EMBO Practical Course on
Computational analysis of protein-protein interactions:
Sequences, networks and diseases

5th May - 10th June 2018
Rome, Italy

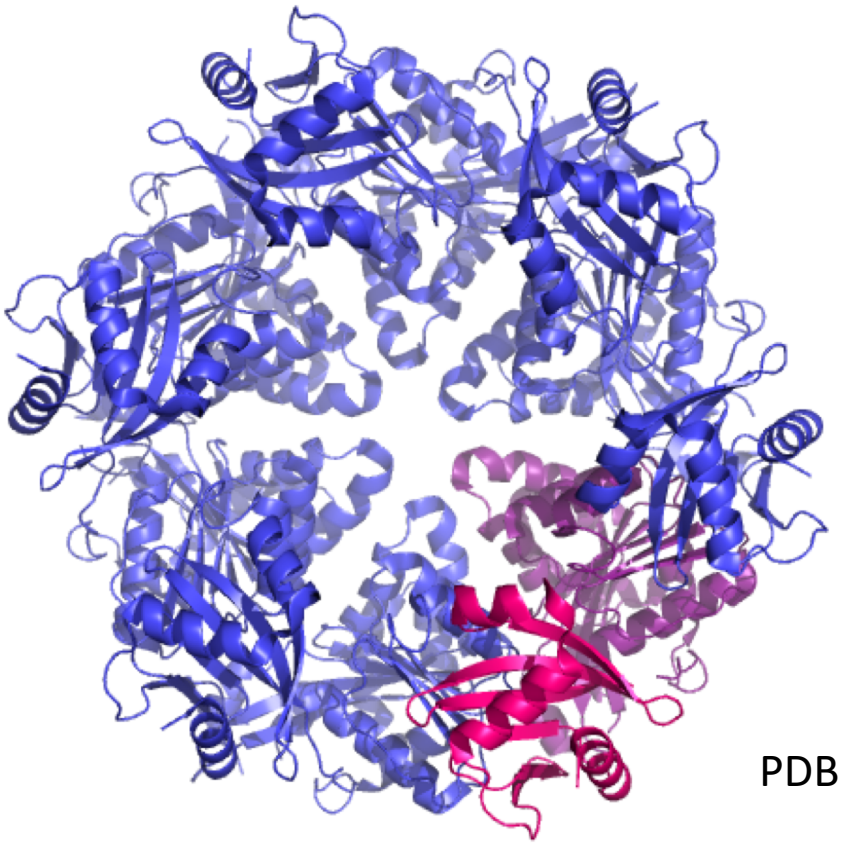
PPI 3D complex prediction and interfaces

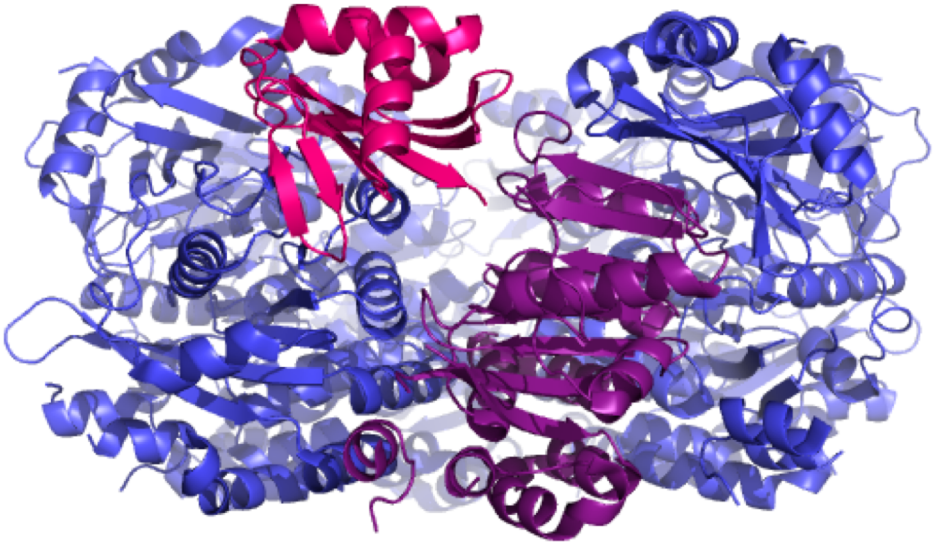
(What we talk about when we talk about docking)

Allegra Via
Institute of Molecular Biology and Pathology
National Research Council, Italy

