Information-driven modeling of biomolecular complexes

Prof. Alexandre M.J.J. Bonvin
Bijvoet Center for Biomolecular Research
Faculty of Science, Utrecht University
the Netherlands
a.m.j.j.bonvin@uu.nl

Universiteit Utrecht

[Faculty of Science Chemistry]

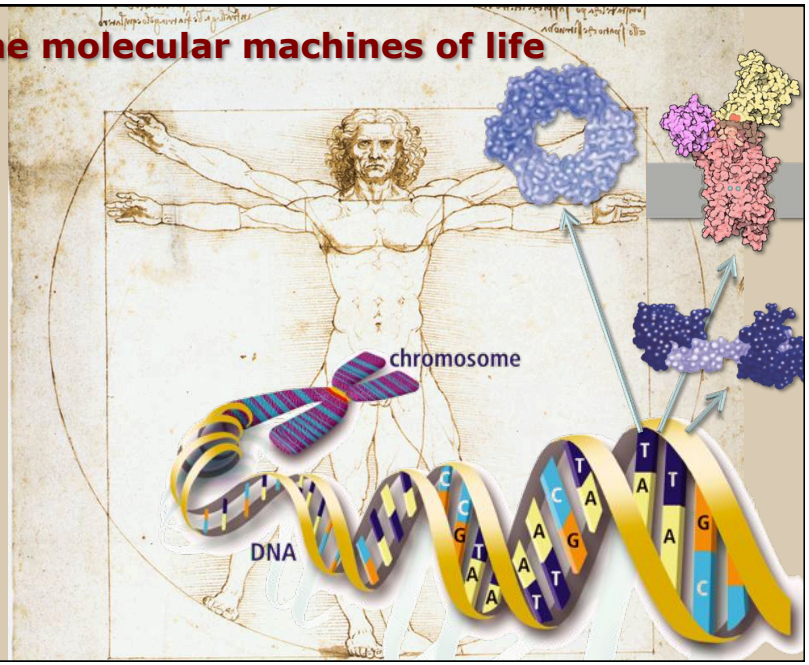
Overview

- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Multiple choices...
- Challenges
- Conclusions & perspectives

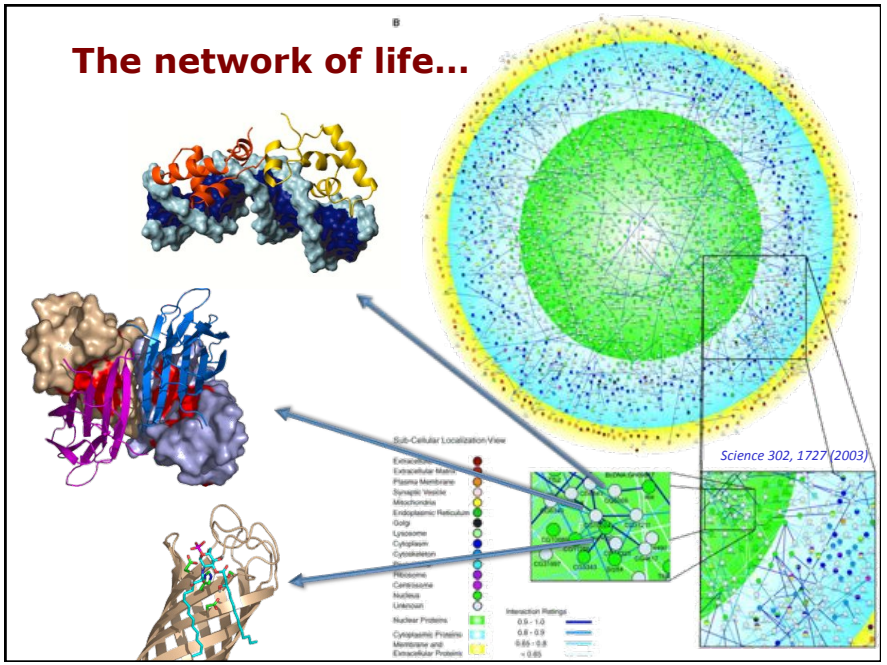
Universiteit Utrecht

[Faculty of Science Chemistry]

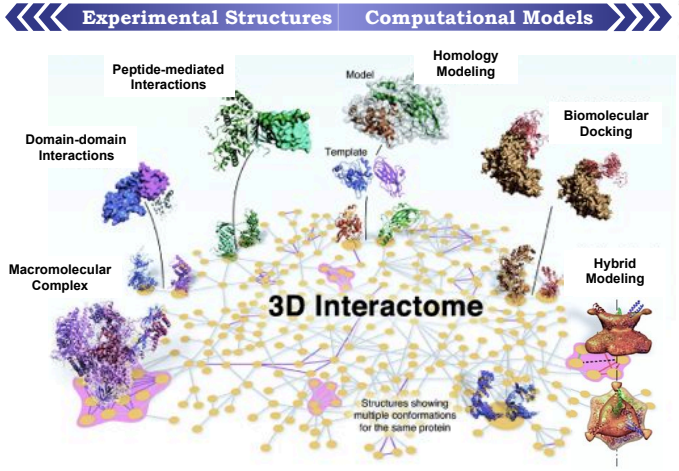
The molecular machines of life



The network of life...



