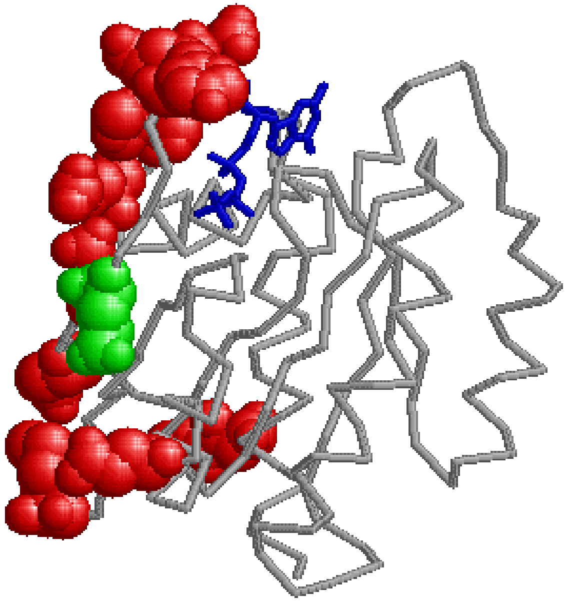


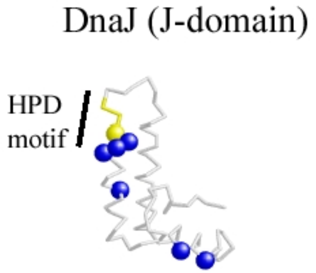
av.= 73%
Number of Proteins
0
10
20
30
40
50
60
70
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
Accuracy (Q2)

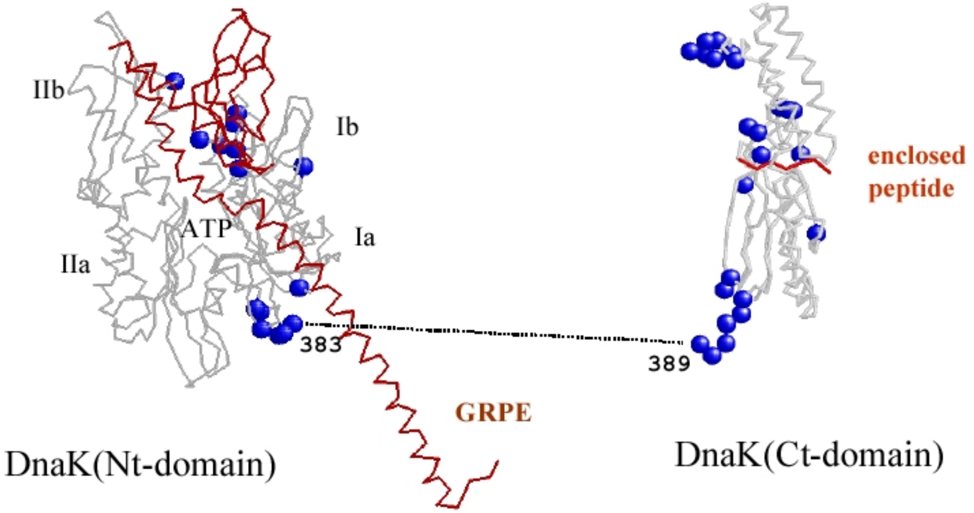


Accuracy
Database
Accuracy / Database Frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
9
Reliability Index









• Piero Fariselli, Florencio Pazos, Alfonso Valencia & Rita Casadio (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem.* **269(5):** 1356-1361.

Prediction of interaction regions
Methods based on structural features

Geometrical docking

Smith, G.R., Sternberg, M.J.E. (2002).
Prediction of protein-protein interactions by docking methods.
Curr Opin Struct Biol. **12:**28-35.

Halperin, I., Ma, B., Wolfson, H., Nussinov, R. (2002).
Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*. **47:**409-443.

Backbone conformation

Watson, J.D., Milner-White, E.J.
A novel main-chain anion binding site in proteins: the nest.
A particular combination of phi, psi values in succesive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions.
J Mol Biol 2002 315:171-182

H-bonds characteristics

Fernández, A., Scheraga, H.A.
Insufficiently dehydrated hydrogen bonds as determinants of protein interactions.
Proc Natl Acad Sci USA 2003 100:113-118

Kortemme, T., Morozov, A.V., Baker, D.
An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes.
Journal of Molecular Biology 2003 326:1239-1259

Stability

Luque, I., Freire, E.
Structural stability of binding sites: consequences for binding affinity and allosteric effects.
Proteins 2000 S4:63-71

Disordered regions

Tompa, P.
Intrinsically unstructured proteins.
Trends Biochem Sci 2002 **27:**527-533

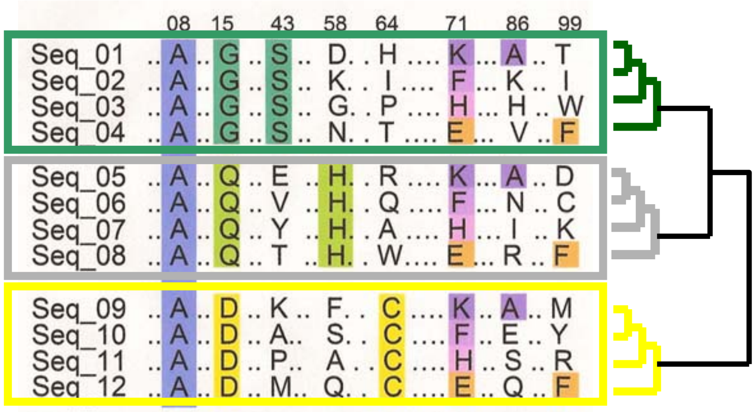
Uversky, V.N.
Natively unfolded proteins: A point where biology waits for physics.
Protein Sci 2002 **11:**739-756

Clefts

Laskowski, R.A., Luscombe, N.M., Swindells, M.B., Thornton, J.M. (1996).
Protein clefts in molecular recognition and function.
*Protein Science***. 5:**2438-2452.

....

Unsupervised methods (phylogeny-based)



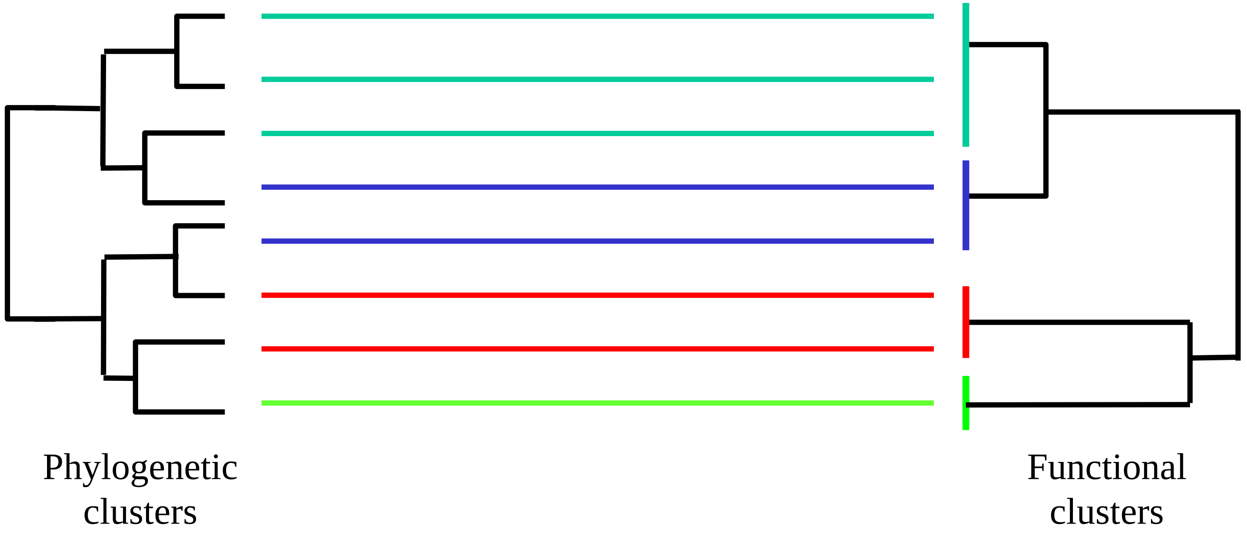
```
         08  15   43   58   64     71   86   99
Seq_01 ..A..G..S.. D. . H .... K... A .. T
Seq_02 ..A..G..S.. K. . I .... F... K .. I
Seq_03 ..A..G..S.. G. . P .... H... H .. W
Seq_04 ..A..G..S.. N. . T .... E... V .. F
Seq_05 ..A..Q.. E ..H. . R .... K... A .. D
Seq_06 ..A..Q.. V ..H. . Q .... F... N .. C
Seq_07 ..A..Q.. Y ..H. . A .... H... I .. K
Seq_08 ..A..Q.. T ..H. . W.... E... R .. F
Seq_09 ..A..D.. K.. F. . C .... K.. A.. M
Seq_10 ..A..D.. A.. S. . C .... F.. E .. Y
Seq_11 ..A..D.. P.. A . . C .... H.. S .. R
Seq_12 ..A..D.. M.. Q. . C .... E.. Q .. F
```

- Assumption: the sub-functional classification coincides with the one implicit in the phylogeny

- Proper for most of the cases, according to the accepted scenario of divergent evolution to function

- Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J.E. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, **311**, 395-408.
- Andrade, M.A., Casari, G., Sander, C. and Valencia, A. (1997) Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol Cybern*, **76**, 441-450.
- Armon, A., Graur, D. and Ben-Tal, N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, **307**, 447-463.
- Bickel, P.J., Kechris, K.J., Spector, P.C., Wedemayer, G.J. and Glazer, A.N. (2002) Finding important sites in protein sequences. *Proc Natl Acad Sci USA*, **99**, 14764-14771.
- Casari, G., Sander, C., Valencia, A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol*, **2**, 171-178.
- del Sol Mesa, A., Pazos, F. and Valencia, A. (2003) Automatic Methods for Predicting Functionally Important Residues. *J Mol Biol*, **326**, 1289-1302.
- Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. and Thornton, J.M. (2006) A method for localizing ligand binding pockets in protein structures. *Proteins.*, **62**, 479-488.
- Kinoshita, K. and Ota, M. (2005) P-cats: prediction of catalytic residues in proteins from their tertiary structures. *Bioinformatics*, **21**, 3570-3571.
- La, D., Sutch, B. and Livesay, D.R. (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins*, **58**, 309-320.
- Landgraf, R., Xenarios, I. and Eisenberg, D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol*, **307**, 1487-1502.
- Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) An Evolutionary Trace method defines binding surfaces common to protein families. *J Mol Biol*, **257**, 342-358.
- Livingstone, C.D. and Barton, G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, **6**, 645-756.
- Yu, G.X., Park, B.H., Chandramohan, P., Munavalli, R., Geist, A. and Samatova, N.F. (2005) In silico discovery of enzyme-substrate specificity-determining clusters. *J Mol Biol*, **352**, 1105-1117.