Friday, 9th Novembre 2018

Hexameric PilT motor protein complex from *P. Aeruginosa*

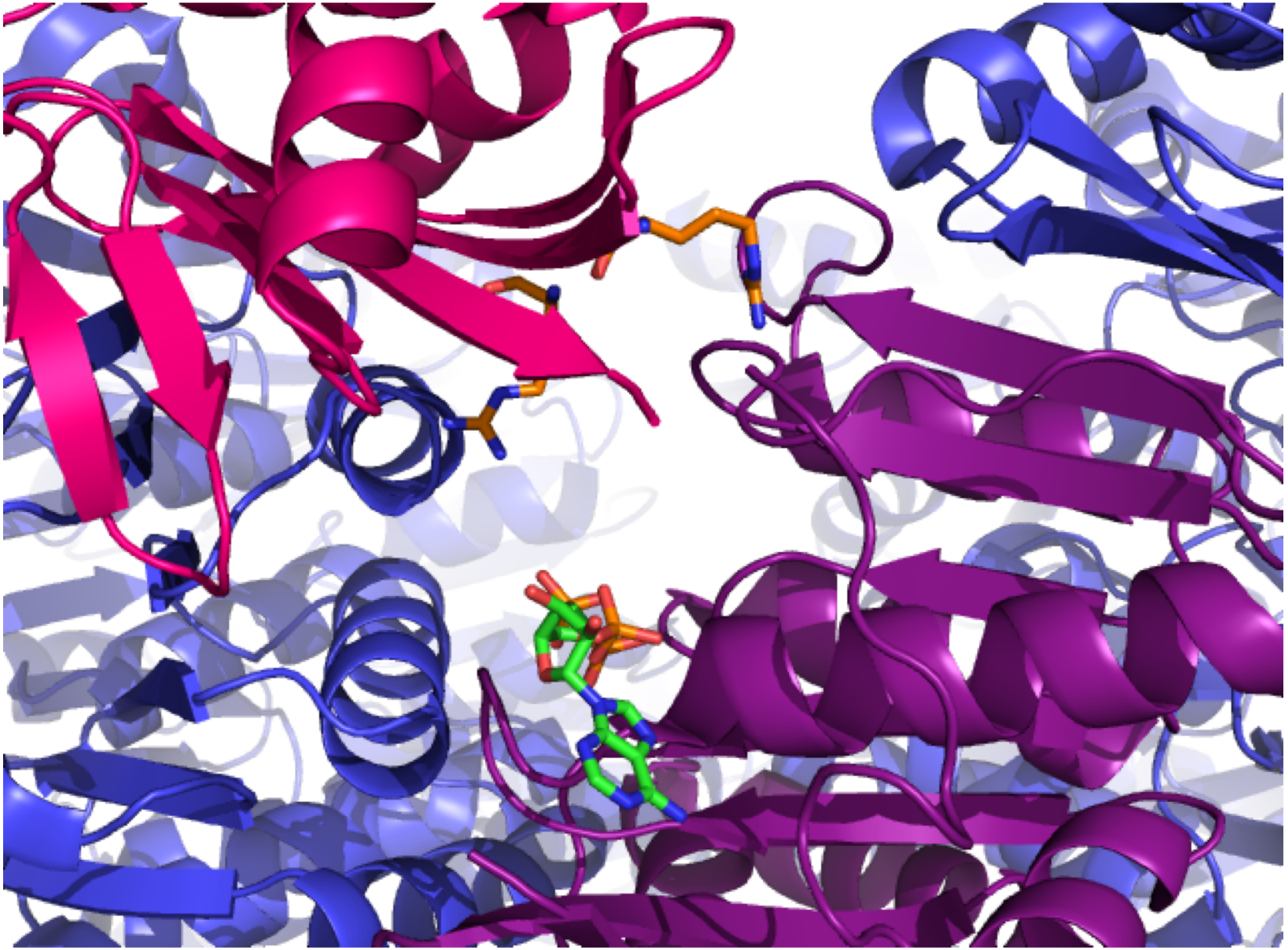




PDB: 3JVV

http://www.ebi.ac.uk/pdbe/about/news/powerful-little-motor

Why predicting protein complexes?



Interface of the PilT domains with arginine residues shown in orange sticks and ATP analog shown in green

Molecular docking

Molecular docking: an optimisation problem

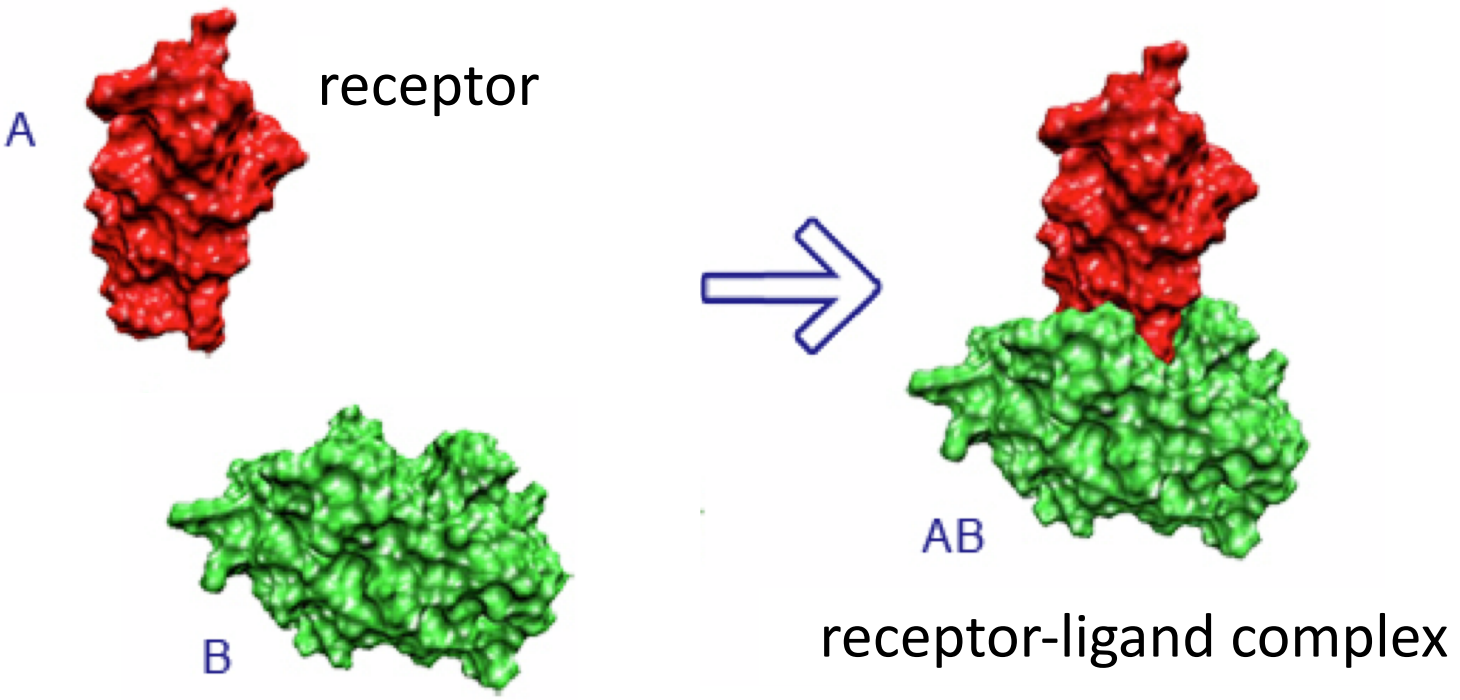
DOCKING = put a ligand in the pocket (= best torsions) + maximise its affinity

Step 1: Pose generation → Conformational space search

Step 2: Scoring → Ranking of potential solutions

The docking problem

Given the atomic coordinates of two molecules, predict their "correct" bound association



Which ligands?