Adding the 3rd dimension

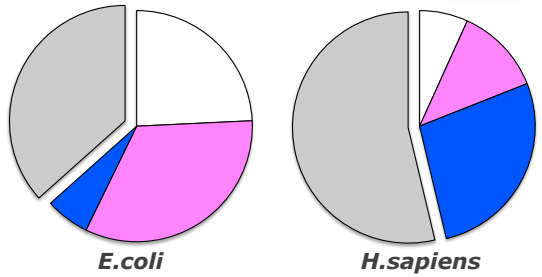


Stein *et al. Curr Op Struct Biol. 2011*

Structural coverage of proteomes

Unique proteins in proteomes

- ~4,300 proteins in *E.coli*
- ~20,300 proteins in *H.sapiens*



- with complete structures
- with complete models
- with partial models
- without structural data

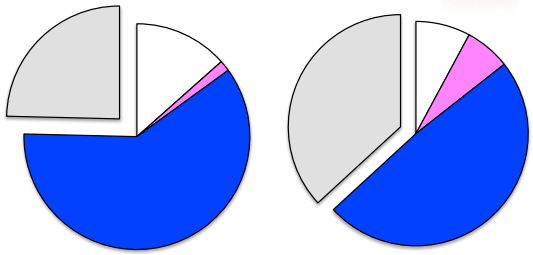
Statistics from Interactome3D (2013-01)

Mosca *et al.* Nature Methods 2013

Structural coverage of interactomes

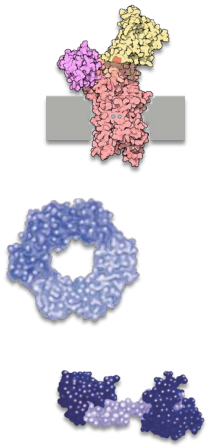
Unique interactions in interactomes

- ~7,500 binary interactions in *E.coli*
- ~44,900 binary interactions in *H.sapiens*



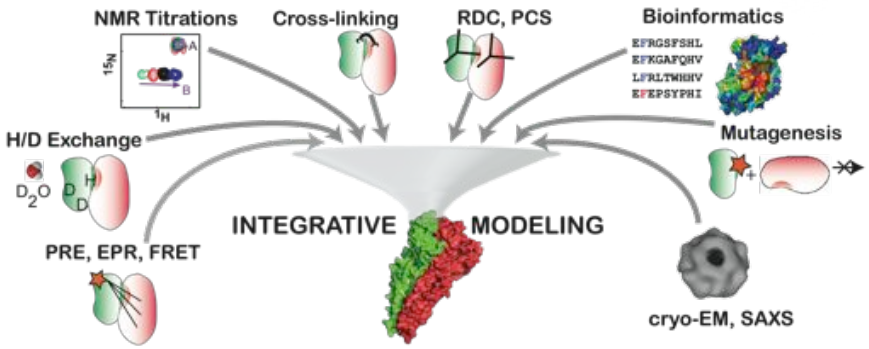
- with complete structures
- with partial (domain-domain) or complete models
- with structures for the interactors (suitable for docking)
- without structural data

What can we learn from 3D structures (models) of complexes?



- Models provide structural insight into **function and mechanism** of action
- Models can drive and **guide experimental studies**
- Models can help **understand and rationalize** the effect of disease-related mutations
- Models provide a starting point for **drug design**

What is Integrative Modeling?