Function-phylogeny disagreement **(?)**



Main circumstances of potential disagreement

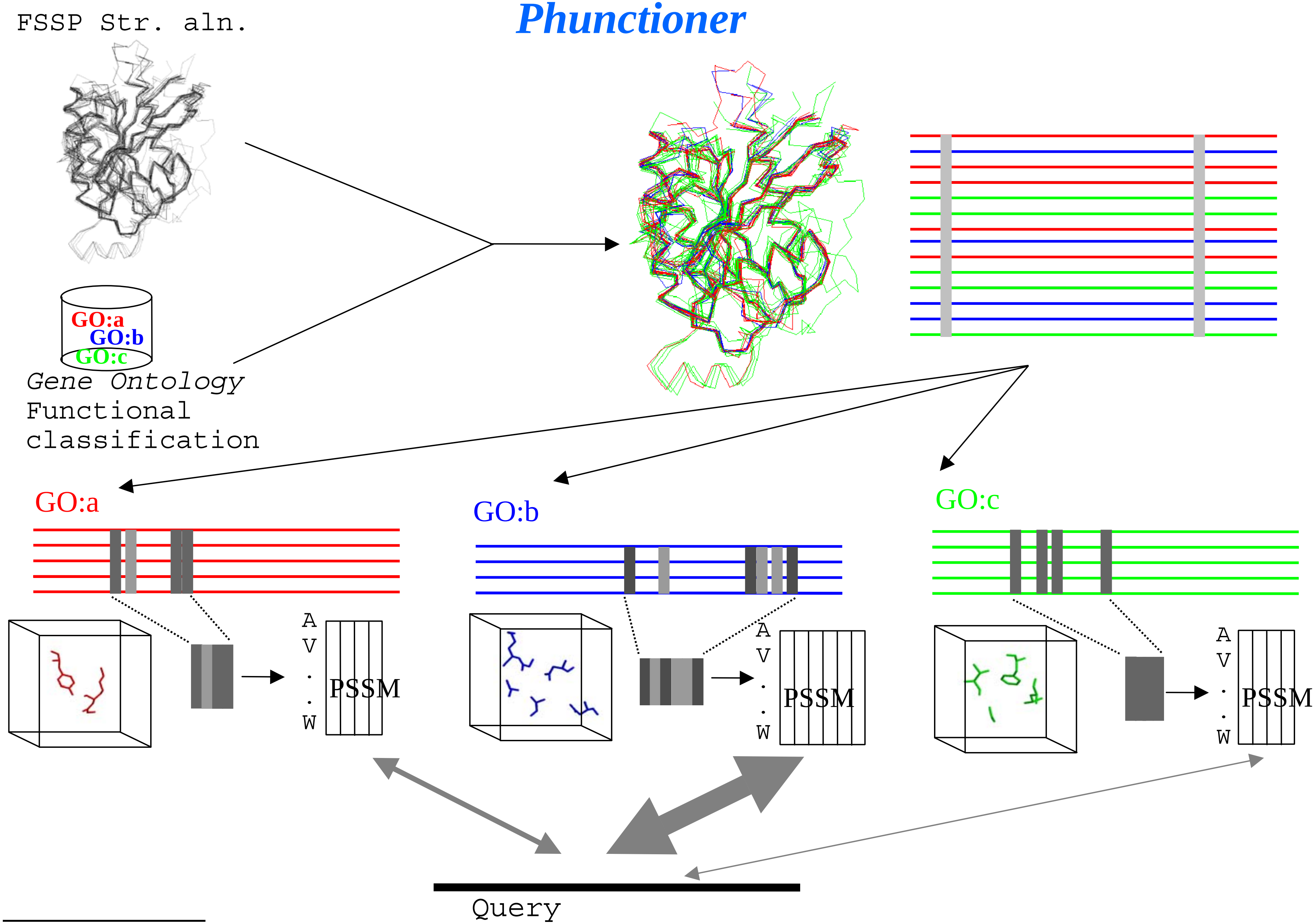
- Many functional and structural features in a protein family that push together its evolution but only one phylogeny can be observed. The observed phylogeny arises from many different independent (to some extent) functional constraints. The specific divergence due to a function can be masked within this “composite phylogeny”

- Structural alignments linking distant proteins

- Details of molecular function may evolve convergently

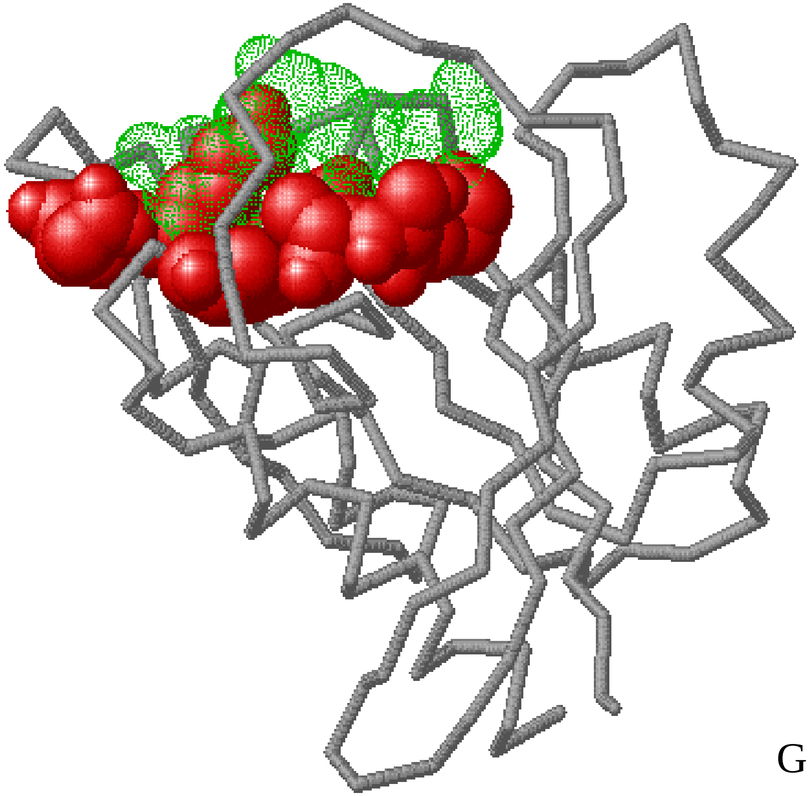
- …

Phunctioner



Pazos, F. and Sternberg, M.J.E. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, **101**, 14754–14759.