Protein-ligand interactions

- elemental ions
- small molecules
- peptides
- macromolecules

Three classes of docking

Protein-small molecule

Lot of flexibility
Binding sites tend to be small and deep

Protein-peptide

8-10 amino acids
Floppy backbones

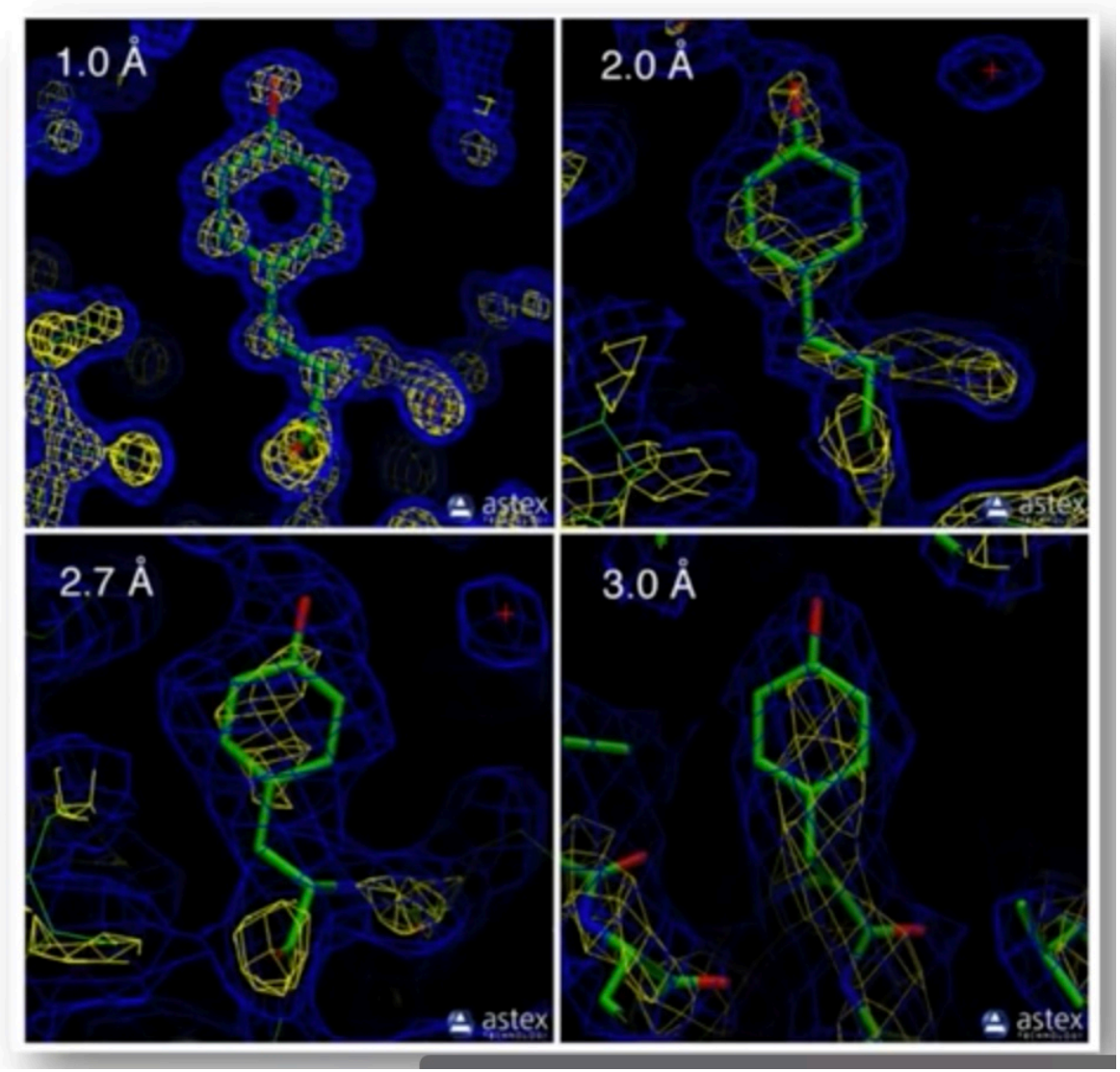
Protein-protein

Stable backbones
Binding interfaces tend to be large and flat

Where do we start?

- Ideally: a GOOD crystal structure

- Ideally: a GOOD crystal structure

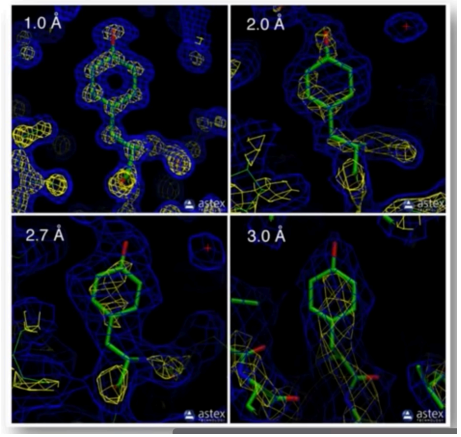


- Ideally: a GOOD crystal structure
- Usually worse: a homology model

- Ideally: a GOOD crystal structure
- Usually worse: a homology model

Please remember: an X-RAY/PDB structure is also just a «modelled» snapshot by a crystallographer

