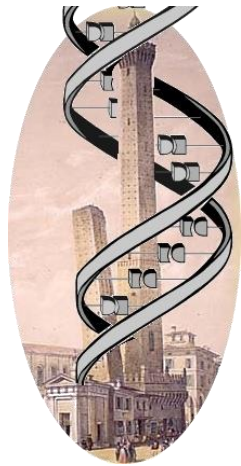




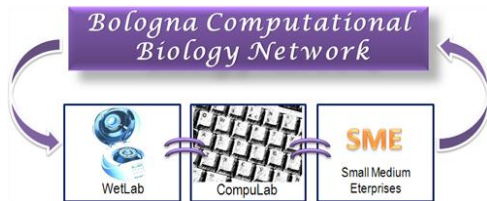
ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA



# The multidimensional problem of protein-protein interactions

Rita Casadio

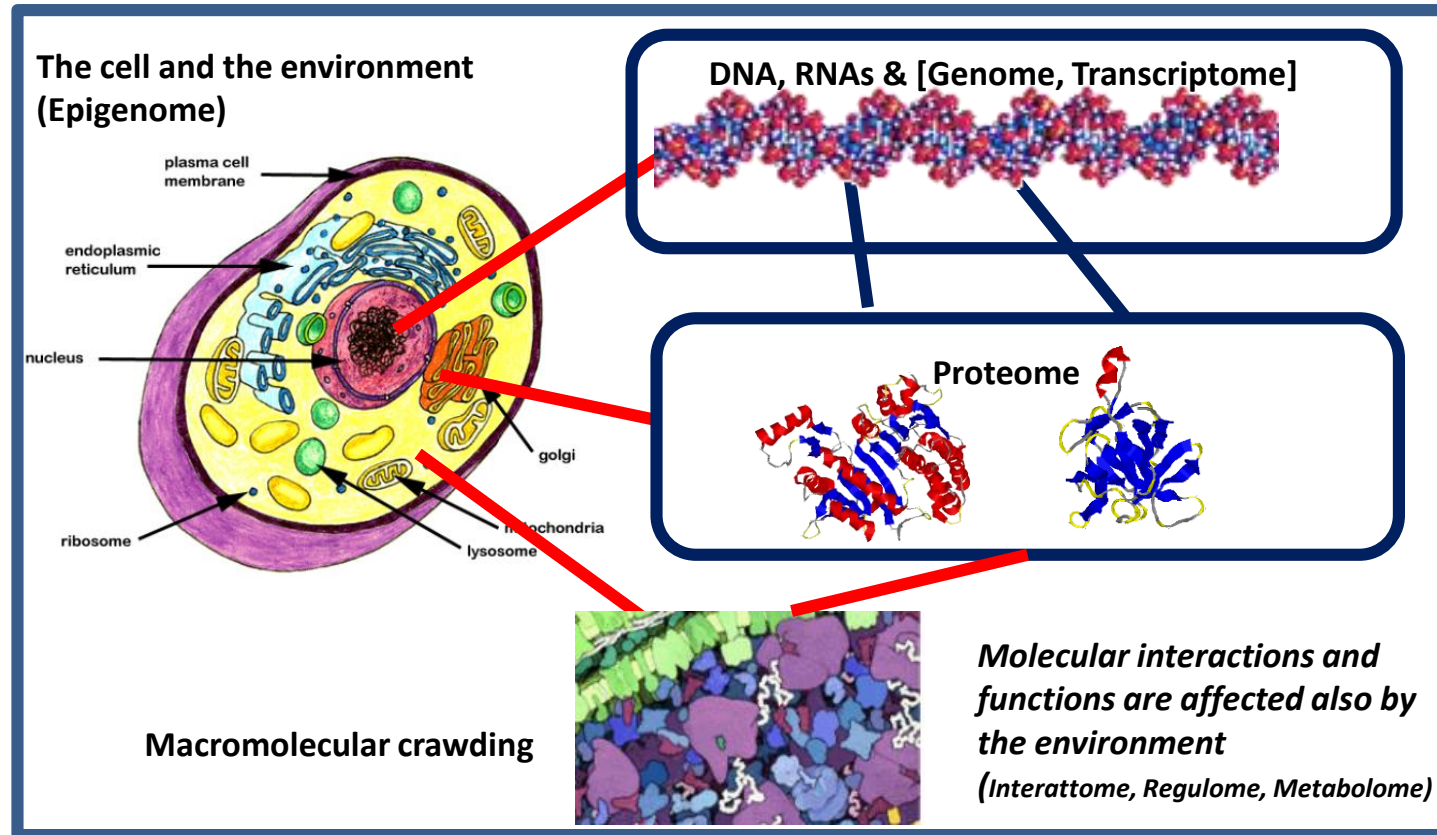
**BIOCOMPUTING GROUP**  
University of Bologna, Italy





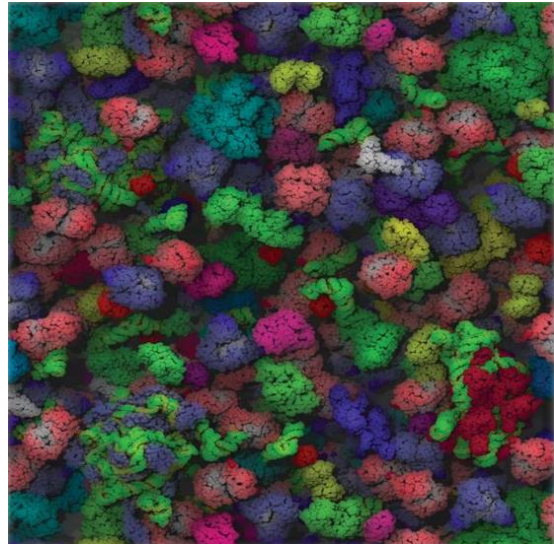
# The ingredients of biological complexity at the cell level

*From genes to proteins, their interaction and the interplay with the environment*





# LIFE IS CROWDED: Macromolecular crowding is under-appreciated, including protein phase separation



**The Crowded Cell:** An atomically detailed model of the crowded *E. coli* cytoplasm, including the 50 most abundant macromolecules. RNA is shown as green and yellow. Reprinted from: McGuffee SR, Elcock AH (2010) Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput Biol* 6(3): e1000694.

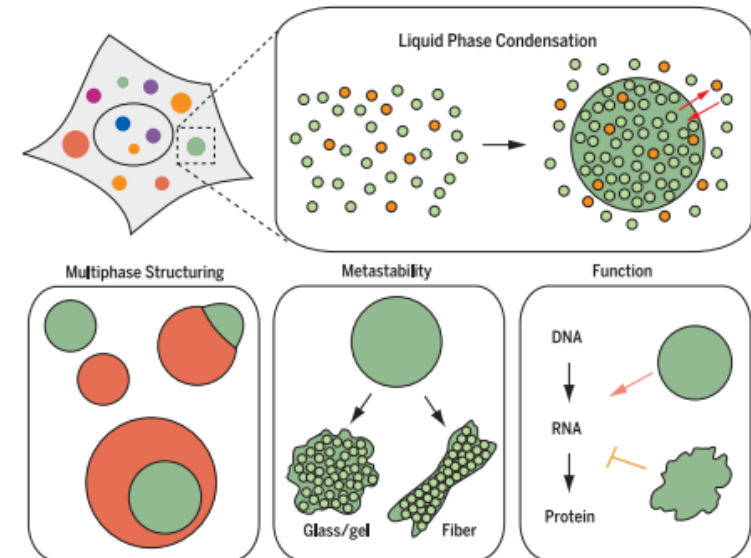
## REVIEW SUMMARY

### CELLULAR BIOPHYSICS

# Liquid phase condensation in cell physiology and disease

Yongdae Shin and Clifford P. Brangwynne\*

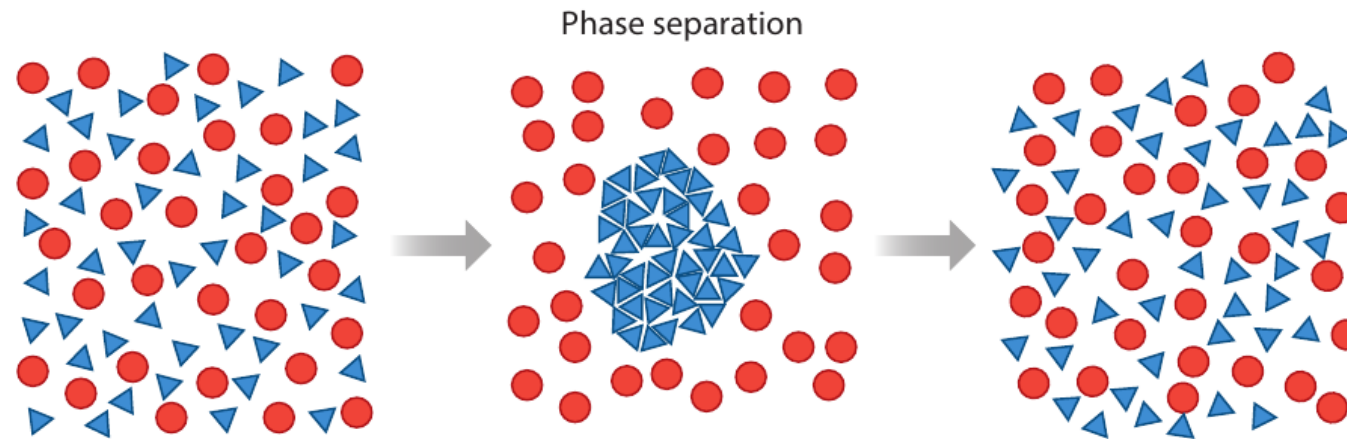
Science 357, 1253 (2017)



Liquid phase condensation: An emerging paradigm of cellular organization. Living cells



## PROTEIN PHASE SEPARATION



*Simplified representation of the dynamics of protein phase separation*

*Membraneless organelles are common to several types of cells working under physiological conditions*



## Two important initiatives for deciphering the human proteomes:

### The Protein Atlas:

**THE HUMAN PROTEOME**

The Power of Proteins. The human genome consists of approximately 25,000 protein-coding genes. If DNA can be equated with the blueprint for a house, then proteins can be thought of as the bricks and mortar, plumbing and paint—essentially everything that makes up the house. This poster summarizes the multiple ongoing antibody- and transcriptome-based proteome projects and where in the human body the research is focused. For more detailed information, visit [www.proteinatlas.org](http://www.proteinatlas.org)

**THE HUMAN PROTEIN ATLAS**  
The Human Protein Atlas is a multi-organ and tissue-specific atlas of all human proteins. It is a multi-organ and tissue-specific atlas of all human proteins. It is a multi-organ and tissue-specific atlas of all human proteins.

**THE TISSUE-SPECIFIC PROTEOMICS**  
The expression of all human protein-coding genes has been measured in human tissues. The expression of all human protein-coding genes has been measured in human tissues.

**THE SECRETED AND MEMBRANE PROTEOMES**  
Secreted and membrane proteins play major roles in many biological and pathological processes. They are secreted and membrane proteins.

**THE EXOME PROTEOME**  
The expression of a variety of protein isoforms in each cell defines the structure and function of the human proteome and thereby controls cellular and organismal functions.

**THE CANCER PROTEOMES**  
Over 200 genes have been implicated in the

**THE HUMAN PROTEIN ATLAS**  
www.proteinatlas.org  
posters.sciencemag.org/humanproteome

The protein data covers 15313 genes (78%) for which there are available antibodies. The mRNA expression data is derived from deep sequencing of RNA (RNA-seq) from 37 different normal tissue types



<https://www.proteinatlas.org/humanproteome/tissue>

### The Human Proteome Project:

#### HUMAN PROTEOME PROJECT (HPP)



The Human Proteome Project (HPP) is an international project organized by the Human Proteome Organization (HUPRO) that aims to revolutionize our understanding of the human proteome via a coordinated effort by many research laboratories around the world. It is designed to map the entire human proteome in a systematic effort using currently available and emerging techniques. Completion of this project will enhance understanding of human biology at the cellular level and lay a foundation for development of diagnostic, prognostic, therapeutic, and preventive medical applications.

> More on the HPP

#### HPP PROGRESS TO DATE (PHASE I)

February 2019

**19,823**  
PREDICTED GENOME-CODING PROTEINS  
(neXtProt PE1+ PE2 + PE3 + PE4)



**17,694**  
FOUND PROTEINS  
(neXtProt PE1)



**2,129**  
MISSING PROTEINS  
(neXtProt PE2 + PE3 + PE4)

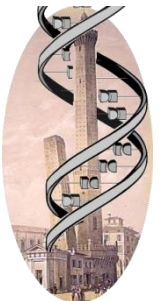
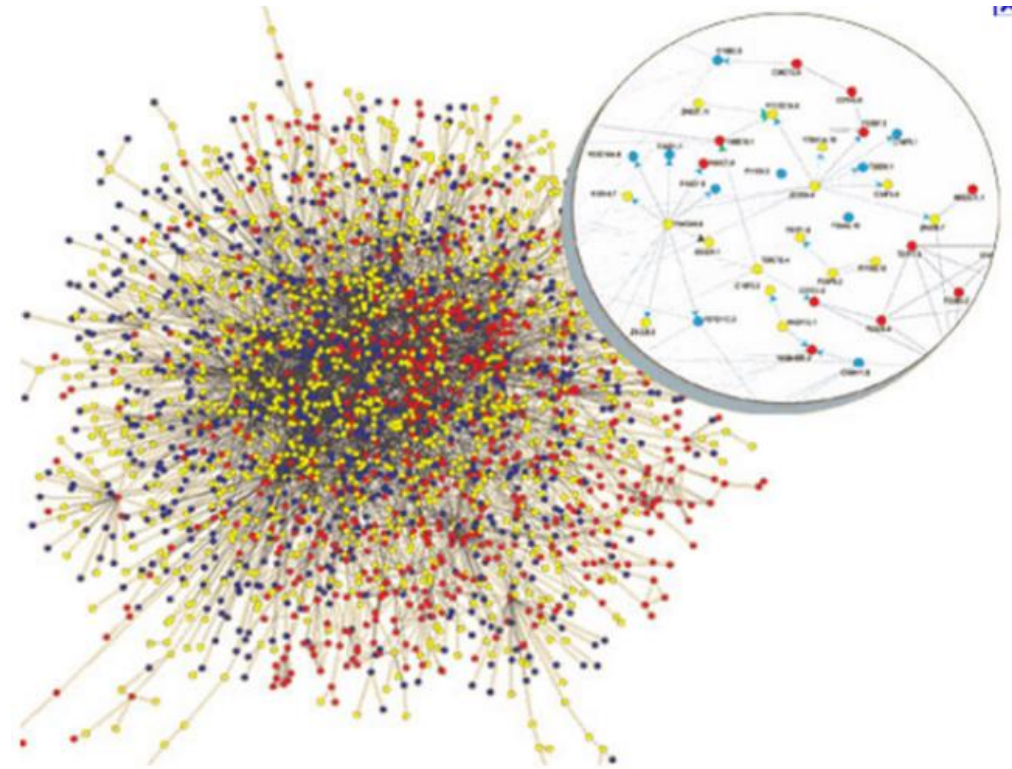