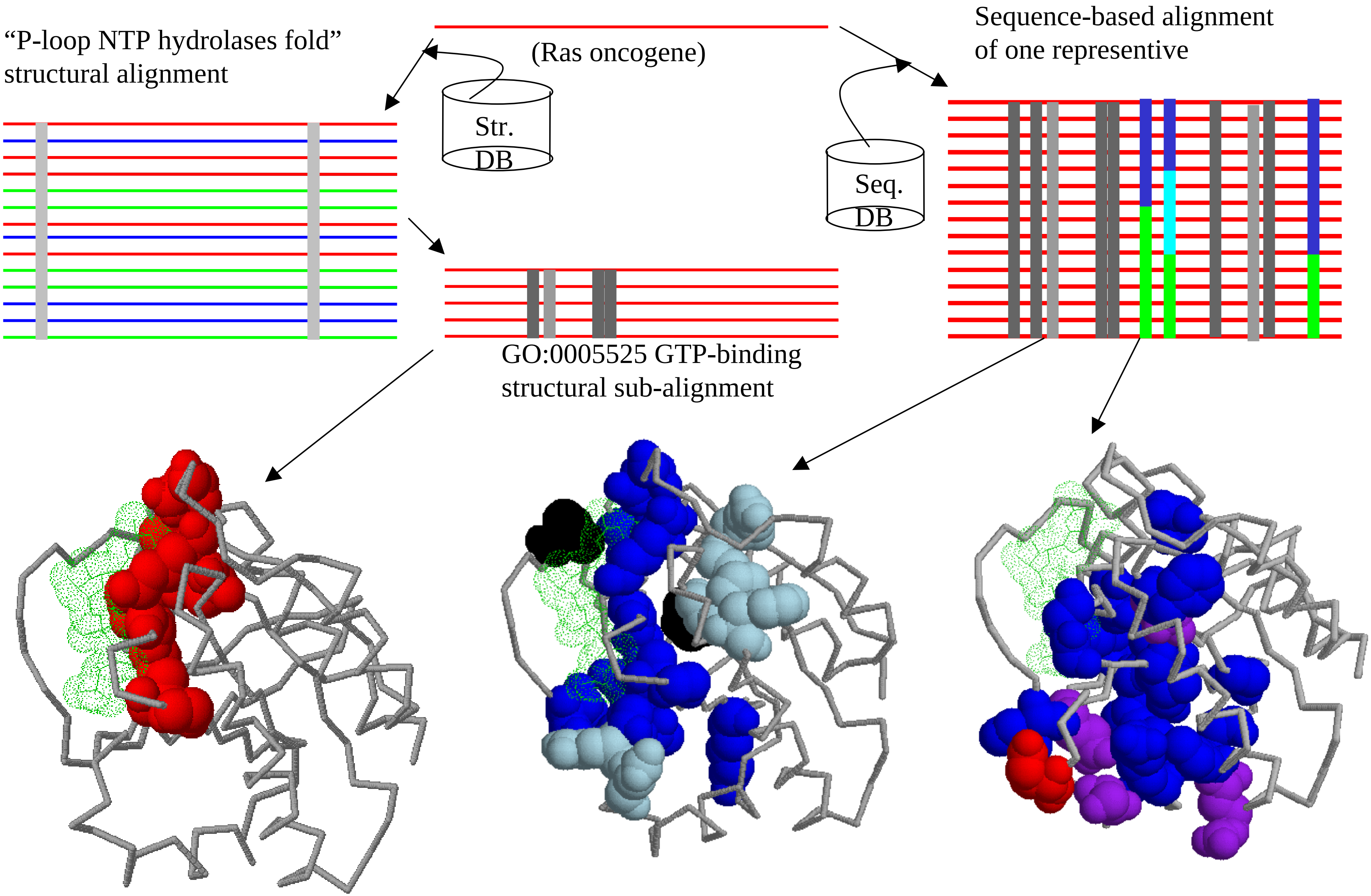
Locating Functional Residues



GTP binding (GO:0005525)

Pazos, F. and Sternberg, M.J.E. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, **101**, 14754–14759.

Locating Functional Residues



“P-loop NTP hydrolases fold” structural alignment

(Ras oncogene)

Str. DB

Seq. DB

Sequence-based alignment of one representive

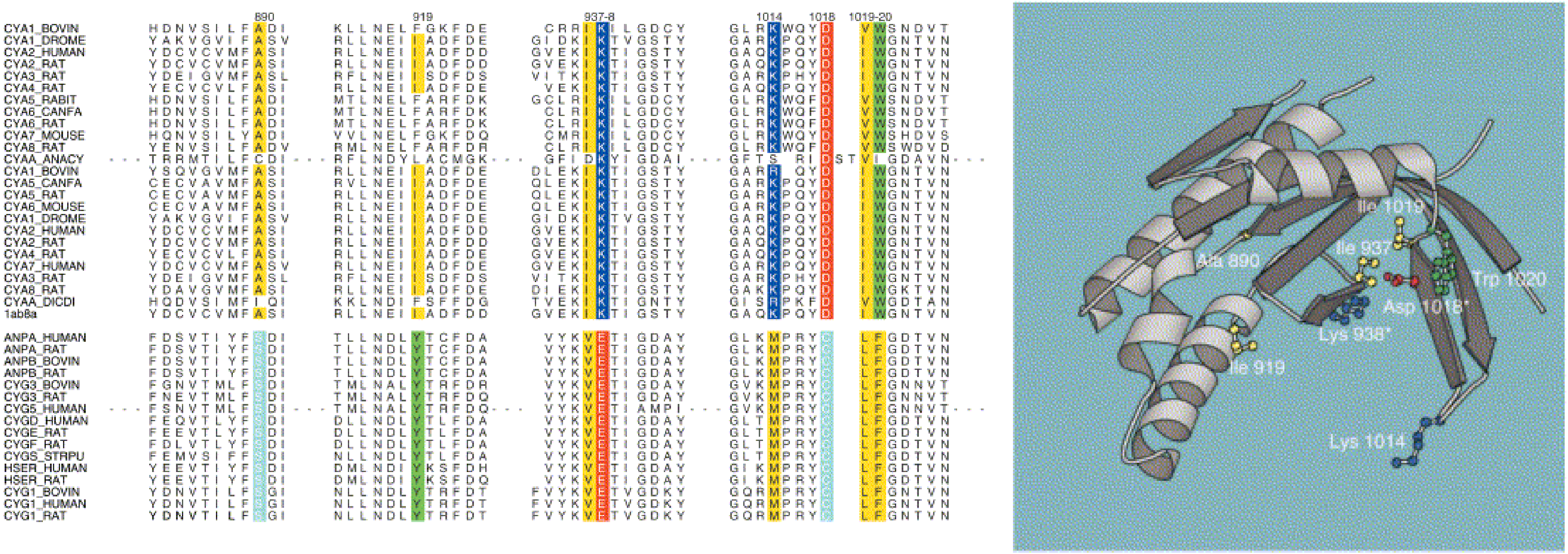
GO:0005525 GTP-binding structural sub-alignment

Phunctioner

Conservation

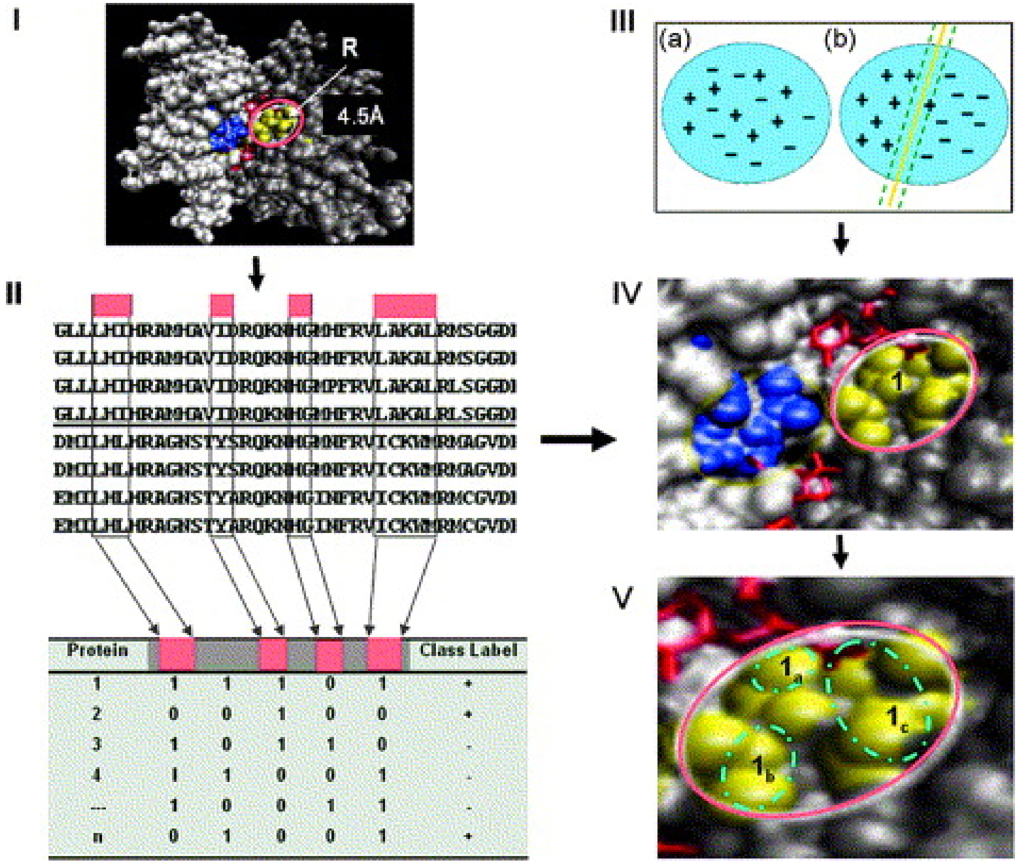
Family-dependent conservation

Functional subtypes not based on sequence
Supervised methods



Hannenhalli, S.S. and Russell, R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol*, **303**, 61-76.

Functional subtypes not based on sequence
SPR



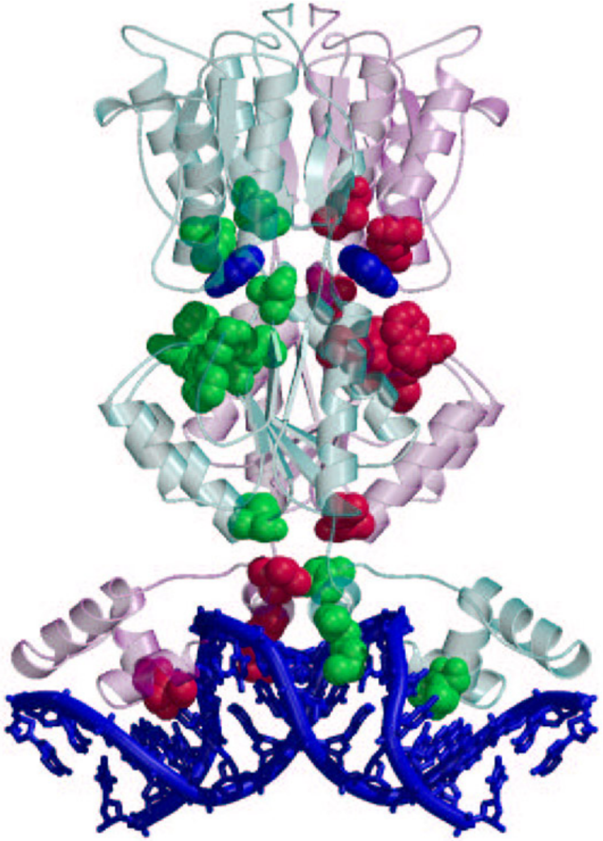
```
GLLLHIHRAMHAVIDRQKNHGMHFRVLAKALRMSGGDI
GLLLHIHRAMHAVIDRQKNHGMHFRVLAKALRMSGGDI
GLLLHIHRAMHAVIDRQKNHGMPFRVLAKALRLSGGDI
GLLLHIHRAMHAVIDRQKNHGMHFRVLAKALRLSGGDI
DMILHLHRAGNSTYSRQKNHGMNFRVICKWMRMAGVDI
DMILHLHRAGNSTYSRQKNHGMNFRVICKWMRMAGVDI
EMILHLHRAGNSTYARQKNHGINFRVICKWMRMCGVDI
EMILHLHRAGNSTYARQKNHGINFRVICKWMRMCGVDI
```

Protein						Class Label
1	1	1	1	0	1	+
2	0	0	1	0	0	+
3	1	0	1	1	0	-
4	1	1	0	0	1	-
---	1	0	0	1	1	-
n	0	1	0	0	1	+

Yu, G.X., Park, B.H., Chandramohan, P., Munavalli, R., Geist, A. and Samatova, N.F. (2005) In silico discovery of enzyme-substrate specificity-determining residue clusters. *J Mol Biol*, **352**, 1105-1117.