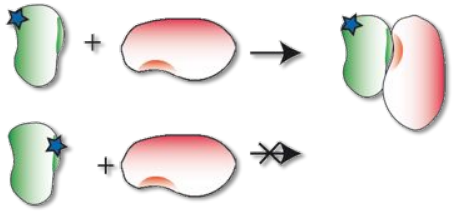
Related reviews

- Halperin *et al.* (2002) **Principles of docking: an overview of search algorithms and a guide to scoring functions.** *PROTEINS: Struc. Funct. & Genetics* **47,** 409-443.
- **Special issues of *PROTEINS:*** (2003) (2005) (2007) (2010) and (2103), which are **dedicated to CAPRI.**
- van Dijk ADJ, Boelens R and Bonvin AMJJ (2005). **Data-driven docking for the study of biomolecular complexes**. *FEBS Journal* **272** 293-312.
- de Vries SJ and Bonvin AMJJ (2008). **How proteins get in touch: Interface prediction in the study of biomolecular complexes**. *Curr. Pept. and Prot. Research* **9**, 394-406.
- Melquiond ASJ, Karaca E, Kastritis PL and Bonvin AMJJ (2012). **Next challenges in protein-protein docking: From proteome to interactome and beyond**. *WIREs Computational Molecular Science* **2**, 642-651 (2012).
- Karaca E and Bonvin AMJJ (2013). **Advances in integrated modelling of biomolecular complexes**. *Methods*, **59**, 372-381 (2013).

Overview

- Introduction
- **Information sources**
- General aspects of docking
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Multiple choices
- Challenges
- Conclusions & perspectives

Experimental sources: mutagenesis



Advantages/disadvantages

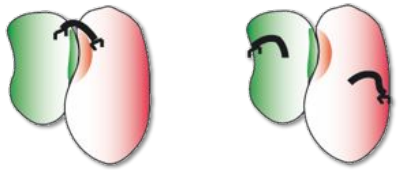
+ Residue level information

- Loss of native structure should be checked

Detection

- **Binding assays**
- **Surface plasmon resonance**
- **Mass spectrometry**
- **Yeast two hybrid**
- **Phage display libraries, ...**

Experimental sources:
cross-linking and other chemical modifications



Advantages/disadvantages

+ Distance information between linker residues
- Cross-linking reaction problematic
- Detection difficult

Detection

- Mass spectrometry

Experimental sources:
H/D exchange

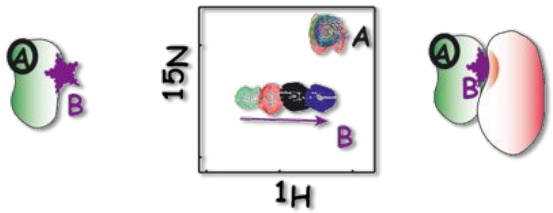
Advantages/disadvantages

+ Residue information
- Direct vs indirect effects
- Labeling needed for NMR

Detection

- Mass spectrometry
- NMR ^{15}N HSQC

Experimental sources:
NMR chemical shift perturbations



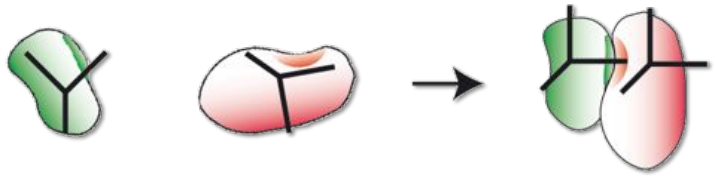
Advantages/disadvantages

+ Residue/atomic level
+ No need for assignment if combined with a.a. selective labeling
- Direct vs indirect effects
- Labeling needed

Detection

- NMR ^{15}N or ^{13}C HSQC

Experimental sources:
NMR orientational data (RDCs, relaxation)



Advantages/disadvantages

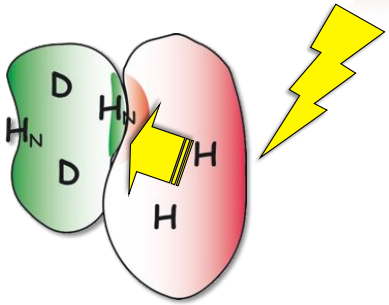
+ Atomic level
- Labeling needed

Detection

- NMR

Experimental sources:
NMR saturation transfer

Amide protons at interface
are saturated
==> intensity decrease



Advantages/disadvantages

+ Residue/atomic level

+ No need for assignment if combined with a.a. selective labeling

- Labeling (including deuteration) needed

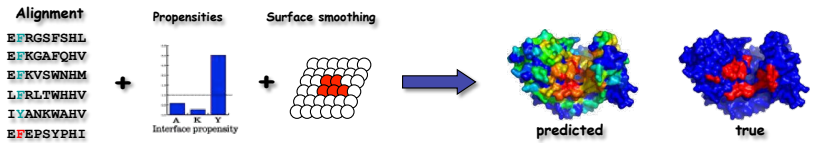
Other potential experimental sources

- **Paramagnetic probes in combination with NMR**
- **Cryo-electron microscopy or tomography and small angle X-ray scattering (SAXS) ==> shape information**

- **Fluorescence quenching**
- **Fluorescence resonance energy transfer (FRET)**
- **Infrared spectroscopy combined with specific labeling**
- **…**

