Relevant features to predict protein-protein interaction sites

Piero Fariselli

Biocomputing group

- University of Bologna, Italy -

www.biocomp.unibo.it

Two different types of data

Protein-Protein Interactions (Physical)

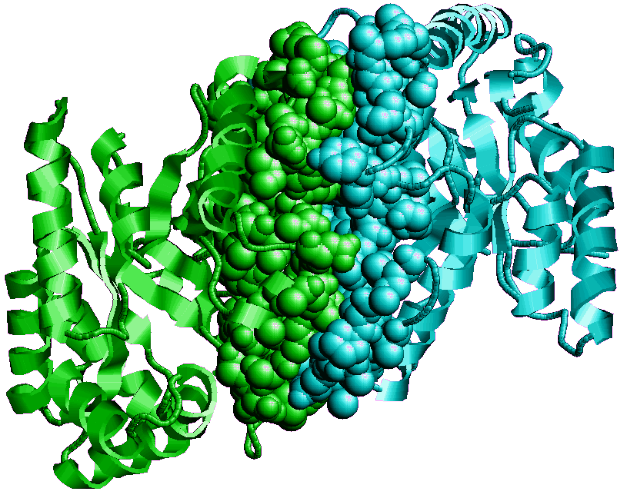
or

Protein-Protein Association Networks

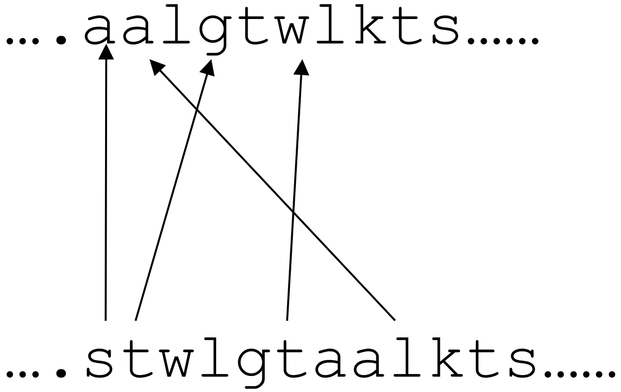
Sequence or structure ?

Atomic level

Interacting structures



Sequence level



….aalgtwlkts……

….stwlgtaalkts……

Atomic level

+ Exact location

Atomic description

- Availability of the 3D coordinates

Sequence level

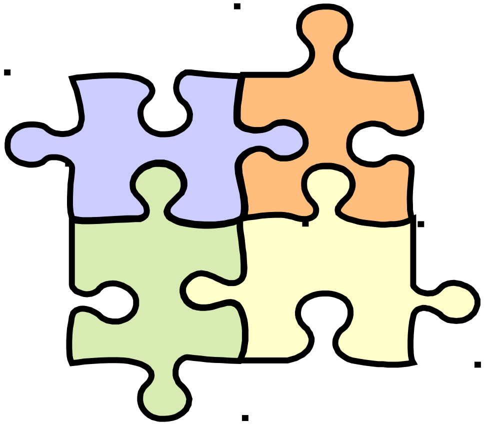
+ Whole genome computation

- No exact location

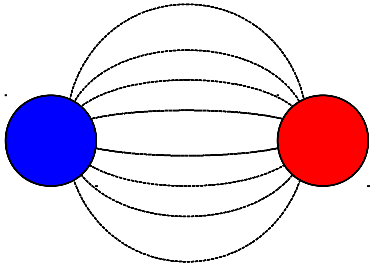
No atomic description

Structural level

Geometric criteria :
e.g. surface complementarity



Physical principles:
electrostatics, hydrophobicity



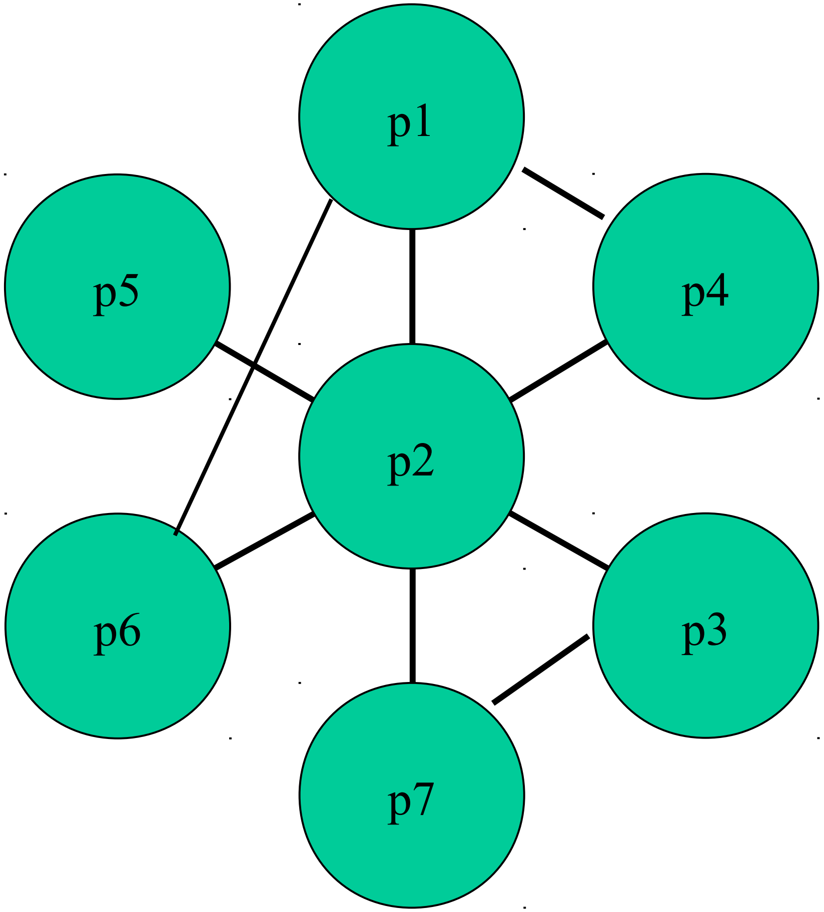
Three major problems

1. Protein-Protein interaction networks:
   given a set of proteins,
   predict the possible partners

2. Docking:
   given a pairs of proteins, known to interact,
   predict the geometry of the complex

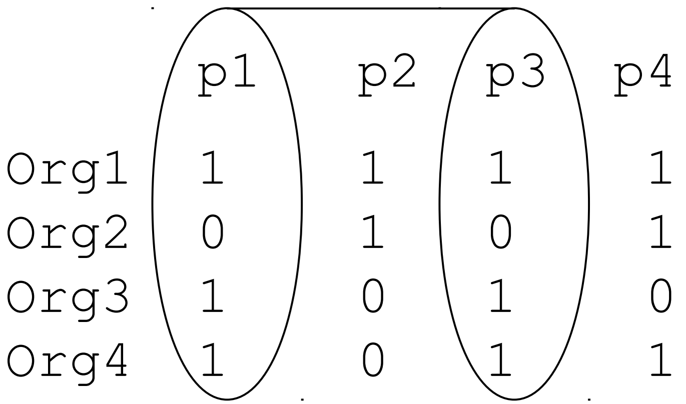
3. Protein-interaction sites:
   given a single protein,
   predict possible interacting regions

Protein-Protein interactions as undirected graph



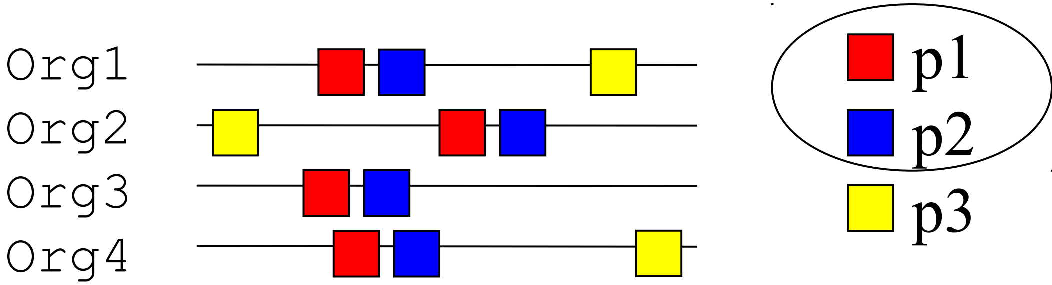
Sequence level (Historical approaches)

Phylogenetic profile

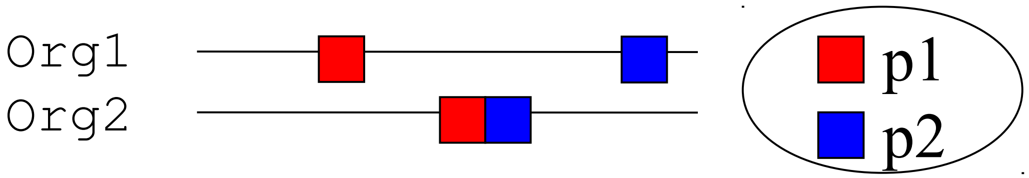


	p1	p2	p3	p4
Org1	1	1	1	1
Org2	0	1	0	1
Org3	1	0	1	0
Org4	1	0	1	1

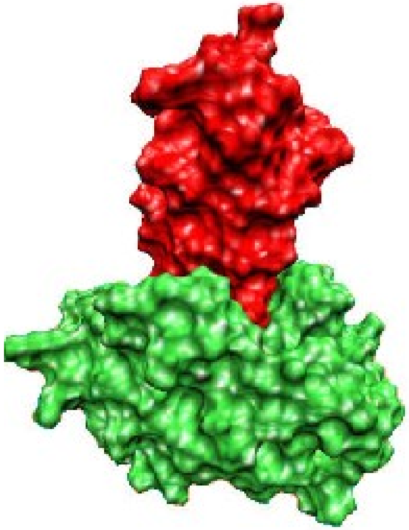
Gene neighborhood



Gene fusion

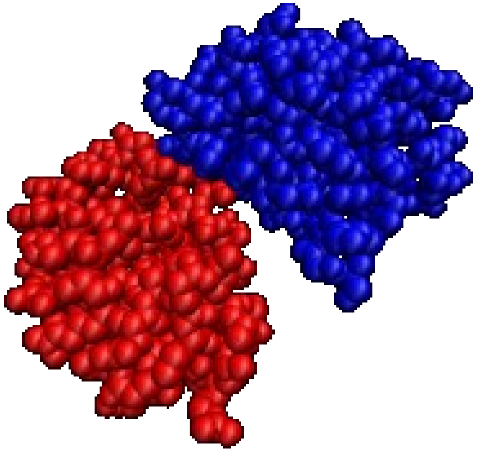


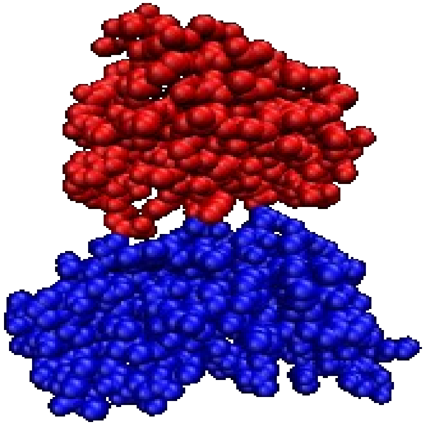
Docking

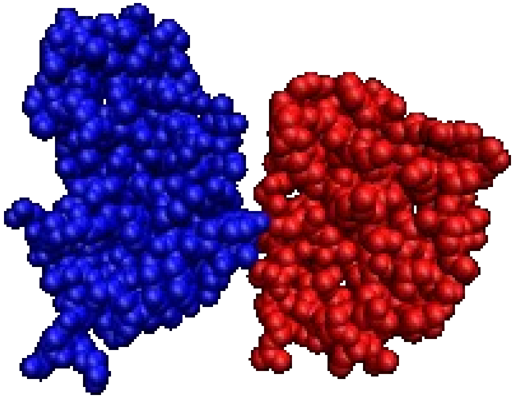


Why is this difficult?

- # of possible conformations are astronomical
  - thousands of degrees of freedom (DOF)
- Free energy changes are small
  - Below the accuracy of our energy functions
- Molecules are flexible
  - alter each other's structure as they interact







Some techniques

- **Surface representation**, that efficiently represents the docking surface and identifies the regions of interest (cavities and protrusions)
    - Connolly surface
    - Lenhoff technique
    - Kuntz et al. Clustered-Spheres
    - Alpha shapes
- **Surface matching** that matches surfaces to optimize a binding score:
    - Geometric Hashing

Surface Matching

- Find the transformation (rotation + translation) that will maximize the number of matching surface points from the receptor and the ligand

First satisfy steric constraints…

- Find the best fit of the receptor and ligand using only geometrical constraints

… then use energy calculations to refine the docking

- Select the fit that has the minimum energy

Some Examples of Docking Programs

- DOCK (I. D. Kuntz, UCSF)
- FTDOCK (Gabb, Jackson, Sternberg)
- AutoDOCK (A. Olson, Scripps)
- RosettaDOCK (Baker, U Wash., Gray, JHU)

CAPRI (Critical Assessment of PRediction of Interactions)

- Several editions (Fourth 2009-2013)
- Blind experiment. (CAPRI such as CASP).
- Some points
- (rosettadesigngroup.com/blog/535/capri-state-of-protein-protein-docking/):

Easy targets are easy

- – No dramatic improvement since the last evaluation
- – Hierarchical approach for docking

Problem 2-3: prediction of the interacting surfaces using correlated mutations

Correlated mutations

N sequences

Multiple sequence alignment

```
1 MVKGPGLYTDIGKKARDLLYKDYHSDKKFTISTYSPTGVAITSS
2 MVKGPGLYSDIGKRARDLLYRDYQSDHKFTLTTYTANGVAITST
3 MVKGPGLYTEIGKKARDLLYRDYQGDQKFSVTTYSSTGVAITTT
```

i *j*

M = *N*·(*N*-1)/2 couples

```
1 MVKGPGLYTDIGKKARDLLYKDYHSDKKFTISTYSPTGVAITSS
2 MVKGPGLYSDIGKRARDLLYRDYQSDHKFTLTTYTANGVAITST

1 MVKGPGLYTDIGKKARDLLYKDYHSDKKFTISTYSPTGVAITSS
3 MVKGPGLYTEIGKKARDLLYRDYQGDQKFSVTTYSSTGVAITTT

2 MVKGPGLYSDIGKRARDLLYRDYQSDHKFTLTTYTANGVAITST
3 MVKGPGLYTEIGKKARDLLYRDYQGDQKFSVTTYSSTGVAITTT
```

S : McLachlan substitution matrix

M-valued vectors:

$$V_i \begin{bmatrix} \boldsymbol{S}(T;S) \\ \boldsymbol{S}(T;T) \\ \boldsymbol{S}(S;T) \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{S}(I;L) \\ \boldsymbol{S}(I;V) \\ \boldsymbol{S}(L;V) \end{bmatrix} V_j$$

Correlation:

$$C_{ij}=\frac{1}{M}\cdot\sum_{k=1}^{M}\frac{\left(V_i(k)-\langle V_i\rangle\right)\cdot\left(V_j(k)-\langle V_j\rangle\right)}{\sigma\left(V_i\right)\cdot\sigma\left(V_j\right)}$$