<https://www.hupo.org/human-proteome-project>

# From the proteome to the interactome.... Protein-protein interaction networks

- Play a major role in generating the complexity of cellular processes
- If perturbed can lead to impairment of biological functions and to disease
- Crucial in host-pathogen communication

*Yeast two-hybrid*  
*Affinity purification + mass spec*



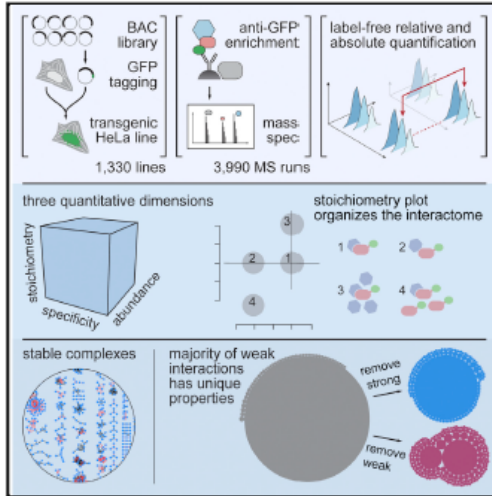


# Sampling the human interactomic space.....

Cell

## A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances

### Graphical Abstract



### Highlights

- Human interactome dataset connecting 5,400 proteins with 28,500 interactions
- Three quantitative dimensions measure specificities, stoichiometries, and abundances
- Stable complexes are rare but stand out by a signature of balanced stoichiometries
- Weak interactions dominate the network and have critical topological properties



Hein et al., 2015, Cell 163, 712–723  
 October 22, 2015 ©2015 Elsevier Inc.  
<http://dx.doi.org/10.1016/j.cell.2015.09.053>

CellPress

Article



NIH Public Access

Author Manuscript

*Nat Biotechnol.* Author manuscript; available in PMC 2013 July 11.

Published in final edited form as:

*Nat Biotechnol.* ; 30(2): 159–164. doi:10.1038/nbt.2106.

## Three-dimensional reconstruction of protein networks provides insight into human genetic disease

Xiujuan Wang<sup>1,2,\*</sup>, Xiaomu Wei<sup>2,3,\*</sup>, Bram Thijssen<sup>4,\*</sup>, Jishnu Das<sup>1,2,\*</sup>, Steven M Lipkin<sup>3</sup>, and Haiyuan Yu<sup>1,2,†</sup>

<sup>1</sup>Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA <sup>2</sup>Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA <sup>3</sup>Department of Medicine, Weill Cornell College of Medicine, New York, NY 10021, USA <sup>4</sup>Department of Bioinformatics, Maastricht University, 6200 MD Maastricht, The Netherlands

### Abstract

In an effort to understand molecular mechanisms of human disease and to determine genes responsible, we systematically examine relationships between 3,949 genes, 62,663 mutations and 3,453 associated disorders within the framework of a three-dimensional structurally resolved human interactome, consisting of 4,222 high-quality binary protein-protein interactions with their



Cell 159, 1212–1226, November 20, 2014

Resource

## A Proteome-Scale Map of the Human Interactome Network

Thomas Rolland,<sup>1,2,19</sup> Murat Taşan,<sup>1,3,4,5,19</sup> Benoit Charlotiaux,<sup>1,2,19</sup> Samuel J. Pevzner,<sup>1,2,6,7,19</sup> Quan Zhong,<sup>1,2,8,19</sup> Nidhi Sahni,<sup>1,2,19</sup> Song Yi,<sup>1,2,19</sup> Irma Lemmens,<sup>9</sup> Celia Fontanillo,<sup>10</sup> Roberto Mosca,<sup>11</sup> Atanas Kamburov,<sup>1,2</sup> Susan D. Ghiassian,<sup>1,12</sup> Xiping Yang,<sup>1,2</sup> Lila Ghamsari,<sup>1,2</sup> Dawit Balcha,<sup>1,2</sup> Bridget E. Begg,<sup>1,2</sup> Pascal Braun,<sup>1,2</sup> Marc Brehme,<sup>1,2</sup> Martin P. Broly,<sup>1,2</sup> Anne-Ruxandra Carvunis,<sup>1,2</sup> Dan Convery-Zupan,<sup>1,2</sup> Roser Corominas,<sup>13</sup> Jasmin Coulombe-Huntington,<sup>1,14</sup> Elizabeth Dann,<sup>1,2</sup> Matija Dreze,<sup>1,2</sup> Amélie Dricot,<sup>1,2</sup> Changyu Fan,<sup>1,2</sup> Eric Franzosa,<sup>1,14</sup> Fana Gebreab,<sup>1,2</sup> Bryan J. Gutierrez,<sup>1,2</sup> Madeleine F. Hardy,<sup>1,2</sup> Mike Jin,<sup>1,2</sup> Shuli Kang,<sup>13</sup> Ruth Kirov,<sup>1,2</sup> Guan Ning Lin,<sup>13</sup> Katja Luck,<sup>1,2</sup> Andrew MacWilliams,<sup>1,2</sup> Jörg Menche,<sup>1,2</sup> Ryan R. Murray,<sup>1,2</sup> Alexandre Palagi,<sup>1,2</sup> Matthew M. Poulin,<sup>1,2</sup> Xavier Rambout,<sup>1,2,15</sup> John Rasla,<sup>1,2</sup> Patrick Reichert,<sup>1,2</sup> Viviana Romero,<sup>1,2</sup> Elen Ruysinck,<sup>9</sup> Julie M. Sahalie,<sup>1,2</sup> Annemarie Scholz,<sup>1,2</sup> Akash A. Shah,<sup>1,2</sup> Amitabh Sharma,<sup>1,2</sup> Yun Shen,<sup>1,2</sup> Kerstin Spirohn,<sup>1,2</sup> Stanley Tam,<sup>1,2</sup> Alexander O. Tejada,<sup>1,2</sup> Shelly A. Trigg,<sup>1,2</sup> Jean-Claude Twizere,<sup>1,2,15</sup> Kerwin Vega,<sup>1,2</sup> Jennifer Walsh,<sup>1,2</sup> Michael E. Cusick,<sup>1,2</sup> Yu Xia,<sup>1,14</sup> Albert-László Barabási,<sup>1,14,16</sup> Lilia M. Iakoucheva,<sup>13</sup> Patrick Aloy,<sup>1,17</sup> Javier De Las Rivas,<sup>10</sup> Jan Tavernier,<sup>9</sup> Michael A. Calderwood,<sup>1,2,20</sup> David E. Hill,<sup>1,2,20</sup> Tong Hao,<sup>1,2,20</sup> Frederick P. Roth,<sup>1,3,4,5,18,\*</sup> and Marc Vidal<sup>1,2,\*</sup>

# Comparing two recently released human interactomes....



## Size of two experimental human interactomes recently released

	<u>HuRI</u> <sup>1</sup>	BioPlex2.0 <sup>2</sup>
Number of proteins*	8,470	10,844
Number of interactions	51,907	55,498

<sup>1</sup> Database downloaded from <http://www.interactome-atlas.org> on July 27, 2019

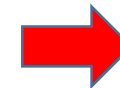
<sup>2</sup> Database downloaded from <https://bioplex.hms.harvard.edu> on July 27, 2019

\* Splicing isoforms are collapsed

Reference: 23,423 coding genes of GRCh38.p12

## Overlap between HuRI and BioPlex

	<u>HuRI</u>	BioPlex2.0
Number of shared proteins	4,827 (56.9%)	4,827 (44.5%)
Number of interactions among shared proteins	16,133 (31.1%)	12,610 (22.7%)
Number of shared interactions	829 (5.1%)	829 (6.6%)



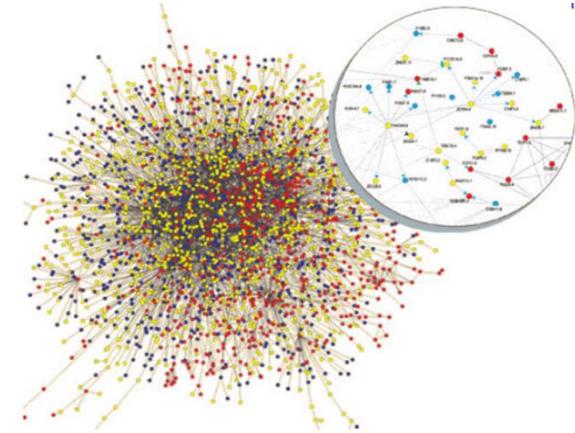
**1128 proteins with shared interactions**





## Open issues of the interactomic space:

- Data quality and reproducibility
- Coverage of the proteome
- Tissue specificity
- Subcellular localisation
- Co-expression



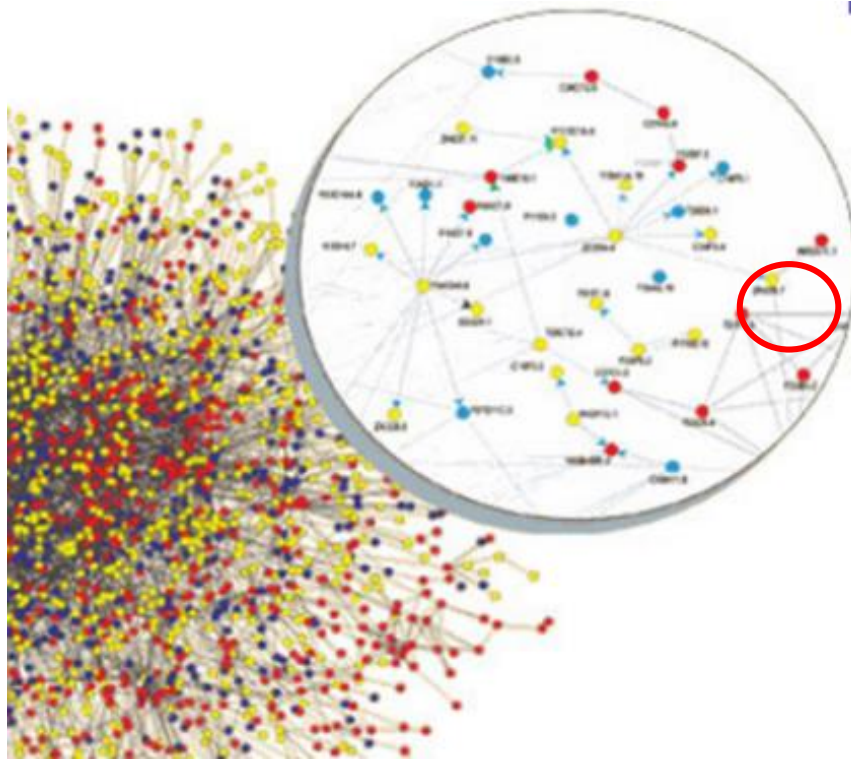


# To which extent do we know the properties of the different nodes/edges?

**System scale**



**Atomic scale**

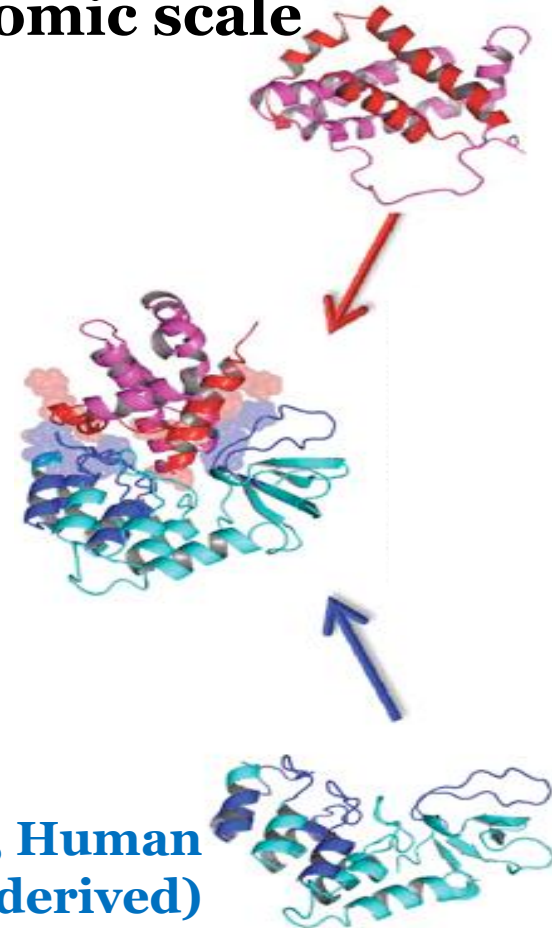


**IntAct, Human**

**287,042 binary physical interactions**  
**28,423 proteins involved**

**Interactome3D, Human  
(PDB derived)**

**5,575 resolved interactions**  
**1,843 proteins involved**



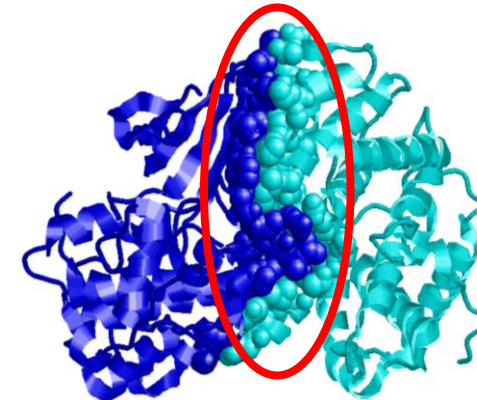


# Leveraging available structural information

## Prediction of protein interaction sites

Identifying residues in interaction patches in a protein,  
without knowing the interaction partner

starting from monomer structure

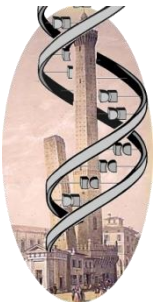
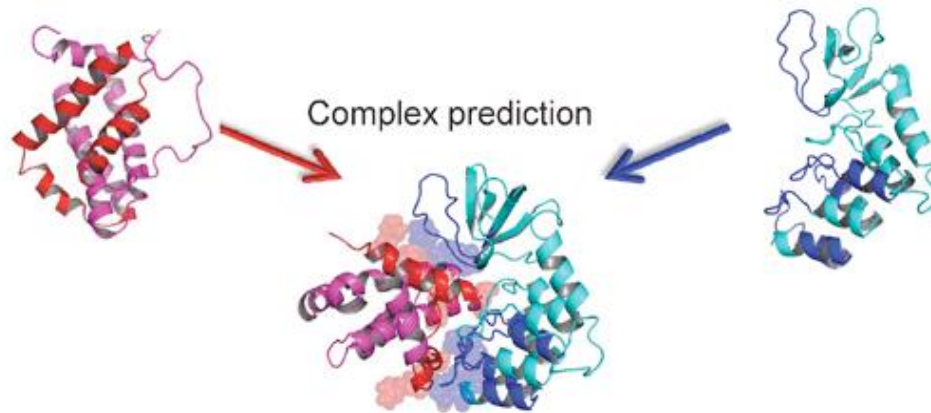


starting from monomer sequence

```
>sp|Q196Z6  
MDAVKHPVSSLAPHRLGKGFYFMVNSKQI IKKFYNKSFLECSATQETSTTPHDRDNMDQA  
KLC SLVFEIFKKQTRLLAHLVLEEAIDLNNDLLFAVIYFNDEAILNSLVRHLYKYKPYC  
DFTVRAQELDLVVHLDLGHCIDRLKSFIDPDTACFVLC SNLSQLTGLNCLKRVLKHKI IQ  
KSYHLYLLLKTSHKVQQQWDPVHVVEKYVTKRMVSYALTDNNPLLLAIVLDRLLVKLP  
DDFAPLIVGIIESNRFKVECLPTLLQYHNRVKT TTKPIRI
```

