

# Analysis and Prediction of Protein Complex

Laboratory of Bioinformatics I  
Module 2

April 16, 2019

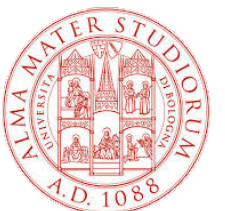
Emidio Capriotti

<http://biofold.org/>



**Biomolecules**  
**Folding and**  
**Disease**

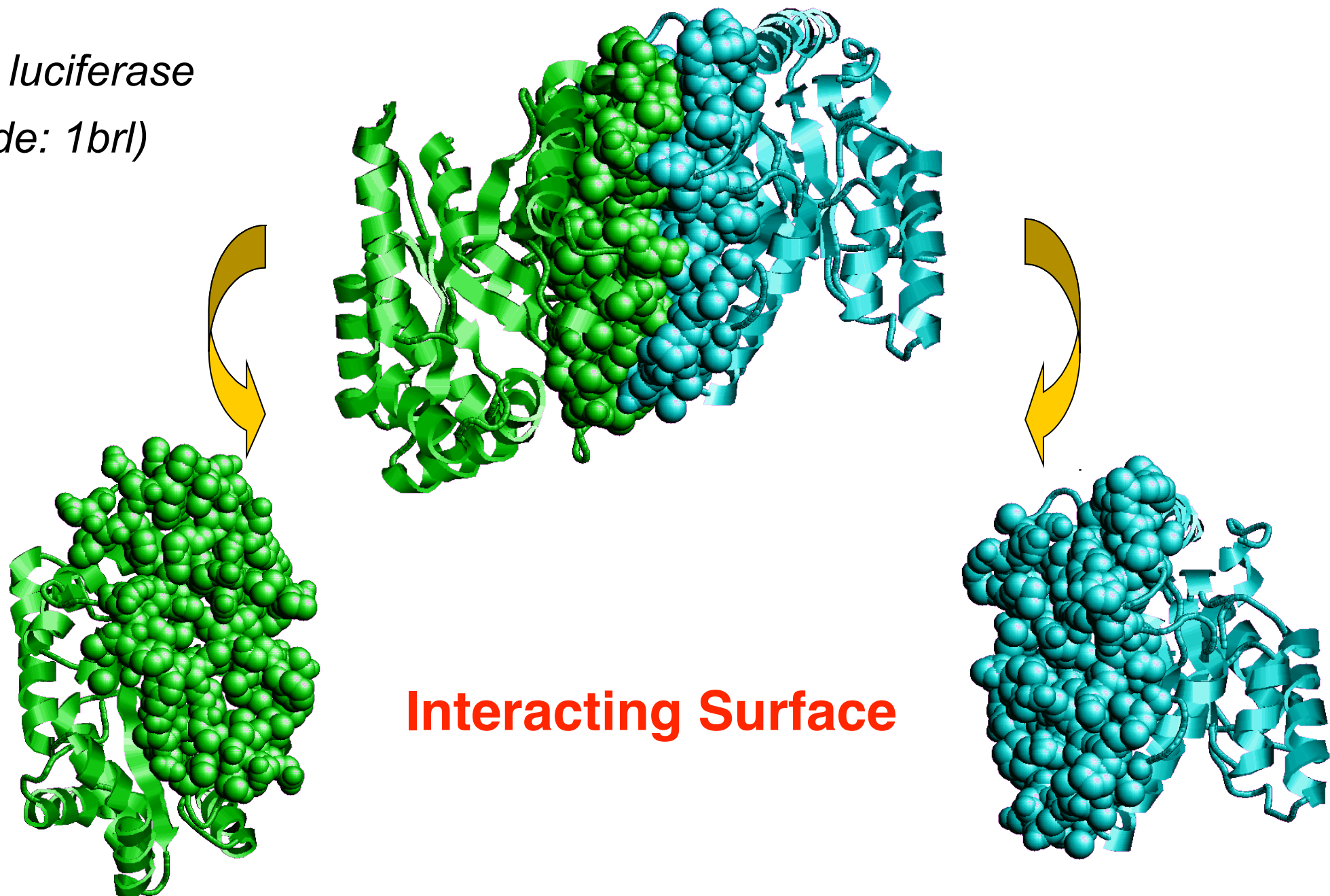
Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna



# Interacting surface

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

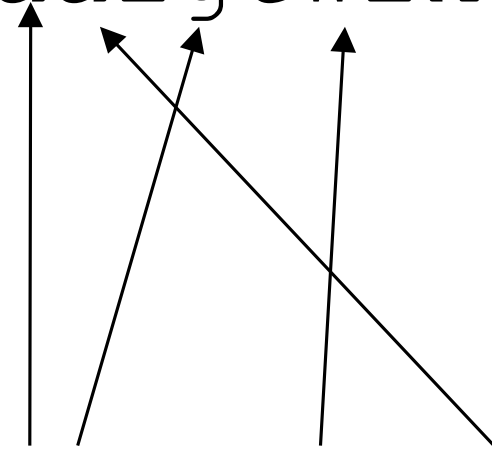
*Bacterial luciferase*  
(PDB code: 1brl)