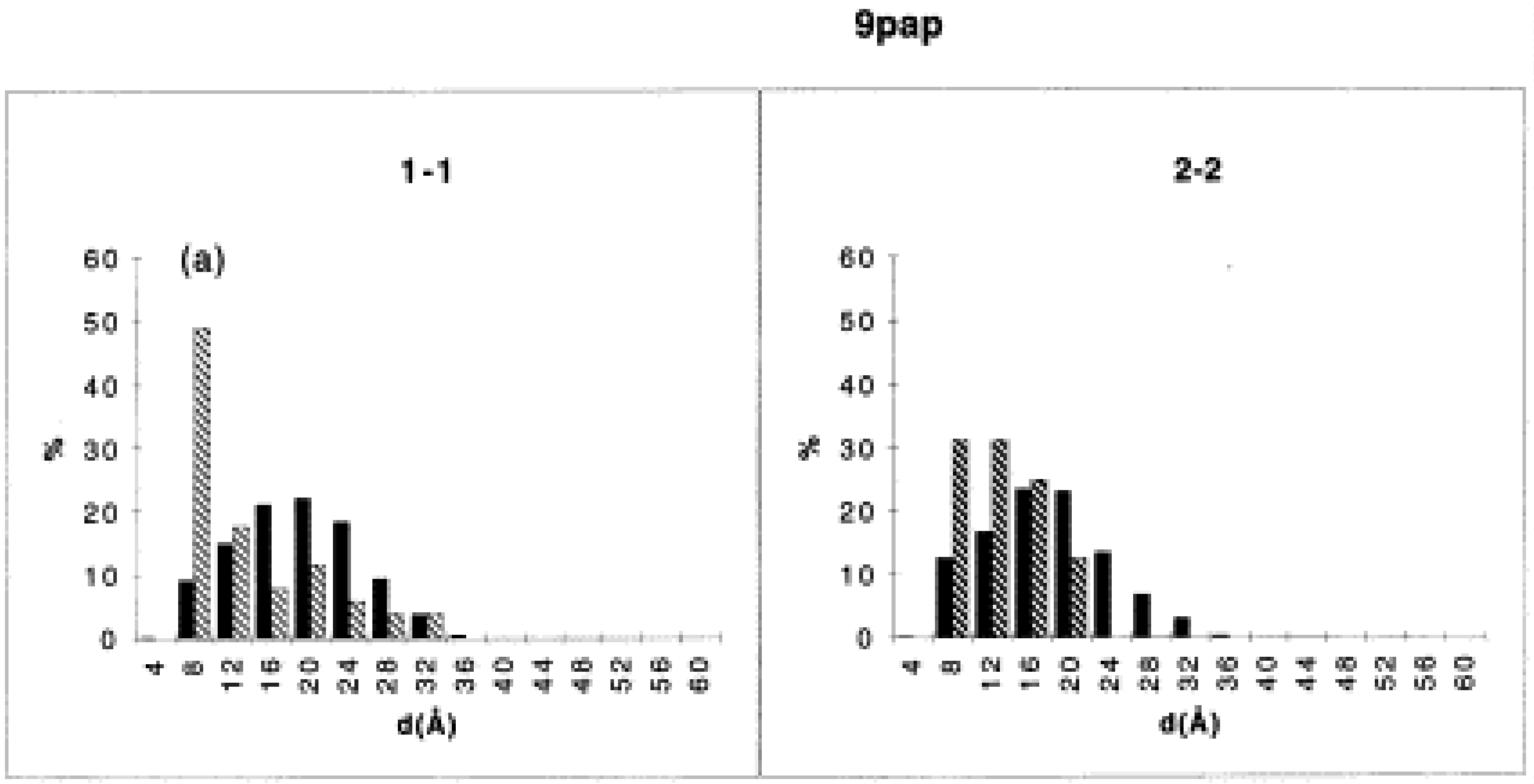


Figure 1. Bar diagrams comparing the proportions of pairs of residues at different distances. Distributions are represented for all residues (filled bars) and for correlated pairs of residues (hatched bars) in papain (9pap). (a) Distances between pairs in the ...

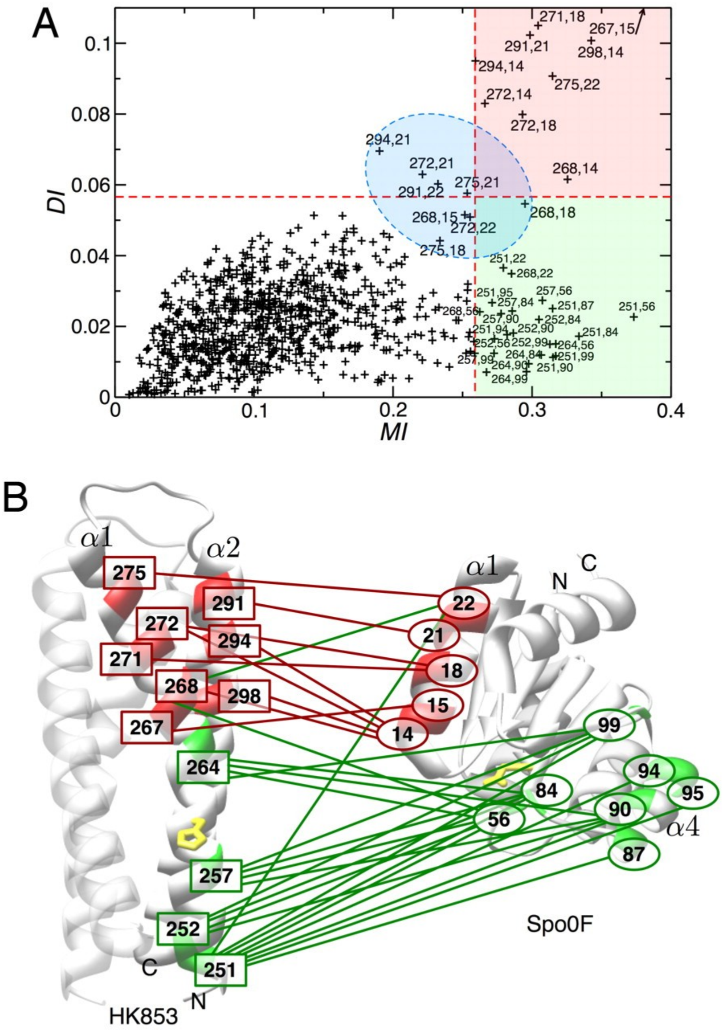
Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, Alfonso Valencia

Correlated mutations contain information about protein-protein interaction 1

Journal of Molecular Biology, Volume 271, Issue 4, 1997, 511–523

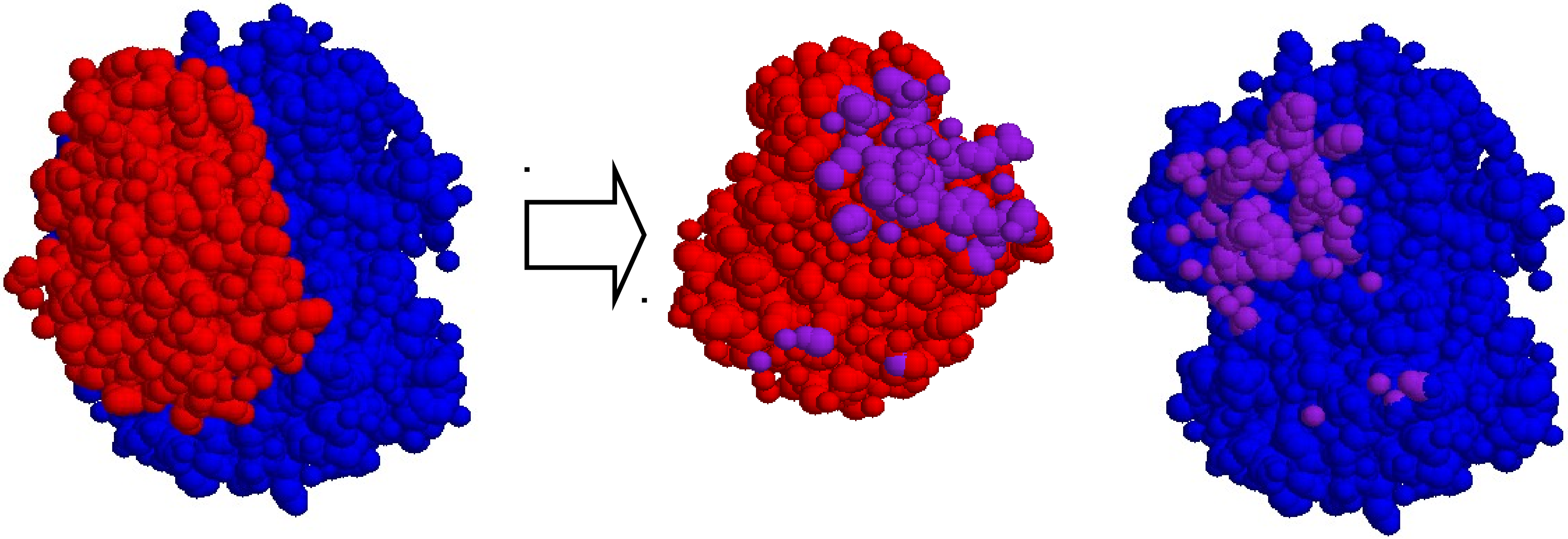
http://dx.doi.org/10.1006/jmbi.1997.1198

The combined covariance/message-passing approach detects 2 groups of correlated pairs.



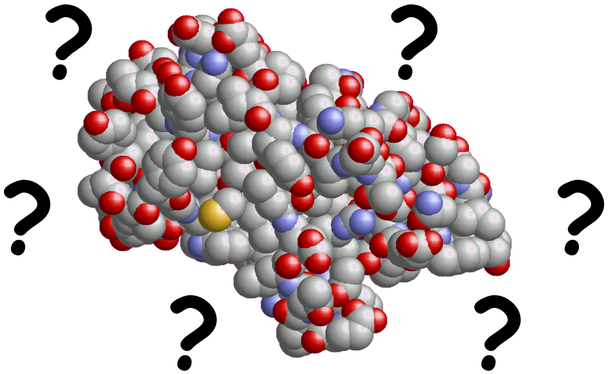
Martin Weigt et al. PNAS 2009;106:67-72

Prediction of protein-interaction sites



Prediction of Protein Interaction sites(*): Zen-Dock view

What is the sound of one hand clapping?

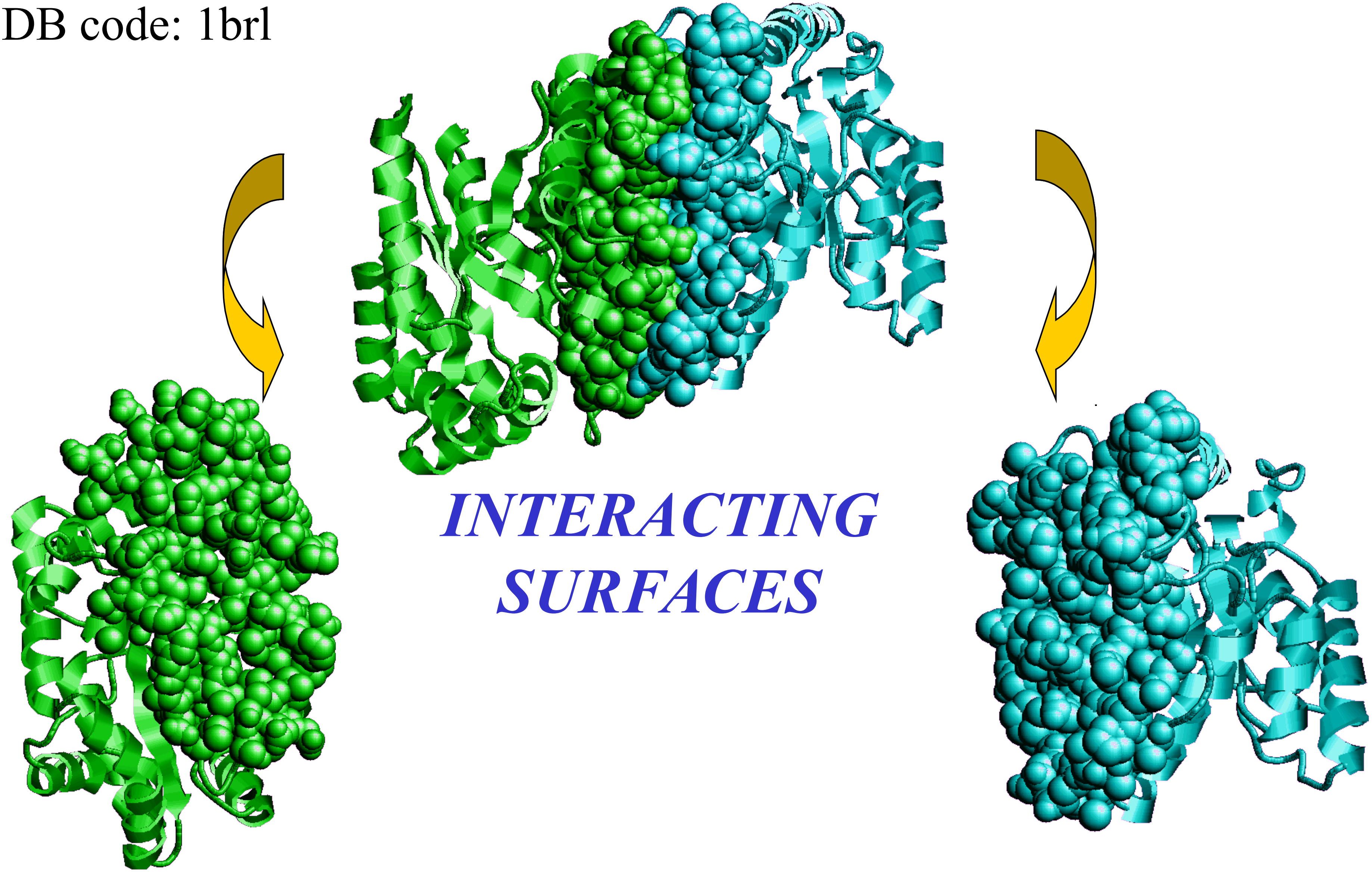


Can we predict IF and WHERE
a protein interact without knowing the
protein partners?