Predicting interaction surfaces

- **In the absence of any experimental information (other than the unbound 3D structures) we can try to predict interfaces from sequence information?**
- **WHISCY:**
  WHat Information does Surface Conservation Yield?



http://www.nmr.chem.uu.nl/whiscy

De Vries, van Dijk Bonvin. *Proteins* 2006

Predicting interaction surfaces

- Several other approaches have been described:
  - HSSP **(Sander & Schneider, 1993)**
  - Evolutionary trace **(Lichtarge *et al.*, 1996)**
  - Correlated mutations **(Pazos *et al.*, 1996)**
  - ConsSurf **(Armon *et al.*, 2001)**
  - Neural network **(Zhou & Shan, 2001) (Fariselli *et al.*, 2002)**
  - Rate4Site **(Pupko *et al.*, 2002)**
  - ProMate **(Neuvirth *et al.*, 2004)**
  - PPI-PRED **(Bradford & Westhead, 2005)**
  - PPISP **(Chen & Zhou, 2005)**
  - PINUP **(Liang et al., 2006)**
  - SPPIDER **(Kufareva et al, 2007)**
  - PIER **(Porolo & Meller, 2007)**
  - SVM method **(Dong et al., 2007)**
  - … and many more since then

  - Our recent meta-server: **CPORT (de Vries & Bonvin, 2011)**

See review article (de Vries & Bonvin 2008)

Interface prediction servers

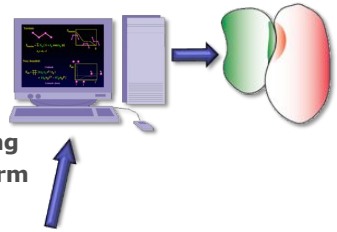
- PPISP (Zhou & Shan,2001; Chen & Zhou, 2005)
  http://pipe.scs.fsu.edu/ppisp.html
- ProMate (Neuvirth et al., 2004)
  http://bioportal.weizmann.ac.il/promate
- WHISCY (De Vries et al., 2005)
  http://www.nmr.chem.uu.nl/whiscy
- PINUP (Liang et al., 2006)
  http://sparks.informatics.iupui.edu/PINUP
- PIER (Kufareva et al., 2006)
  http://abagyan.scripps.edu/PIER
- SPPIDER (Porollo & Meller, 2007)
  http://sppider.cchmc.org

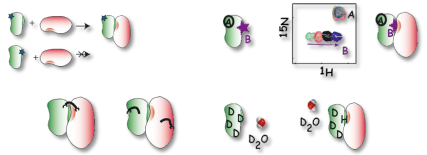
Consensus interface prediction (CPORT)
haddock.chem.uu.nl/services/CPORT

CPORT webserver

Combining experimental or predicted data with docking

- ***a posteriori*: data-filtered docking**
  - **Use standard docking approach**
  - **Filter/rescore solutions**
- ***a priori*: data-directed docking**
  - **Include data directly in the docking by adding an additional energy term or limiting the search space**

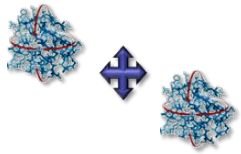




Overview

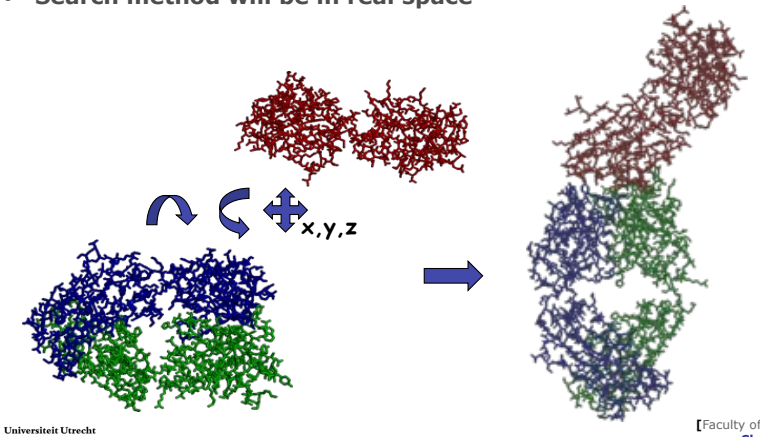
Docking

- **Choices to be made in docking:**
  - Representation of the system
  - Sampling method:
    - 3 rotations and 3 translations
    - Internal degrees of freedom?
  - Scoring
  - Flexibility, conformational changes?
  - Use experimental information?