# Prediction features

## Protein Sequence

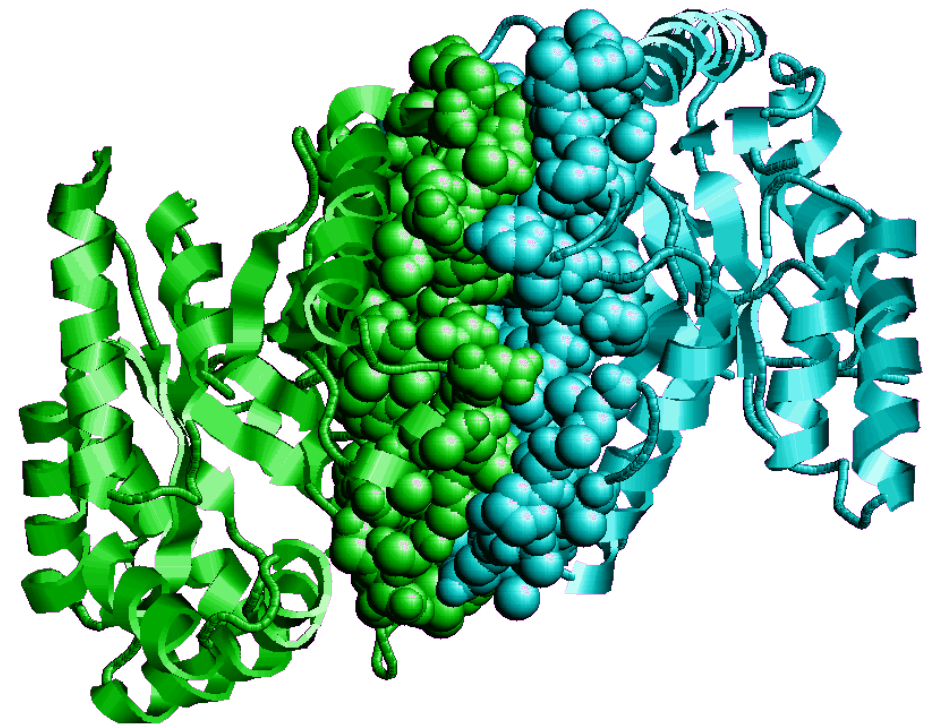
...aalgtwlkts.....  
...stwlgtaal kts.....



**+ Whole genome computation**

**- No exact location, No atomic description**

## Protein Structure



**+ Exact location Atomic description**

**- Availability of the 3D coordinates**

# Three major problems

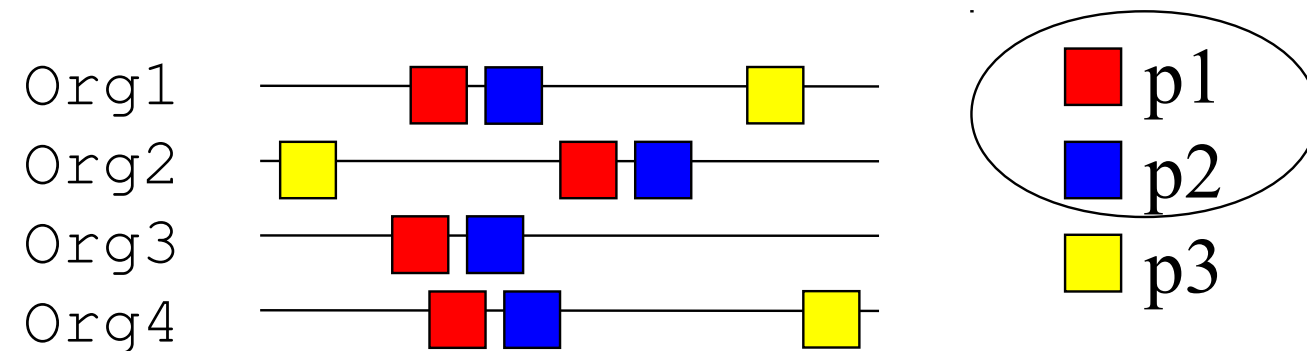
- **Protein-Protein interaction networks:** given a set of proteins, predict the possible partners
- **Docking:** given a pairs of proteins, known to interact, predict the geometry of the complex
- **Protein-interaction sites:** given a single protein, predict possible interacting regions

# Sequence-based methods

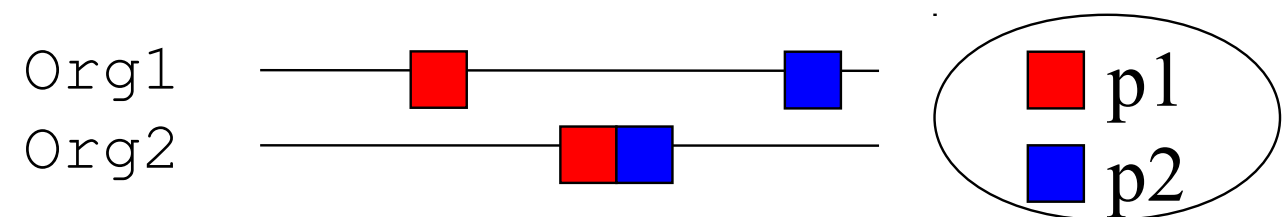
**Phylogenetic Profiling:** interacting proteins should co-evolve and should have orthologs in closely related species.