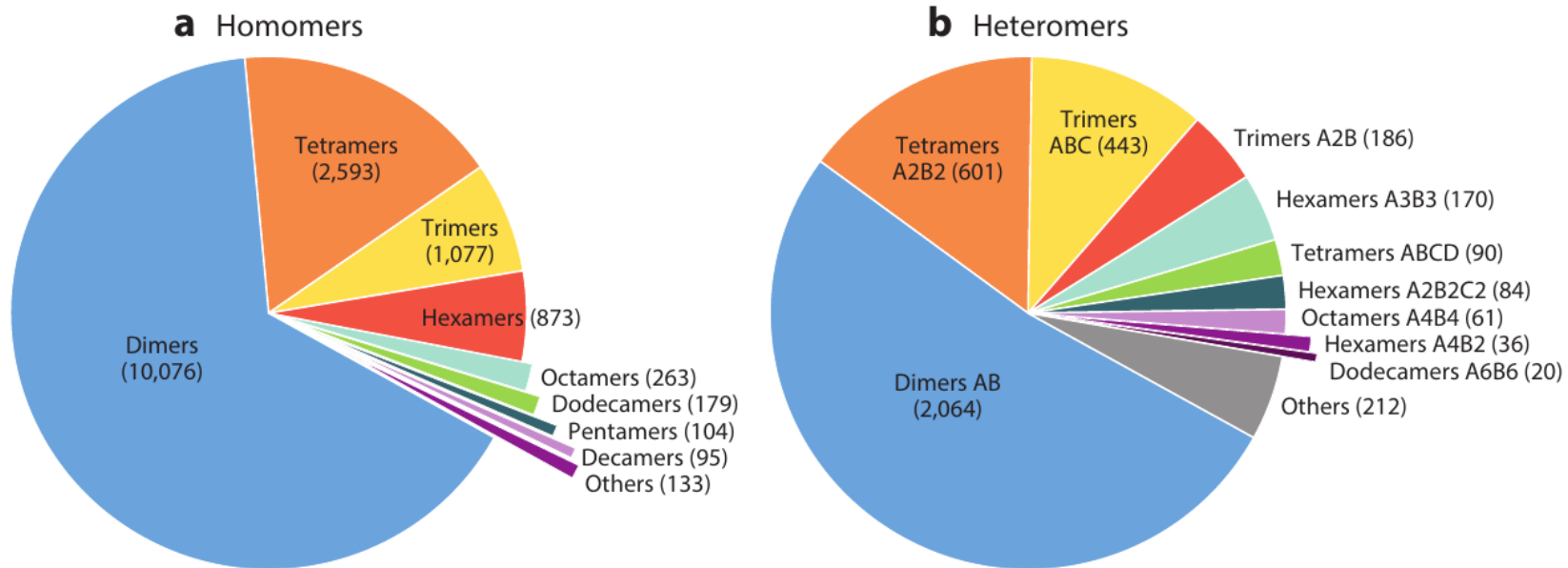
# Why interaction sites are important?

- **Characterization of protein function**
  - Interaction site  $\leftrightarrow$  functional site
  - Disease-related variations often occur at interaction sites
- **Discovery of novel interaction patches**
- **Improvement of docking methods**
  - By reducing the number of possible conformations

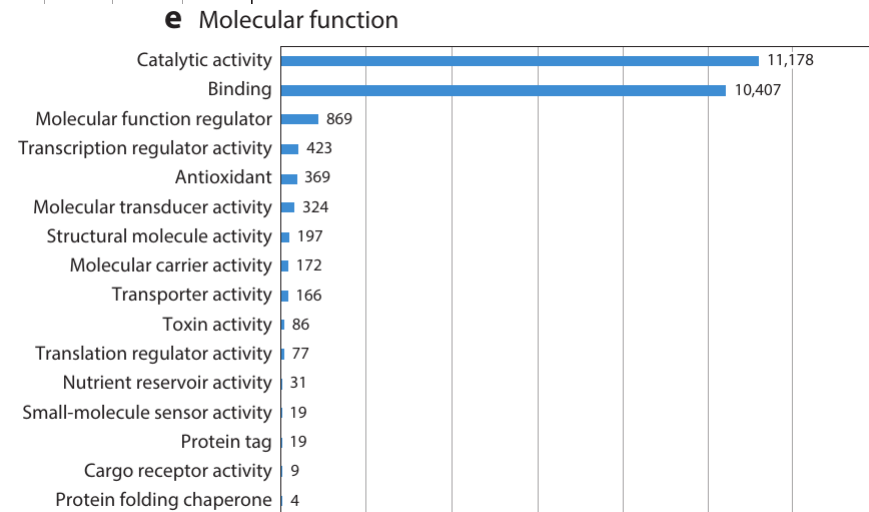
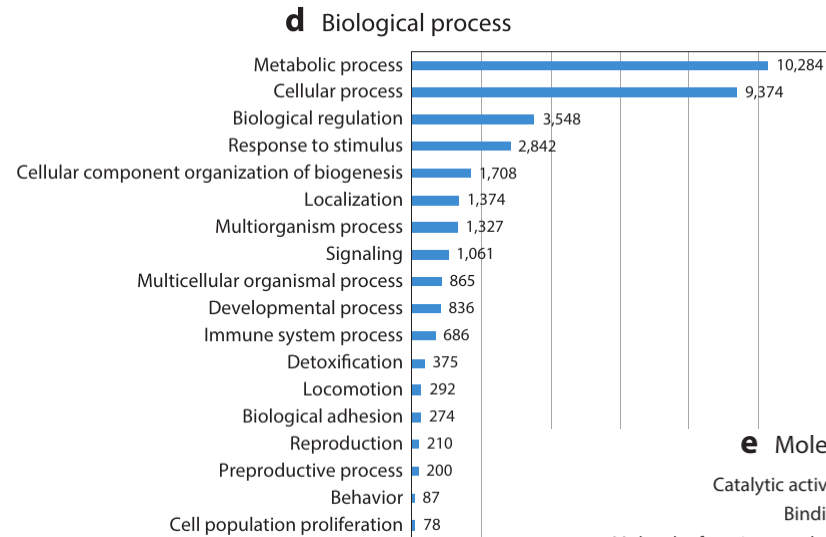
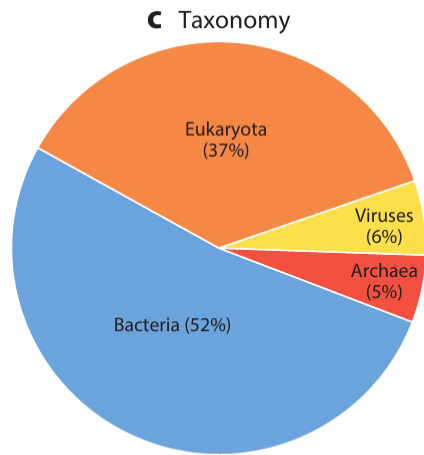


# Some Statistics.....from the PDB

**WHICH PROTEIN COMPLEXES ? (about 67.000 in the PDB as to July 2019)**  
*About 1/3 of the PDB*



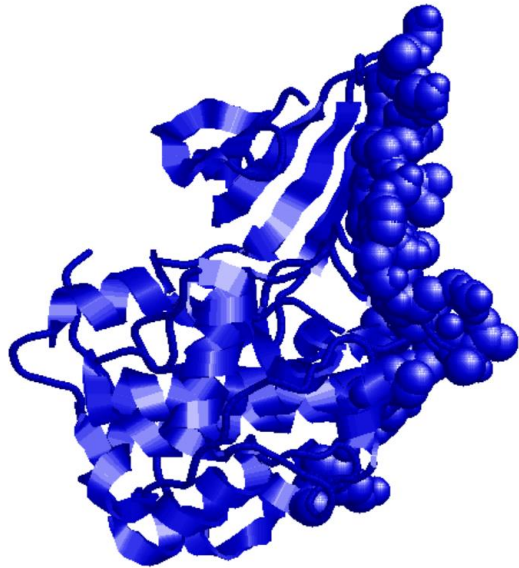
# Some Statistics.....from the PDB



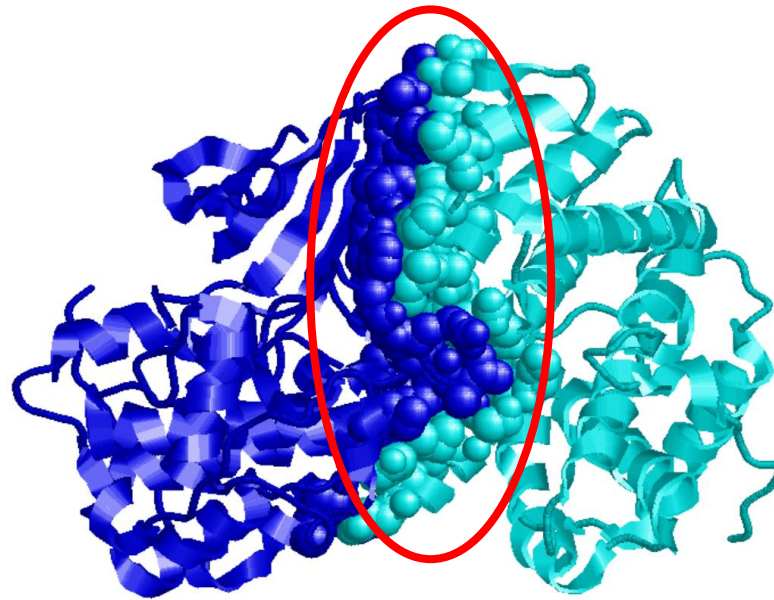
**TAXONOMY & GO terms  
of Protein Complexes**

*(Caption abt)*

# Definition of interaction surfaces



*Cyclin-dependent kinase 2*



*PDB: 5IF1*

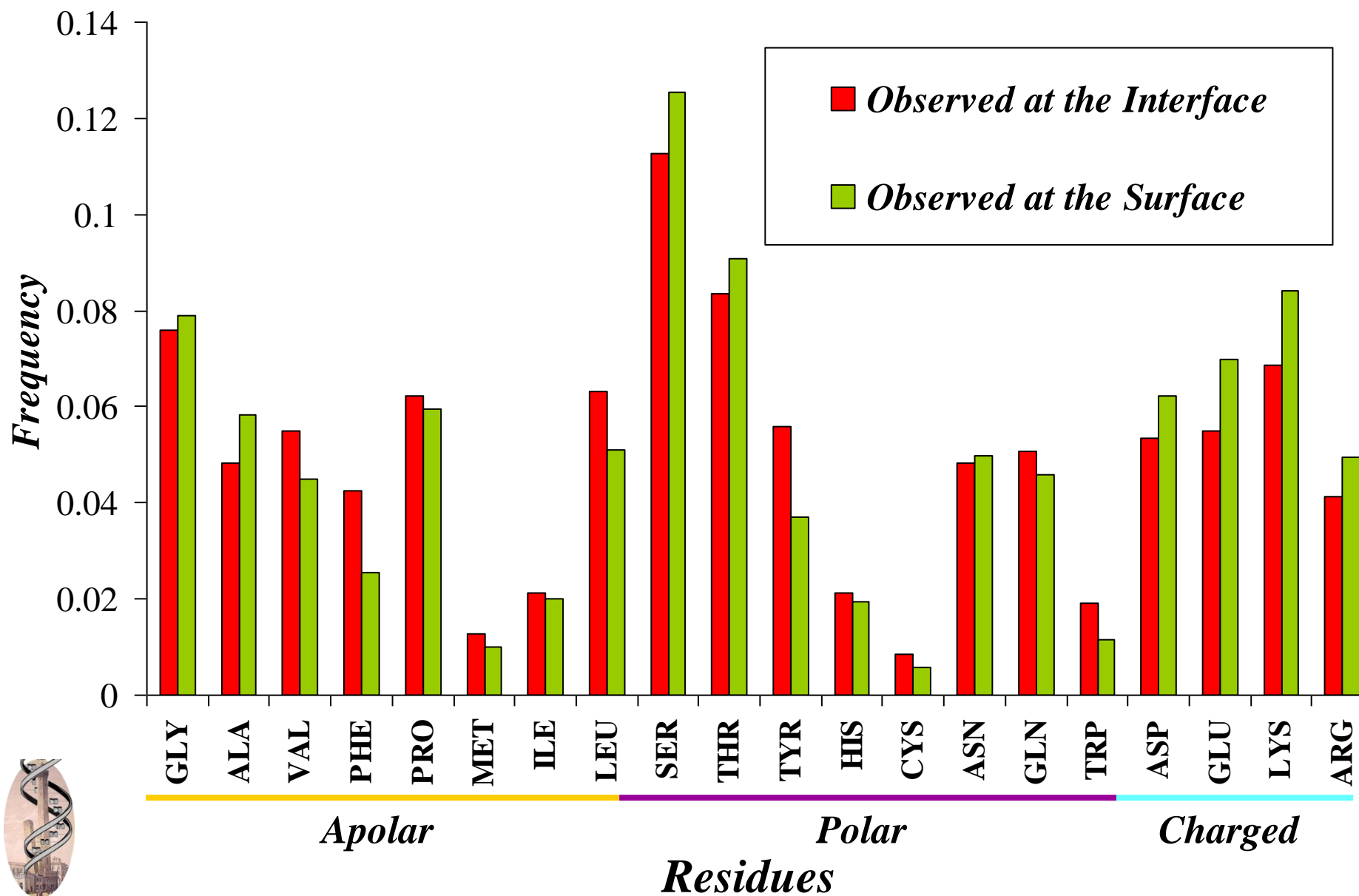


*Cyclin-A2*

## **Definition:**

The set of any residue showing a Difference in Accessible Surface Area (DASA  $\geq 1 \text{ \AA}^2$ ) between the monomer and the complex