(*) *Alfonso Valencia's idea*

Interacting Proteins

Bacterial luciferase (*Vibrio harveyi*)
PDB code: 1brl



INTERACTING SURFACES

Definition of interacting interface

- ***Difference in Accessible Surface Area (ASA) between monomers and complex***

- ***Round Patches (as 1 but with smooth contour)***

- ***Distance between CA-atoms (e.g. 1.2nm)***

- ***All interactions (using all the available interacting chains)***

- ***Pairwise interactions (using only the largest interacting surface)***

How measure the performance?

$$Q_2 = \frac{\text{correct predictions}}{\text{total predictions}} = \frac{p+n}{N}$$

$$Q(x) = \frac{\text{correct predictions in class } x}{\text{total observations in class } x} = \frac{p}{p+u}$$

$$P(x) = \frac{\text{correct predictions in class } x}{\text{total predictions in class } x} = \frac{p}{p+o}$$

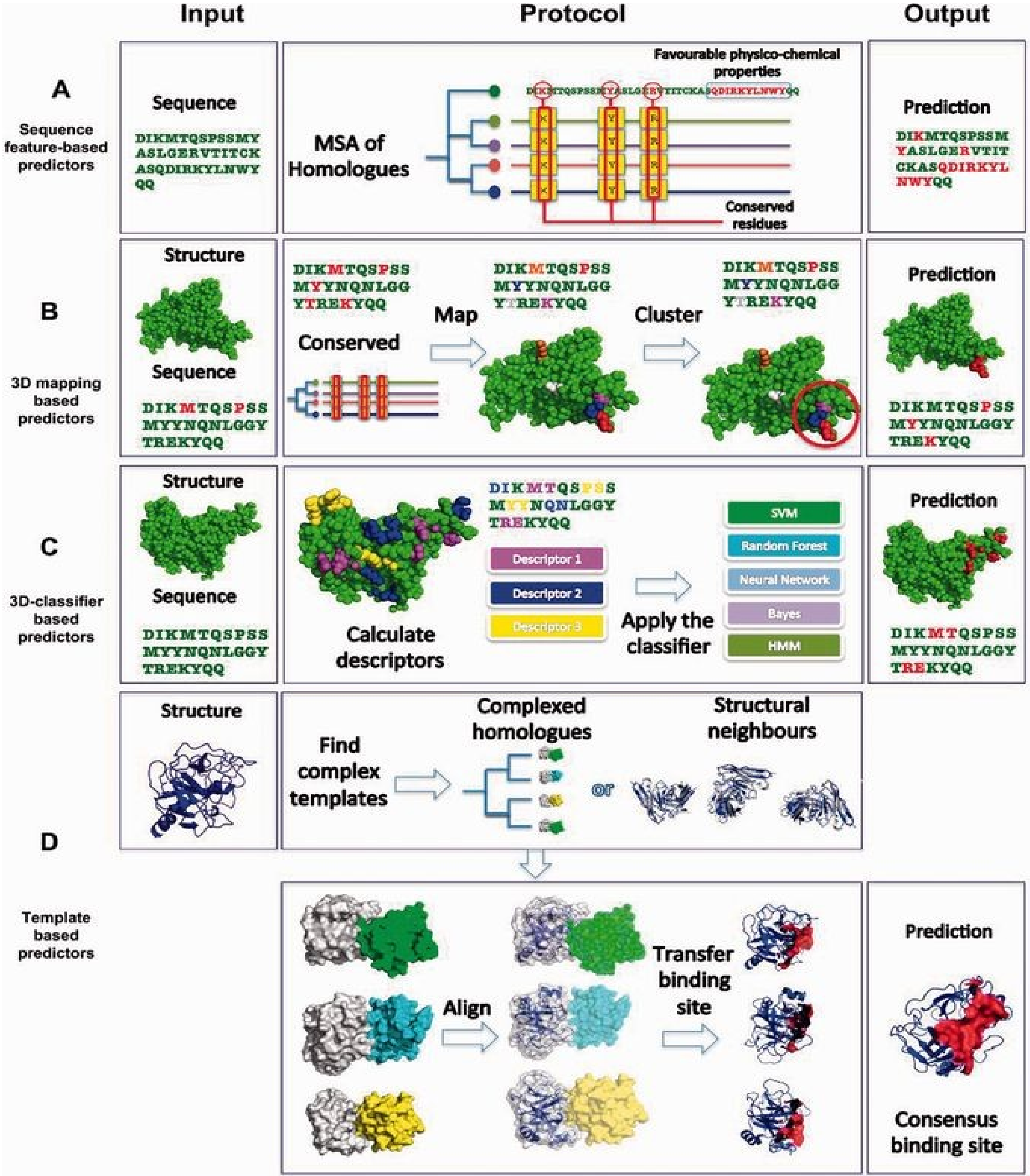
$$C = \textbf{Correlation index} = \frac{p \cdot n - o \cdot u}{[(p+o)\cdot(p+u)\cdot(n+o)\cdot(n+u)]^{1/2}}$$

Legend:

p = true positives, n = true negatives

u = false negatives, o = false positives

Classification of existing protein interface prediction methods.



What are the features, on the protein surface, that indicate possible protein-protein interaction?

Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272: 133–43

Features:

- solvation potential,
- residue interface propensity,
- *hydrophobicity*,
- *planarity*,
- Protrusion,
- accessible surface area

Homodimers easier to predict
(more hydrophobic and with flatter surfaces)

Relevant characteristics of PPI surfaces

- ***Shape, chemical affinity and flexibility***

- ***Presence of charged and polar residues***

- ***Average hydrophobicity mainly in homodimers***

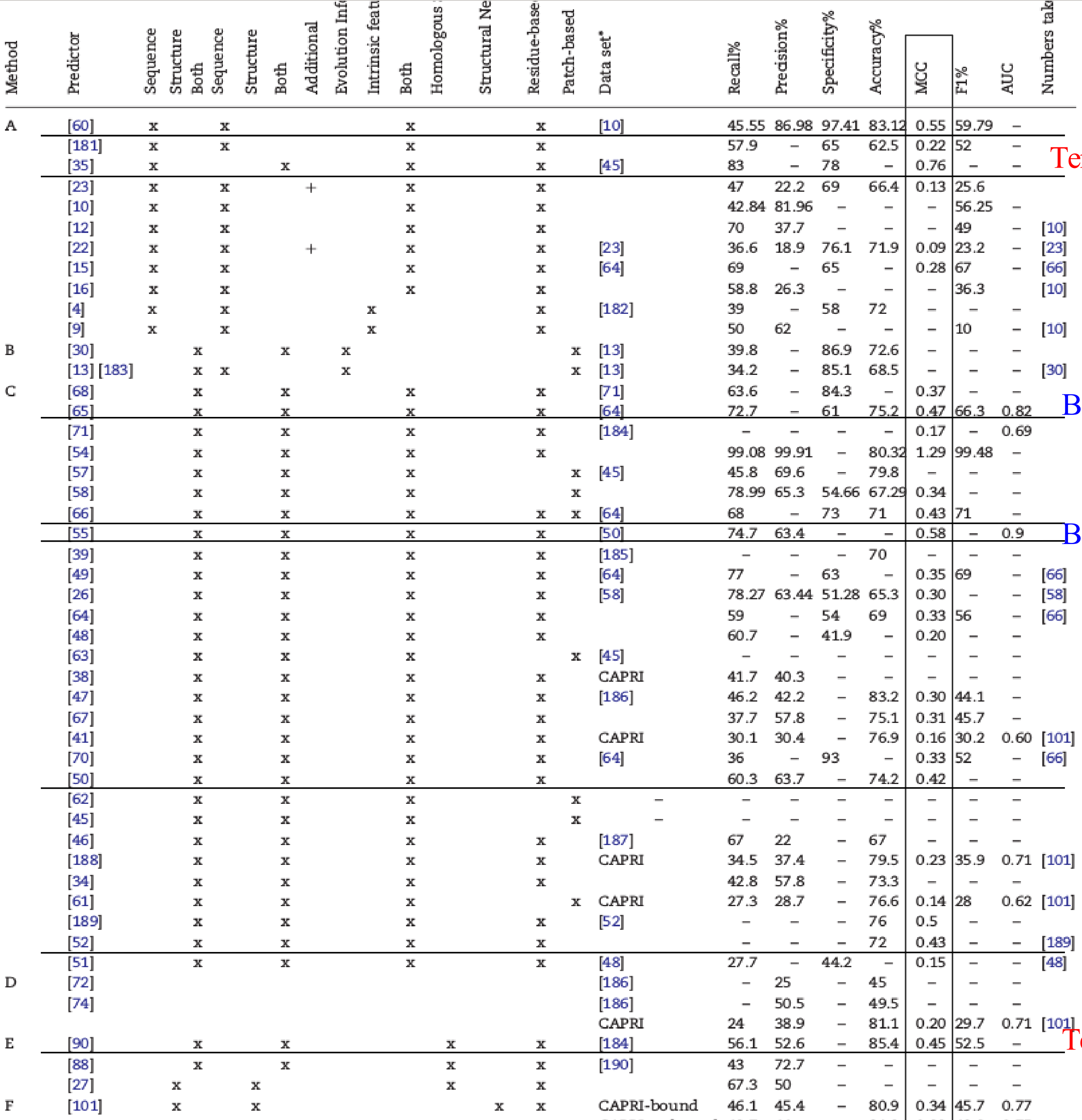
- ***Residue composition***

- ***Presence of hot-spot residues***

- ***Composition difference in different types of interaction sites (hetero/homo-obligomer, hetero/homo-transient complexes)***

However, not so easy..

Esmaielbeiki et al. Brief. in Bioinf., 2015, 1–15



Method	Predictor	Sequence	Structure	Both	Sequence	Structure	Both	Additional	Evolution Inf	Intrinsic featu	Both	Homologous	Structural Ne	Residue-base	Patch-based	Data set*	Recall%	Precision%	Specificity%	Accuracy%	MCC	F1%	AUC	Numbers tak
A	[60]	x			x						x			x		[10]	45.55	86.98	97.41	83.12	0.55	59.79	–	
	[181]	x			x						x			x			57.9	–	65	62.5	0.22	52	–	
	[35]	x					x				x			x		[45]	83	–	78	–	0.76	–	–	
	[23]	x			x			+			x			x			47	22.2	69	66.4	0.13	25.6		
	[10]	x			x						x			x			42.84	81.96	–	–	–	56.25	–	
	[12]	x			x						x			x			70	37.7	–	–	–	49	–	[10]
	[22]	x			x			+			x			x		[23]	36.6	18.9	76.1	71.9	0.09	23.2	–	[23]
	[15]	x			x						x			x		[64]	69	–	65	–	0.28	67	–	[66]
	[16]	x			x						x			x			58.8	26.3	–	–	–	36.3		[10]
	[4]	x			x					x				x		[182]	39	–	58	72	–	–	–	
	[9]	x			x					x				x			50	62	–	–	–	10	–	[10]
B	[30]			x			x		x						x	[13]	39.8	–	86.9	72.6	–	–	–	
	[13] [183]			x	x				x						x	[13]	34.2	–	85.1	68.5	–	–	–	[30]
C	[68]			x			x				x			x		[71]	63.6	–	84.3	–	0.37	–	–	
	[65]			x			x				x			x		[64]	72.7	–	61	75.2	0.47	66.3	0.82	
	[71]			x			x				x			x		[184]	–	–	–	–	0.17	–	0.69	
	[54]			x			x				x			x			99.08	99.91	–	80.32	1.29	99.48	–	
	[57]			x			x				x				x	[45]	45.8	69.6	–	79.8	–	–	–	
	[58]			x			x				x				x		78.99	65.3	54.66	67.29	0.34	–	–	
	[66]			x			x				x			x	x	[64]	68	–	73	71	0.43	71	–	
	[55]			x			x				x			x		[50]	74.7	63.4	–	–	0.58	–	0.9	
	[39]			x			x				x			x		[185]	–	–	–	70	–	–	–	
	[49]			x			x				x			x		[64]	77	–	63	–	0.35	69	–	[66]
	[26]			x			x				x			x		[58]	78.27	63.44	51.28	65.3	0.30	–	–	[58]
	[64]			x			x				x			x			59	–	54	69	0.33	56	–	[66]
	[48]			x			x				x			x			60.7	–	41.9	–	0.20	–	–	
	[63]			x			x				x				x	[45]	–	–	–	–	–	–	–	
	[38]			x			x				x			x		CAPRI	41.7	40.3	–	–	–	–	–	
	[47]			x			x				x			x		[186]	46.2	42.2	–	83.2	0.30	44.1	–	
	[67]			x			x				x			x			37.7	57.8	–	75.1	0.31	45.7	–	
	[41]			x			x				x			x		CAPRI	30.1	30.4	–	76.9	0.16	30.2	0.60	[101]
	[70]			x			x				x			x		[64]	36	–	93	–	0.33	52	–	[66]
	[50]			x			x				x			x			60.3	63.7	–	74.2	0.42	–	–	
	[62]			x			x				x				x	–	–	–	–	–	–	–	–	
	[45]			x			x				x				x	–	–	–	–	–	–	–	–	
	[46]			x			x				x			x		[187]	67	22	–	67	–	–	–	
	[188]			x			x				x			x		CAPRI	34.5	37.4	–	79.5	0.23	35.9	0.71	[101]
	[34]			x			x				x			x			42.8	57.8	–	73.3	–	–	–	
	[61]			x			x				x				x	CAPRI	27.3	28.7	–	76.6	0.14	28	0.62	[101]
	[189]			x			x				x			x		[52]	–	–	–	76	0.5	–	–	
	[52]			x			x				x			x			–	–	–	72	0.43	–	–	[189]
	[51]			x			x				x			x		[48]	27.7	–	44.2	–	0.15	–	–	[48]
D	[72]															[186]	–	25	–	45	–	–	–	
	[74]															[186]	–	50.5	–	49.5	–	–	–	
																CAPRI	24	38.9	–	81.1	0.20	29.7	0.71	[101]
E	[90]			x			x					x		x		[184]	56.1	52.6	–	85.4	0.45	52.5	–	
	[88]			x			x					x		x		[190]	43	72.7	–	–	–	–	–	
	[27]		x			x						x		x			67.3	50	–	–	–	–	–	
F	[101]		x			x							x	x		CAPRI-bound	46.1	45.4	–	80.9	0.34	45.7	0.77	

Template

Best set

Best set

Template

First attempts

Zhou, H.X. & Shan, Y. -Prediction of protein interaction sites from sequence proofile and residue neighbor list-.
***Proteins** 44:336-343 (2001)*

Fariselli P, Pazos F, Valencia A, Casadio R -Prediction of protein--protein interaction sites in heterocomplexes with neural networks-
***Eur J Biochem** 269:1356-1361 (2002)*

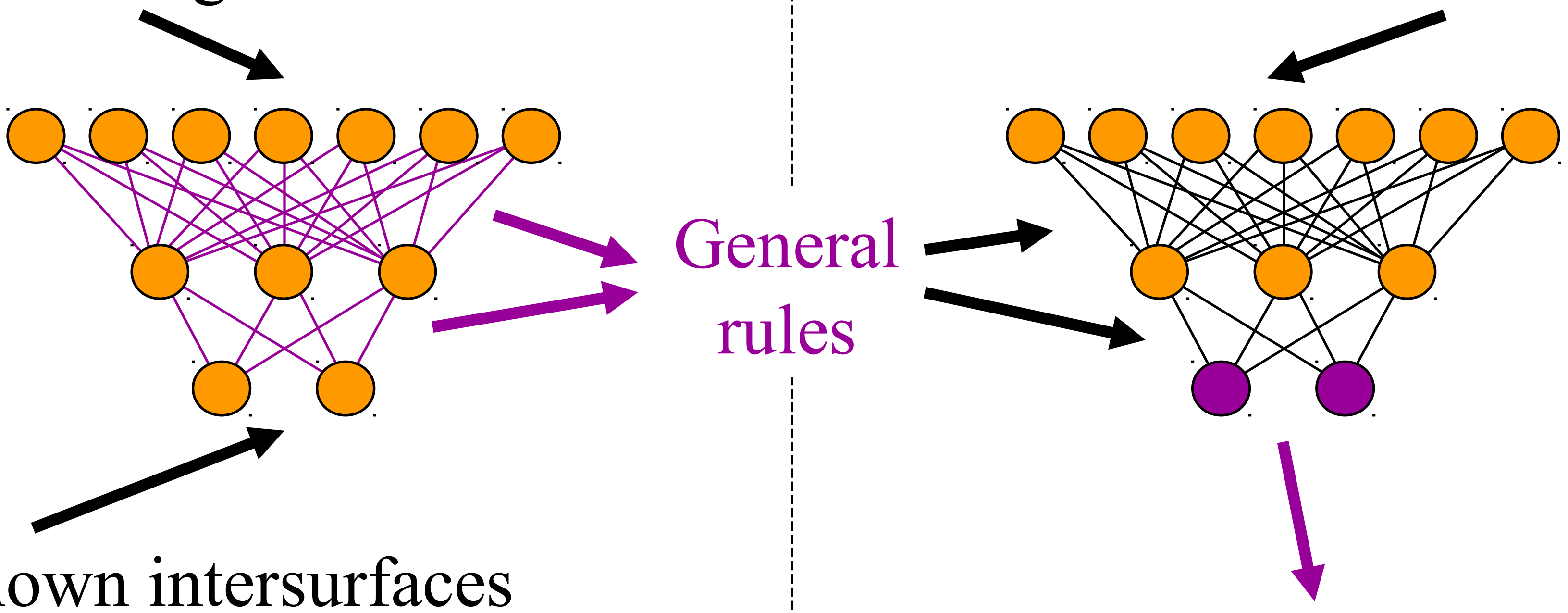
A type of problem solvable with a machine learning approach