there is human intervention… not the holy experimental truth



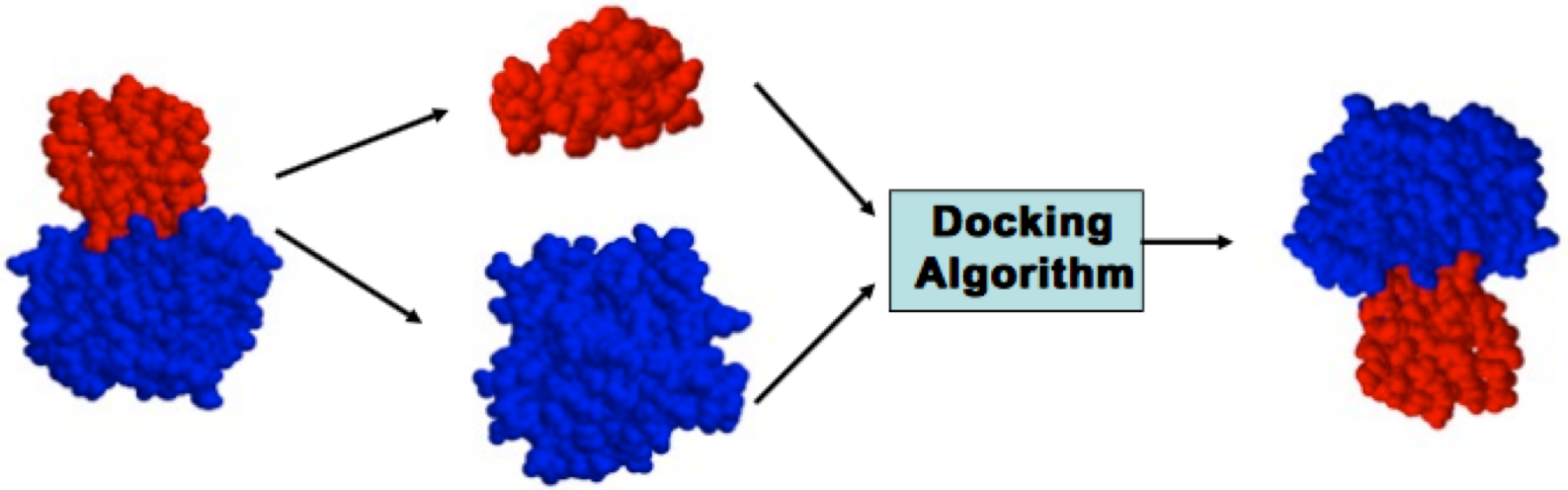
The PDB educational portal

https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction

- Ideally: a GOOD crystal structure
- Usually worse: a homology model
- A ligand, usually: the co-crystallised molecule

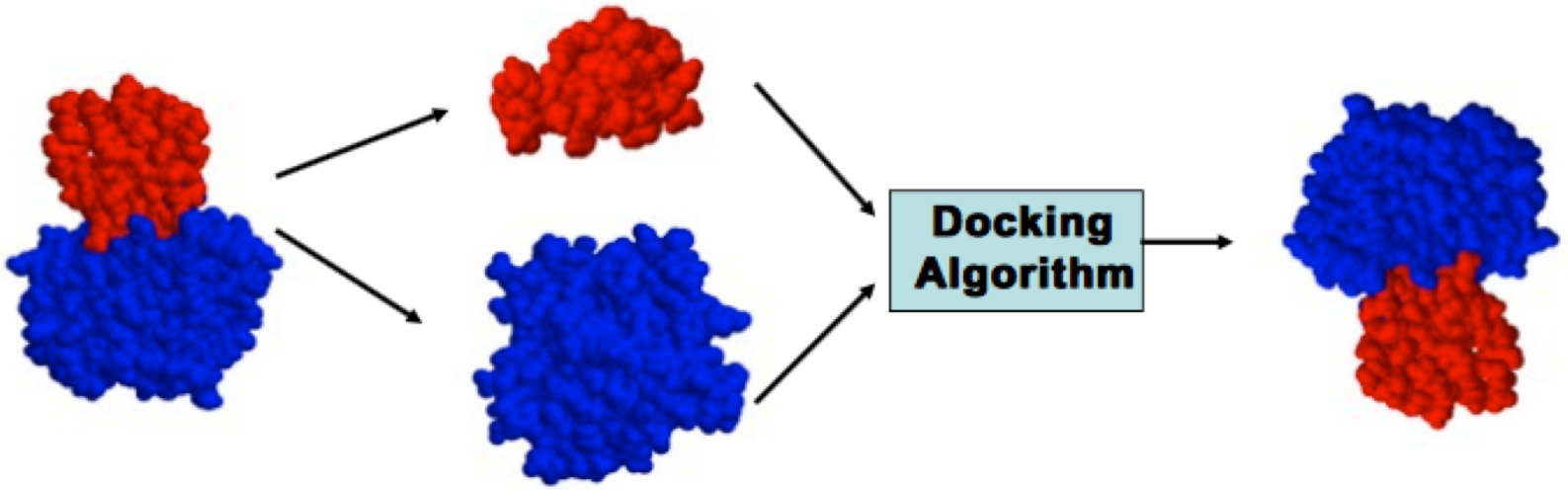
- Ideally: a GOOD crystal structure
- Usually worse: a homology model
- A ligand, usually: the co-crystallised molecule

Bound docking: artificial separation of receptor and ligand
Goal: reconstruct the native complex