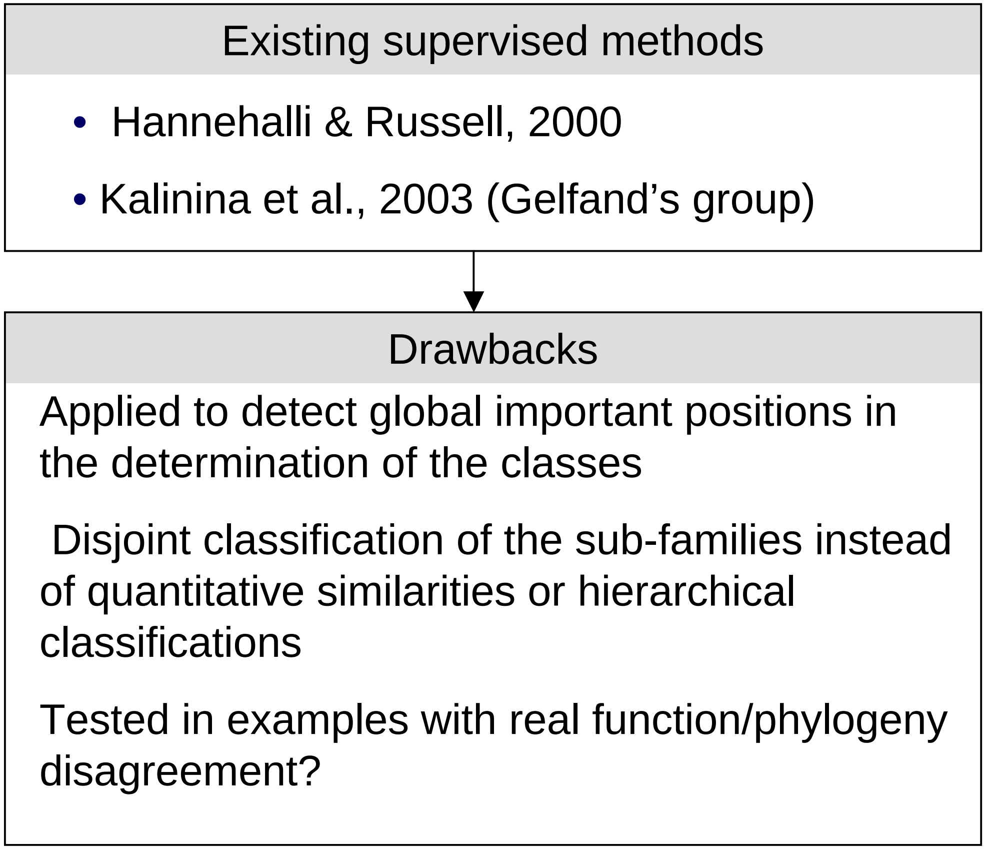
SDPred

$$I_i = \sum_{\substack{x=1..20\\ y=1..Y}} f_i(x,y)\log\frac{f_i(x,y)}{f_i(x)f(y)} \quad (1)$$



Mirny, L.A. and Gelfand, M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol*, **321**, 7-20.

Supervised methods
Based on an external functional classification



Existing supervised methods

- Hannehalli & Russell, 2000
- Kalinina et al., 2003 (Gelfand’s group)

Drawbacks

Applied to detect global important positions in the determination of the classes

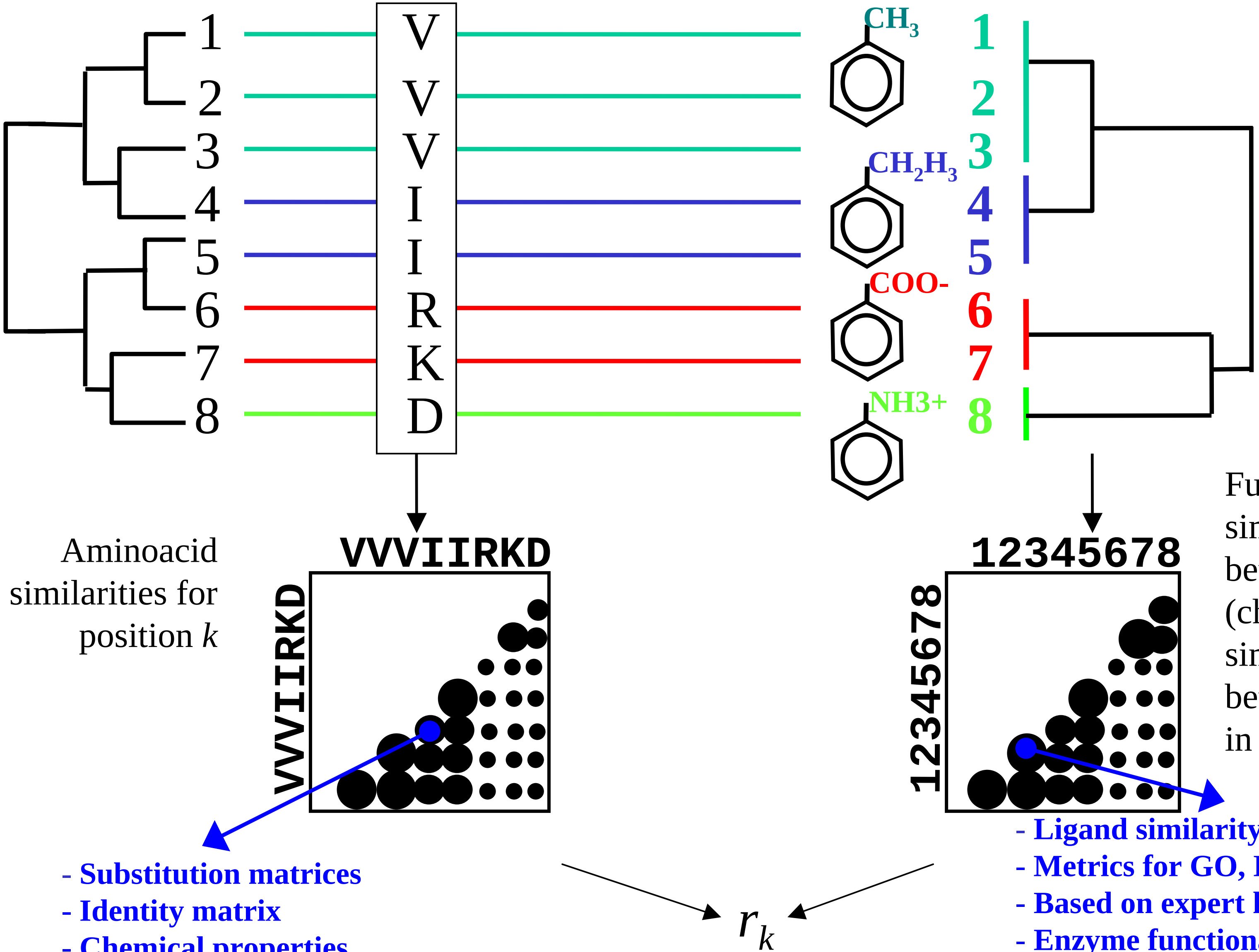
Disjoint classification of the sub-families instead of quantitative similarities or hierarchical classifications

Tested in examples with real function/phylogeny disagreement?

Hannenhalli, S.S. and Russell, R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol*, **303**, 61-76.

Kalinina, O.V., Mironov, A.A., Gelfand, M.S. and Rakhmaninova, A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*, **13**, 443-456.