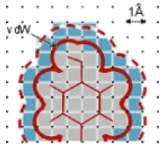
Explicit representation of the system

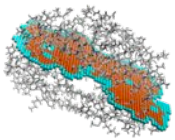
- x,y,z, coordinates of each atom for both molecules
- Search method will be in real space

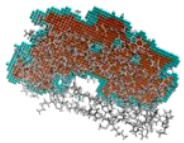


Grid-based representation of the system

- Discretise of the 3D structure of a protein onto a grid



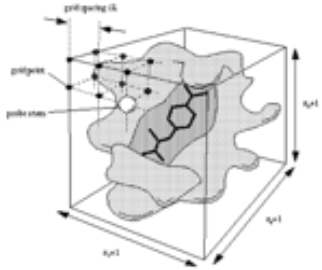




  - "Shape representation" of the protein (source: bigger / Krippahl)
  - Resolution defined by grid spacing
  - Docking will require to match the shapes ("geometric matching")
  - Search in real or Fourier space

Mixed representations of the system

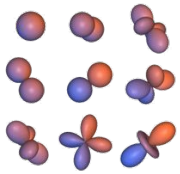
- Ligand and/or part of the interacting region is explicitly represented
- Remaining of structure is mapped onto a grid
- Interaction explicit atoms <-> grid
- E.g. AutoDock, ICM

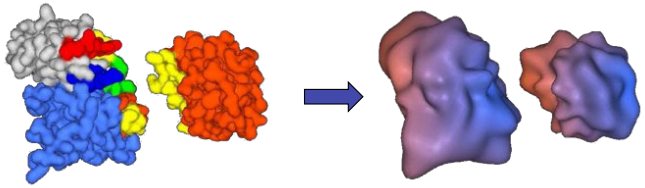


Surface representation of the system: spherical harmonics

- Surface of protein described by an expansion of spherical harmonics, e.g.

$$r(\theta,\phi) = \sum_{l=0}^{15} \sum_{m=-1}^{1} a_{lm} \psi_{lm}(\theta,\phi)$$

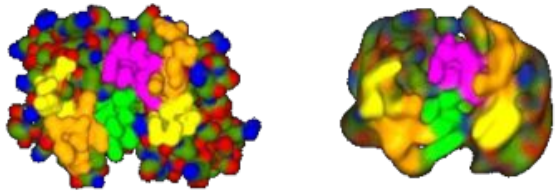




(source: HEX / Richie)

Surface representation of the system: spherical harmonics

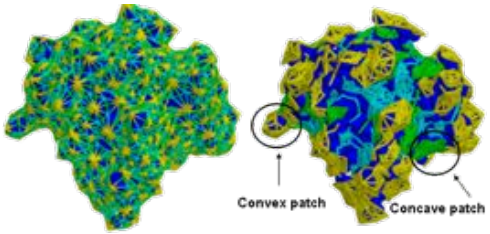
- By varying the number of terms in the expansion the resolution can be tuned



(source: HEX / Richie)

Surface representation of the system: surface patches

- Molecular shape representation: identify relevant "puzzle" pieces from the surface (e.g. convex or concave patches)
- Try to find mathing patches (geometric hashing)
- E.g.: PatchDock (Nussinov & Wolfson)



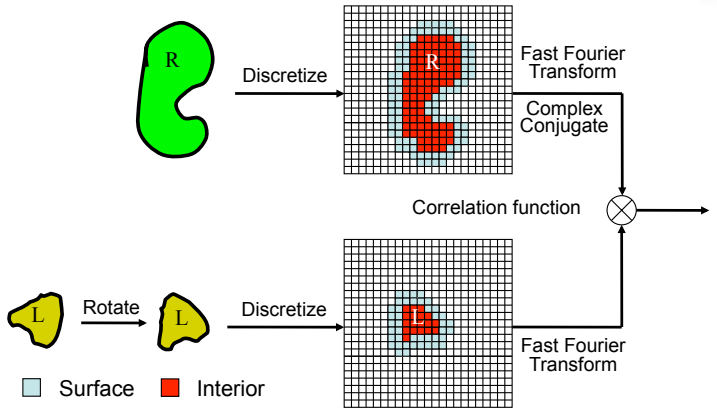
(source: PatchDock / Nussinov & Wolfson)