	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:** interacting proteins and co-evolving homologs tend to have close genomic locations.

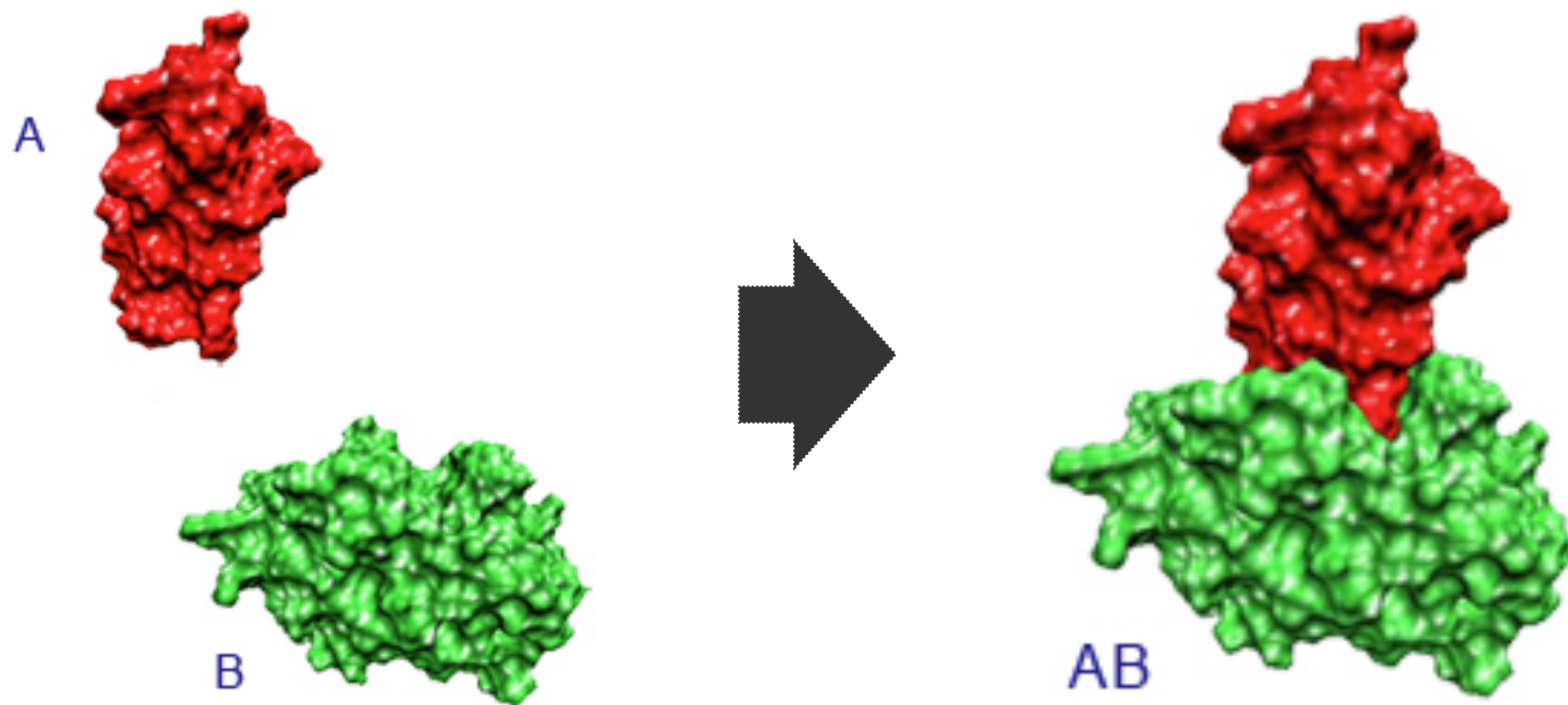


**Gene Fusion:** two proteins that interact tend to have homologs in other genomes that are fused into a unique protein



# Protein Docking

- Computational schemes that aims to **find the “best” matching between two molecules**, a **receptor** and a **ligand**
- The molecular docking problem can be defined as follows: **given the atomic coordinates of two molecules, predict their “correct” bound association**





# Protein-Protein docking

- Used to **model the quaternary structure of complexes** formed by two or more interacting proteins
- It is the “**gold standard**” for prediction of PPIs
- It used to **predict if two proteins interact** and also how the interaction takes place ("mode" of binding)
- It is **computationally very challenging** and thus very unlikely to be applied for high throughput purposes.

# What we can learn?

- Do proteins A (receptor) and B (ligand) bind *in vivo*?