Xdet



Aminoacid similarities for position *k*

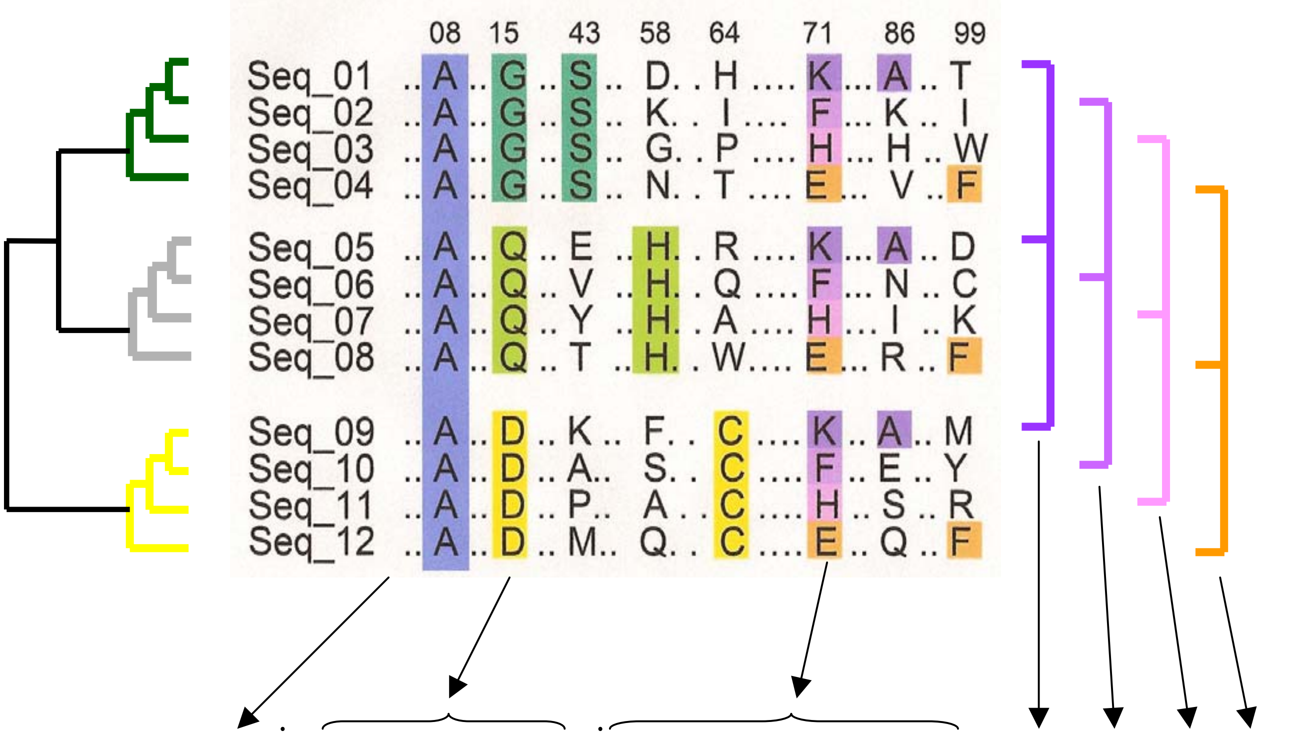
- Substitution matrices
- Identity matrix
- Chemical properties
- ...

Functional similarities between proteins (chemical similarities between ligands in this example)

- Ligand similarity
- Metrics for GO, EC, ...
- Based on expert knowledge
- Enzyme functional parameters (Kcat, Km, ...)
- Binary
- ...

Pazos, F., Rausell, A. and Valencia, A. (2006) Phylogeny-independent detection of functional residues. *Bioinformatics*., **22**, 1440-1448.

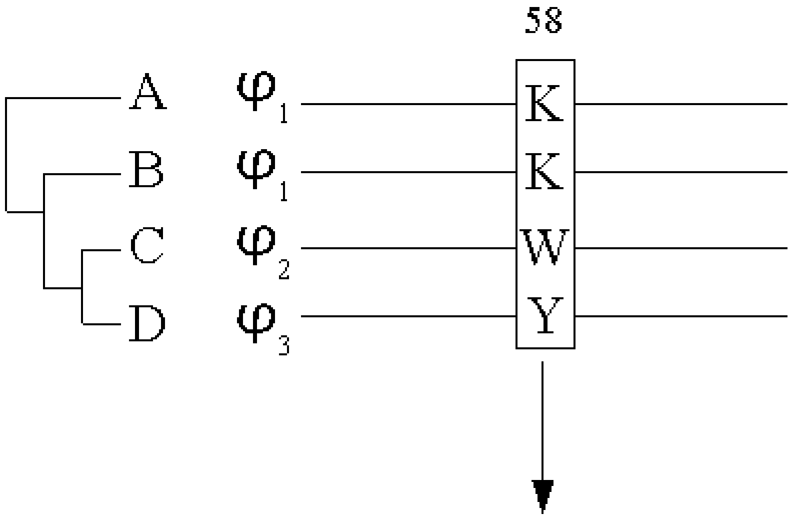
MCdet



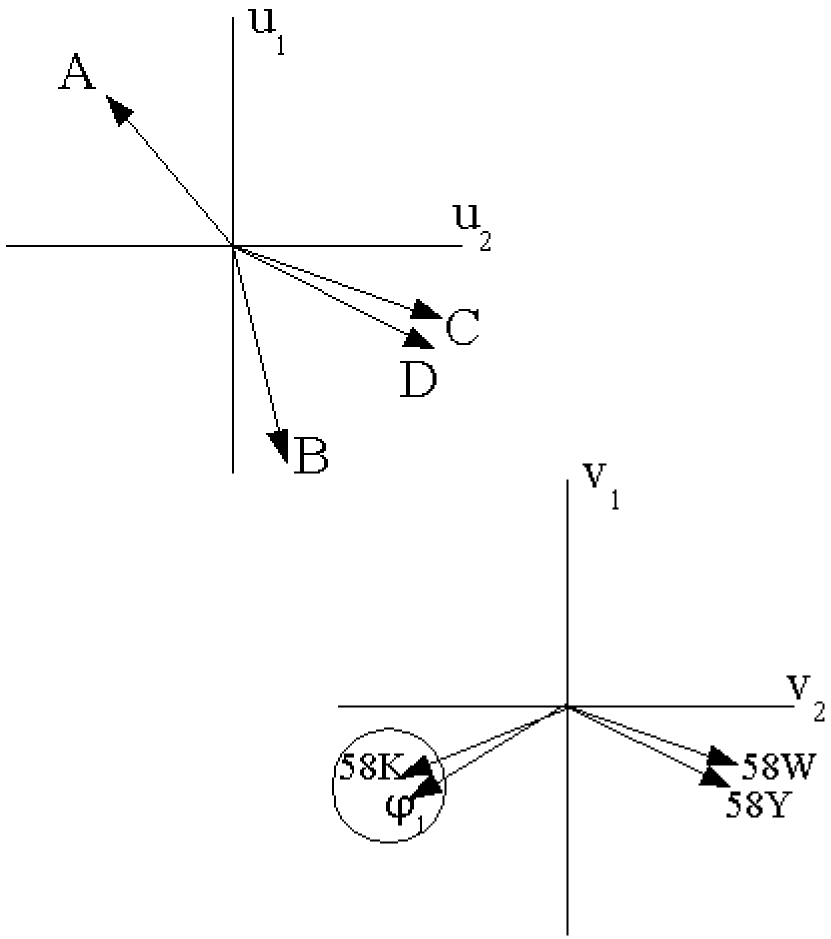
	08	15	43	58	64	71	86	99
Seq_01	A	G	S	D	H	K	A	T
Seq_02	A	G	S	K	I	F	K	I
Seq_03	A	G	S	G	P	H	H	W
Seq_04	A	G	S	N	T	E	V	F
Seq_05	A	Q	E	H	R	K	A	D
Seq_06	A	Q	V	H	Q	F	N	C
Seq_07	A	Q	Y	H	A	H	I	K
Seq_08	A	Q	T	H	W	E	R	F
Seq_09	A	D	K	F	C	K	A	M
Seq_10	A	D	A	S	C	F	E	Y
Seq_11	A	D	P	A	C	H	S	R
Seq_12	A	D	M	Q	C	E	Q	F

	8A	15G	15Q	15D	71K	71F	71H	71E	FX	FY	FZ	FW
Seq_01	1	1	0	0	1	0	0	0	1	0	0	0
Seq_02	1	1	0	0	0	1	0	0	0	1	0	0
Seq_03	1	1	0	0	0	0	1	0	0	0	1	0
Seq_04	1	1	0	0	0	0	0	1	0	0	0	1
Seq_05	1	0	1	0	1	0	0	0	1	0	0	0
Seq_06	1	0	1	0	0	1	0	0	0	1	0	0
Seq_07	1	0	1	0	0	0	1	0	0	0	1	0
Seq_08	1	0	1	0	0	0	0	1	0	0	0	1
Seq_09	1	0	0	1	1	0	0	0	1	0	0	0
Seq_10	1	0	0	1	0	1	0	0	0	1	0	0
Seq_11	1	0	0	1	0	0	1	0	0	0	1	0
Seq_12	1	0	0	1	0	0	0	1	0	0	0	1