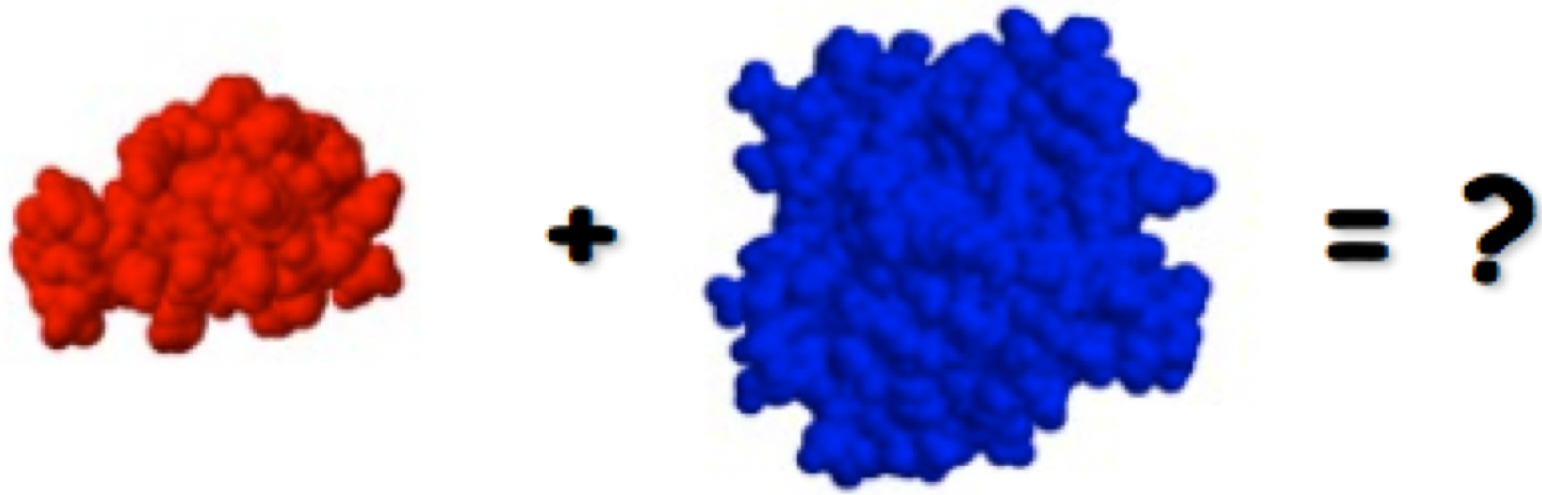
- Ideally: a GOOD crystal structure
- Usually worse: a homology model
- A ligand, usually: the co-crystallised molecule

Bound docking: artificial separation of receptor and ligand
Goal: reconstruct the native complex



- No conformational changes are involved
- Used as first test to validate the algorithm

«Unbound» or «predictive» modelling



reconstruct a complex using the unbound structures of the receptor and the ligand

Halperin et al. Proteins, 2002

An "unbound" structure may be:

- **native**
- **pseudo-native**
- **modelled**

Native: free in solution, in its uncomplexed state

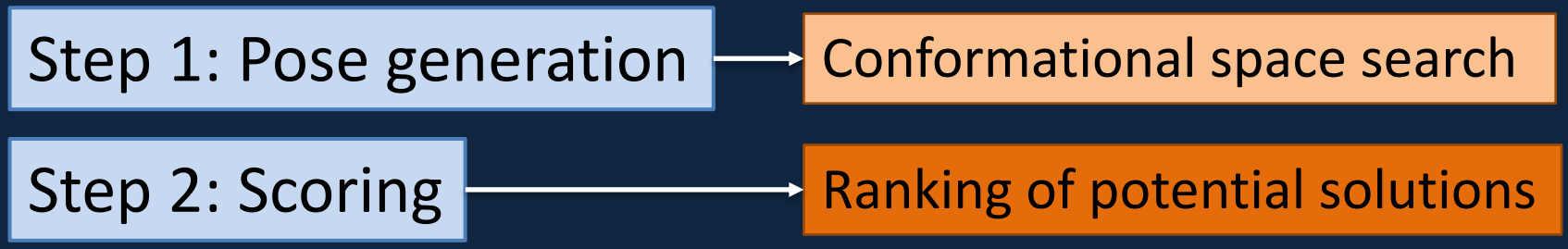
Pseudo-native: structure complexed with a molecule different from the one used for the docking

Halperin et al. Proteins, 2002

Unbound docking is far more complex than bound docking

What are the issues?

- conformational changes (side-chains and backbone movements)
- experimental errors in the structures
- reliability of models


Step 1: Pose generation
Conformational space search
Step 2: Scoring
Ranking of potential solutions

Step 1: Pose generation → Conformational space search

OR

Getting the ligand into the pocket

Translation & **R**otation of ligand needs to be performed

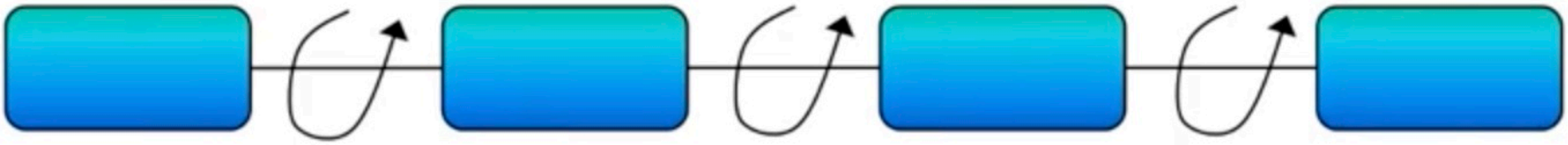
=> An optimisation problem in T and R space

Bond distances & angles in principle known BUT torsions will have to adapt to put the ligand into the pocket

How the confromation will look like?

=> An optimisation problem in T space

Conformational space search: torsions will kill us!



100 valid angles => 100^3 = 1.000.000 points to evaluate it

Let's assume: 1 sec per evaluation…
1.000.000/60 = 16.666 mins = 278 hrs = 11 days

Rigid vs flexible docking

- **Rigid body** - highly simplistic model - receptor & ligand two rigid solid bodies
- **Semi-flexible** - ligand flexible, receptor rigid
- **Flexible docking** – both molecules flexible, BUT flexibility is limited or simplified

Rigid vs flexible docking

- **Rigid body** - highly simplistic model - receptor & ligand two rigid solid bodies
- S**emi-flexible** - ligand flexible, receptor rigid
- **Flexible docking** – both molecules flexible, BUT flexibility is limited or simplified

Rigid vs flexible docking

- **Rigid body** - highly simplistic model - receptor & ligand two rigid solid bodies
- S**emi-flexible** - ligand flexible, receptor rigid
- **Flexible docking** – both molecules flexible, BUT flexibility is limited or simplified

Docking algorithms

- **Rigid docking**
  - fast → can explore the entire receptor and ligand surfaces
  - Less accurate
  - flexibility = "soft" belt into which atoms can penetrate

- **Flexible docking**
  - Slower
  - More accurate
  - Can model side-chain/backbone flexibility
  - highly reliable but too slow for extensive ligand docking

Conformational space search

- **Efficient search algorithm** – fast and effective in covering the relevant conformational space
- **Computationally difficult** - there are many ways to put two molecules together
- **Goal**: locate the most stable state (global minimum) in the energy landscape

Search algorithms

1) **scan of the entire solution space in a predefined systematic manner**

2) **gradual guided progression through solution space** Only part of the solution space is searched, or fitting solutions* are generated.