- Available sets of data (known examples): the interacting structures and the corresponding surfaces

- No simple first principle- or model-based solution

Tools out of machine learning approaches:

Neural Networks



Training

Interacting structures

Known intersurfaces

General rules

Prediction

New structure

Prediction

The data base generation

***The SPIN-PP data base* (*)**

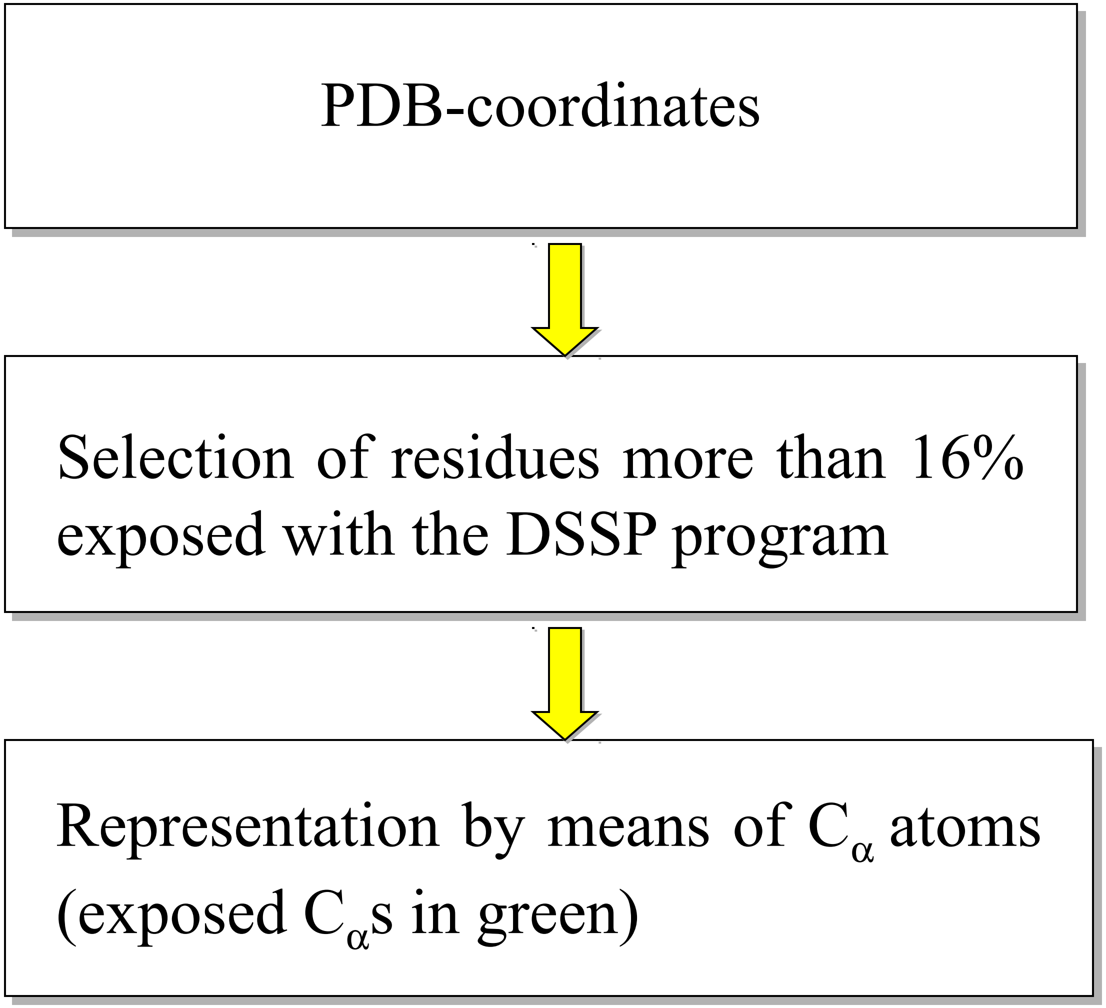
(Honig B columbia.edu)

1. **Only heterocomplexes**
2. **No proteases, no membrane proteins, no small molecules**
3. **Sequence identity ≤ 30%**

226 interacting protomers

(*) created by Florencio Pazos 2000-2001

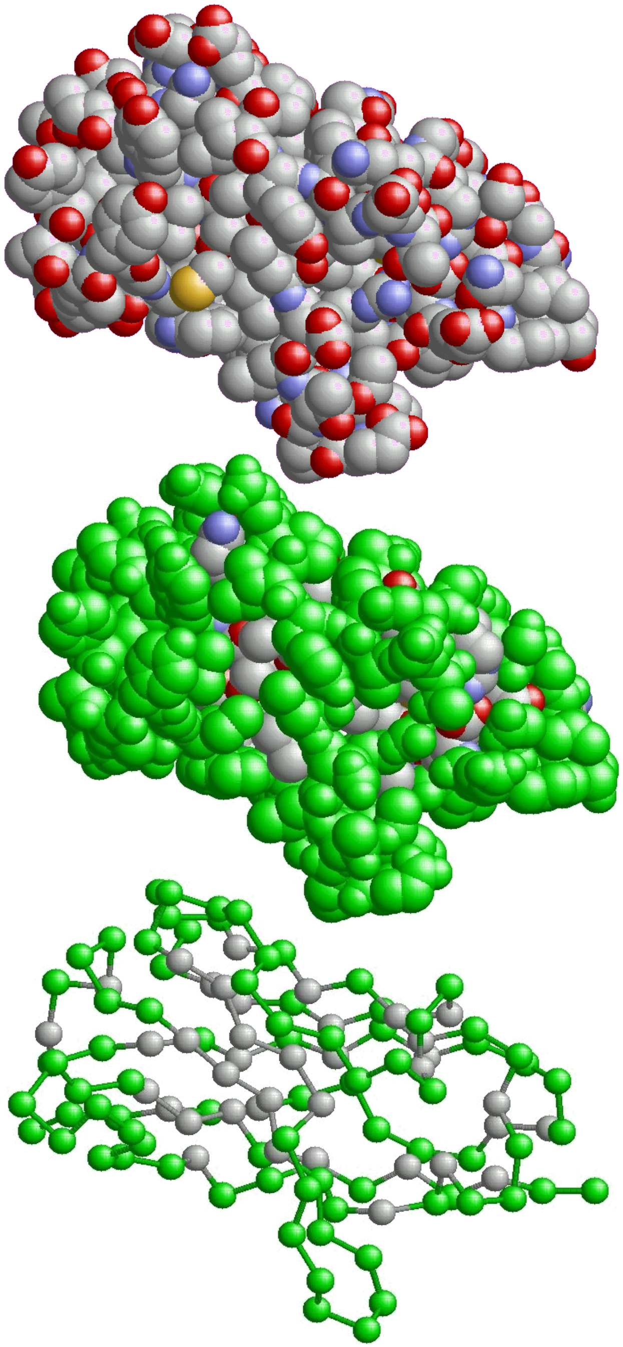
The Protein Representation



PDB-coordinates

Selection of residues more than 16% exposed with the DSSP program

Representation by means of C_α atoms (exposed C_αs in green)

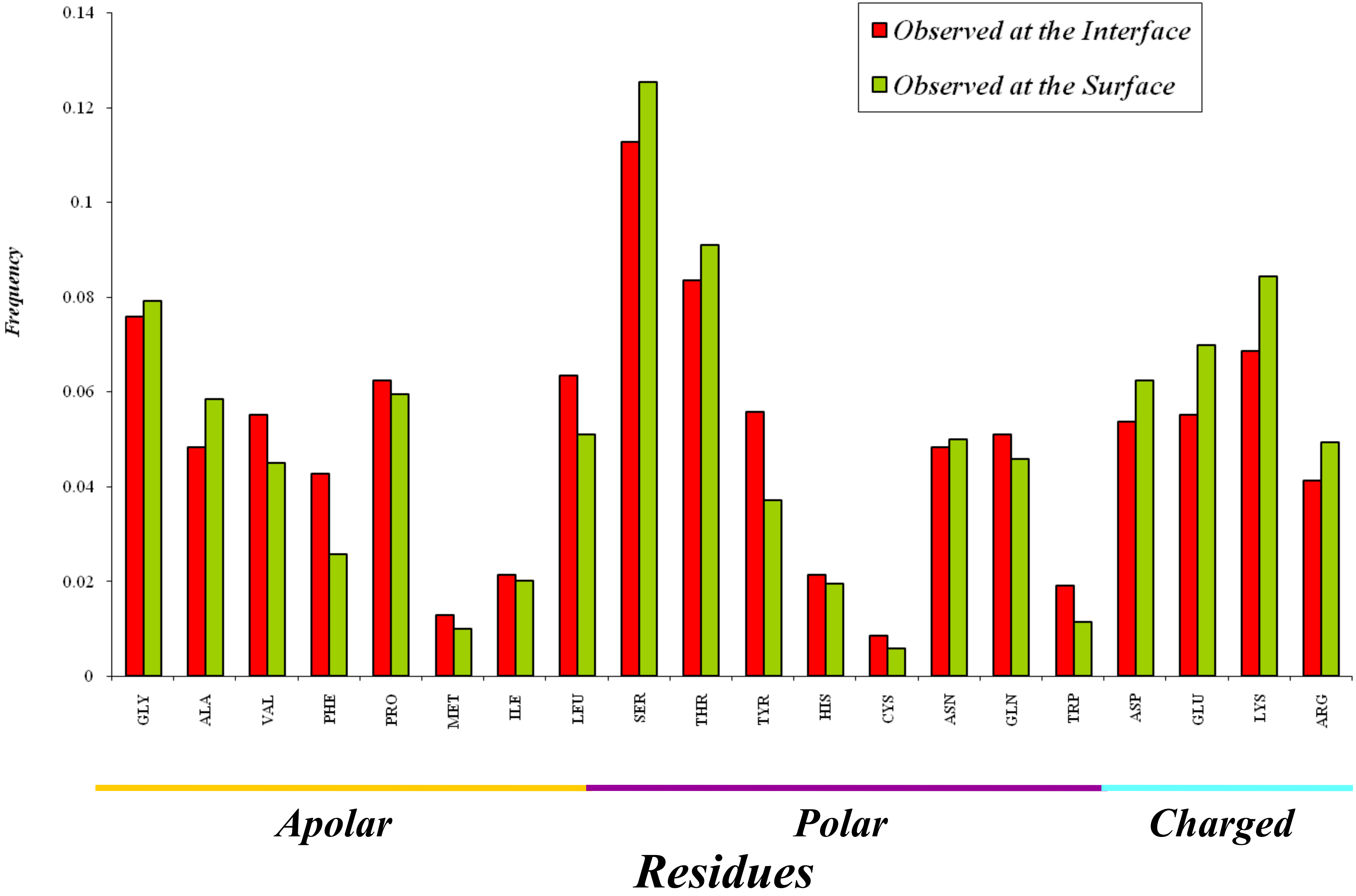


The data base

226 protein monomers

Total residues:	**67,847**
Exposed (> 16%):	**31,910**
Interaction sites:	**12,764**

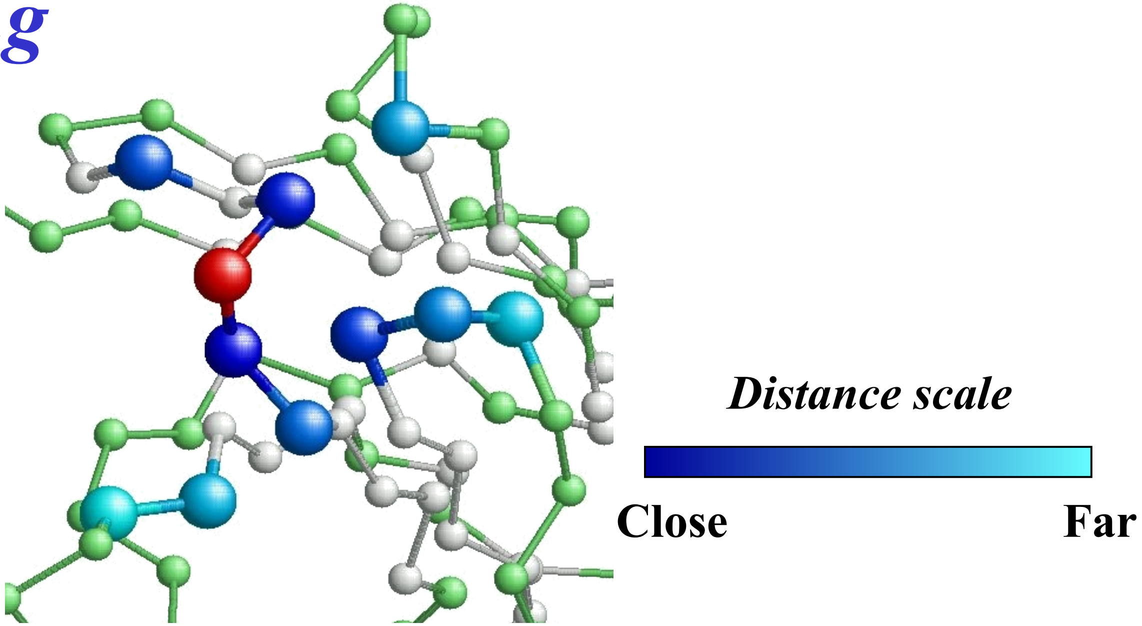
Distributions of apolar, polar and charged residues



The Input Coding

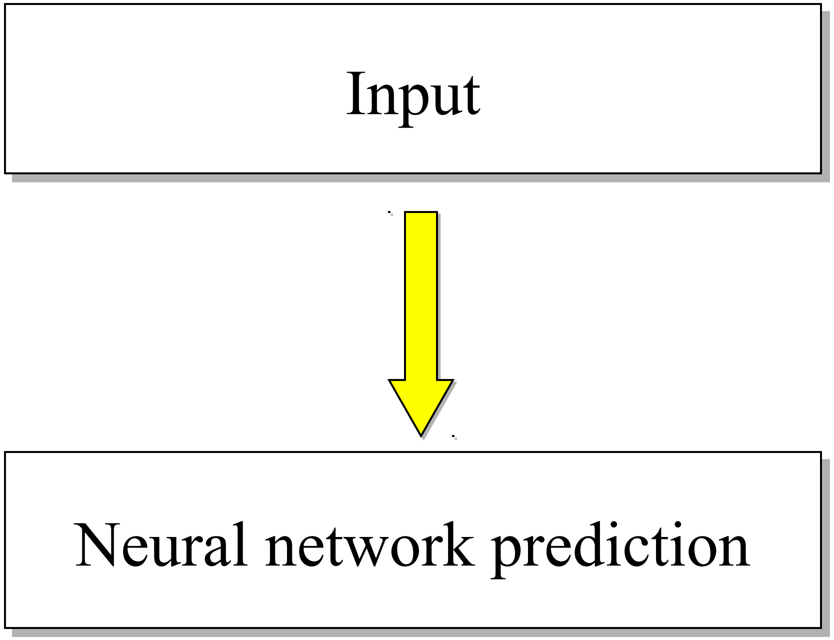
For each exposed C_α (in red) the 10 closest exposed C_αs are selected within 1.2 nm (in blue)

For each selected C_α the corresponding column of the sequence profile is extracted



A	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	70	0	0	0	0	60	0	0	0	0	20	0	0	0	10
E	0	0	0	0	0	0	0	0	0	0	0	0	70	0	0	0	10
F	0	0	0	10	0	33	0	0	0	0	0	0	0	0	0	0	0
G	10	0	30	0	30	0	100	0	0	0	0	50	0	0	0	0	30
H	0	0	0	0	10	0	0	10	30	0	0	0	0	0	0	0	0
K	0	40	0	0	0	0	0	0	10	100	70	0	0	0	0	100	0
I	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	30
M	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0
N	0	0	0	0	10	0	0	0	0	0	30	10	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	40	0	0	0	30	0	0	0	0	0	0	0	0
R	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	33	0	0	0	0	0	0	10	10	0	0	0
T	20	0	0	0	0	33	0	0	0	0	0	30	0	30	100	0	0
V	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	20
W	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	70	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0