## If they do bind:

- What is the spatial configuration they adopt in their bound state?
- What is the structure of the protein complex (**near-native structure**) in atomic details ?
- How strong or weak is their interaction (which types of interactions are present)?
- What is the orientation that maximises the interaction, minimizing the energy of the complex?

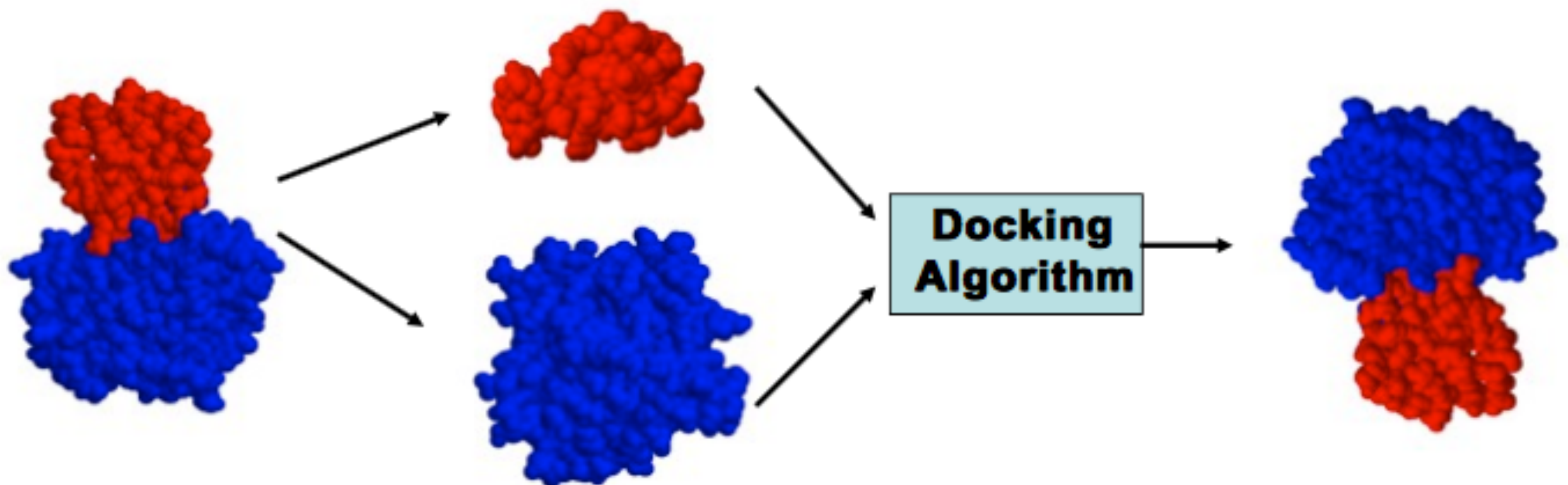
## If they don't bind:

- Would they bind if there was a mutation?



# Bound docking

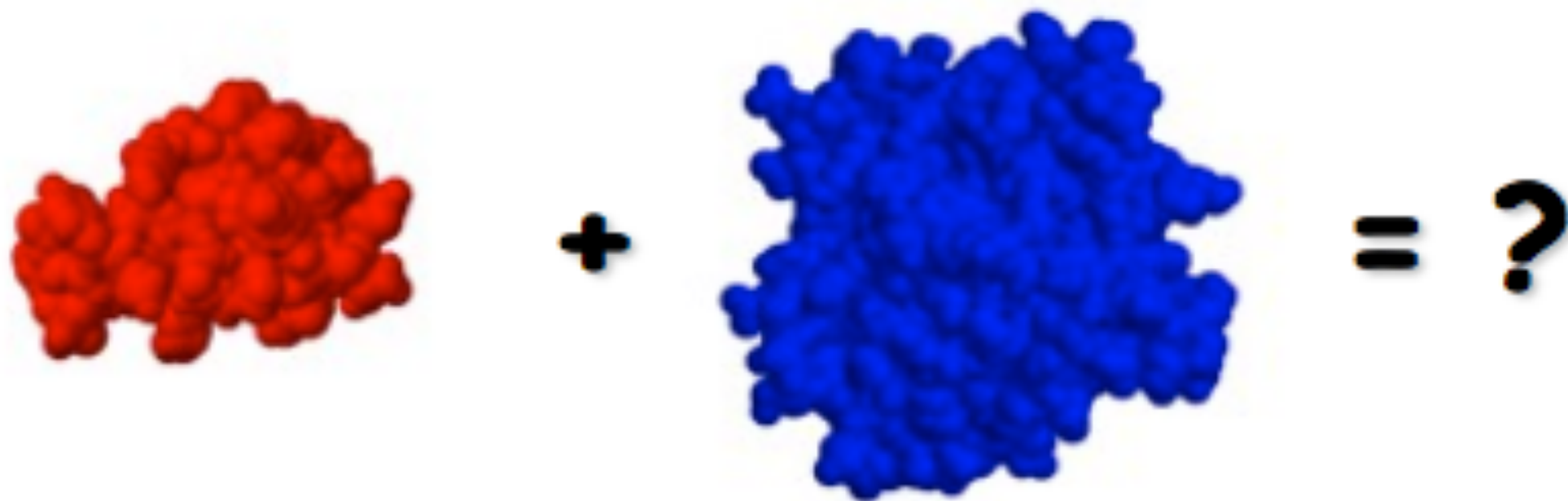
- Reconstruct a complex using the bound structures of the receptor and the ligand.
- After artificial separation of the receptor and the ligand, the goal is to reconstruct the native complex