3) data-driven docking – uses the available information about binding site/interface residues

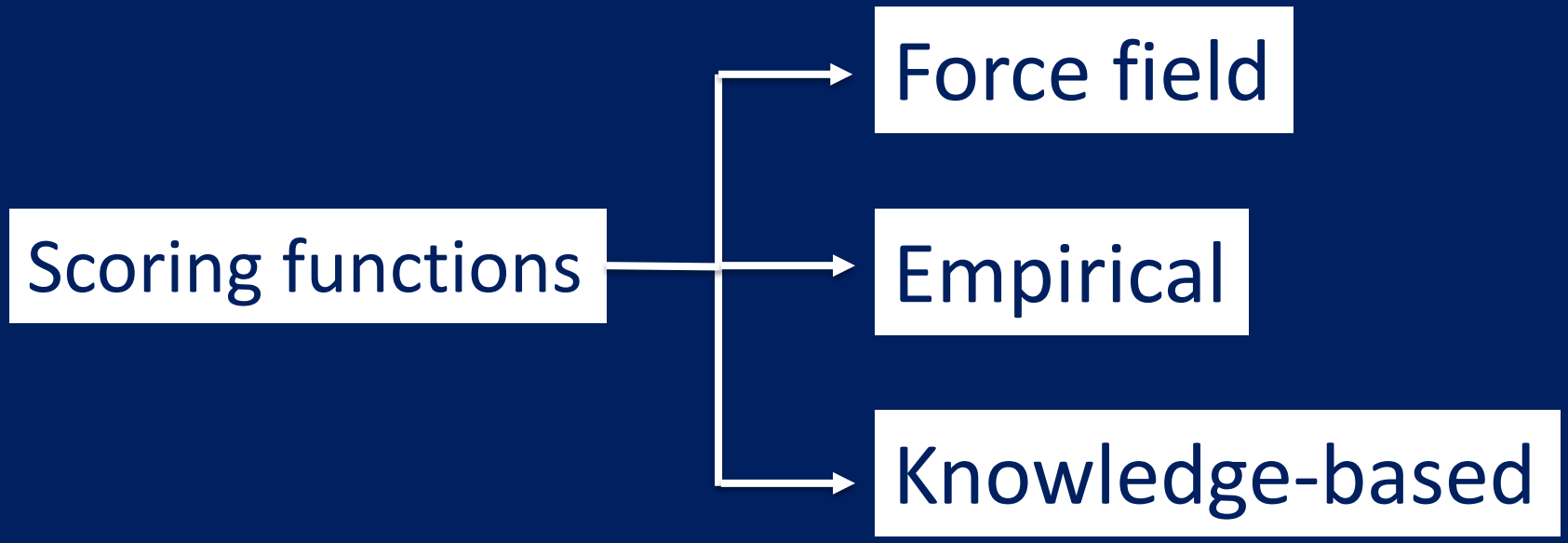
* Solutions meeting pre-defined criteria

A search algorithm may produce an immense number of solutions (10^9)

Ranking of potential solutions

Scoring functions

Prediction of the interaction energy / binding affinity of each conformer of the ligand docked


Scoring functions
Force field
Empirical
Knowledge-based

Force field-based scoring functions

Force field is a set of parameters that define the **potential energy** of a system

They are based on molecular mechanics and estimate the interaction energy between receptor and lignad

$$E_{coul}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

$$E_{vdW}(r) = \sum_{i=1}^{N} \sum_{j=1}^{N} 4\varepsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

Empirical-based scoring functions

- They have simple energy terms and are faster
- *Hydrogen bonds, ionic interactions, hydrophobic interactions, aromatic interactions, number of rotable bonds*
- The binding affinity is usually measured as empirical (weighted) sum of all such interactions

$$\Delta G = \sum_i Wi.\Delta Gi$$

Ideally, the best search algorithms and scoring schemes should be combined

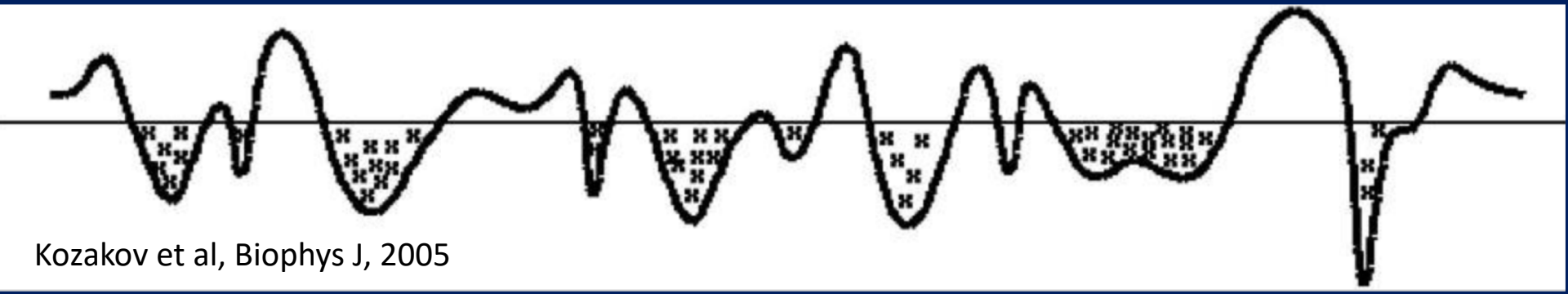
Two-stage ranking:

1. fast scoring to rapidly scan possible solutions and obtain initial "good" candidates – **mostly geometric criteria**
2. Followed by more advanced methods to further discriminate the limited conformations **- energy criteria**

Binding site information (known or predicted) may be included in scoring

Clustering of solutions

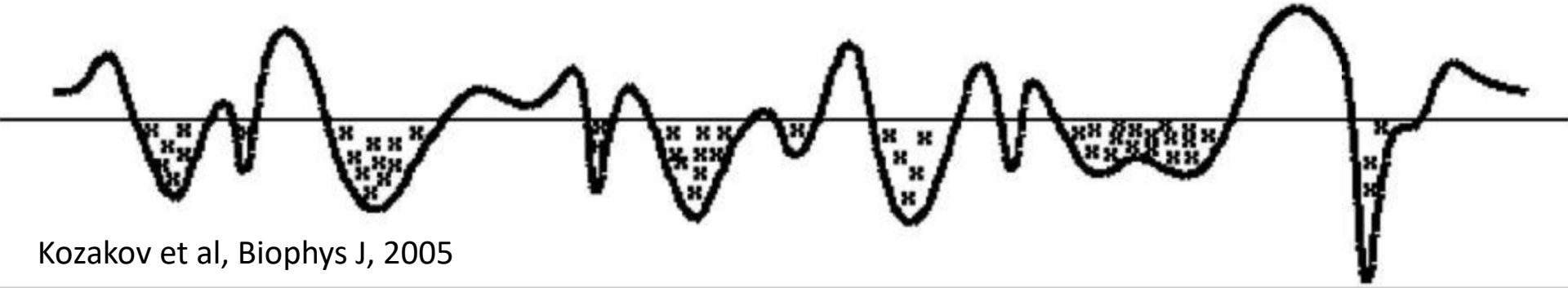
Events that occur in clusters are probably not random



Kozakov et al, Biophys J, 2005

Clustering of solutions

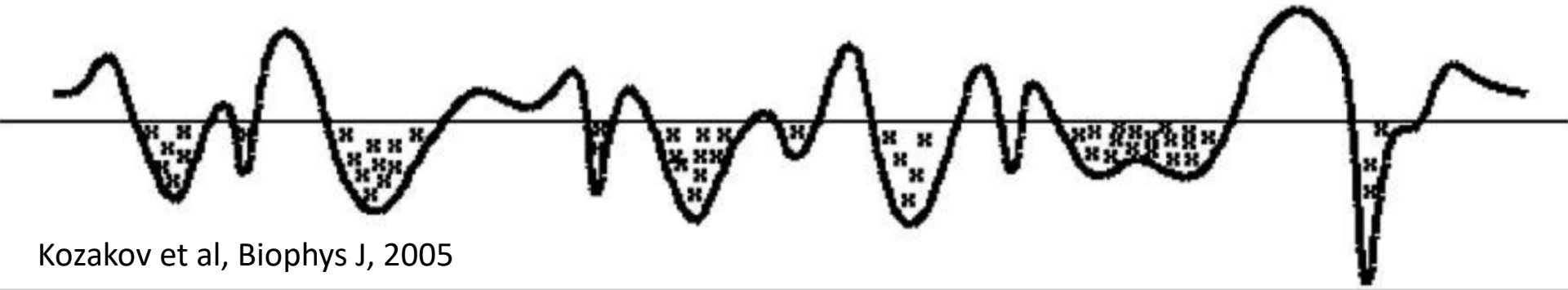
- The cluster with the largest number of low-energy structures is typically the native fold, the center of the most populated cluster being a structure near the native binding site



Kozakov et al, Biophys J, 2005

Clustering of solutions

Looking for large clusters is a major tool of finding near-native conformations



Kozakov et al, Biophys J, 2005

Challenges & limitations 1

Accurate prediction of ligand-receptor binding energy is still a challenge

Challenges & limitations 2

The process of choosing an appropriate scoring function and algorithm for a specific receptor and ligand is tricky

Challenges & limitations 3

- Taking protein flexibility into account is not easy
- Proteins are flexible and may undergo even large conformational changes upon binding

Challenges & limitations 4

Desolvation penalty and conformational energy costs are not addressed clearly in most of the scoring functions available

Challenges & limitations 5

Results are generally poor with weakly interacting proteins

Challenges & limitations 6

Accurate interaction energies are too complicated to compute

Conclusions

- The *molecular docking problem* is far from being solved, BUT…

Conclusions

- The *molecular docking problem* is far from being solved, BUT…
- If the conformational change is limited to surface side-chain atoms, rigid body algorithms have been remarkably successful

Conclusions

- The *molecular docking problem* is far from being solved, BUT…
- If the conformational change is limited to surface side-chain atoms, rigid body algorithms have been remarkably successful
- Integration of experimental information produces reliable results

Conclusions

- The *molecular docking problem* is far from being solved, BUT…
- If the conformational change is limited to surface side-chain atoms, rigid body algorithms have been remarkably successful
- Integration of experimental information produces reliable results
- Relatively easy for enzyme-inhibitor complexes

Conclusions

- The *molecular docking problem* is far from being solved, BUT…
- If the conformational change is limited to surface side-chain atoms, rigid body algorithms have been remarkably successful
- Integration of experimental information produces reliable results
- Relatively easy for enzyme-inhibitor complexes
- Sometimes good results with antigen-antibody pairs