Overview

Systematic search

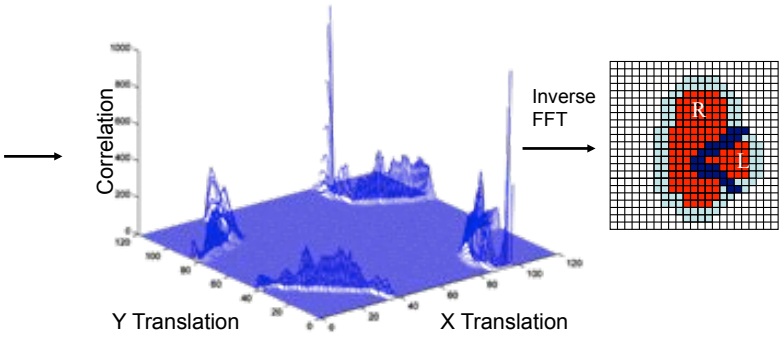
- Sample rotations (3) and translations (3)
- For each orientation calculate a score
- Can be very time consuming depending on scoring function
- Translational search often carried out in (2D or 3D) Fourier space by convolution of the grids
- Examples:
  - FFT methods: Z-DOCK, GRAMM, FTDOCK…
  - Direct search: Bigger (uses fast boolean operations)

Protein Docking Using FFT



(source Rong Chen, Boston University)

Protein Docking Using FFT



(source Rong Chen, Boston University)

Systematic search

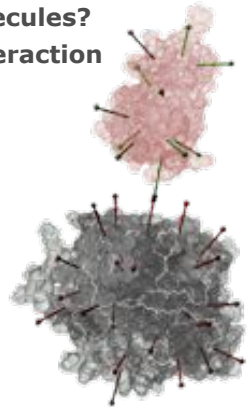
- Search can be carried out stepwise:
  - from low to high resolution
  - from crude to more sophisticated scoring
- A decreasing number of solutions is kept at each stages
- Final solutions often further refined (EM, MD…)

"Energy-driven" search methods

- Conformational search techniques aiming at minimizing some kind of energy function (e.g. VdW, electrostatic…):
  - Energy minimization
  - Molecular dynamics
  - Brownian dynamics
  - Monte-Carlo methods
  - Genetic algorithms
  - …
- Often combined with some simulated annealing scheme

"Energy-driven" search methods

- Still require some sampling of starting conditions:
  - How to position molecules?
  - Should be within interaction (attraction) range
  - E.g. "anchor points" in ICM (Abagyan)



Sample all combinations and for each several rotations

(source Fernando-Recio *et al* J.Mol.Biol (2004) 304:843)

Overview

Dealing with flexibility

- Flexibility makes the docking problem harder!
  - Increased number of degrees of freedom
  - Scoring more difficult
- Difficult to predict a-priori conformational changes
- Current docking methodology can mainly deal with small conformational changes
- Treatment of flexibility depends on the chosen representation of the system and the search method

Dealing with flexibility: "soft docking"

- Deal with small conformational changes (e.g. side-chain rotations) by **allowing overlap** in the (rigid-body) docking

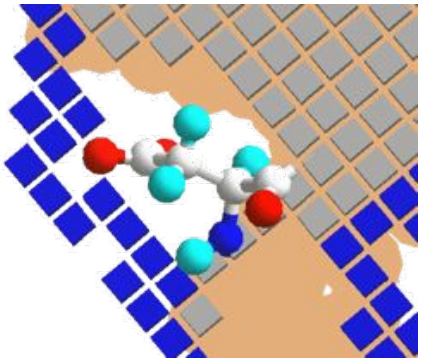
hard vs soft-rigid docking

- "Implicit" flexibility
- Solutions will require refinement to remove bumps

Dealing with flexibility: "soft docking"

- Implementation example in a grid-based method (bigger)

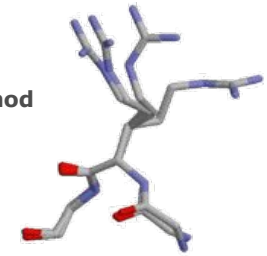
Core grid points corresponding to a flexible side-chain are empty
==> no core overlap during docking



(source: bigger / Krippahl)

Dealing with flexibility: docking from ensembles of conformations

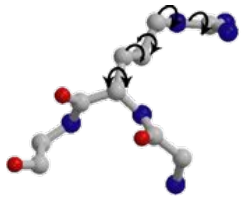
- Instead of using a single starting structure **use an ensemble corresponding to static snapshots** of various conformations, e.g.
  - from NMR
  - from MD or other conformational sampling method



- Applicable both for rigid and flexible docking

Explicit flexibility in docking

- Only for explicit representation of systems, i.e. not for grid- or surface-based methods



Side-chains only

Both side-chains and backbone

- Increases computational costs
- Often only introduced in later refinement stages

Overview

- Introduction
- Information sources
- General aspects of docking
  - Representation of the system
  - Search methods
  - Dealing with flexibility
  - Scoring
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Multiple choices…
- Challenges
- Conclusions & perspectives