# Distributions of apolar, polar and charged residues





**Protein-protein interaction is a biophysical complex phenomenon governed by:**

- 1) Shape**
- 2) Chemical complementarity**
- 3) Flexibility**
- 4) Residue specific composition (less charged, more hydrophobic than in the solvent accessible surface of the protein)**



**Hydrophobic interactions, weak electrostatic interactions, Van der Waals interactions...**

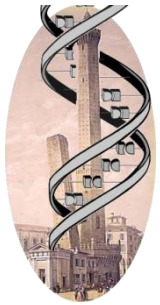
# Properties of interaction surfaces (>5000)

<b>Feature</b>	<b>Method/Program/Source</b>
<b><i>Residue propensity</i></b>	AA frequency tables
<b><i>Physico-chemical properties</i></b>	AAINDEX, hydropathy scales, propensities
<b><i>Residue conservation</i></b>	PSI-BLAST, Jackhmmer, HHBlits
<b><i>Curvature</i></b>	Coleman method, SurfRace
<b><i>Depth and protrusion indexes</i></b>	PSAIA
<b><i>Solvent exposure</i></b>	DSSP, PSAIA, NACCESS
<b><i>Secondary structure</i></b>	DSSP, STRIDE
<b><i>B-Factors</i></b>	Curated from PDB
<b><i>Electrostatic potentials</i></b>	APBS, FoldX, DelPhi
<b><i>Energy of solvation</i></b>	APBS

***No major emerging feature***



***Information is in complex feature combinations***



# A sample of implemented methods

Method	Category	Based on	Reference
<b>Gallet et al., 2000</b>	Sequence	AA frequency tables	Gallet et al., 2000
<b>Ofran and Rost, 2003</b>	Sequence	Sequence profile + Neural networks	Ofran and Rost, 2003
<b>Chen and Li, 2010</b>	Sequence	Hydrophobicity and sequence profiles + SVM	Chen and Li, 2010
<b>SPPIDER</b>	Structure	Predicted solvent accessibility fingerprint + SVM	Porollo and Meller, 2007
<b>cons-PPISP</b>	Structure	Structural features + Consensus Neural Networks	Chen et al, 2005
<b>ProMate</b>	Structure	Structural property histograms and patch refinement	Neuvirth et al, 2004
<b>PresCont</b>	Structure	Evolutionary information+structural features+SVM	Zeliner et al, 2012
<b>PredUS</b>	Template + Structure	Structural neighbors transfer+SVM refinement	Zhang et al, 2001
<b>PrISE</b>	Template	Local surface similar based on structural element distributions	Jordan et al, 2012
<b>HomPPI</b>	Template	Homologous sequence-based transfer	Xue et al, 2011

# Nowadays: a classification of available computational approaches

- **Template-based methods**

- Exploit sequence/structure similarity to transfer interaction sites from known structural templates
- **Good accuracy**
- **Requires templates of the interacting complex**

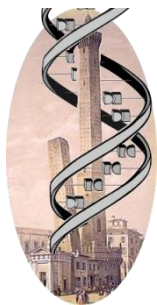
- **Machine-learning approaches**

- *Sequence-based predictors*

- Extract information from monomer sequence
- **Broad applicability**
- **Low accuracy**

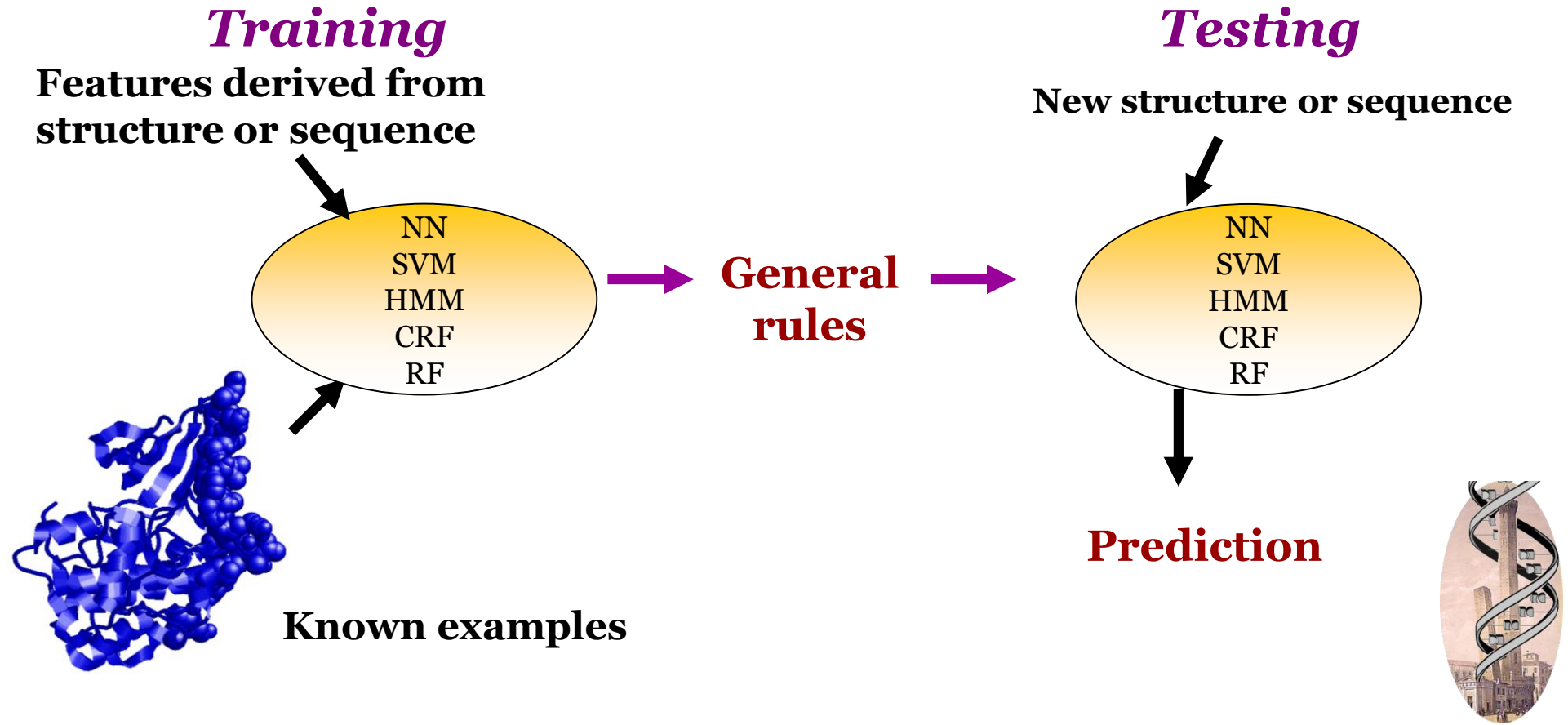
- *Structure-based predictors*

- Extract information from both sequence and structure of the monomer
- **More accurate than sequence-based methods**
- **Requires structural model of the monomer**



# Machine learning methods

- **Problem:** label each residue as interacting or not
- **Methods:** Neural Networks (NN), Support Vector Machines (SVM), Hidden Markov Models (HMMs), Conditional Random Fields (CRF), Random Forests (RF)



# Predictors @ Bologna Biocomputing Group

[www.biocomp.unibo.it/predictors](http://www.biocomp.unibo.it/predictors)

- **ISPRED<sub>1</sub> (Fariselli et al., 2001, 2002)**
  - Structure-derived features
  - Method: Artificial Neural Networks (ANNs)
  - Input features: sequence profiles computed from multiple sequence alignments
- **ISPRED<sub>2</sub>, ISPRED<sub>3</sub> (Savojardo et al., 2011)**
  - Structure-derived features
  - Method: Hidden Markov Support Vector Machines (HM-SVMs)
  - Input features: sequence profiles + solvent accessibility
- **ISPRED<sub>4</sub> (Savojardo et al., 2017)**
  - Structure-derived features
  - Method: SVM+Grammatical-Restrained Hidden CRF (Fariselli et al., 2009)
  - Input features: extended feature set to encode each surface residue
- **ISPRED-SEQ (Savojardo et al., 2020, in preparation)**
  - Sequence-derived features
  - Method: Deep learning
  - Input features: extended feature set to encode each residue



## Prediction of protein–protein interaction sites in heterocomplexes with neural networks

Piero Fariselli<sup>1</sup>, Florencio Pazos<sup>2</sup>, Alfonso Valencia<sup>2</sup> and Rita Casadio<sup>1</sup>

<sup>1</sup>CIRB and Department of Biology, University of Bologna via Irnerio, Bologna, Italy; <sup>2</sup>Protein Design Group, CNB-CSIC Cantoblanco, Madrid, Spain

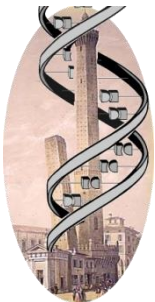
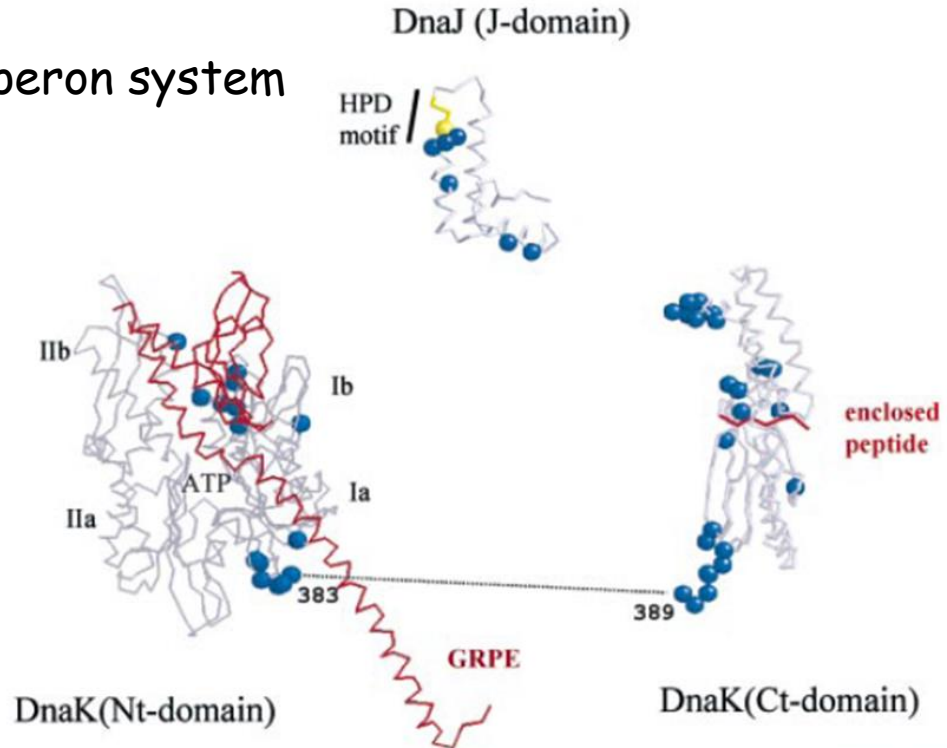
In this paper we address the problem of extracting features relevant for predicting protein–protein interaction sites from the three-dimensional structures of protein complexes. Our approach is based on information about evolutionary conservation and surface disposition. We implement a neural network based system, which uses a cross validation procedure and allows the correct detection of 73% of the residues involved in protein interactions in a selected database comprising 226 heterodimers. Our analysis confirms that the chemico-physical properties of residues are difficult to distinguish from the surface. However neural network representation of the interacting

are sufficient to generalize over the different features of the contact patches and to predict whether a residue in the protein surface is or is not in contact. By using a blind test, we report the prediction of the surface interacting sites of three structural components of the DnaK molecular chaperone system, and find close agreement with previously published experimental results. We propose that the predictor can significantly complement results from structural and functional proteomics.

## ISPRED & the role of experimental validation

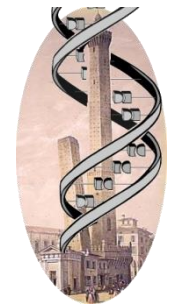
### The DnaK molecular chaperon system

**Fig. 4. Prediction of the interacting surface for the three structural components of the DnaK molecular chaperone system.** The structures of DnaK N-terminal and C-terminal domains, that has been determined separately (PDB codes 1dkg and 1dkx, respectively), are shown at the bottom. The structure of the DnaJ J-domain (PDB code 1xbl) is shown at the top. CA carbons of residues predicted at the putative interfaces by the neural network are shown as spheres depicted in blue. The peptide fragment (enclosed in the DnaK Ct-domain) and the nucleotide exchange factor GrpE protein (co-crystallised with the DnaK Nt-domain) are shown in red colour with thick backbone. The DnaJ conserved HPD motif is shown in yellow.