- No conformational changes are involved
- **Used to validate the algorithm**

# Predictive docking

- Schemes that attempt to reconstruct a complex using the unbound structures of the receptor and the ligand
- An "unbound" structure maybe a **native** structure, a **pseudo-native** structure, or a **modelled** structure
- **Native**: free in solution, in its uncomplexed state
- **Pseudo-native**: structure complexed with a molecule different from the one used for the docking



# Why it is difficult?

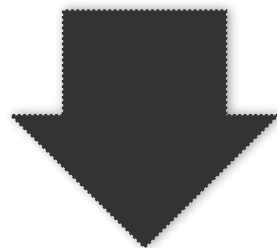
- # of possible conformations are astronomical
  - thousands of degrees of freedom (DOF)
- Free energy changes are small
  - Below the accuracy of our energy functions
- Molecules are flexible
  - alter each other's structure as they interact

# Main docking steps

Representation of the system



Conformational space search

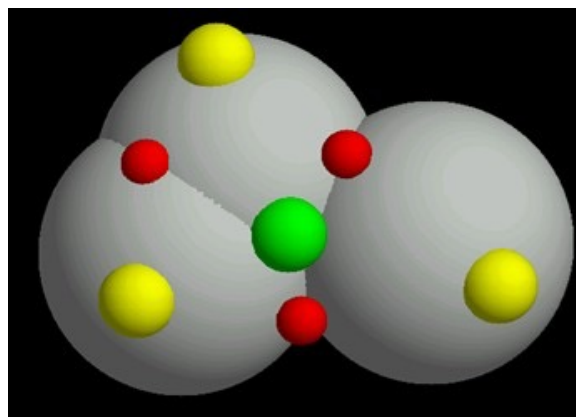
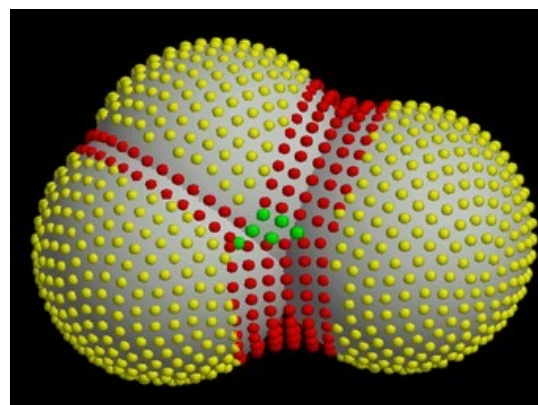


Ranking of potential solutions

# Systems representation

- Docking essentially simulates the interaction of the protein surface
- How do we define a protein surface?
  - Mathematical models (e.g. geometrical shape descriptors, a grid)
  - Static or dynamic treatment of the protein frame (rigid vs flexible)
- The choice of the system (surface) representation decides the types of conformational search algorithms, and the ways to rank potential solutions