Scoring

- **The holy grail in docking!**
- **Depends on the representation of the system and treatment of flexibility**
- **Depends on the type of complexes**
  - **e.g. antibody-antigen might behave differently than enzyme-inhibitors complexes**

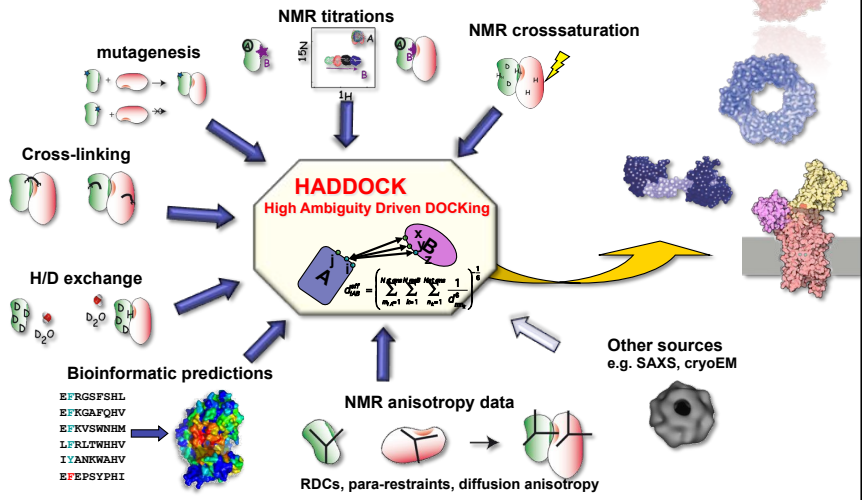
Scoring

- **Score is often a combination of various (empirical) terms such as**
  - **Intermolecular van der Waals energy**
  - **Intermolecular electrostatic energy**
  - **Hydrogen bonding**
  - **Buried surface area**
  - **Desolvation energy**
  - **Entropy loss**
  - **Amino-acid interface propensities**
  - **Statistical potentials such as pairwise residue contact matrices**
  - **…**
- **Experimental filters sometimes applied *a posteriori* if data available (e.g. NMR chemical shift perturbations, mutagenesis,..)**

Overview

- Introduction
- Information sources
- General aspects of docking
- **Information-driven docking with HADDOCK**
- Incorporating biophysical data into docking
- Multiple choices…
- Challenges
- Conclusions & perspectives

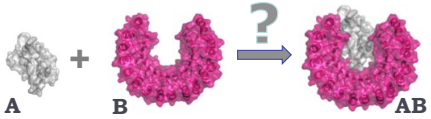
Data-driven HADDOCKing



Dominguez, Boelens & Bonvin. *JACS* 125, 173 (2003).

HADDOCK 2.1

- **High Ambiguity-Driven DOCKing**
  - **Different types of molecules** (up to 6 simultaneously)
    - **Proteins, Peptides, Nucleic Acids, Small molecules**
  - **Experimental (or predicted) low or high resolution data**
    - **E.g. Interface, Distances, Orientation, …**



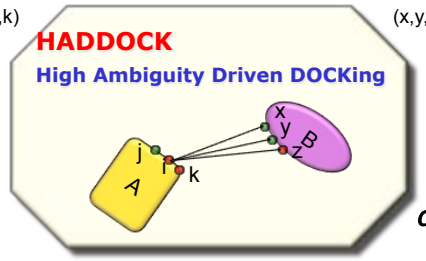
Web-server:	http://www.haddocking.org
Code:	http://www.nmr.chem.uu.nl/haddock/

de Vries *et al. Nat Prot. 2010*

Data-driven docking with HADDOCK

List of interface residues for protein A (i,j,k)

List of interface residues for protein B (x,y,z)



Effective distance d_{iAB}^{eff} calculated as

$$d_{iAB}^{eff} = \left(\sum_{m_i=1}^{Nat\,ams_A} \sum_{k=1}^{N\,resB} \sum_{n_k=1}^{Nat\,ams_B} \frac{1}{d_{mn_k}^{6}} \right)^{-\frac{1}{6}}$$

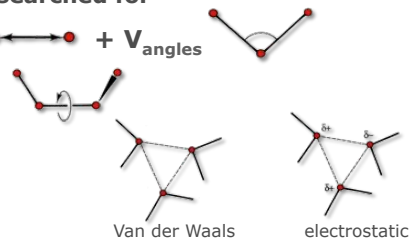
Ambiguous Interaction Restraint: (Nilges & Brunger 1991)

a residue must make contact with any residue from the other list

Different fraction of restraints (typically 50%) randomly deleted for each docking trial to deal with inaccuracies and errors in the information used