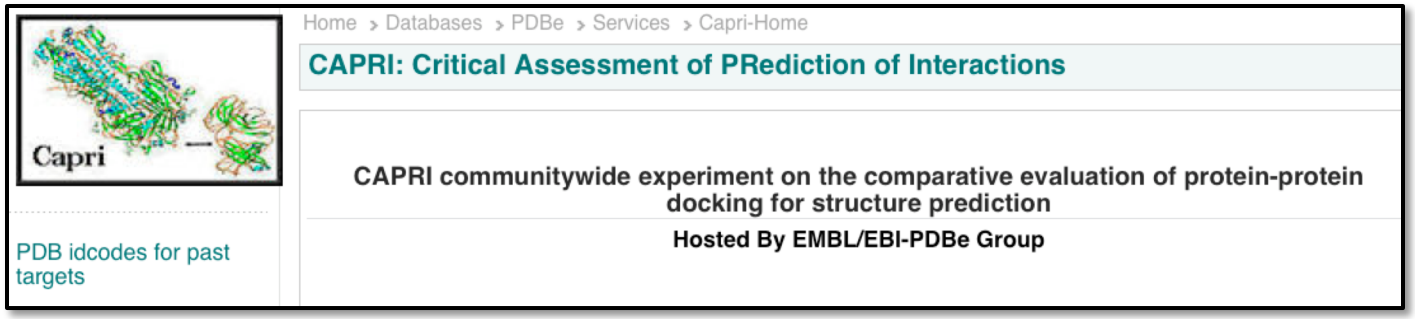
How can you choose a docking program?

- **Protein-Ligand docking** : AutoDock, DOCK, Gold, Glide, FlexX, Fred, MOE, Surflex
- **Protein-Protein docking** : ClusPro, ZDOCK, GRAMM-X, RosettaDock, DOT
- **Protein-Nucleic acid docking** : ParaDock, HADDOCK, YASARA DOCK, DOT

How can you choose a docking program?

- Who is the receptor?
- Who is the ligand?
- Is it free?
- Is it difficult to use?
- How does it perform?
- Search algorithm?
- Scoring function?

CAPRI: **C**ritical **A**ssessment of **PR**otein **I**nteractions



http://www.ebi.ac.uk/msd-srv/capri/

- 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: **C**ritical **A**ssessment of **PR**otein **I**nteractions

- The CAPRI experiment is double-blind
  - submitters do not know the solved structure
  - the assessors do not know the correspondence between a submission and the identity of its creator
- International meetings
- *Proteins: Structure, Function, and Bioinformatics*

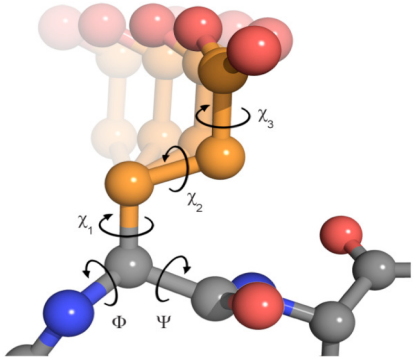
Special CAPRI Issue: Fifth Meeting on the Critical Assessment of Predicted Interactions

December 2013

CAPRI drives the community to develop new techniques

Side-chain flexibility

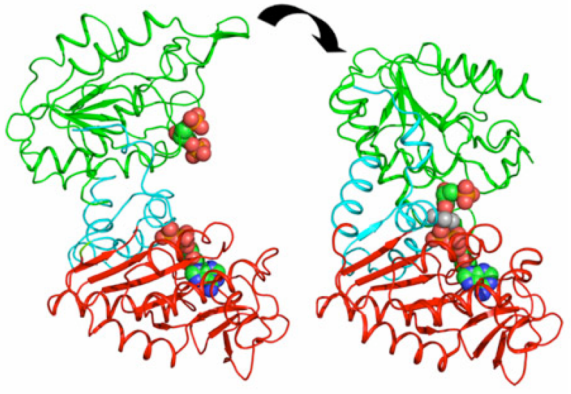
First rounds the big challenge was to model correctly side-chain conformations



Dihedral angles in glutamate

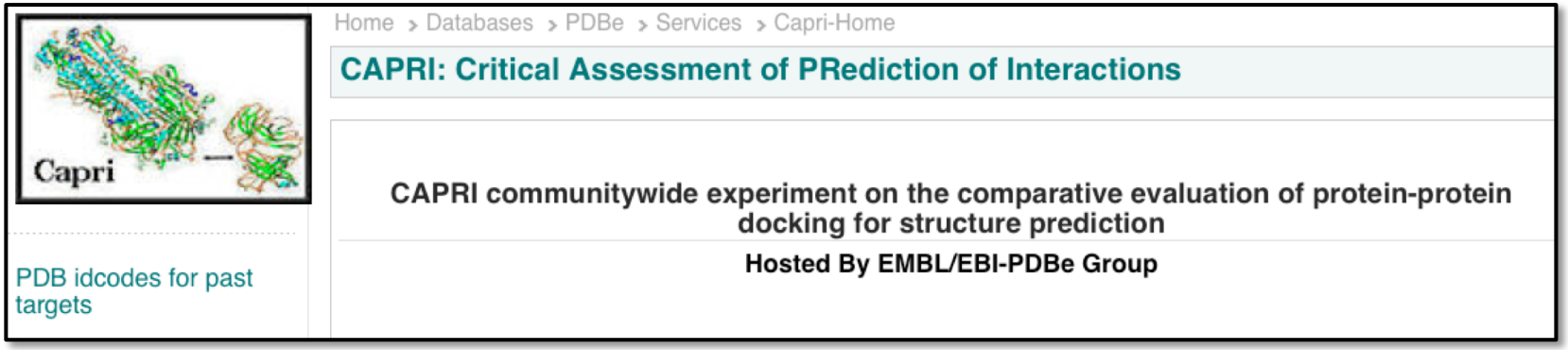
Domain movements

Today the development is focused on backbone flexibilty



Domain movements in PGK catalysis.
J. Biol. Chem. (2011) 286, 14040-14048

The clustering application to protein-protein docking was first introduced in CAPRI by Camacho & Gatchell,D. (2003)



CAPRI experiment … who is the winner ??

There is no official ranking and in many cases the differences between the different algorithms are not huge

CAPRI experiment … who is the winner ??

- **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

- **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