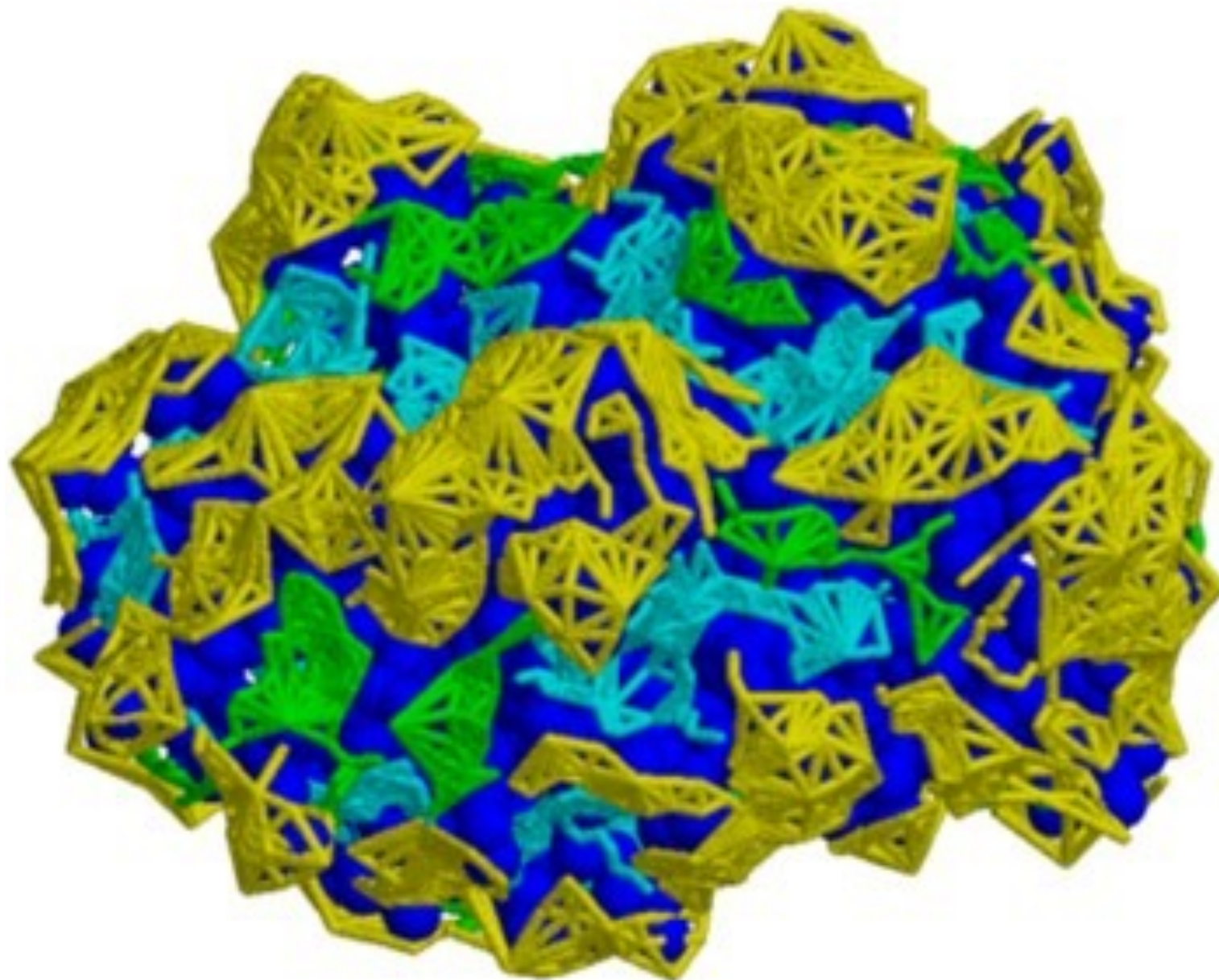
## Surface representation





# Patch detection

- Divide the surface into connected, non-intersecting, equal sized patches of critical points with similar curvature



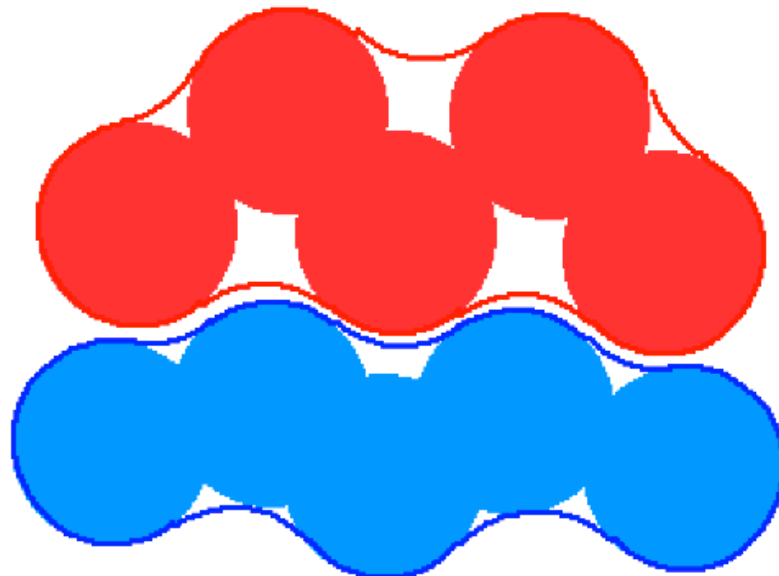
**Yellow:** knob patches  
**Cyan:** hole patches  
**Green:** flat patches  
**Blue:** protein