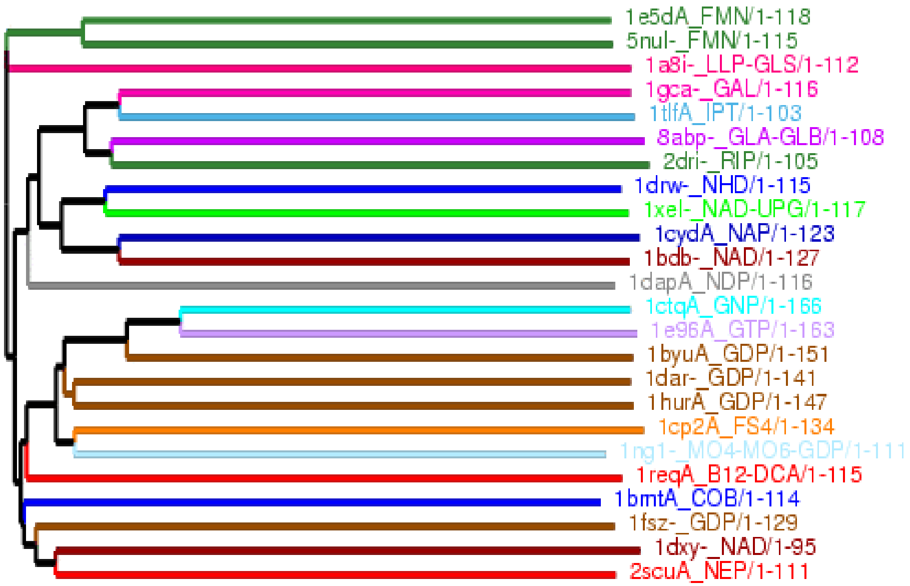
MCdet

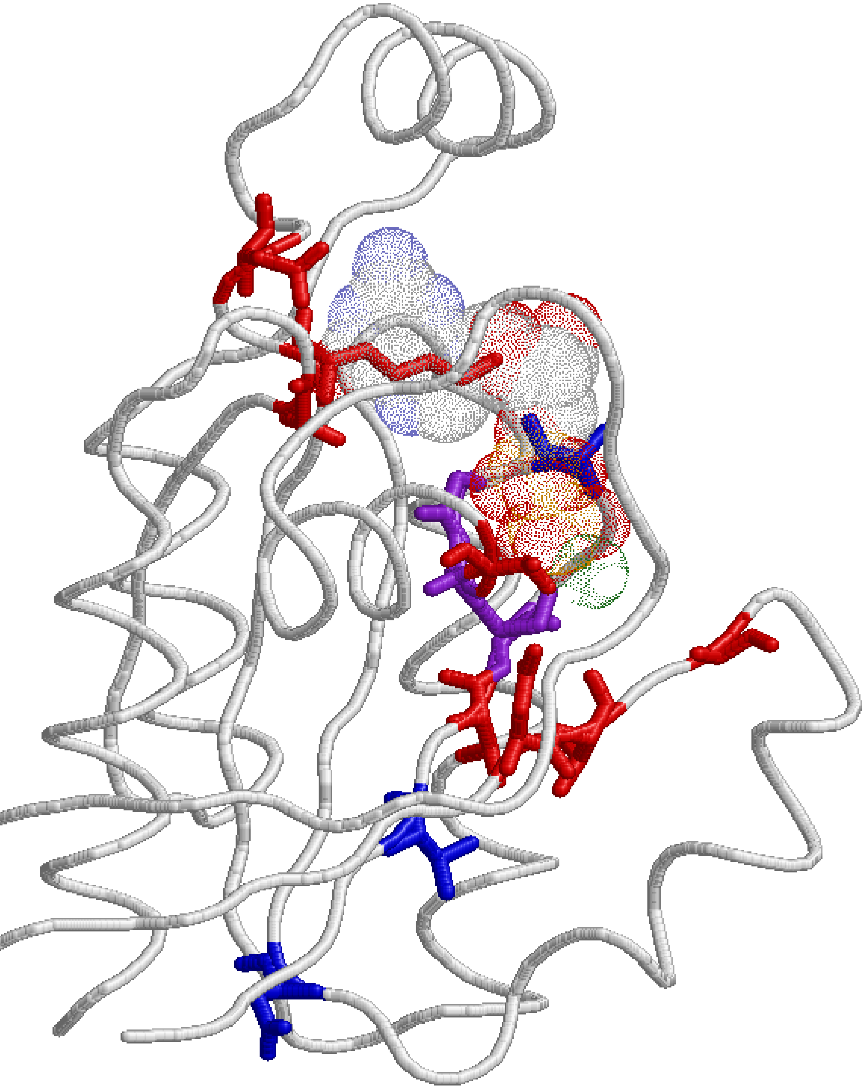


	...	58K	58W	58Y	...	φ_1
A	...	1	0	0	...	1
B	...	1	0	0	...	1
C	...	0	1	0	...	0
D	...	0	0	1	...	0



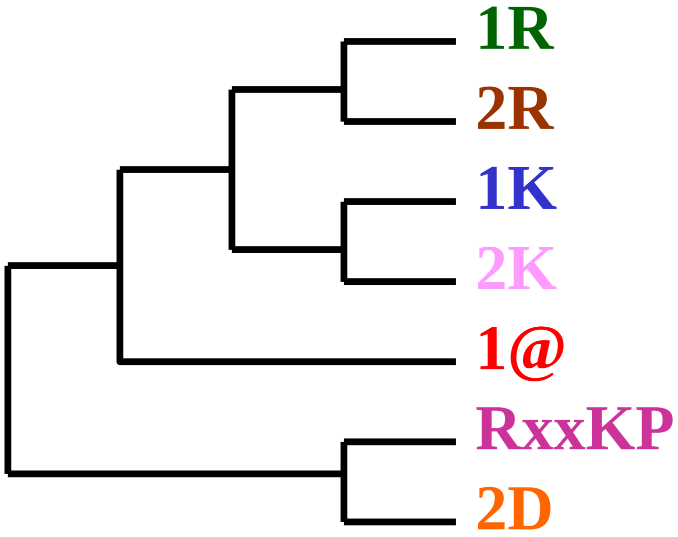
Ras-p21 structural homologs

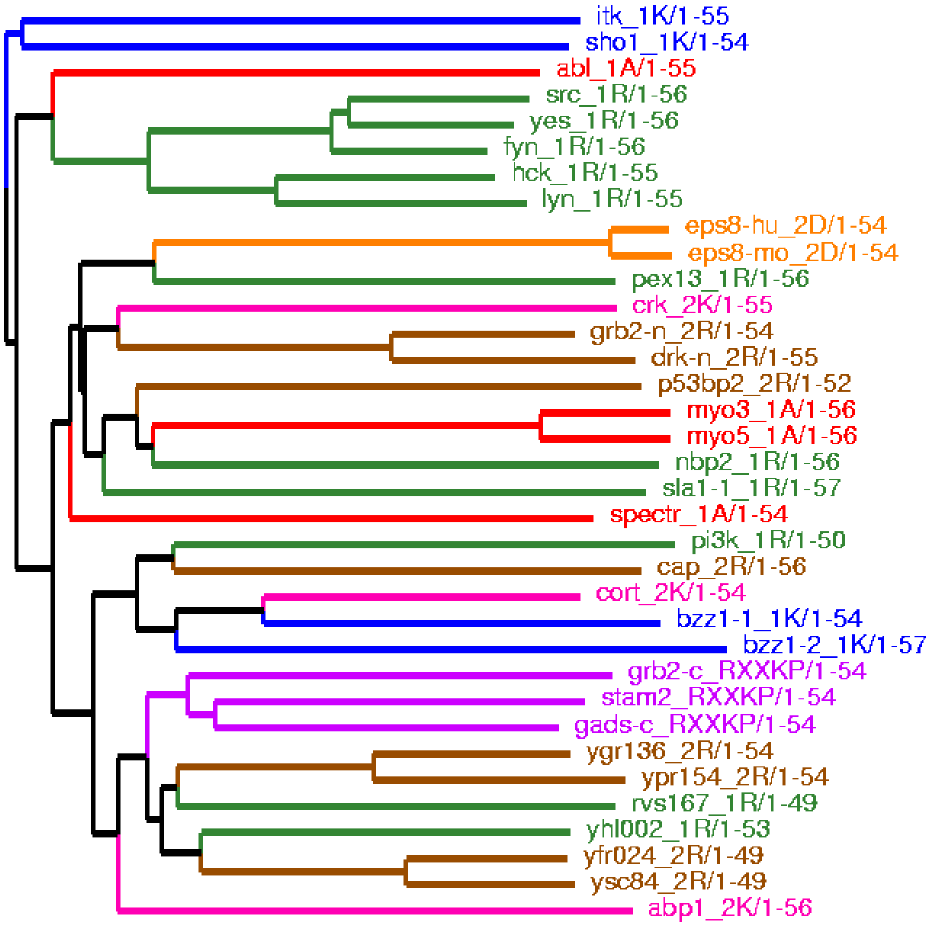




Function-phylogeny disagreement	medium
Reason	remote homology
Functional similarities (*Xdet*)	chemical similarity between ligands (Tanimoto coeficient)