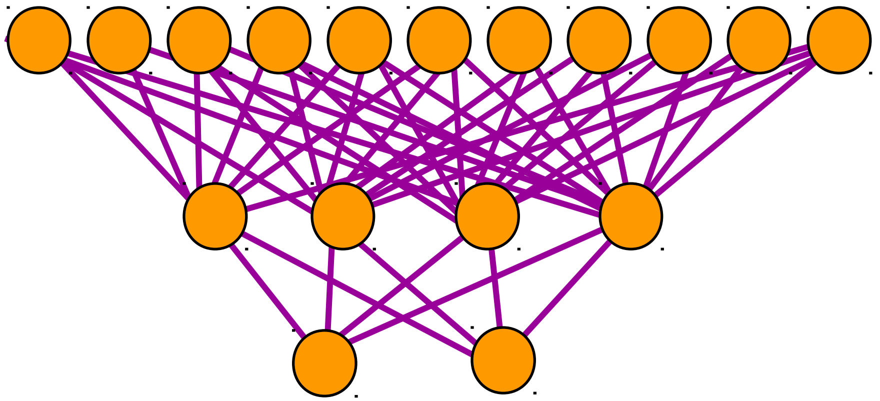
The Predictor



Input

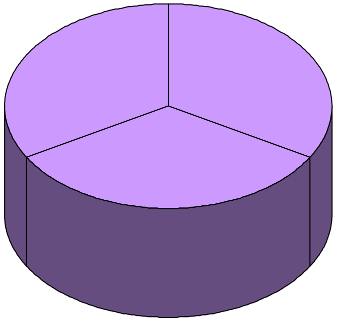
Neural network prediction

0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
60	0	0	0	0	20	10	0	0	70	0
0	0	0	0	0	70	10	0	0	0	0
0	0	0	0	10	0	0	0	33	0	0
0	0	0	30	0	0	30	0	0	30	10
10	0	0	10	0	0	0	0	0	0	0
0	100	0	0	0	0	0	40	0	0	0
0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	30	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	10	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	40	0	0	0	0	0	0	0
0	0	0	0	0	0	0	50	0	0	0
0	0	0	0	0	10	0	0	33	0	0
0	0	100	0	0	0	0	0	33	0	20
0	0	0	10	0	0	20	0	0	0	0
0	0	0	0	0	0	0	10	0	0	0
0	0	0	0	90	0	0	0	0	0	70



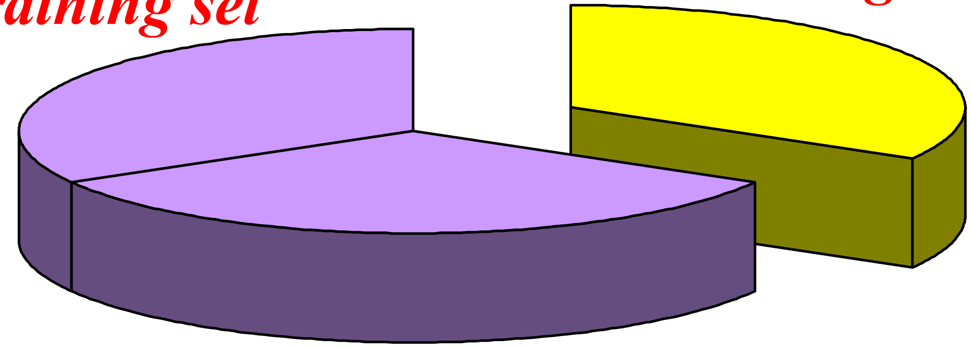
The cross validation procedure

Protein set

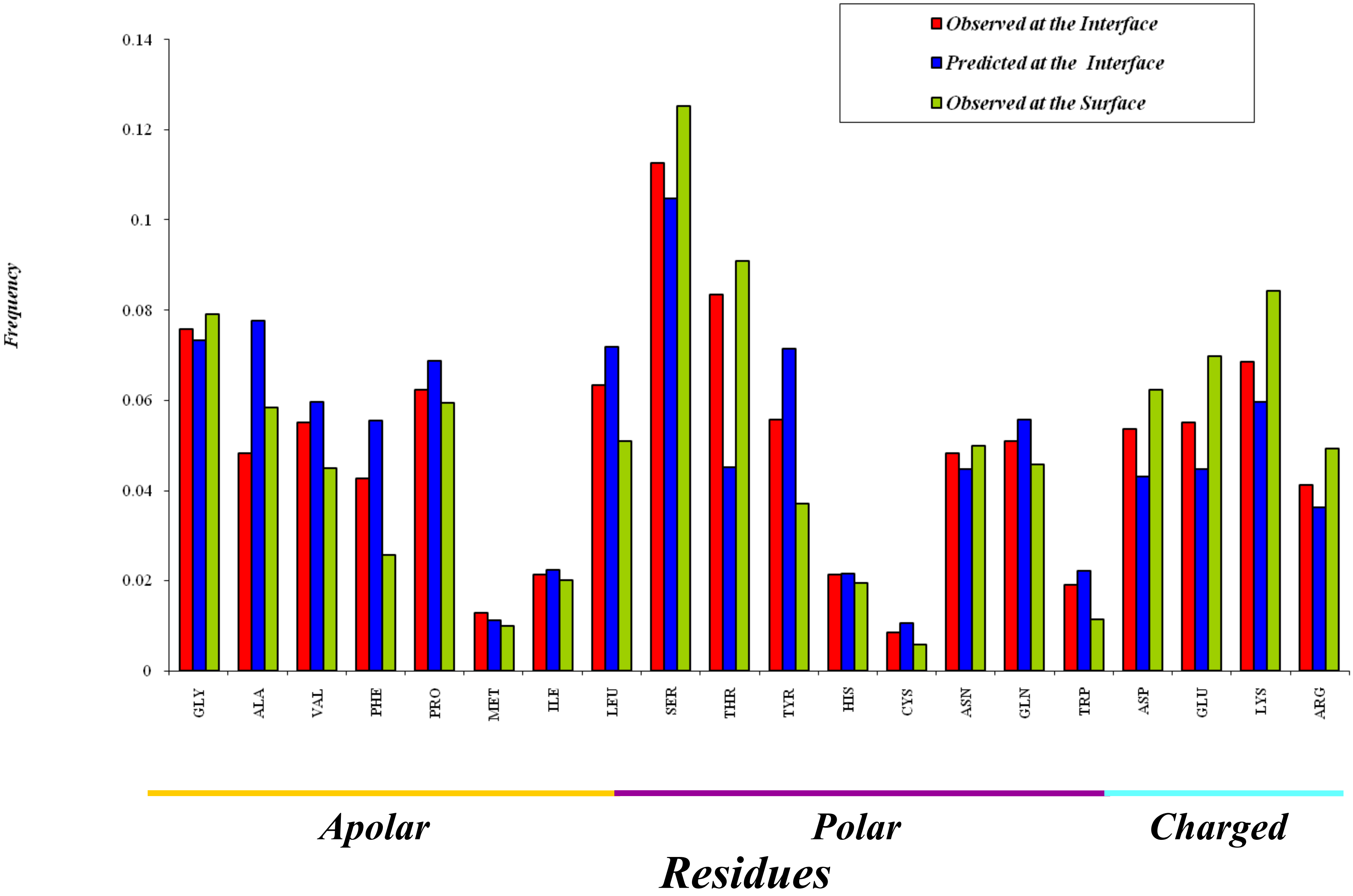


Training set

Testing set



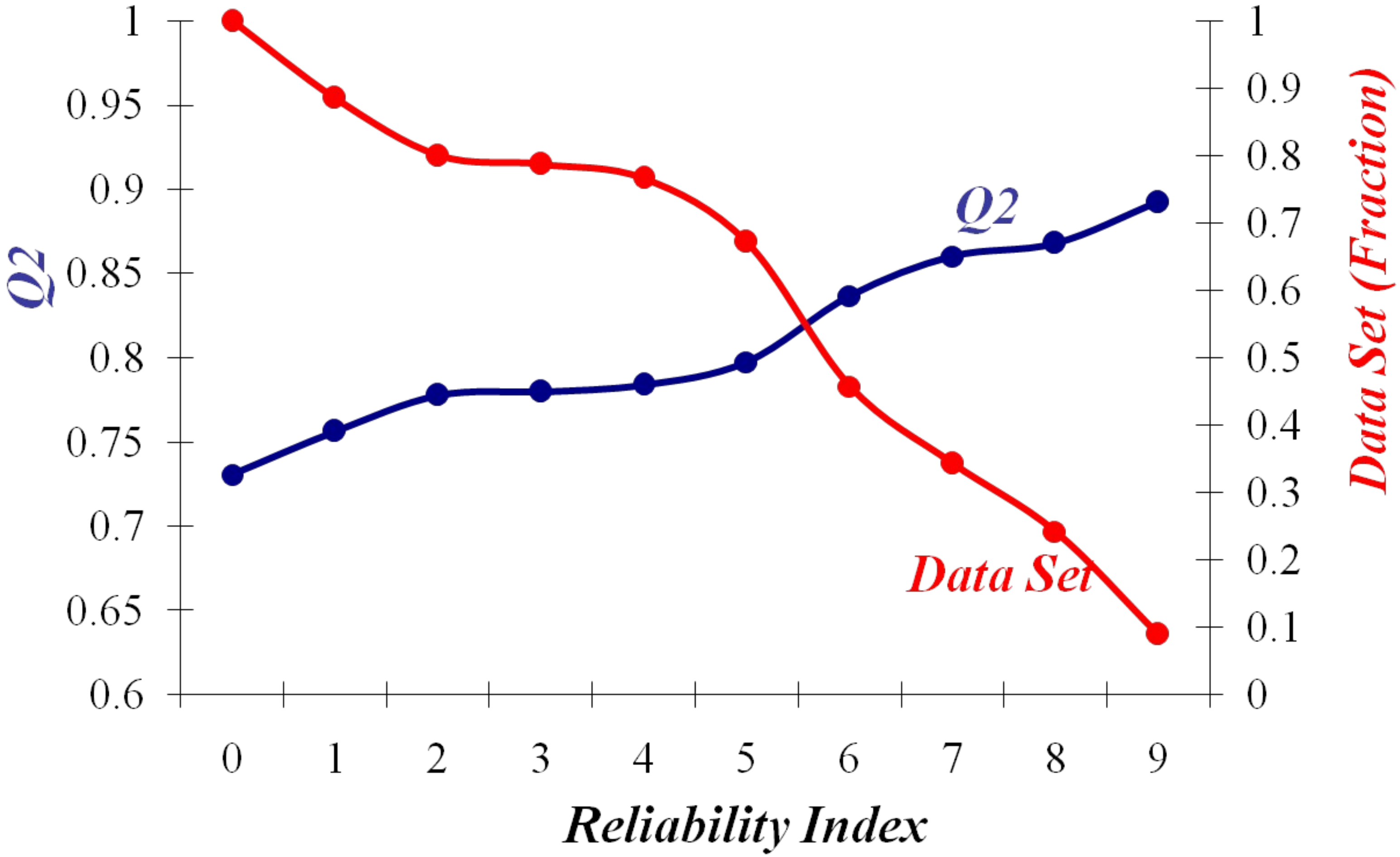
Distributions of apolar, polar and charged residues



Method performance

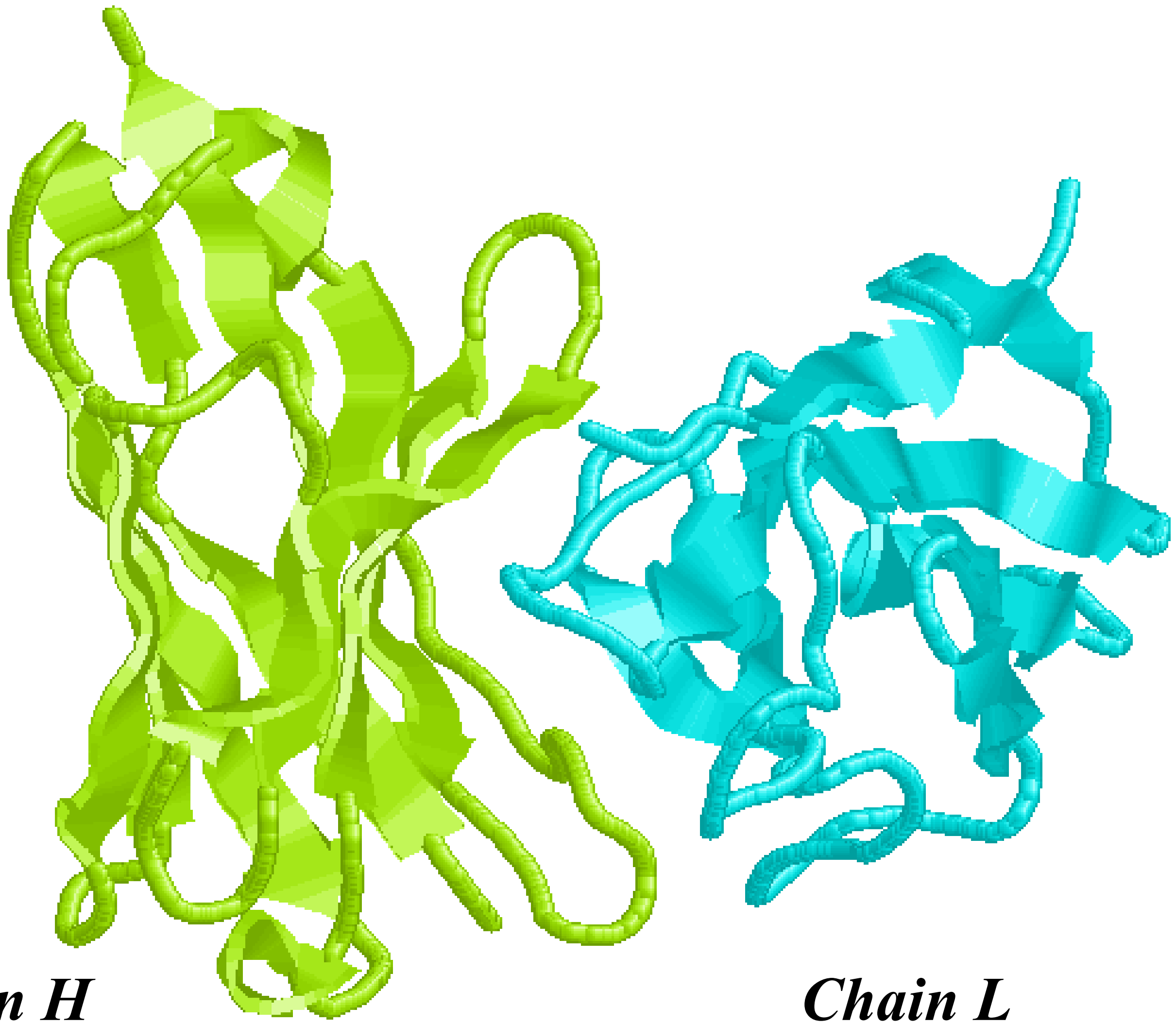
Q2	MCC	P(i)	Q(i)	P(n)	Q(n)
0.73	0.43	0.72	0.56	0.73	0.85