# Molecular recognition

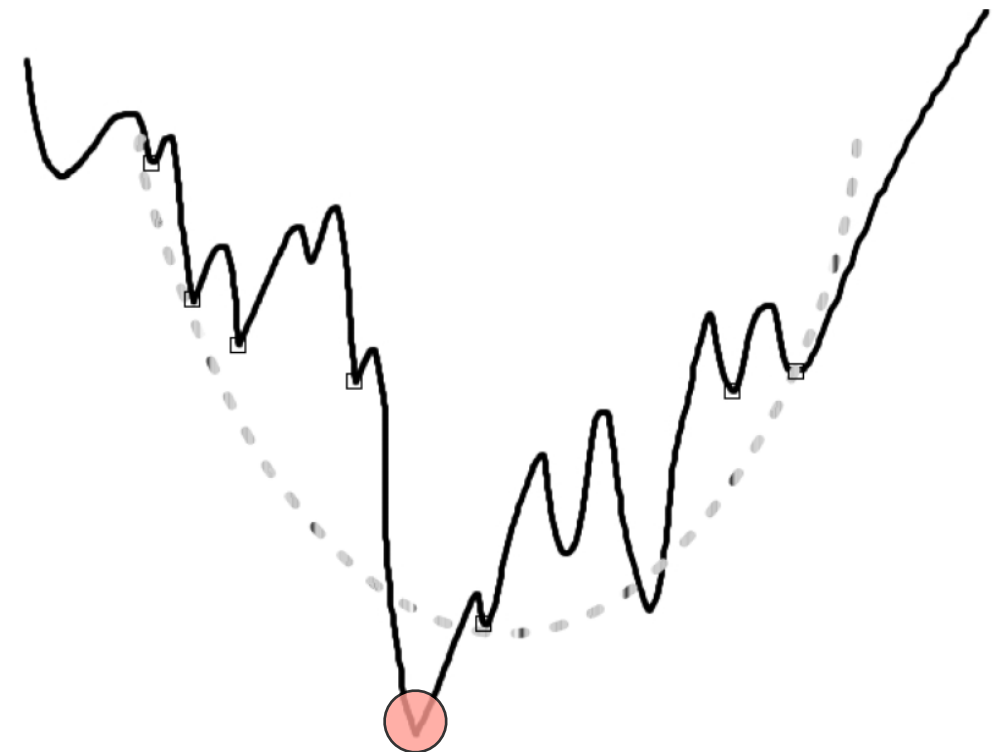
- Van der Waals
- Electrostatics
- Hydrophobic contacts
- Hydrogen bonds
- Salt bridges

All interactions act at short ranges → surface complementarity is needed for tight binding



# Conformational space

- Efficient search algorithm
- Speed and effectiveness in covering the relevant conformational space
- Computationally difficult - there are many ways to put two molecules together (3 translational + 3 rotational degrees of freedom)
- **Goal:** locate the most stable state (global minimum) in the energy landscape



# Docking types

- **Rigid body** is a highly simplistic model that regards the two proteins as two rigid solid bodies
  - fast → can explore the entire receptor and ligand surfaces
  - Less accurate
  - flexibility = "soft" belt into which atoms can penetrate
- The **semi-flexible** model is asymmetric; one of the molecules is considered flexible, while the receptor is regarded as rigid
- **Flexible** docking. Both molecules are considered flexible, though flexibility is limited or simplified
  - Slower
  - More accurate
  - Can model side-chain/backbone flexibility
  - highly reliable but too slow for extensive ligand docking

# Minimization protocols

- scan of the entire solution space in a **predefined systematic** manner  
e.g., complete searches of all orientations between two rigid molecules by systematically rotating and translating one molecule about the other
- a **gradual guided progression** through solution space. Only part of the solution space is searched, or fitting solutions are generated.  
e.g., Monte Carlo, simulated annealing, molecular dynamics (MD), and evolutionary algorithms.
- **Data-driven docking**  
it uses the available information about binding site/interface residues .

# Scoring the predictions

- A search algorithm may produce a large number of solutions ( $\sim 10^9$ )
- **Goal:** discriminate between "correct" native solutions, i.e., with **low RMSD from the crystal structure** and others within reasonable computation time
- **Good scoring function:** fast enough to allow its application to a large number of potential solutions
  - effectively discriminates between native and non-native docked conformations
  - should include and appropriately weight all the energetic ingredients.

# Scoring parameters

- Geometric complementarity - how to score complementarity is strongly coupled with the surface representation.
- Intermolecular overlap – tolerance to slight interface clashes and penalty for protein interior clashes (surface "belt" of nonpenalised penetration area)
- Intra-molecular overlap – when backbone flexibility is taken into account
- Hydrogen bonding
- Contact area: total interactions =  $hh + pp + hp$  (h = hydrophobic, p = polar)
- Pairwise aa and atom-atom contacts – empirical term derived from observed statistical frequency of aa contacts in X-ray proteins
- Electrostatic interactions and solvation energy