SH3 domains

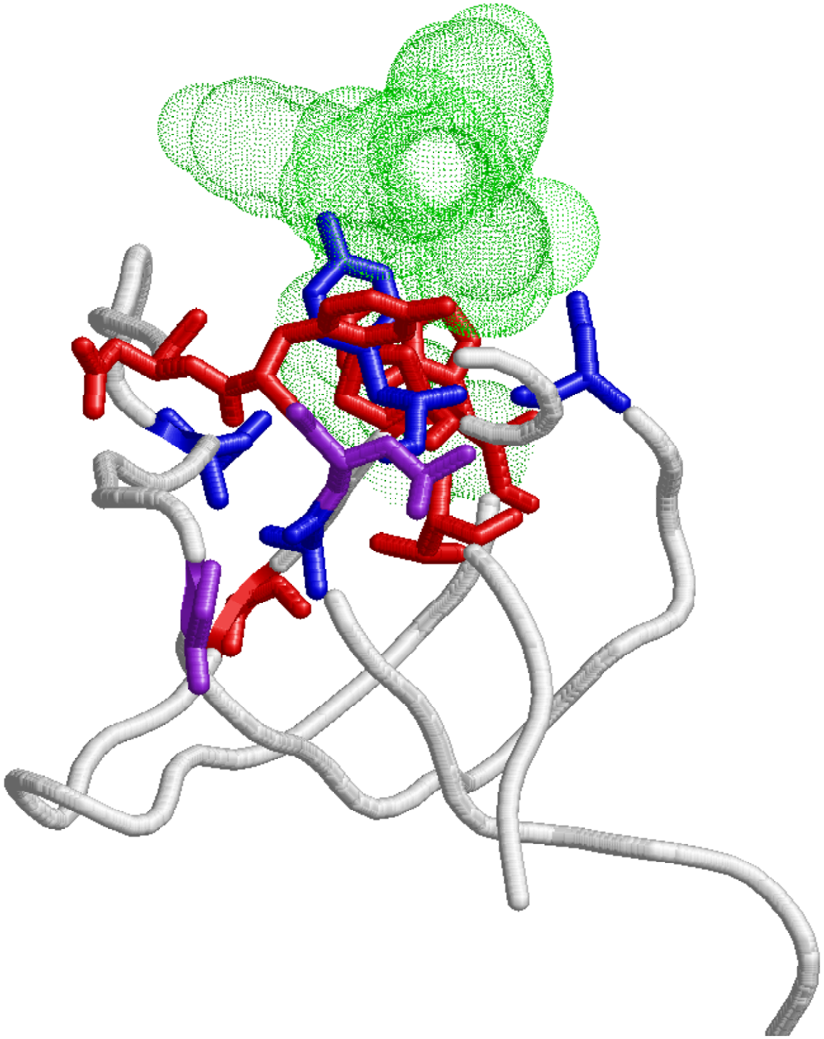


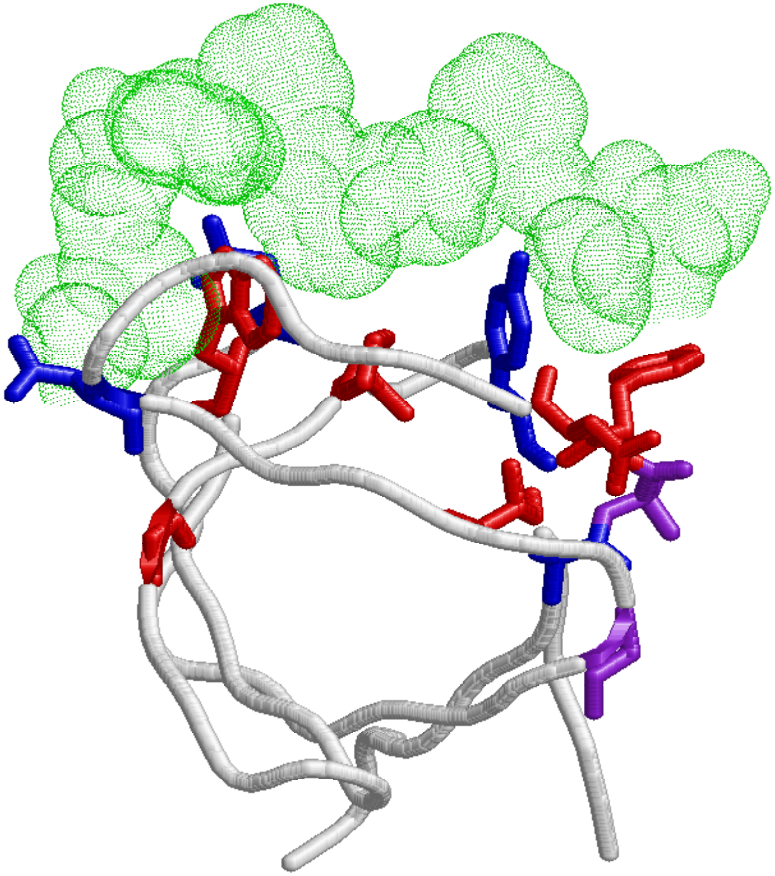


Function-phylogeny disagreement	high
Reason	complex human-based functional deffinition, remote homology
Functional similarities (*Xdet*)	quantified from the function hierarchy

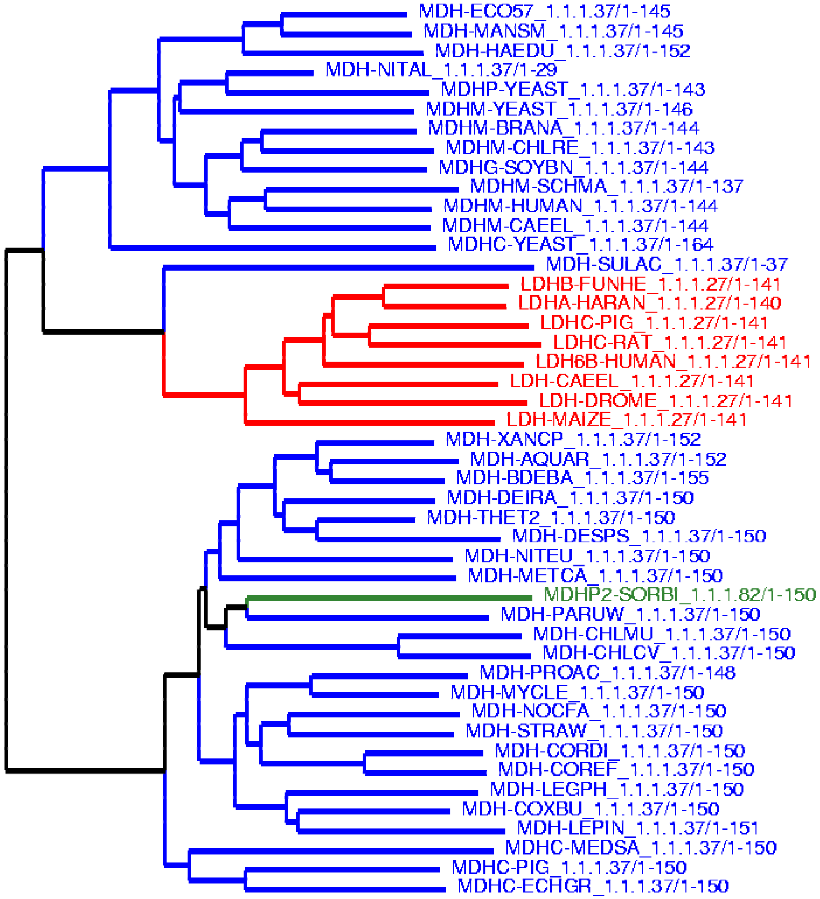
Cesareni, G., Panni, S., Nardelli, G. and Castagnoli, L. (2002) Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett.*, **513**, 38-44.