Q2 accuracy score as a function of the reliability index of the prediction



Mouse monoclonal antibody fragment Fv4155

PDB code: 1BFV Resolution: 0.21 nm

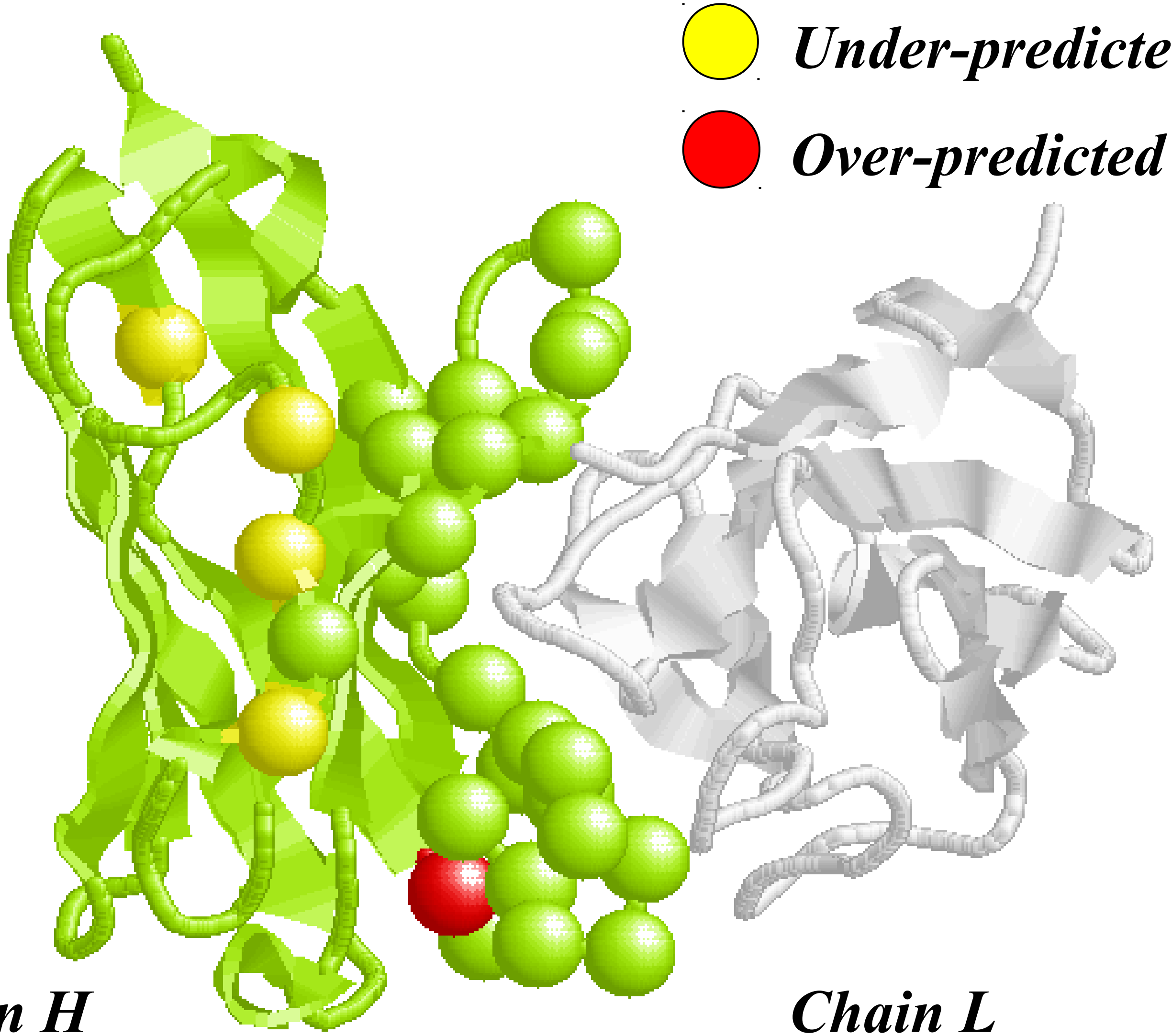


Prediction of protein-protein interaction sites

🟢 *Correctly predicted*

🟡 *Under-predicted*

🔴 *Over-predicted*



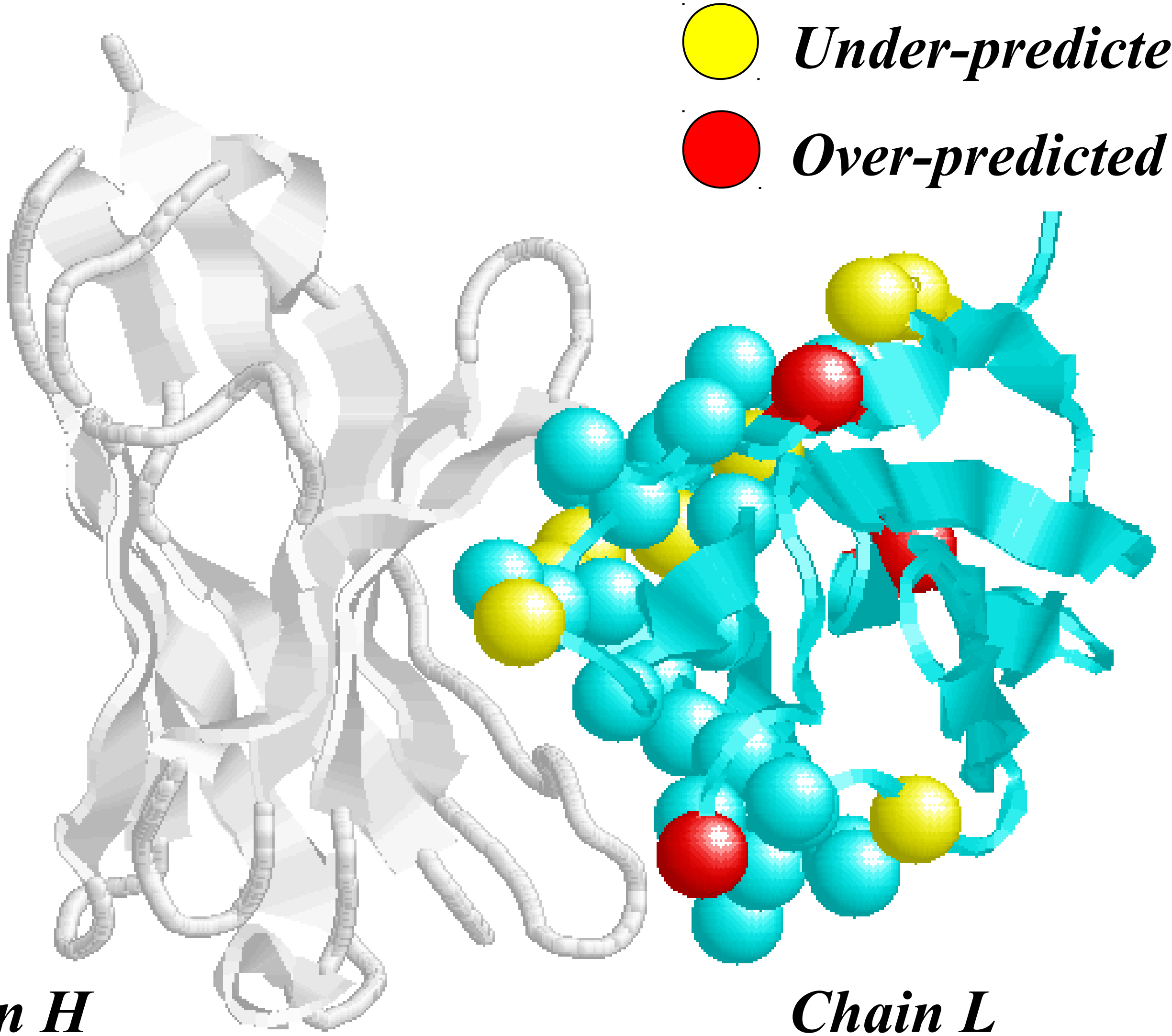
Q_2= 90 %

Prediction of protein-protein interaction sites

🔵 *Correctly predicted*

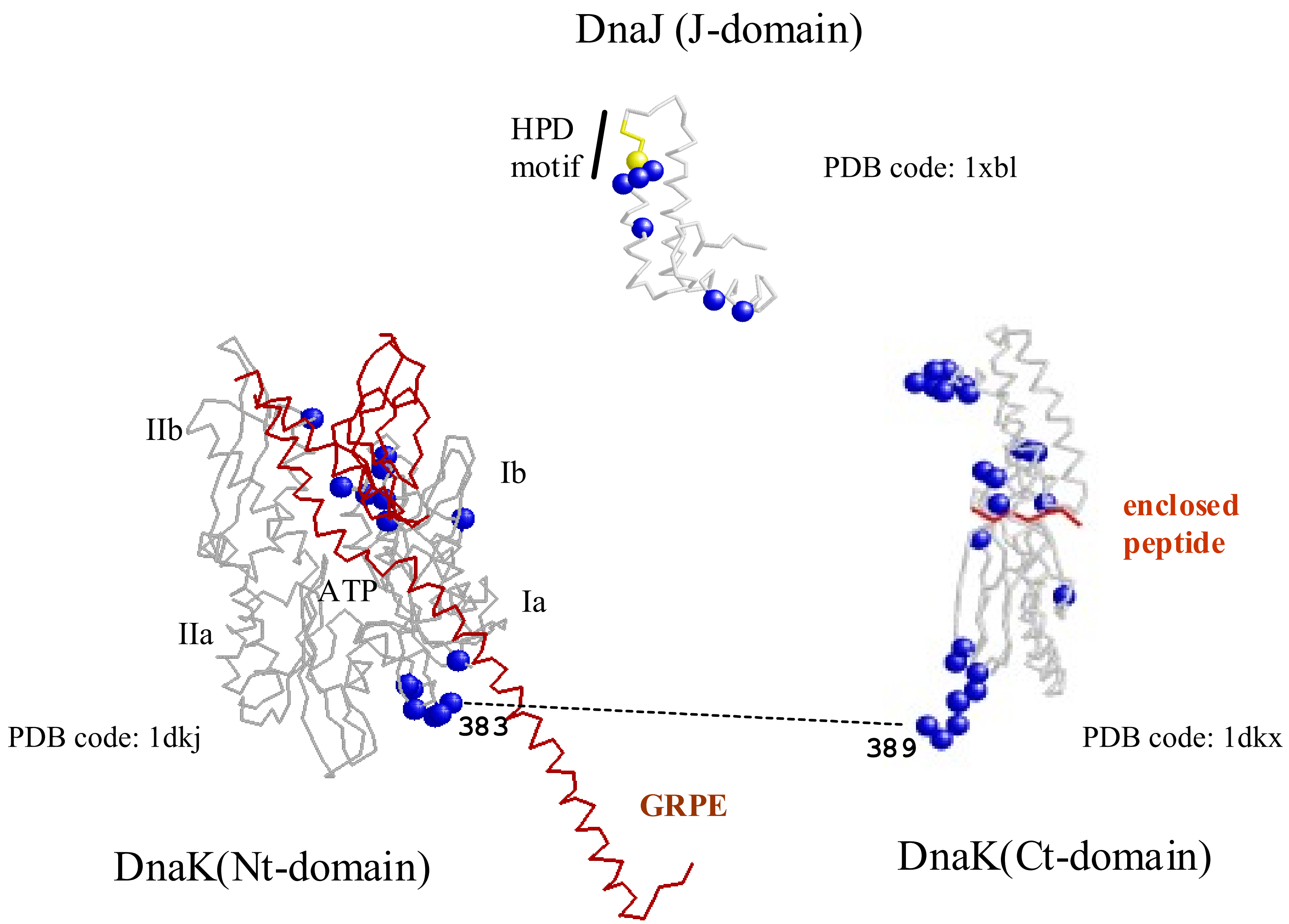
🟡 *Under-predicted*

🔴 *Over-predicted*



Q_2= 87 %

Prediction of protein-protein interaction sites in DnaK molecular chaperone system



Improvement with “Patch Smoothing” (*)

$$F(i) = \boldsymbol{\Sigma}_{j=0..N}\ w(i,j)\ \ O(R_i(j\ ))\ /\ [\boldsymbol{\Sigma}_{j=0..N}\, w(i,j)]$$

where $w(i,j)$ is a weight associated to the neighbor j of i and $O(k)$ is the network output.

Weighting schemes:

i) **Uniform**: $w^U(i,j) = 1 \quad \textit{for all } j$

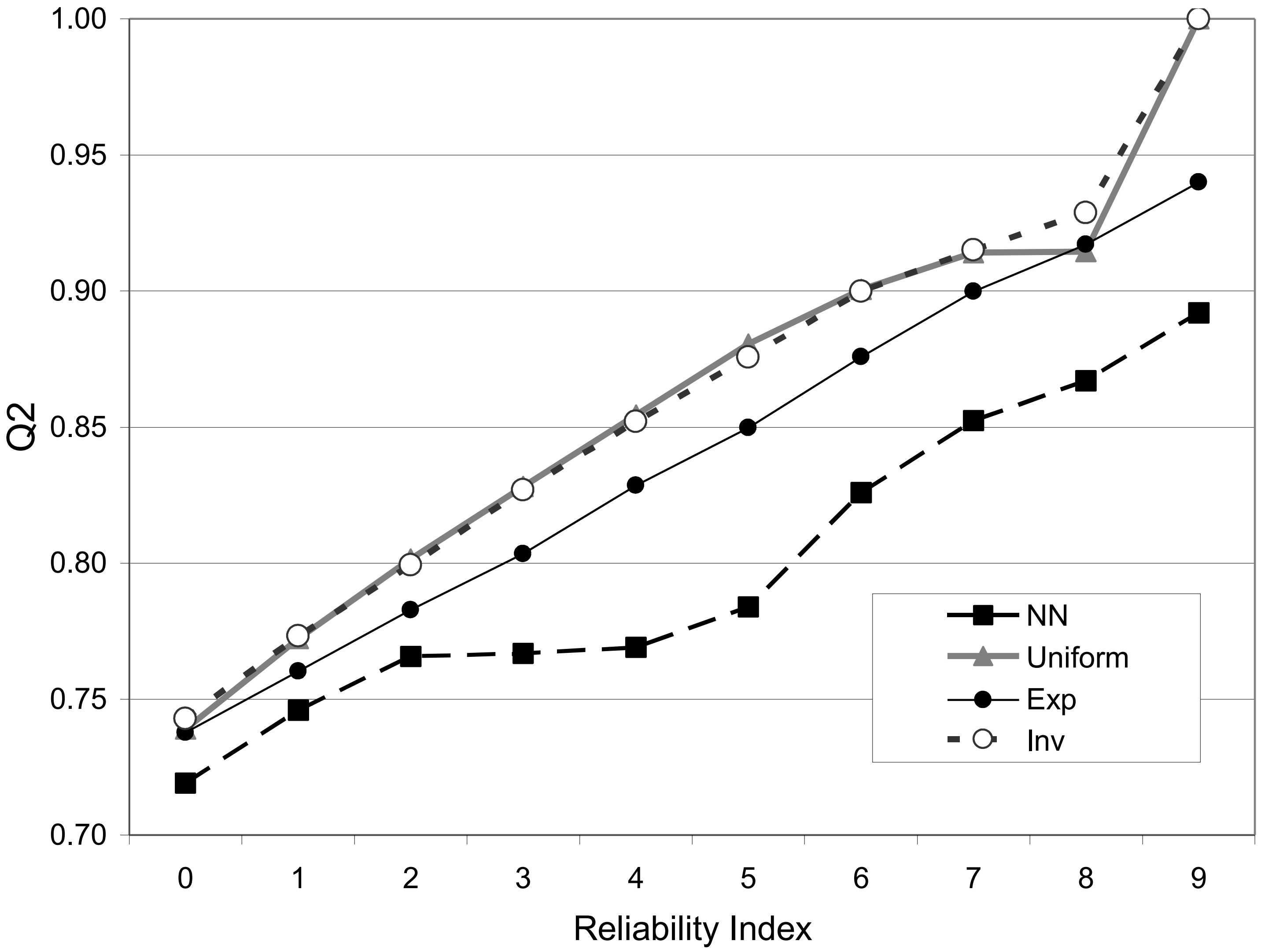
ii) **Exp**: $w^E(i,j) = exp\ [-d(i\ ,\ R_i(j))]$

iii) **Inv:** $w^I(i,j) = 1/\,[\mathrm{d}(i\ ,\ R_i(j))\ (1\text{-}\delta\ (0,\ j))+\ \delta(0,\ j)\ ]$

d(i , j) = Euclidean distance

Fariselli et al. 2003

Q2 scores of the Neural network (*NN*) and smoothing algorithms (*Uniform, Exp, Inv*) as a function of the reliability index (*Rel*) of the prediction.