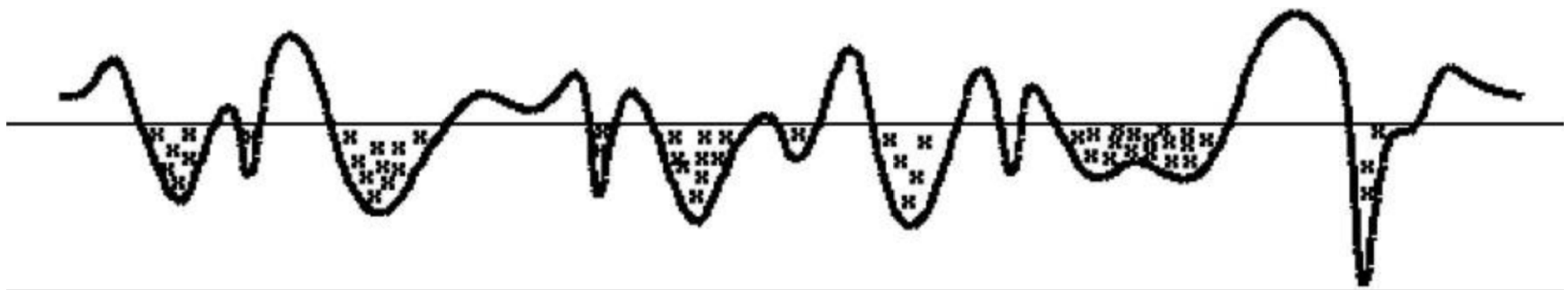


# Knowledge-based scores

- Knowledge of the **location of the binding site** on one or both proteins drastically reduces the number of possible solutions
- Knowledge of the **specific binding site residues** reduces the search space even further
- Info about active site residues: site directed mutagenesis, chemical cross-linking, phylogenetic data
- Sometimes the binding site can be predicted
- For some families the major binding sites are known in advance (e.g. serine proteases and immunoglobulins)

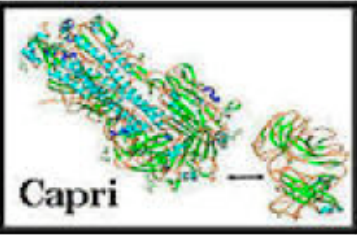
# Prediction clustering

- **Events that occur in clusters are probably not random**
- The cluster with the largest number of low-energy structures is typically the native fold, the center of the most populated cluster being a structure near the native binding site
- Looking for large clusters is a major tool of finding near-native conformations



# CAPRI Experiments

- CAPRI is a community-wide experiment in modelling the molecular structure of protein complexes
- CAPRI is a **blind prediction experiment** aimed at testing the performance of protein docking methods
- Rounds take place about every six months
- Each round contains between one and six target protein–protein complexes whose structures have been recently determined experimentally
- Targets are unpublished crystal or NMR structures of complexes, whose coordinates are held privately by the assessors, with the co-operation of the structural biologists who determined them
- The atomic coordinates of the two proteins are given to groups for prediction



Capri

PDB idcodes for past targets

Home > Databases > PDBe > Services > Capri-Home

## CAPRI: Critical Assessment of PRediction of Interactions

CAPRI communitywide experiment on the comparative evaluation of protein-protein docking for structure prediction

Hosted By EMBL/EBI-PDBe Group

# Conclusions (-)

- The *molecular docking* **problem** is far from being solved
- It is difficult to find very specific properties of protein-protein interfaces
- Results are generally **poor with weakly interacting proteins**
- Proteins are flexible and may undergo even **large conformational changes upon binding**
- Exhaustive space searches provide **too many conformations**
- Accurate **interaction energies** are too complicated to compute
- For most complexes the **highest ranked structures are still false positives** (high RMSD from the complex)
- No efficient method for **reliable discrimination between correct solutions and FPs** is currently available, in particular if the binding site is unknown
- Many FPs displaying **good surface** complementarity are **far from the native complex**

# Conclusions (+)

- If the conformational change is limited to surface side-chain atoms, **rigid body algorithms have been remarkably successful**, even in absence of knowledge of the binding site
- Side-chain flexibility can be handled via a "soft" tolerance belt"
- Docking in steps" is a promising strategy: Initial rigid-body, entire surface algorithm followed by a dynamic method overcoming energy barriers
- **Integration of experimental information** produces reliable docking results
- Relatively **easy for enzyme-inhibitor complexes**
- Sometimes **good results with antigen-antibody pairs**

# Some methods

- **HADDOCK** (software/web server).  
<http://haddock.chem.uu.nl>
- **CLUSPRO** (software/web server)  
<http://cluspro.bu.edu>
- **ICM-pro** (desktop-modeling environment)  
[http://www.molsoft.com/protein\\_protein\\_docking.html](http://www.molsoft.com/protein_protein_docking.html)
- **ROSETTADOCK** (software/web server)  
<http://graylab.jhu.edu/docking/rosetta/>
- <http://rosettadock.graylab.jhu.edu/submit>
- **GRAMM-X** (web server)  
<http://vakser.bioinformatics.ku.edu/resources/gramm/grammx>
- **PATCHDOCK/FIREDOCK** (software/web server)  
<http://bioinfo3d.cs.tau.ac.il/PatchDock/>
- **HEX** (software/web server)  
<http://hexserver.loria.fr>



# Exercise

Download the DSSP file of the **Bacterial luciferase** (*Vibrio harveyi*) from the PDB (code: 1BRL)

- Generate the **DSSP** file for the protein complex and the isolated chains A and B
- Calculate the total **solvent accessible area** of the complex and isolated chains and calculate the surface of interaction for both chains.
- Given the size of the binding surface **what kind of protein interaction** it is expected?
- Find the **residue at the interface** and calculate the **variation of relative solvent accessible area**. Which residue are buried in the interacting surface?

Chain = col 12, AA = col 14, SS = col 17, Acc: cols 36-38, Phi: cols 104-109, Psi: cols 110-115