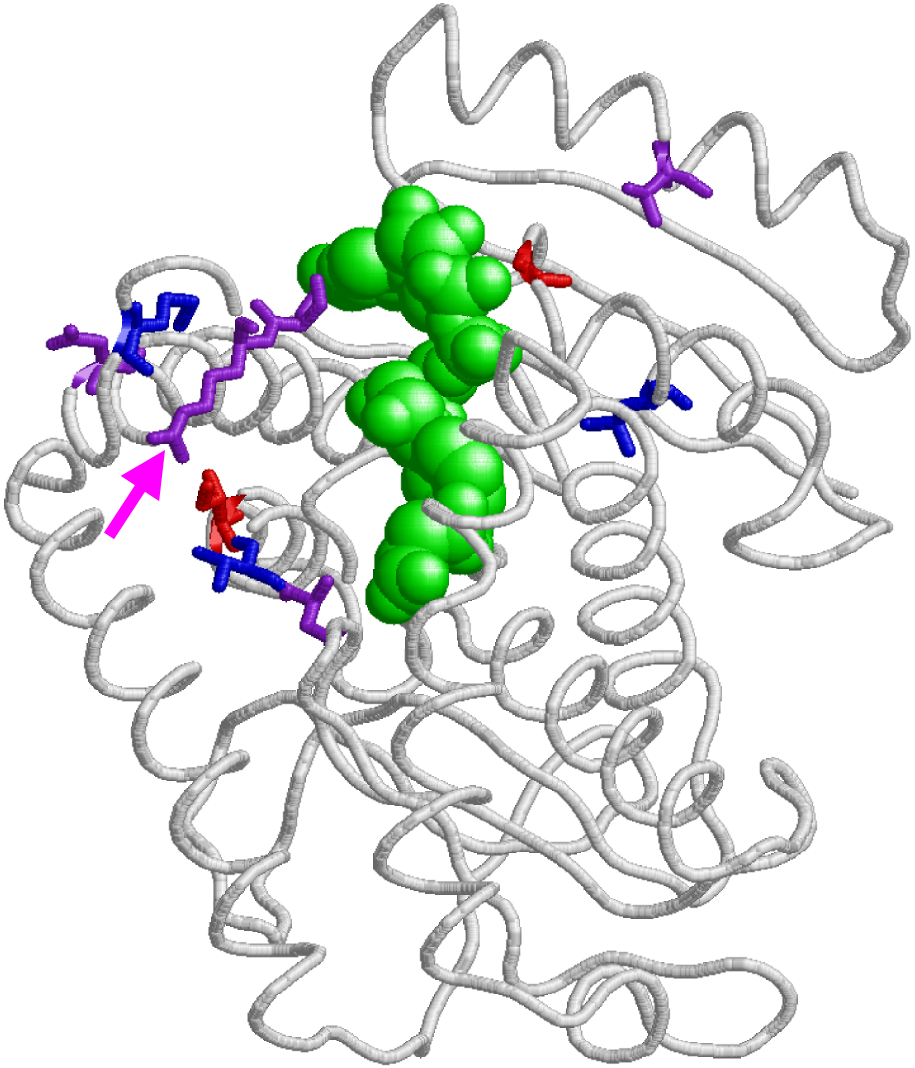
SH3 domains





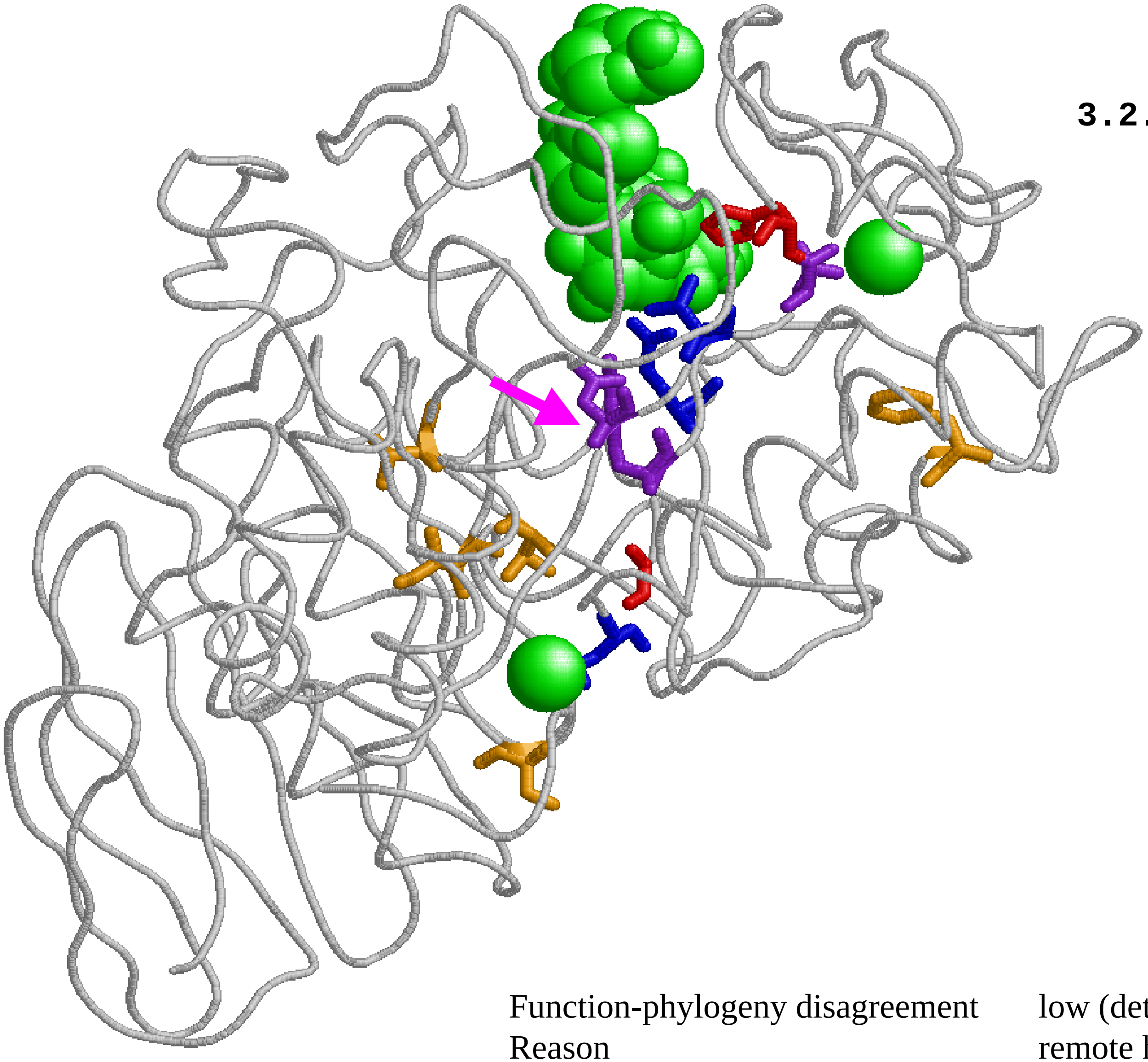
Lactate/malate dehydrogenases





Function-phylogeny disagreement	high
Reason	convergent evolution (?)
Functional similarities (*Xdet*)	binary (0/1)

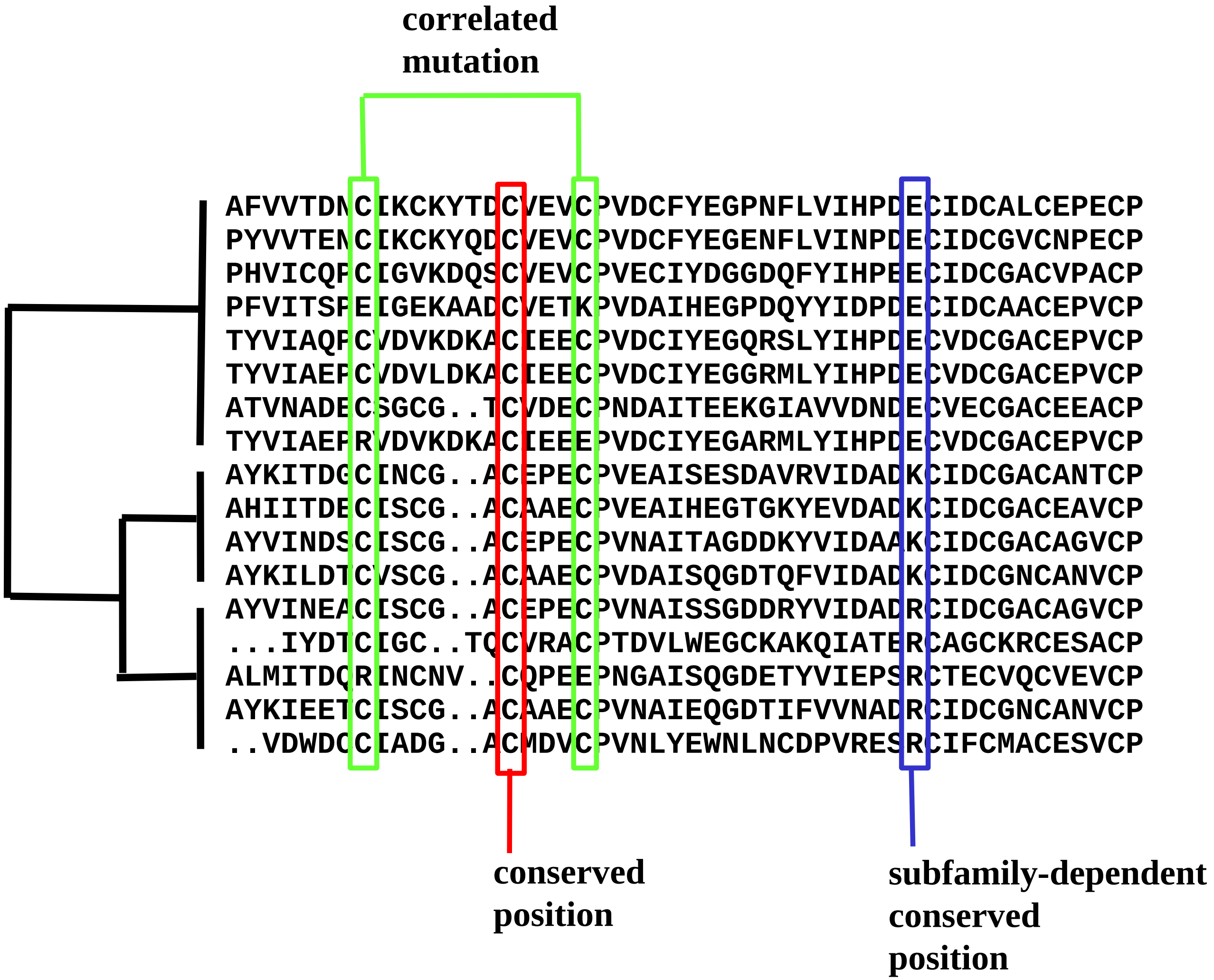
TIM-barrel hydrolases



3.2.1.*

Function-phylogeny disagreement	low (detailed distances)
Reason	remote homology
Functional similarities (*Xdet*)	binary (0/1)

Prediction of functional regions
Sequence-based methods

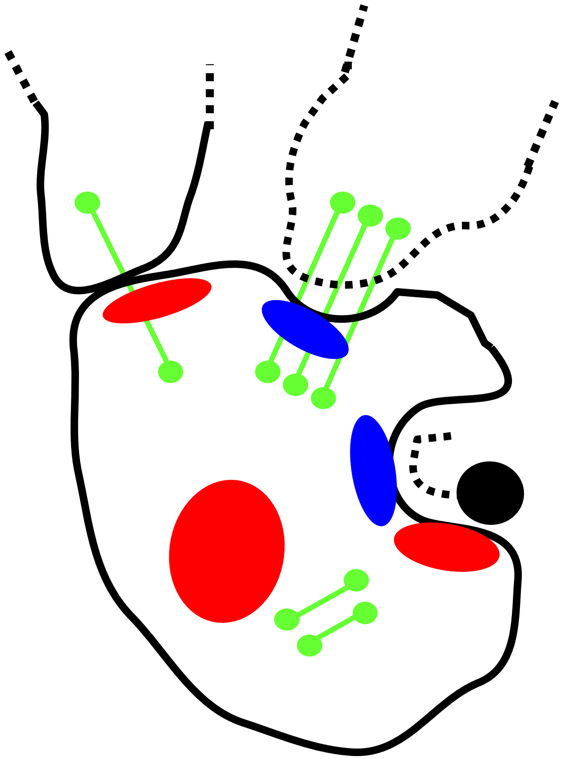


correlated mutation

```
AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECP
PYVVTENCIKCKYQDCVEVCPVDCFYEGENFLVINPDECIDCGVCNPECP
PHVICQPCIGVKDQSCVEVCPVECIYDGGDQFYIHPEECIDCGACVPACP
PFVITSPEIGEKAADCVETKPVDAIHEGPDQYYIDPDECIDCAACEPVCP
TYVIAQPCVDVKDKACIEECPVDCIYEGQRSLYIHPDECVDCGACEPVCP
TYVIAEPCVDVLDKACIEECPVDCIYEGGRMLYIHPDECVDCGACEPVCP
ATVNADECSGCG..TCVDECPNDAITEEKGIAVVDNDECVECGACEEACP
TYVIAEPRVDVKDKACIEEEPVDCIYEGARMLYIHPDECVDCGACEPVCP
AYKITDGCINCG..ACEPECPVEAISESDAVRVIDADKCIDCGACANTCP
AHIITDECISCG..ACAAECPVEAIHEGTGKYEVDADKCIDCGACEAVCP
AYVINDSCISCG..ACEPECPVNAITAGDDKYVIDAAKCIDCGACAGVCP
AYKILDTCVSCG..ACAAECPVDAISQGDTQFVIDADKCIDCGNCANVCP
AYVINEACISCG..ACEPECPVNAISSGDDRYVIDADRCIDCGACAGVCP
...IYDTCIGC..TQCVRACPTDVLWEGCKAKQIATERCAGCKRCESACP
ALMITDQRINCNV..CQPEEPNGAISQGDETYVIEPSRCTECVQCVEVCP
AYKIEETCISCG..ACAAECPVNAIEQGDTIFVVNADRCIDCGNCANVCP
..VDWDCCIADG..ACMDVCPVNLYEWNLNCDPVRESRCIFCMACESVCP
```

conserved position

subfamily-dependent conserved position



Pazos, F. and Bang, J.-W. (2006) Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics*, **1**, 15-23.