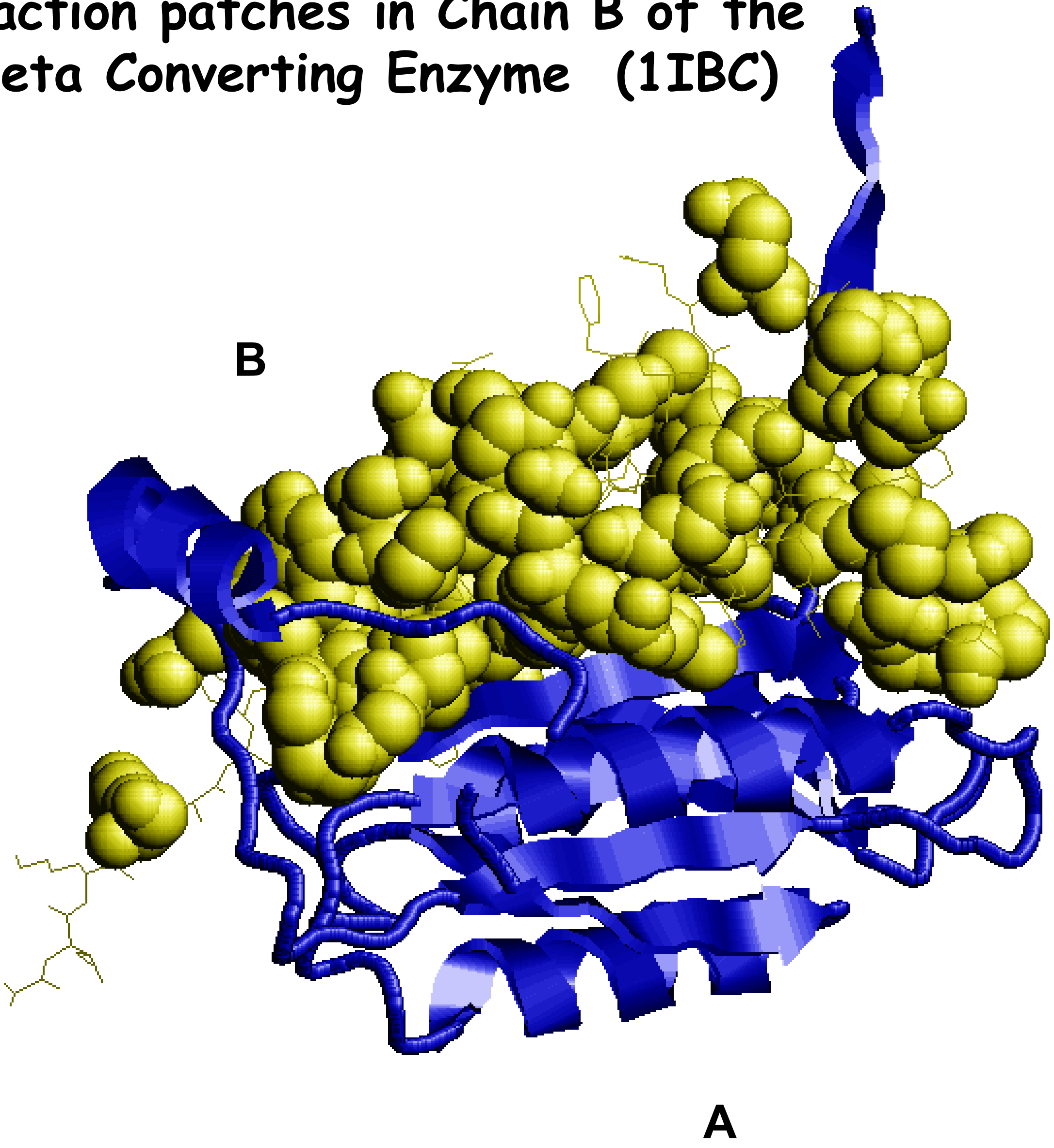


Scoring the efficiency of the neural network-based predictor

			Interaction site (i)		***Non interaction site (ni)***	
	Q_2	***C***	***P(i)***	***Q(i)***	***P(ni)***	***Q(ni)***
ISPRED	**0.73**	**0.43**	**0.72**	**0.56**	**0.73**	**0.85**
ISPRED+ FILTER	**0.76**	**0.50**	**0.75**	**0.59**	**0.75**	**0.87**

ISPRED predicting interaction patches in Chain B of the Inhibited Interleukin-1beta Converting Enzyme (1IBC)

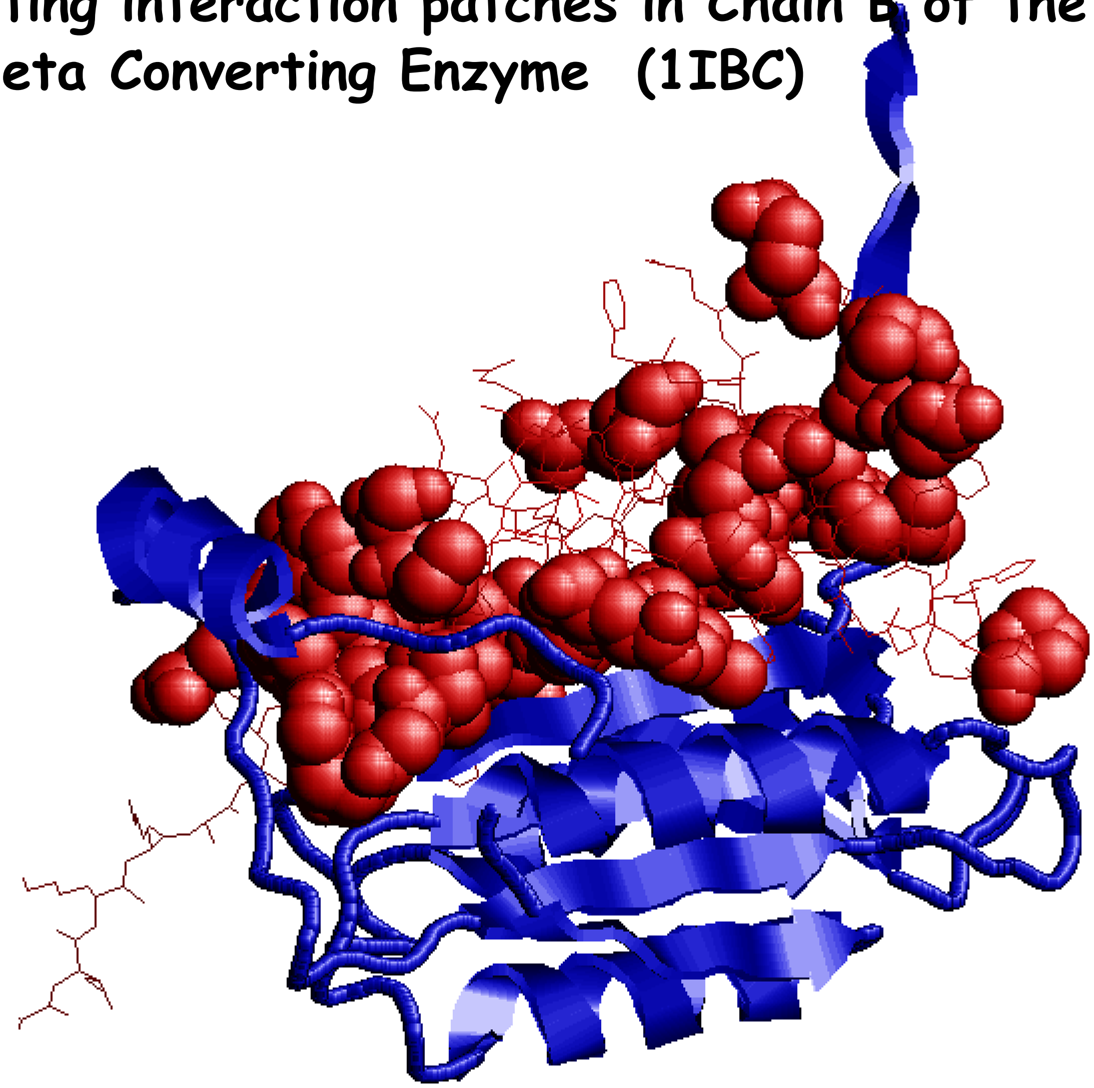
Q2=72%



Source: **HOMO SAPIENS; synthetic construct**

ISPRED + Filter predicting interaction patches in Chain B of the Inhibited Interleukin-1 beta Converting Enzyme (1IBC)

Q2=80%

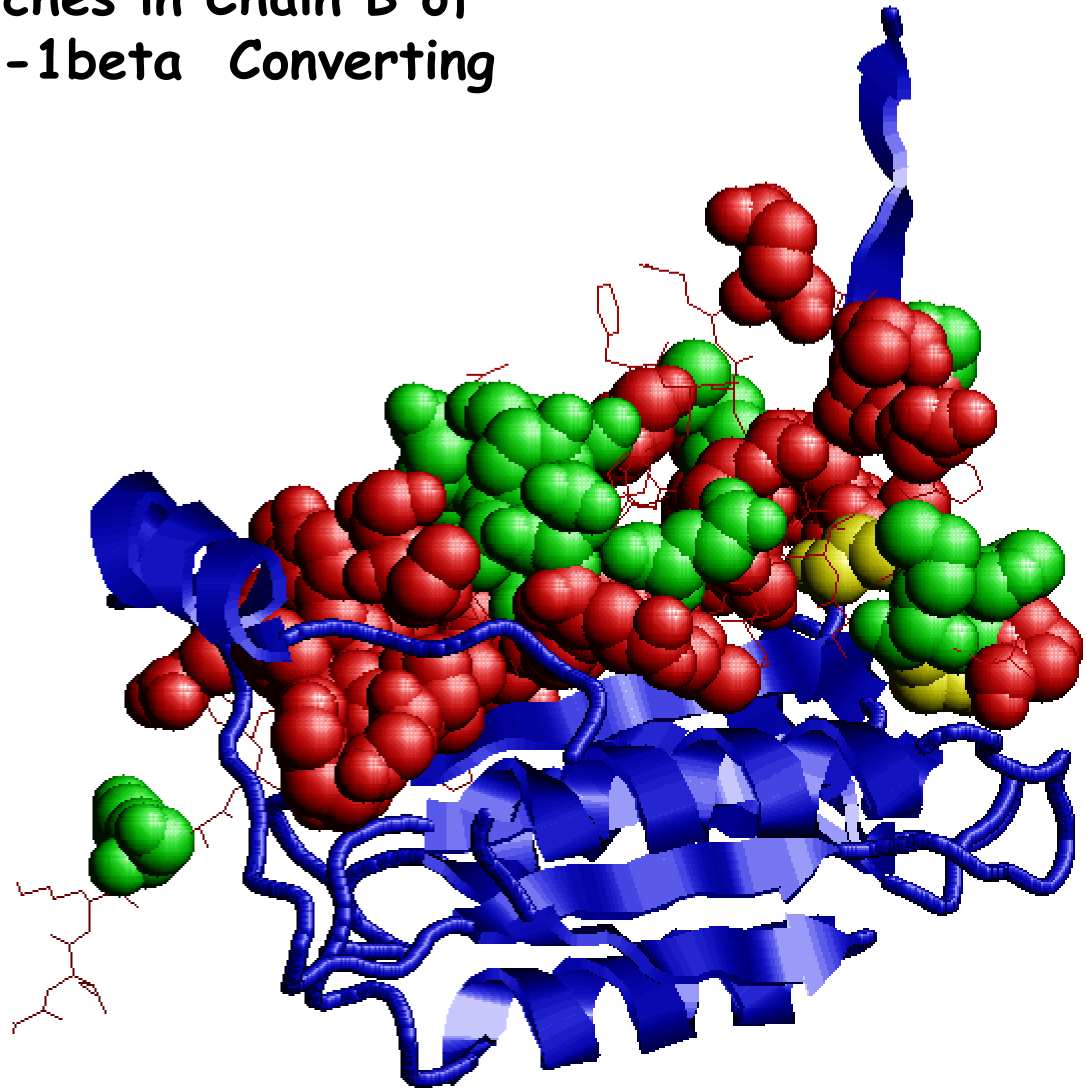


Predicted interaction patches in Chain B of the Inhibited Interleukin-1beta Converting Enzyme (1IBC)

- 🟥 Retained by **Filter**
- 🟩 Overpredicted
- 🟨 Underpredicted

Before Filter Q2=72%

With Filter Q2=80%



Problems with our old method

- Some sequence redundancy left in the set
- Data set contained too many antibodies chains

=> Resulted in too optimistic results

More realistic results with this encoding on a different set*:

Q2=64-68% MCC=25-30 %

With filter

Q2=65-70% MCC=28-35 %

* Ezkurdia et al. Brief. Bioinf 2009

Others approaches

More of 40 papers based on:

- Neural networks
- Support Vector Machines
- Random Forests
- Conditional Random Fields
- Hidden Markov Support Vector Machines

Introducing several different kind of information :

- Residue solvent accessibility or protrusion index
- Voronoi representation of the residue environment
- *Differentially conserved residue*
- PDB B-factor
- Hydrophobicity
- *Predicted residue solvent accessibility*
- Secondary structures
- Electrostatic potentials

Two very interesting ideas are:

- The difference between the “Predicted” and the “Observed” residue solvent accessibility[1]
- The usage of Markovian information by means of Hidden-Markov Support Vector Machines[2]

[1] Porollo and Meller Proteins. 2007
[2] Liu et al. BMC Bioinformatics. 2009

Hidden Markov Support Vector Machines

Y. Altun, I. Tsochantaridis, and T. Hofmann, “*Hidden Markov Support Vector Machines*,” ICML, 2003.

Slides taken also from:
“*Structured Output Prediction with Structural Support Vector Machines*”
by Thorsten Joachims

- The predominant formalism for modeling and predicting label sequences has been based on Hidden Markov Models (HMMs) and variations thereof.
- But HMMs have at least three limitations:
  - They are typically trained in a *non-discriminative* manner.
  - The *conditional independence* assumptions are often too restrictive.
  - They are based on *explicit feature representations* and lack the power of kernel-based methods.
- HM-SVMs address all of the above shortcomings, and retaining some of the key advantages of HMMs:
  - The Markov chain dependency structure between labels.
  - An efficient dynamic programming formulation.
- Two crucial ingredients of HM-SVMs:
  - The maximum margin principle
  - A kernel-centric approach to learning non-linear discriminant functions.