# ISPRED4: features

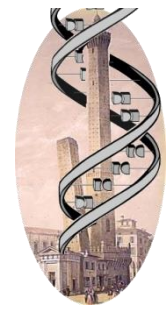
- **Surface residues are determined (relative solvent accessibility  $\geq 20\%$ )**
- **46 descriptors are used to encode each surface residue**
- **Descriptors derived from sequence and averaged on structure:**
  - Sequence profile from MSA (20)
  - Residue conservation and co-evolution from MSA (3)
  - Residue physico-chemical properties (11)
- **Descriptors extracted from structure:**
  - Solvent exposure computed by DSSP (1)
  - Depth and protrusion geometrical indexes (7)
  - Secondary structure (3)
  - Average B-factor (1)



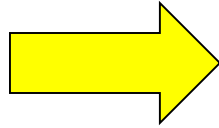
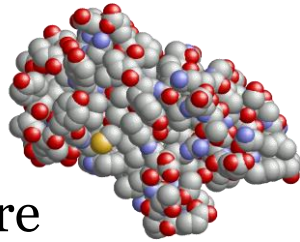




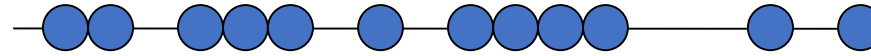
# ISPRED4 workflow



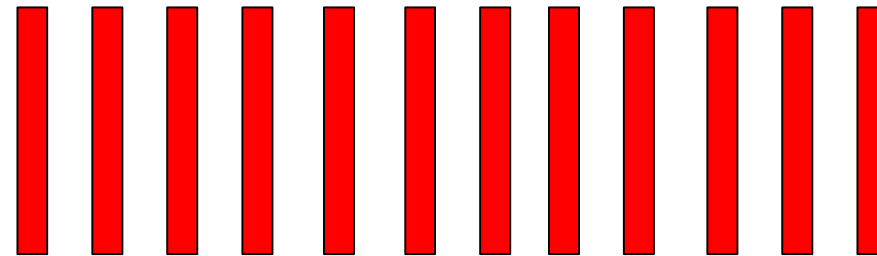
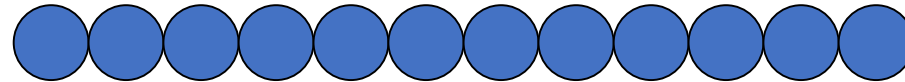
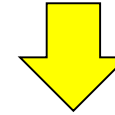
3D  
structure



Protein Sequence

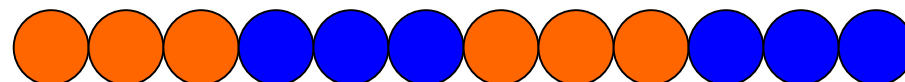
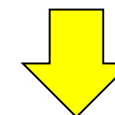


Retain only exposed residues  
(RSA  $\geq$  20%)



Input vectors  
(46 features)

*SVM+GRHCRFs*



 Non-interaction site

 Interaction site

**Predicted classification:**

Savojardo et al., Bioinformatics, 2017  
Savojardo et al., CIBB 2011, LNBI 7548, 2011  
Fariselli et al., AI Mo Biol., 2009

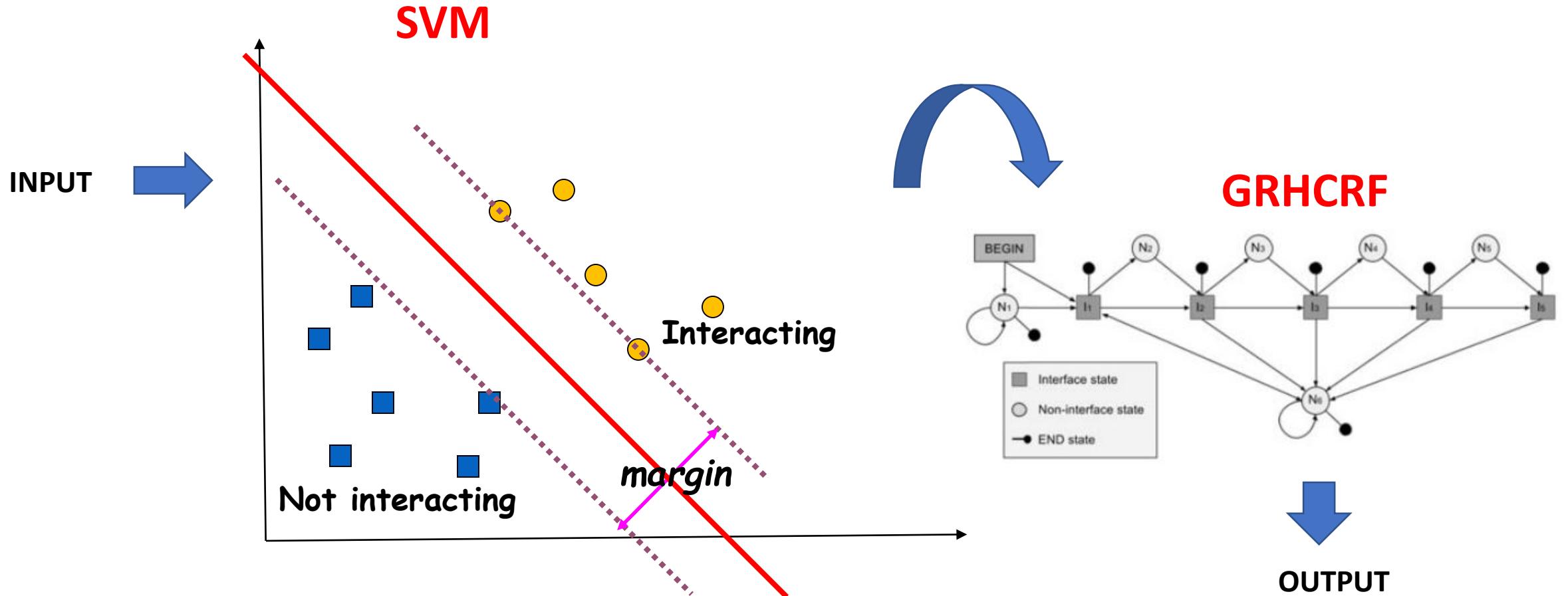


# ISPRED4: Support Vector Machines (SVM) + Grammatical Restrained Hidden Conditional Random Fields (GRHCRFs)

Savojardo et al., Bioinformatics, 2017

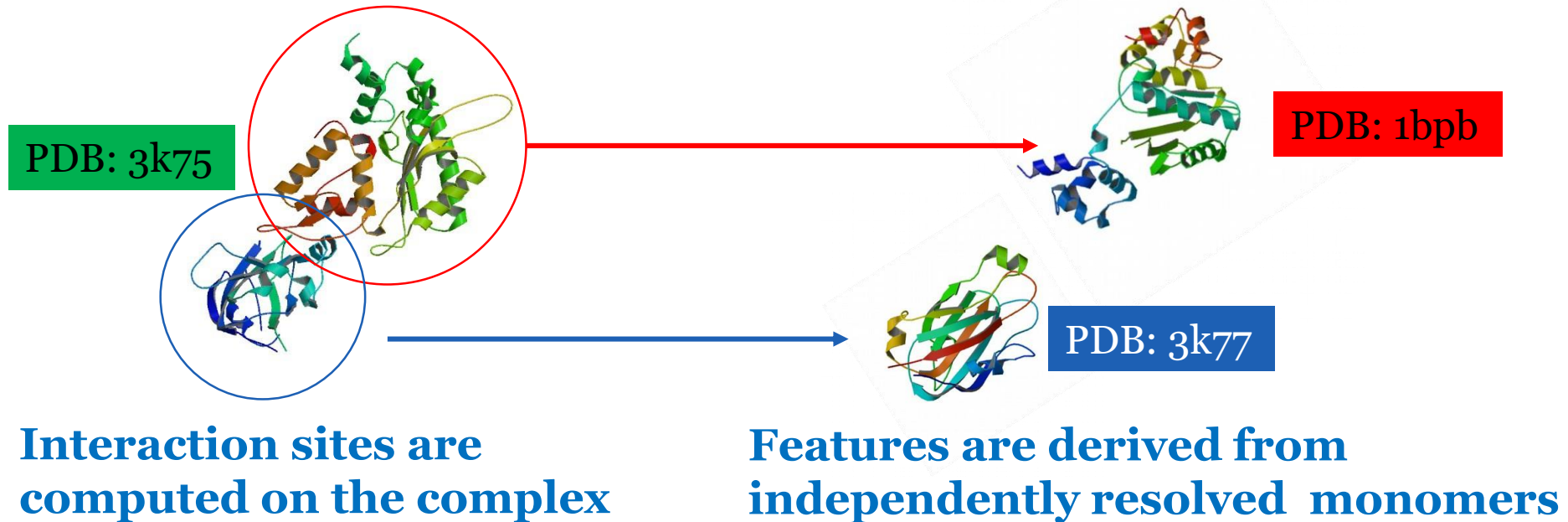
Savojardo et al., CIBB 2011, LNBI 7548, 2011

Fariselli et al., Al Mo Biol., 2009



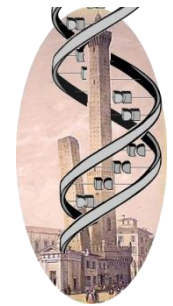
# Training dataset

To avoid biases due to conformational changes upon binding



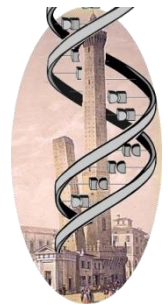
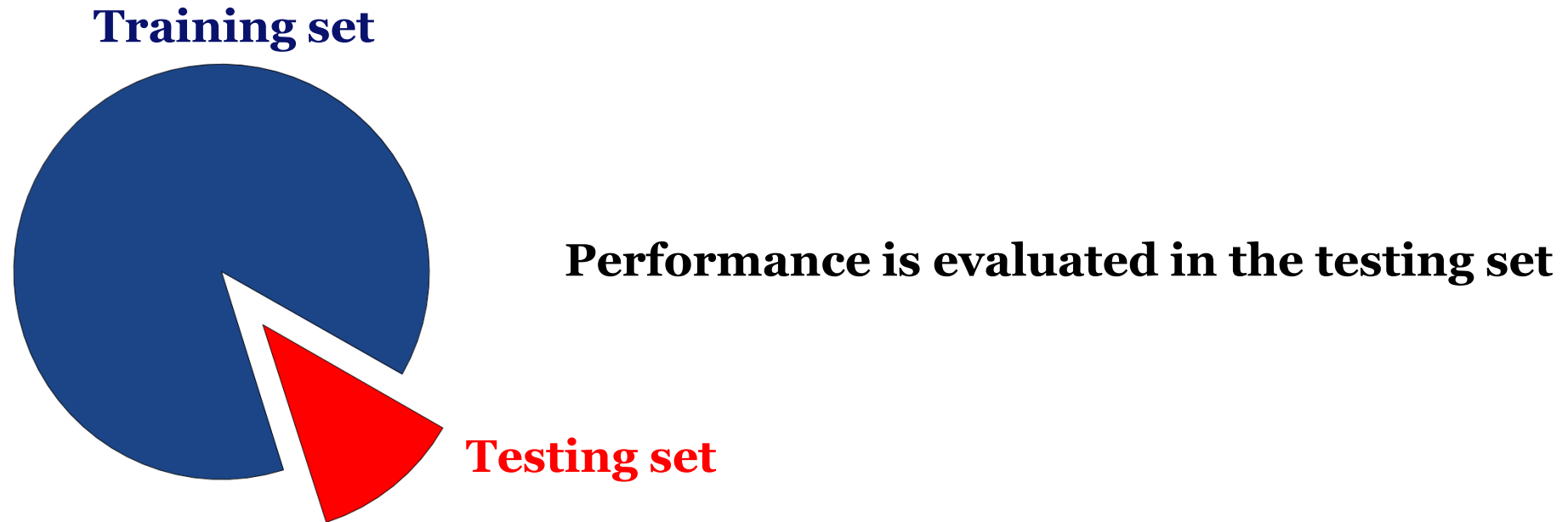
- **TrainStructDB: 151** high-resolution protein complexes derived from the Docking Benchmark v5 => **314** unbound chains

- 67,235 total residues
- 39,046 exposed residues
- 8,649 interaction sites 30,397 non interaction sites



# Evaluation procedure

- **10 Fold cross-validation procedure on TrainStructDB**



# Scoring indexes

- **Residue-level scoring measures**

TP: True positives, FN: False negatives

FP: False positives, TN: True negative

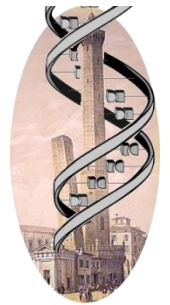
$$\text{Precision(Specificity)} = \frac{TP}{TP + FP}$$

$$\text{Recall(Sensitivity)} = \frac{TP}{TP + FN}$$

$$Q2 = \frac{TP + TN}{TP + TN + FP + FN}$$

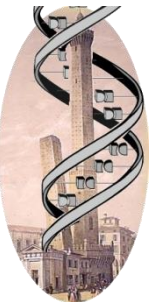
$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$



# CAPRI Blind dataset

- **CAPRI: Critical Assessment of Predicted Interactions**  
(<https://www.ebi.ac.uk/msd-srv/capri/>)
- Targets extracted from past CAPRI experiments (rounds 1-29)
- Only targets sharing < 30% sequence identity with any sequence in the training set
- 22 different bound structures including 29 chains
  - 6,369 total residues
  - 3,613 exposed residues
  - 868 interaction sites                      2,745 non interaction sites



# Performance comparison

## Performance on the TrainStructDB dataset (Cross-validation)

Method	Method type	Precision	Recall	F1	Q2	MCC
ISPRED4	Structure	0.78	0.39	0.52	0.84	0.48
ISPRED3	Structure	0.26	0.80	0.39	0.47	0.16
SPPIDER	Structure	0.39	0.54	0.45	0.72	0.28
cons-PPISP	Structure	0.46	0.27	0.34	0.77	0.23
PredUs	Homology	0.37	0.76	0.50	0.67	0.34
PrISE	Homology	0.42	0.41	0.41	0.83	0.33