ISPRED2*

Savojardo et al., 2012

ISPRED2

Simple idea: *take the best of the available approaches and combine*

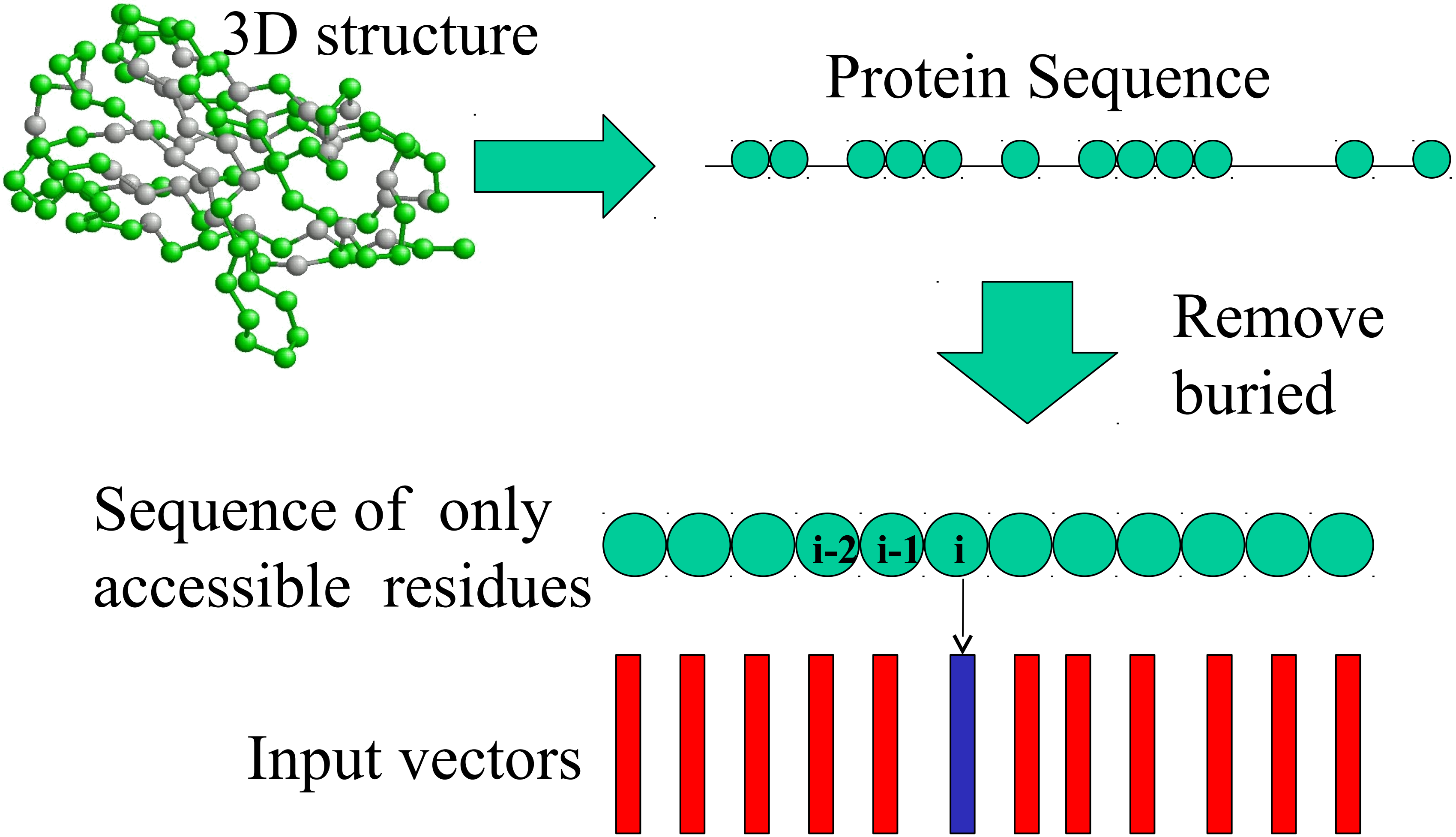
Method:

- Basic Method: HM-SVM
- Smooth the Prediction: spatial average

Input:

- Use sequence profile (but test also PSI-BLAST PSSM)
- Include the difference between predicted and observed residue solvent accessibilty

HM-SVM for predicting PPIs



3D structure

Protein Sequence

Remove buried

Sequence of only accessible residues

Input vectors

ISPRED2

Method:

Hidden Markov Support Vector Machines (HM-SVM)
+ surface smoothing by local average prediction

Input:

- Position Specific Scoring Matrix (as computed by PSI-BLAST –Q)
- Difference between observed and predicted residue solvent accessibility.

DataSet: 1,124 chains with a low level of sequence identity **(**Liu at al. 2009)

- Non-NMR structures with resolution better than 4 Å.
- Protein chains with > 40 residues.
- PQS check to retain only biologically functional complexes and avoid crystal packing ones (Henrick and Thornton, 1998).
- Cross-validation on NCBI BLASTClust subsets (Altschul et al., 1990).

Cross-validation performance of the different methods

Method	Q2 (%)	Sp (%)	Sn(%)	C(%)	Encoding
NN	64	55	77	31	(Profile+RSA)
NN	66	62	84	35	(PSSM+RSA)
NN	69	65	82	39	(PSSM+dSA)
NN	69	65	85	40	<(PSSM+dSA)>
HM-SVM	68	70	65	36	(Profile+RSA)
HM-SVM	70	72	66	40	(PSSM+RSA)
HM-SVM	71	73	67	42	(PSSM+dSA)
HM-SVM	71	73	68	43	<(PSSM+dSA)>

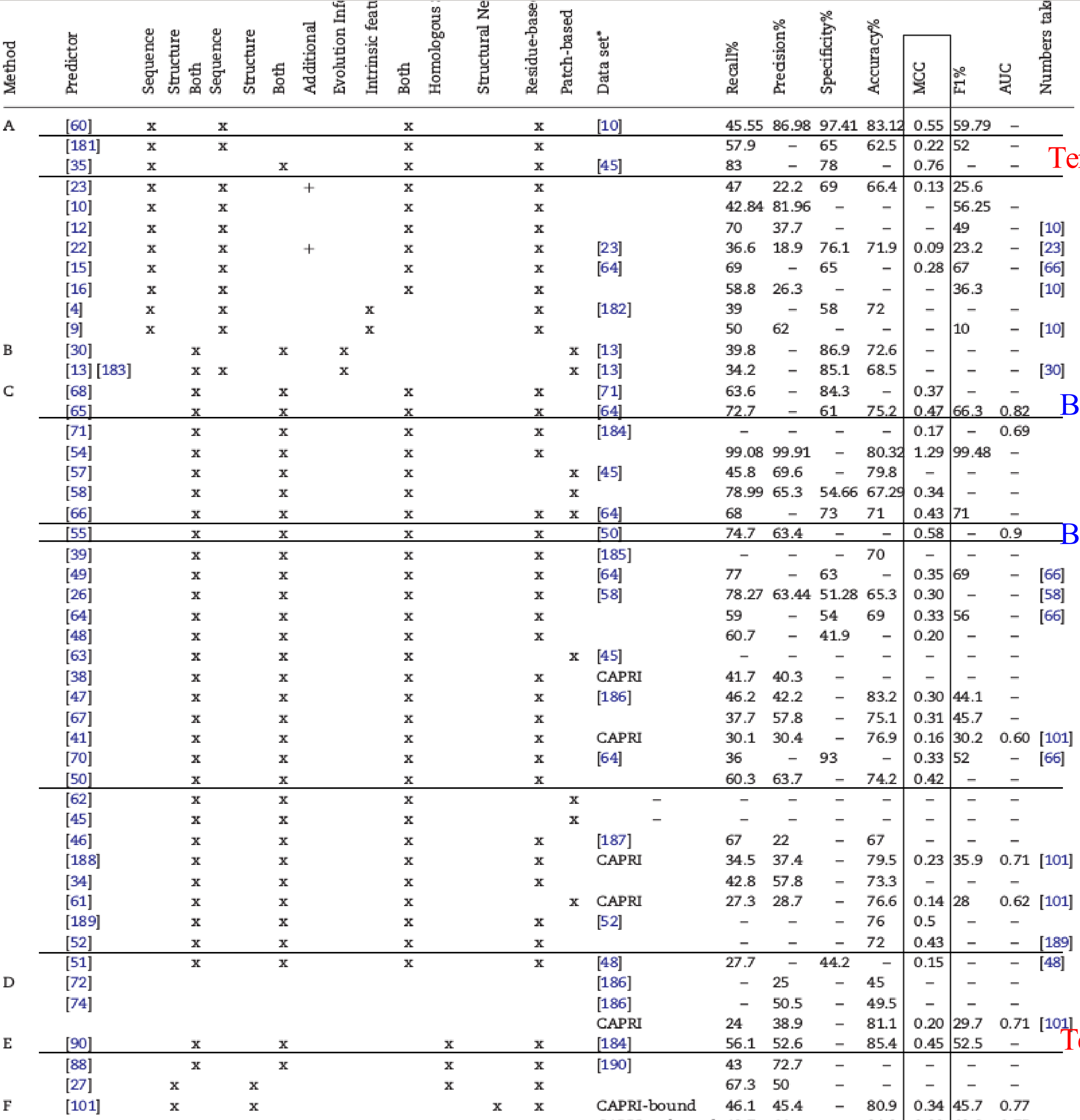
Performance using different definitions of interaction sites.

Interaction sites definition	Q2 (%)	Sp (%)	Sn(%)	C(%)
Liu et al. 2009 *	71	73	67	43
Jones and Thornton 1997	71	73	67	42
Fariselli et al. 2001	71	73	68	43

Comparison with other methods

Method	**Q2 (%)**	**Sp (%)**	**Sn(%)**	**C(%)**	**F1(%)**
Wang et al. 2006	NA	65	69	28	67
Nguyen-Rajapakse 2006	NA	93	36	33	52
Deng et al. 2009	NA	63.	77	35	69
Liu et al. 2009	**69**	**52**	**59**	**33**	**55**
ISPRED2	**71**	**73**	**68**	**43**	**71**

Esmaielbeiki et al. Brief. in Bioinf., 2015, 1–15



Method	Predictor	Sequence	Structure	Both	Sequence	Structure	Both	Additional	Evolution Inf	Intrinsic feat	Both	Homologous	Structural Ne	Residue-base	Patch-based	Data set*	Recall%	Precision%	Specificity%	Accuracy%	MCC	F1%	AUC	Numbers tak
A	[60]	x			x						x			x		[10]	45.55	86.98	97.41	83.12	0.55	59.79	–	
	[181]	x			x						x			x			57.9	–	65	62.5	0.22	52	–	
	[35]	x					x				x			x		[45]	83	–	78	–	0.76	–	–	
	[23]	x			x			+			x			x			47	22.2	69	66.4	0.13	25.6		
	[10]	x			x						x			x			42.84	81.96	–	–	–	56.25	–	
	[12]	x			x						x			x			70	37.7	–	–	–	49	–	[10]
	[22]	x			x			+			x			x		[23]	36.6	18.9	76.1	71.9	0.09	23.2	–	[23]
	[15]	x			x						x			x		[64]	69	–	65	–	0.28	67	–	[66]
	[16]	x			x						x			x			58.8	26.3	–	–	–	36.3		[10]
	[4]	x			x					x				x		[182]	39	–	58	72	–	–	–	
	[9]	x			x					x				x			50	62	–	–	–	10	–	[10]
B	[30]			x			x		x						x	[13]	39.8	–	86.9	72.6	–	–	–	
	[13] [183]			x	x				x						x	[13]	34.2	–	85.1	68.5	–	–	–	[30]
C	[68]			x			x				x			x		[71]	63.6	–	84.3	–	0.37	–	–	
	[65]			x			x				x			x		[64]	72.7	–	61	75.2	0.47	66.3	0.82	
	[71]			x			x				x			x		[184]	–	–	–	–	0.17	–	0.69	
	[54]			x			x				x			x			99.08	99.91	–	80.32	1.29	99.48	–	
	[57]			x			x				x				x	[45]	45.8	69.6	–	79.8	–	–	–	
	[58]			x			x				x				x		78.99	65.3	54.66	67.29	0.34	–	–	
	[66]			x			x				x			x	x	[64]	68	–	73	71	0.43	71	–	
	[55]			x			x				x			x		[50]	74.7	63.4	–	–	0.58	–	0.9	
	[39]			x			x				x			x		[185]	–	–	–	70	–	–	–	
	[49]			x			x				x			x		[64]	77	–	63	–	0.35	69	–	[66]
	[26]			x			x				x			x		[58]	78.27	63.44	51.28	65.3	0.30	–	–	[58]
	[64]			x			x				x			x			59	–	54	69	0.33	56	–	[66]
	[48]			x			x				x			x			60.7	–	41.9	–	0.20	–	–	
	[63]			x			x				x				x	[45]	–	–	–	–	–	–	–	
	[38]			x			x				x			x		CAPRI	41.7	40.3	–	–	–	–	–	
	[47]			x			x				x			x		[186]	46.2	42.2	–	83.2	0.30	44.1	–	
	[67]			x			x				x			x			37.7	57.8	–	75.1	0.31	45.7	–	
	[41]			x			x				x			x		CAPRI	30.1	30.4	–	76.9	0.16	30.2	0.60	[101]
	[70]			x			x				x			x		[64]	36	–	93	–	0.33	52	–	[66]
	[50]			x			x				x			x			60.3	63.7	–	74.2	0.42	–	–	
	[62]			x			x				x				x	–	–	–	–	–	–	–	–	
	[45]			x			x				x				x	–	–	–	–	–	–	–	–	
	[46]			x			x				x			x		[187]	67	22	–	67	–	–	–	
	[188]			x			x				x			x		CAPRI	34.5	37.4	–	79.5	0.23	35.9	0.71	[101]
	[34]			x			x				x			x			42.8	57.8	–	73.3	–	–	–	
	[61]			x			x				x				x	CAPRI	27.3	28.7	–	76.6	0.14	28	0.62	[101]
	[189]			x			x				x			x		[52]	–	–	–	76	0.5	–	–	
	[52]			x			x				x			x			–	–	–	72	0.43	–	–	[189]
	[51]			x			x				x			x		[48]	27.7	–	44.2	–	0.15	–	–	[48]
D	[72]															[186]	–	25	–	45	–	–	–	
	[74]															[186]	–	50.5	–	49.5	–	–	–	
																CAPRI	24	38.9	–	81.1	0.20	29.7	0.71	[101]
E	[90]			x			x					x		x		[184]	56.1	52.6	–	85.4	0.45	52.5	–	
	[88]			x			x					x		x		[190]	43	72.7	–	–	–	–	–	
	[27]		x			x						x		x			67.3	50	–	–	–	–	–	
F	[101]		x			x							x	x		CAPRI-bound	46.1	45.4	–	80.9	0.34	45.7	0.77	

Template

Best set

Best set

Template

A few (trivial) take home messages:

The most reliable prediction methods are “template-based” approaches, but they are limited to the available complexes

The most relevant features for *ab initio* prediction (so far) are:
- evolutionary information (sequence profile, conservation)
- residue solvent accessibility (especially when compared with the predicted)
- local residue environment

Probably, once properly exploited, also play the surface shape and the local geometry may improve the predictive performances.

Hierarchical predictions, such as cascade of methods, neighboring filtering, Markovian properties, possibly Deep Learning, improve the quality of the predictions

A few (trivial) take home messages (cont.):

Prediction from structure is more accurate than from sequence

It is mandatory to use a proper cross-validation procedure with no similarity between training and testing sets, also with respect to other predictors used to derive the features.

The definition of the protein-protein interaction is not relevant.
They are all very correlated and the results appear to be almost independent.

The dataset must be chosen according to the problem to tackle:
general heterocomplexes, homocomplexes, antibody-antigens, proteases, etc.

A really trivial take home messages (cont.):

It is better to be rich, beautiful and healthy,
than poor, ugly and ill !

Thank you very much for your attention.

That's all!

Suggested reading:

Ezkurdia et al. "Progress and challenges in Progress and challenges in predicting protein interfaces". *Brief Bioinform*. 2009;(10):233–246.

Esmaielbeiki et al. "Predicting protein-protein interaction sites". *Brief Bioinform*. 2015 1:1-15, doi: 10.1093/bib/bbv027