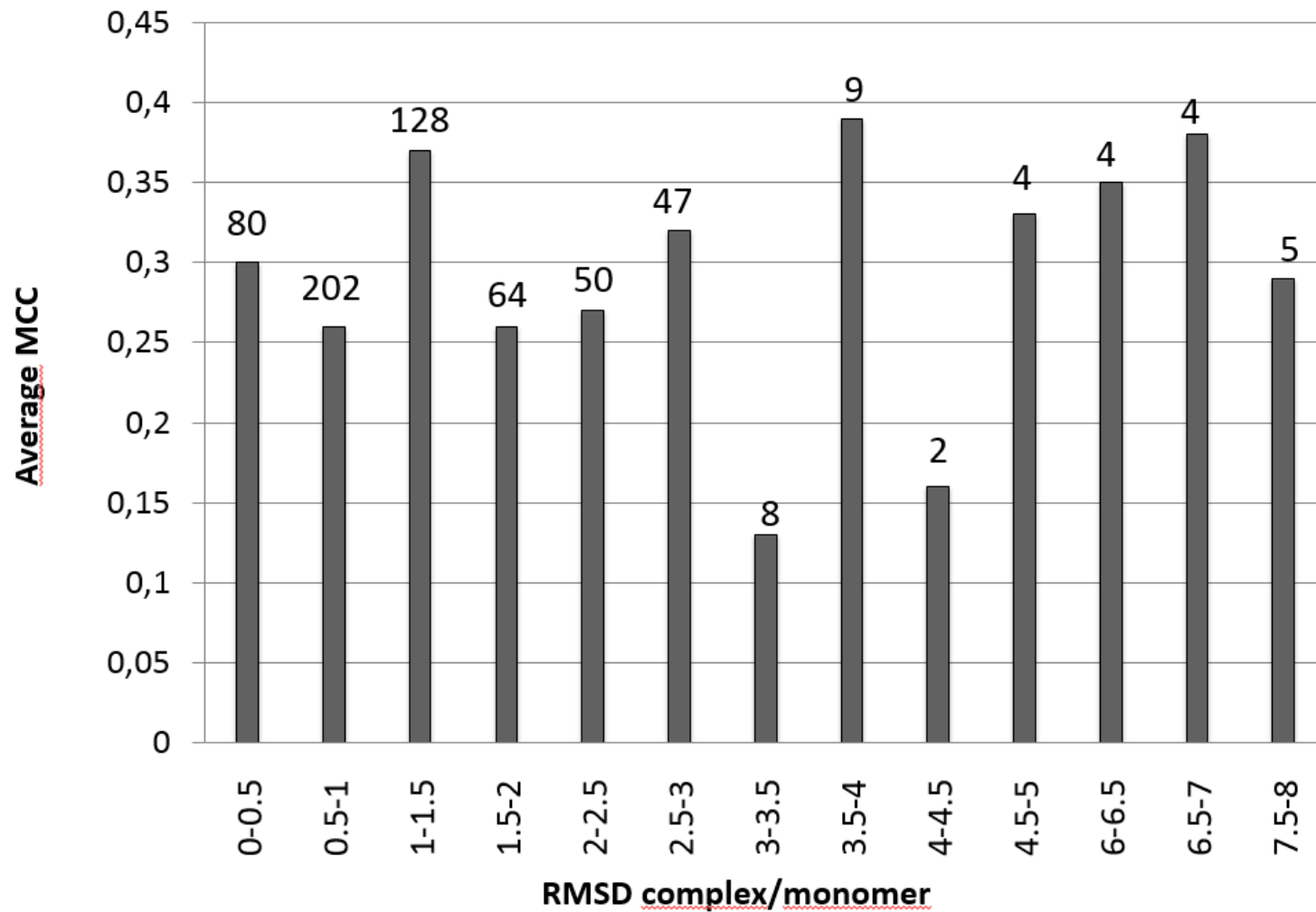
## Performance on the CAPRI blind dataset

Method	Method type	Precision	Recall	F1	Q2	MCC
ISPRED4	Structure	0.60	0.38	0.47	0.67	0.28
ISPRED3	Structure	0.26	0.68	0.38	0.45	0.05
SPPIDER	Structure	0.36	0.39	0.37	0.68	0.16
cons-PPISP	Structure	0.33	0.17	0.22	0.72	0.08
PredUs	Homology	0.38	0.62	0.47	0.67	0.26
PrISE	Homology	0.41	0.36	0.38	0.72	0.21





# ISPREd4 scoring over the Docking Benchmark Version 5

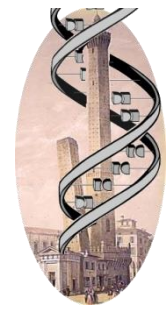


609 monomers/230 complexes as from Docking Benchmark Version 5  
(Vreven et al, J. Mol. Biol 2015, 427)



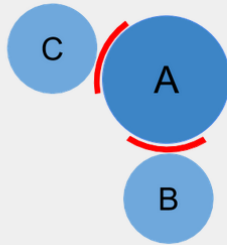
# ISPRED4 website

<http://ispred4.biocomp.unibo.it>



Home SearchJob Software/Datasets References Contact BiocomputingGroup

## Welcome to the ISPRED4 prediction server



ISPRED4 (Interaction Site PREDictor version 4) is a web-server for predicting protein-protein interface residue starting from protein structure. In particular, ISPREDv4 adopts machine-learning methods (SVM+CRF) to predict interaction state of each residue in the protein by extracting several features from the protein sequence and structure.

### Submit a PDB file

To start using ISPRED4 you simply need to upload a protein 3D structure in PDB format and specify the protein chain you want to analyze. Therefore, press the "Start predictor" button to submit your job. The server accepts in input a single protein structure. [Download example input data.](#)

Input PDB file:  Nessun file selezionato

PDB chain:

### Sequence view

Protein id: 5HLU [chain A]  
 Protein length: 152  
 Surface length (RSA>=0.16): 101

Interface  Surface  Buried

```

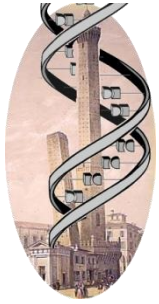
2  L E E G G W S V H R V W A K V H A D V A G R G D I L T R L F K R S H P E R F K H L K K F A R M K A B E D T
62  K R A G V Y E P A L G A I T R K K G R H R A E L K F L A Q S H A C H R R E P R K Y L E P F I S E A T R V L N R H R P
122 P F A R A A G A M H R A I E L F R K E D A R Y K E L G Y G G
    
```

### Detailed report

Prev 1 2 3 4 5 6 7 8 Next

Residue number	Residue type	ASA	RSA	Predicted RSA	Depth	Protrusion	Surface	Interface	Probability
22	A	19	0.18	0.42	0.53	0.51	yes	no	0.21
23	G	7	0.08	-	-	-	no	-	-
24	H	4	0.02	-	-	-	no	-	-
25	G	1	0.01	-	-	-	no	-	-
26	Q	29	0.15	-	-	-	no	-	-
27	D	44	0.27	0.14	0.34	0.33	yes	no	0.27
28	I	2	0.01	-	-	-	no	-	-
29	L	3	0.02	-	-	-	no	-	-
30	I	7	0.04	-	-	-	no	-	-
31	R	104	0.42	0.36	0.36	0.58	yes	yes	0.88
32	L	11	0.07	-	-	-	no	-	-
33	F	3	0.02	-	-	-	no	-	-
34	K	102	0.5	0.31	0.17	0.77	yes	no	0.44
35	S	49	0.38	0.18	0.22	0.79	yes	yes	0.74
36	H	39	0.21	0.16	0.81	0.54	yes	no	0.4
37	P	78	0.57	0.33	0.43	0.86	yes	yes	0.92
38	E	68	0.35	0.25	0.14	0.85	yes	yes	0.7
39	T	6	0.04	-	-	-	no	-	-
40	L	16	0.1	-	-	-	no	-	-
41	E	136	0.7	0.28	0.0	1.2	yes	yes	0.82

COPYRIGHT



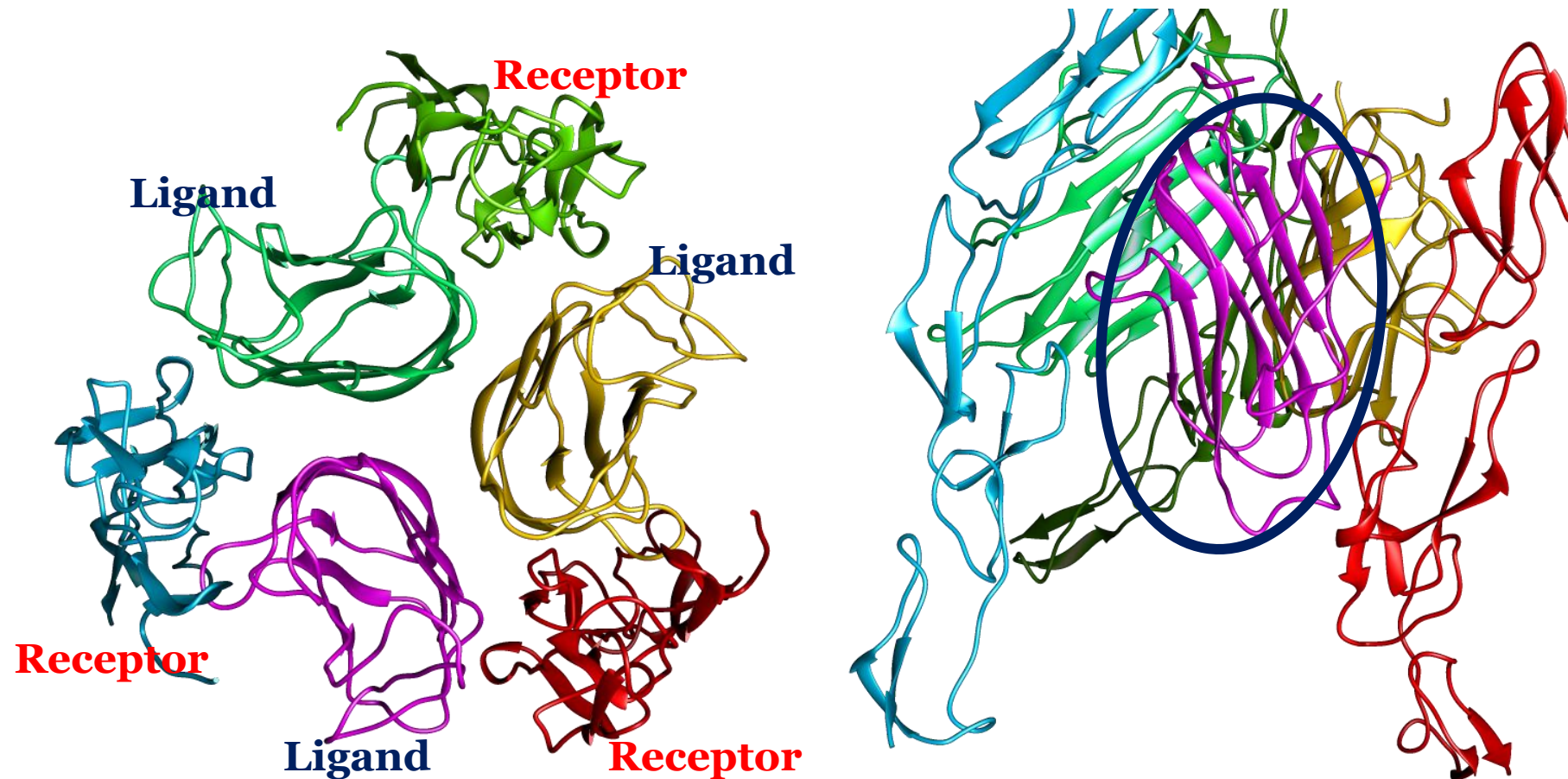
## ISPRED4 at work & applications

# 2HEV: complex between OX40L and OX40 extracytoplasmic domains (2.41 Å)

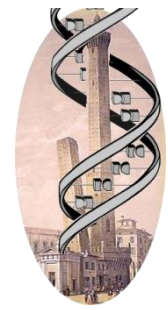


**Tumor necrosis factor ligand superfamily member 4 (TNFSF4)**

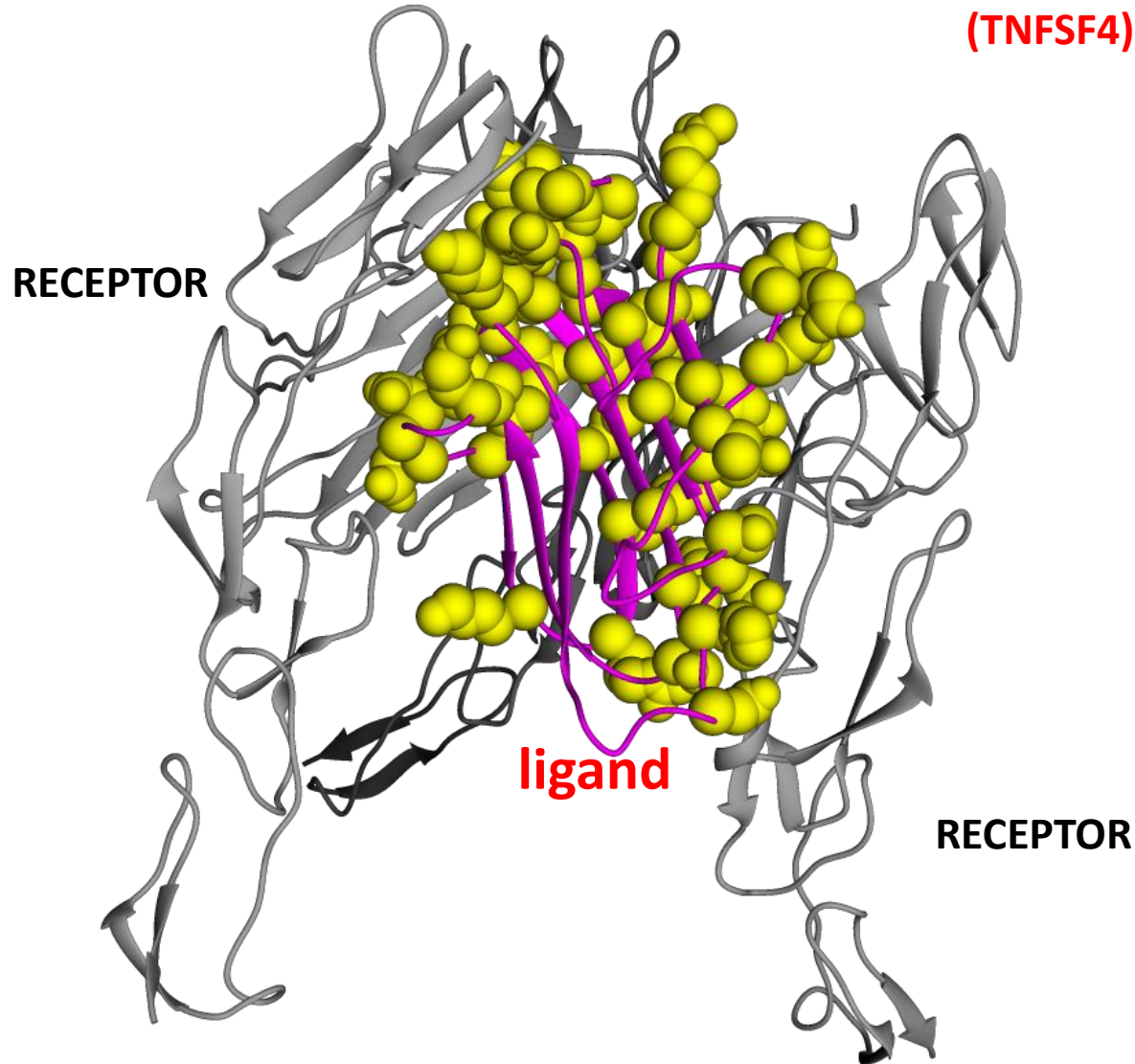
**Tumor necrosis factor receptor superfamily member 4 (TNFRS4)**



# 2HEV: known interaction sites



## Tumor necrosis factor ligand superfamily member 4 (TNFSF4)

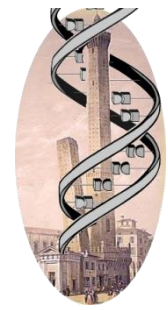


**79 solvent accessible residues  
out of 126**

**● Interaction sites (43)**

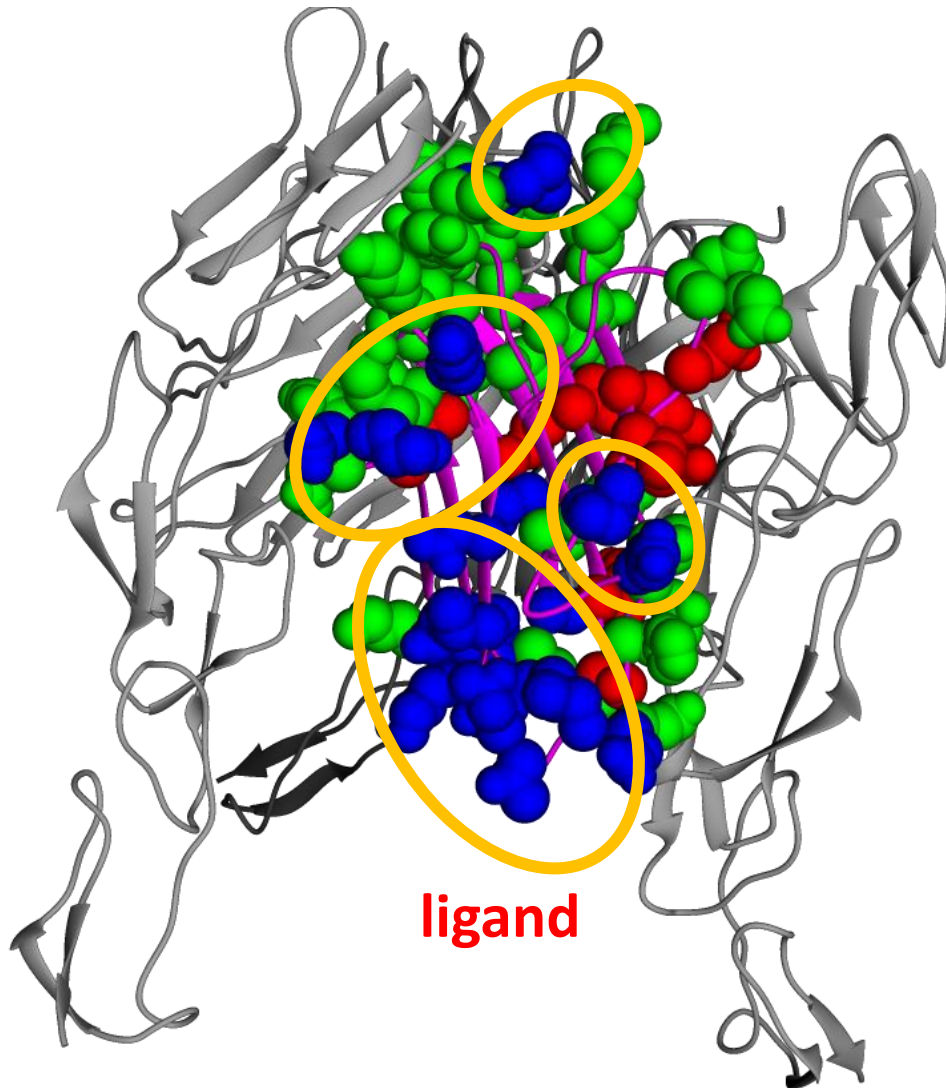
**Both interfaces with the  
receptor units are considered**




# 2HEV: prediction with ISPRED4



## Tumor necrosis factor ligand superfamily member 4

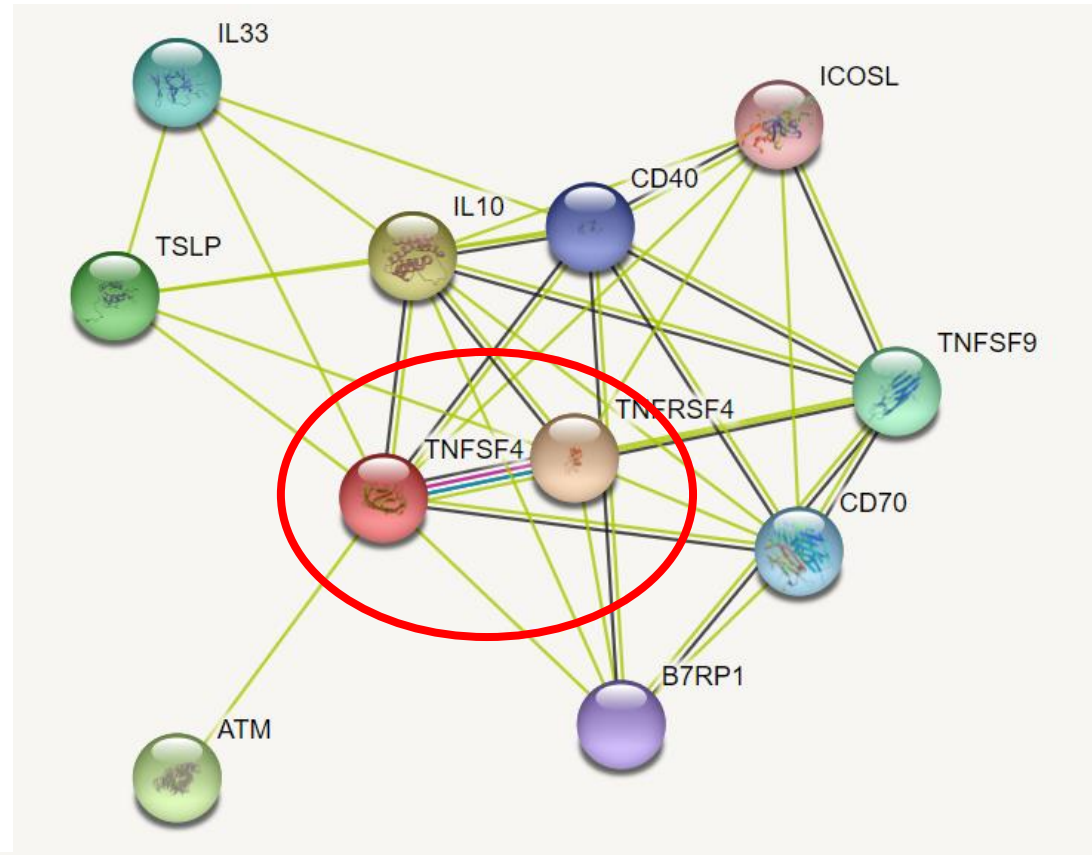
(TNFSF4)



-  Correctly predicted (34/43)
-  False negatives (9/43)
-  False positives (19)  
or new possible interaction patches?

*High precision predictor (0.78)*

# New possible interaction patches for the complex: other interactions reported in PPI networks



Known Interactions

- from curated databases
- experimentally determined

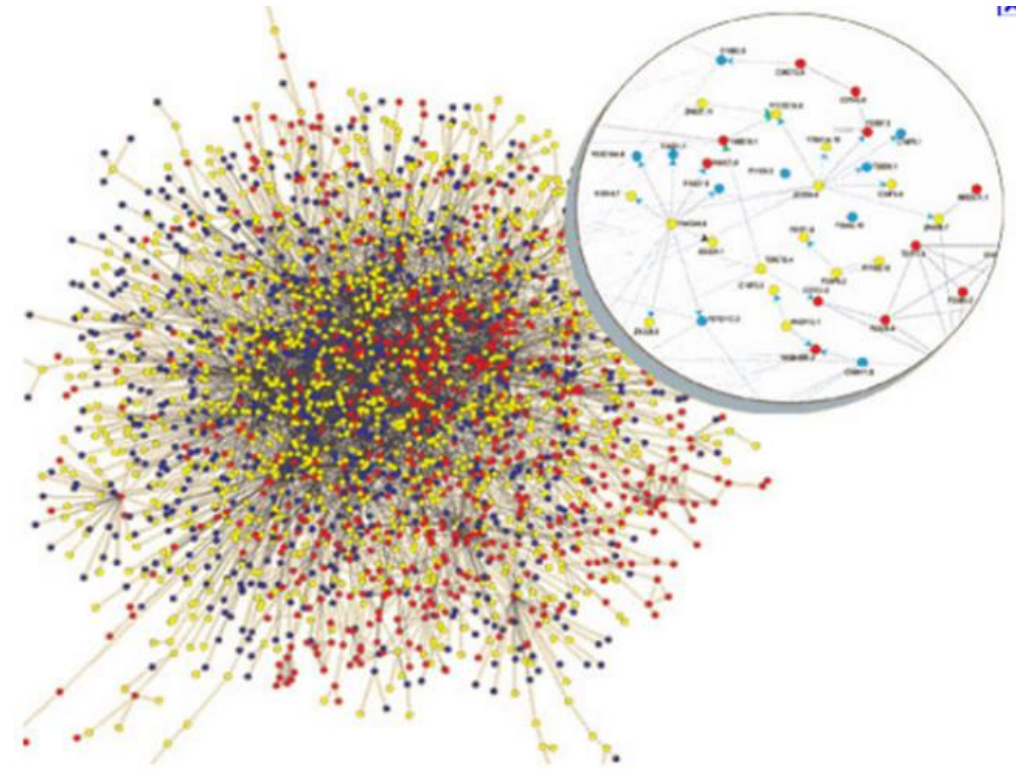
From STRING, <https://string-db.org/>

## Application 2

# Searching for protein interactions sites that are involved in disease related variations

### Materials

- The human protein 3D data base
- Humsavar ([ebi16.uniprot.org/docs/humsavar](http://ebi16.uniprot.org/docs/humsavar))
- ISPRED4



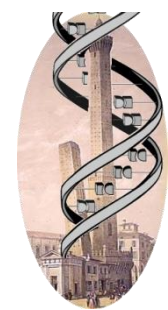


# Interaction sites, variations and diseases

	Disease related	Neutral	Total	No of proteins
<b>Varianted sites on OMIM related proteins</b>	14,103	9,127	23,230	3,804
<b>Mapped on PDB</b>	7,204	3,151	10,355	1,363
<b>Solvent exposed</b>	2,774	2,115	4,889	1,324
<b>Predicted in interaction</b>	1,194	703	1,897	654



	Disease related	Neutral
<b>Interaction/exposed</b>	43%	33%
<b>Interaction site distribution</b>	63%	37%



## Application 3

### Testing hypothesis:

***Are proteins involved in phase transitions endowed with intrinsically disorder regions (IDRs)?***

**e.g. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, et al. 2018. Protein Phase Separation: A New Phase in Cell Biology. Trends Cell Biol. 28:420-435**

### Materials

- **A data base of detected protein-protein interactions ( IntAct and Bio-Grid)**
- **A data base of a database of protein disorder and mobility annotations (MobiDB)**
- **A set of proteins known to form a membrane-less organelle**
- **ISPRED4 structure and/or sequence based**



*IntAct: [www.ebi.ac.uk/intact/](http://www.ebi.ac.uk/intact/); Bio-Grid: [thebiogrid.org/](http://thebiogrid.org/); MobiDB: [mobidb.bio.unipd.it/](http://mobidb.bio.unipd.it/)*

## Application 3

# The membrane-less organelle and its proteins

- *Cajal bodies (CBs) are spherical nuclear bodies of 0.3–1.0  $\mu\text{m}$  in diameter found in the nucleus of proliferative cells like embryonic cells and tumor cells, or metabolically active cells like neurons. CBs are membrane-less organelles and largely consist of proteins and RNA.*
- *25 proteins were present in UniprotKB (September 2019) with a Cajal body Cellular Component annotation. Most of the proteins have only a known sequence.*



# Application 3

## Predicting protein-protein interaction sites and flexible regions of proteins of the Cajal body

<u>UniProt</u>	<u>Gene</u>	<u>Length (#)</u>	<u>PPI (#)</u>	<u>Flexible sites (#)</u>	<u>Flexible PPI (#)</u>	<u>IntAct interactors (#)</u>	<u>BioGRID interactors (#)</u>
P38432	COIL	576	149	244	14	123	110
Q9BUR4	WRAP53	548	166	165	24	41	54
Q16637	SMN1	294	90	145	44	268	213
P55199	ELL	621	64	194	3	36	54
Q06787	FMR1	632	106	243	49	294	84
Q14331	FRG1	258	44	90	13	14	19
Q15020	SART3	963	115	199	4	125	211
Q5JVS0	HABP4	413	112	313	87	25	105
Q5W0Q7	USPL1	1092	136	169	4	24	25
Q6NT76	HMBX1	420	42	151	12	130	70
Q7LC14	DDX46	1021	170	228	55	27	65
Q7ZE							
Q8W							

# RESULTS



## Correlation between the number of interactors, PPI and flexible sites of the Cajal granule proteins

	PPI (#)	Flexible sites (#)	<u>Flexible PPI (#)</u>
<u>IntAct</u>	0.4 *	0.05	0.59 **
<u>BioGRID</u>	0.41 *	0.12	0.59 **

\* Significant at 5%

\*\* Significant at 1%



### Conclusions

- The Cajal body human proteins have a much larger number of interactors than the average (13 and 18 per human protein, respectively in IntAct and in BioGRID).
- The number of interactors per protein moderately correlates with the number of residues in IDRs (Flexible sites)
- Correlation increases if the number of PPI is considered
- Correlation reaches a satisfactory value when the number of residues that can be annotated both as PPI and IDR is considered



**Suggestion: the inherent flexibility of the residues makes it possible to adjust the interacting surface protein to multiple partners**

## *Our predictors in Bologna*

### PREDICTORS (NN, SVM, HMM, GRHCRF)

**BaCellLo** - Balanced subCellular Localization predictor

**BAR+** - Bologna Annotation Resource

**BetAware** - Detection of Prokaryotic outer-membrane betabarrel proteins

**CCHMM** - Predictor of Coiled-Coils Regions in Proteins

**CCHMMPROF** - Predictor of Coiled-Coils Regions in Proteins exploiting evolutionary information

**CORNET** - Predictor of Residue Contacts in Proteins

**DCON** - Predictor of Disulfide Connectivity in Proteins

**DisLocate** - Find Disulfide bonds in Eukaryotes with predicted subcellular Localization

**FT-COMAR** - Fault Tolerance Reconstruction of 3D Structure from Protein Contact Maps

**HIPPIE** - Protease Inhibitor engine

**I-MUTANT** - Neural Network based Predictor of Protein stability Changes upon Single Point Mutation

**I-MUTANT 2.0** - Support Vector Machines based Predictor of Protein stability Changes upon Single Point Mutation from Protein Sequence and Structure

**I-MUTANT Suite** - Support Vector Machines based Predictor of Protein stability Changes of protein variants and of human SNPs

**ISPRED** - Predictor of Protein Interaction Sites

**K-Fold** - Predictor of the Protein Folding Mechanism and Rate

**PhD-SNP** - Support Vector Machines based Predictor of human Deleterious Single Nucleotide Polymorphisms

**PredGPI** - Predictor of GPI-Anchored Proteins

**SNPs&GO** - Predictor of Human Disease-related Mutations in Proteins with Functional Annotations

**SPEPLip** - Predictor of Signal Peptide and Lipoprotein Cleavage Sites in Proteins

**YAP** - Yet Another Alignment Program (Pairwise Sequence Alignment Using Secondary Structures)

### APPLICATION SERVERS

**TRAMPLE**: the transmembrane protein labelling environment

**PONGO**: a web server for multiple predictions of all-alpha transmembrane proteins

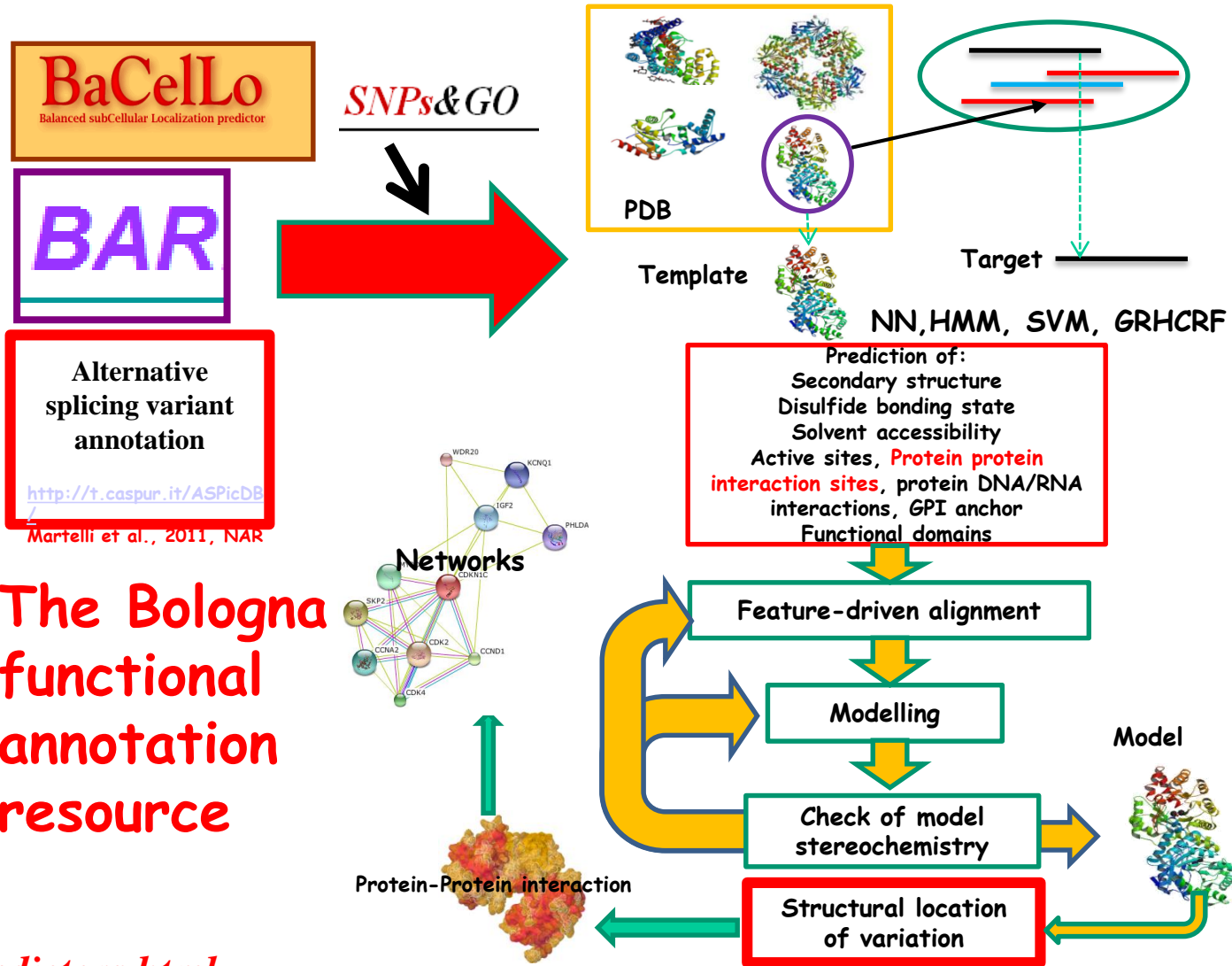
### DATA BASES

**eSLDB** - eukaryotic Subcellular Localization DataBase

**ZenPatches** - Database of predicted protein interaction sites

**DBMFHS** - Data Base of Minimally-Frustrated Helical Segments





# The Bologna functional annotation resource

<http://biocomp.unibo.it/predictors.html>

<https://bio.tools/t?collectionID='Bologna Biocomputing Group'>



# Bologna Biocomputing Group

University of Bologna



Home	Members	Predictors/Databases	Publications	Training	BWS	Activities	Visitors	WebMail
------	---------	----------------------	--------------	----------	-----	------------	----------	---------

## Group Leader

Rita Casadio

**Piero Fariselli**



## Senior Researchers

Emidio Capriotti  
Pietro Di Lena  
Piero Fariselli (PO, Univ. of Torino)  
Pier Luigi Martelli

## PhD Students

Davide Baldazzi  
Giovanni Madeo  
Matteo Manfredi  
Teresa Tavella

**Pier Luigi Martelli**

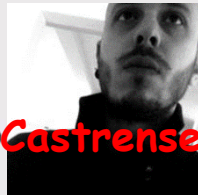


## Former Members

Francesco Aggazio  
Lisa Bartoli  
Samuele Bovo  
Remo Calabrese  
Raffaele Fronza  
Valentina Indio  
Paola Marani  
Ludovica Montanucci  
Andrea Pierleoni  
Damiano Piovesan  
Deepak Rajan  
Priyank Shukla  
Shaline Tiwari

## Young Researchers

Castrense Savojardo **Castrense Savojardo**



## Contract Researchers

Giulia Babbi



**Giulia Babbi**

## External Collaborators

Ivan Rossi (BioDec)

## News and Announcement

**September 16-20, 2019**  
Special Advanced Course on "RNA Analysis" . [Prof. Cedric Notredame](#) -Centro de Regulacio Genomica-Barcelona

**July 9, 2019**  
Workshop del Gruppo "Biologia Computazionale e dei Sistemi" della Societa Italiana di Biochimica a Biologia Molecolare. [Programme](#)

**July 8, 2019**  
Corso Breve su "Interazioni Proteina-Proteina" - Gruppi "Proteine" e "Biologia Computazionale e dei Sistemi" della Societa Italiana di Biochimica a Biologia Molecolare. [Programme](#)

**June 10-14, 2019**  
Special Advanced Course on

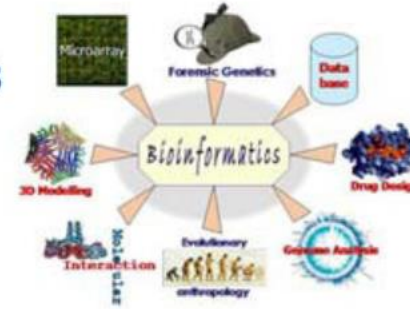




ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

# International Bologna Master in Bioinformatics

## Laurea Magistrale in Bioinformatics



Home
The Master
Programme & Faculty
Application
Fee Waiver/Grants
Useful Info
Erasmus Programme
Timetable/Exam session
Thesis
Final Colloquium
Final Examination
Bologna Winter

Welcome to the web page of the  
**International Bologna Master in Bioinformatics**  
**Laurea Magistrale in Bioinformatics**

**Academic Year 2019/2020**

**Academic Year 2020/2021**

[Bologna Biocomputing Group](#)

[Contact Us](#)

[LM Bioinformatics - UNIBO](#)

## Bologna Winter Schools

**Bologna Biocomputing Group**  
University of Bologna

Home Members Predictors/Databases Publications Training BWS Activities Visitors WebMail

**Bologna Winter School 2020**  
What can we learn from protein structure?

**Bologna Winter School 2019**  
Data Science for Bioinformatics

**Bologna Winter School 2018**  
Big Data and Bioinformatics

**Bologna Winter School 2017**  
Revisiting Bioinformatics Foundations

**Bologna Winter School 2016**  
In Silico Markers for Precision Medicine

**Bologna Winter School 2015**  
NGS data, Bioinformatics and New Molecular Scenarios

**Bologna Winter School 2014**  
Bioinformatics for Biological Complexity

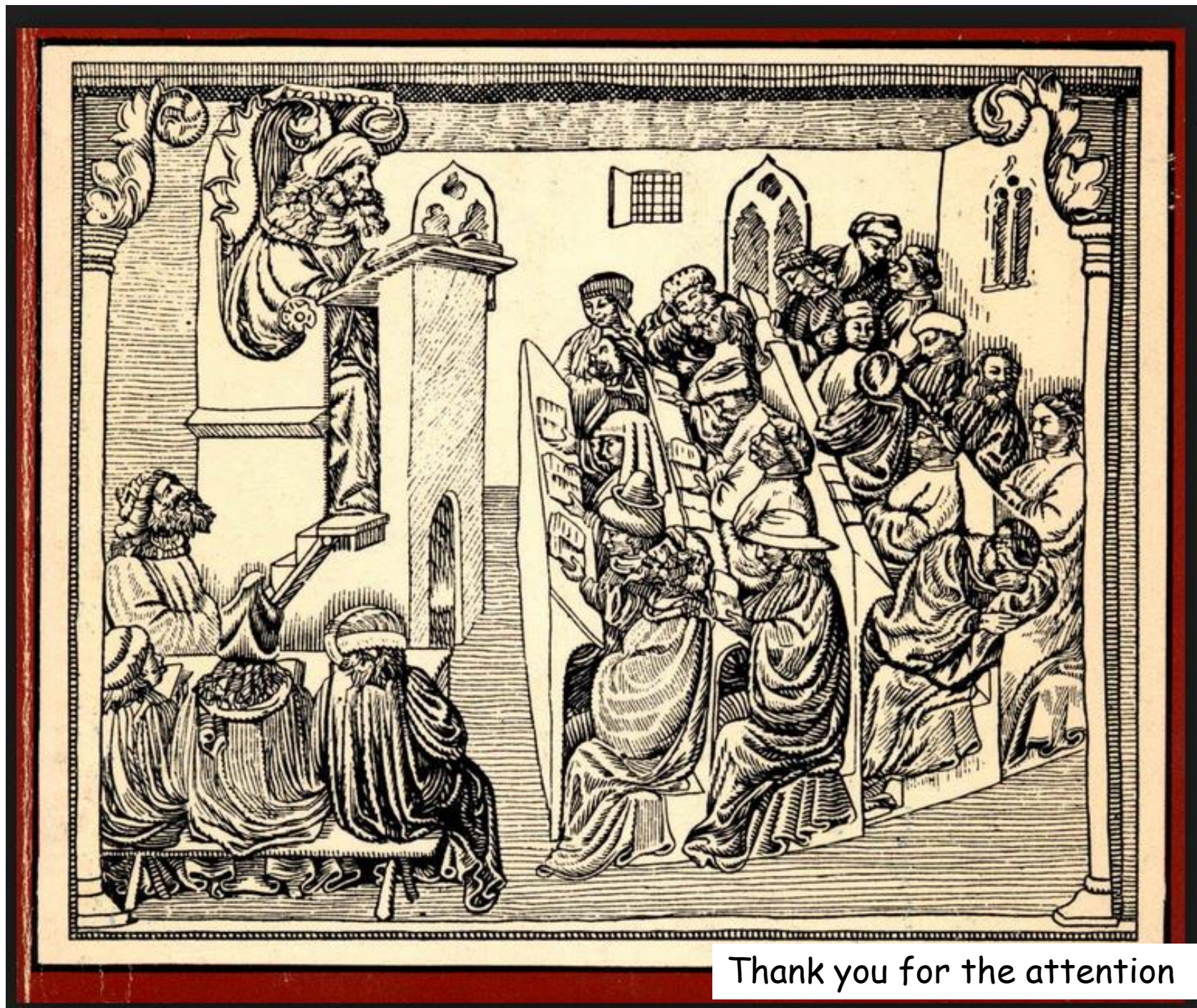
**News and Announcement**

**September 16-20, 2019**  
Special Advanced Course on "RNA Analysis" - Prof. Cedric Notredame - Centro de Regulacio Genomica-Barcelona

**July 9, 2019**  
Workshop del Gruppo "Biologia Computazionale e dei Sistemi" della Societa Italiana di Biochimica a Biologia Molecolare. [Programme](#)

**July 8, 2019**  
Corso Breve su "Interazioni Proteina-Proteina" - Gruppi "Proteine" e "Biologia Computazionale e dei Sistemi" della Societa Italiana di Biochimica a Biologia Molecolare. [Programme](#)

**June 10-14, 2019**



Thank you for the attention

*Bioinformatics*, 33(11), 2017, 1656–1663

doi: 10.1093/bioinformatics/btx044

Advance Access Publication Date: 25 January 2017

Original Paper

OXFORD

---

Structural bioinformatics

## **ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model**

**Castrense Savojardo<sup>1,2,†</sup>, Piero Fariselli<sup>3,†</sup>, Pier Luigi Martelli<sup>1,2,\*</sup> and Rita Casadio<sup>1,2</sup>**

<sup>1</sup>Biocomputing Group, Department of Biological, Geological and Environmental Sciences (BiGeA), University of Bologna, Bologna 40126, Italy, <sup>2</sup>CIG, Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna 40127, Italy and <sup>3</sup>Department of Comparative Biomedicine and Food Science (BCA), University of Padova, Padova 35020, Italy



*Annual Review of Biomedical Data Science*

# Protein–Protein Interaction Methods and Protein Phase Separation

Castrense Savojardo,<sup>1,\*</sup> Pier Luigi Martelli,<sup>1,\*</sup>  
and Rita Casadio<sup>1,2</sup>

**Ann.Rev.Biomed.Data Sci. 2020  
3:89-112**

<sup>1</sup>Biocomputing Group, Department of Pharmacy and Biotechnology and Interdepartmental Center “Luigi Galvani” for Integrated Studies of Bioinformatics, Biophysics, and Biocomplexity, University of Bologna, 40126 Bologna, Italy; email: rita.casadio@unibo.it

<sup>2</sup>Institute of Biomembranes, Bioenergetics, and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), 70126 Bari, Italy