Searching the interaction space in HADDOCK

- **Experimental and/or predicted information is combined with an empirical force field into an energy function whose minimum is searched for**
- $V_{potential} = V_{bonds} + V_{angles}$
  $+ V_{torsion}$
  $+ V_{non\text{-}bonded}$
  $+ V_{exp}$

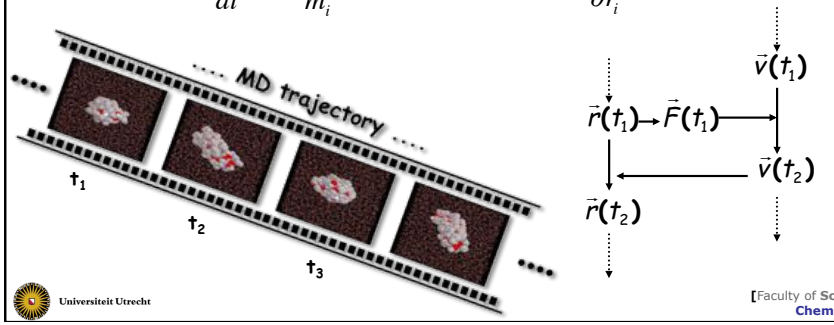


- **Search is performed by a combination of gradient driven energy minimization and molecular dynamics simulations**

Classical mechanics

- Molecular dynamics: generates successive configurations of the system by integrating **Newton's second law**

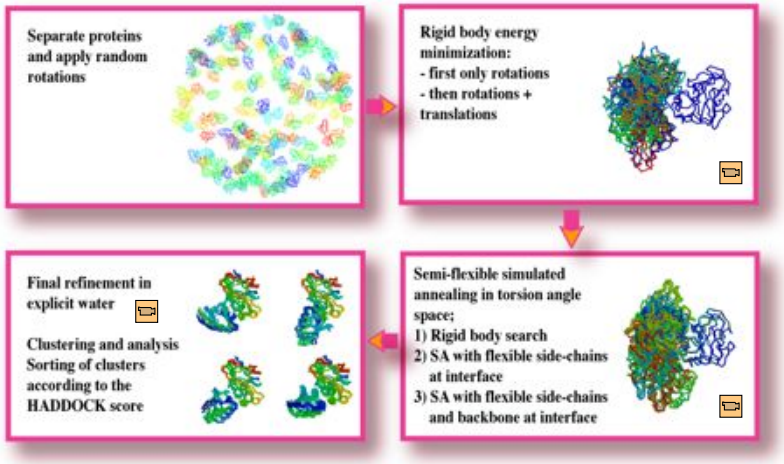
$$\frac{d^2}{dt^2}\vec{r}_i = \frac{\vec{F}_i}{m_i} \quad \text{with} \quad \vec{F}_i = -\frac{\partial V}{\partial \vec{r}_i}$$



Torsion angle dynamics

- dynamics time step dictated by bond stretching: waste of CPU time
- important motions are around torsions
- ~ 3 degrees of freedom per AA (vs $3N_{atom}$ for Cartesian dynamics)
- Available in DYANA, X-PLOR, CNS, X-PLOR-NIH

HADDOCK docking protocol



HADDOCK & Flexibility

- Several levels of flexibility:
- **Implicit:**
  - docking from ensembles of structures
  - Scaling down of intermolecular interactions
- **Explicit:**
  - semi-flexible refinement stage with both side-chain and backbone flexibility during in torsion angle dynamics
  - Final refinement in explicit solvent

Energetics & Scoring

- **OPLS non-bonded parameters** (Jorgensen, *JACS* 110, 1657 (1988))
- **8.5Å non-bonded cutoff, switching function, ε=10**
- **Ranking of based on HADDOCK score defined as:**

Rigid: Score = $0.01\,E_{air} + 0.01\,E_{vdW} + 1.0\,E_{elec} + 1.0\,E_{desolv} - 0.01\,BSA$

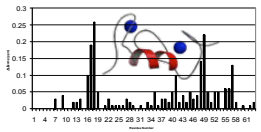
Flexible: Score = $0.1\,E_{air} + 1.0\,E_{vdW} + 1.0\,E_{elec} + 1.0\,E_{desolv} - 0.01\,BSA$

Water: Score = $0.1\,E_{air} + 1.0\,E_{vdW} + 0.2\,E_{elec} + 1.0\,E_{desolv}$

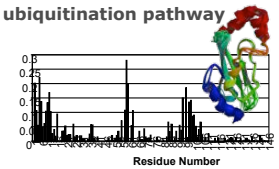
- E_{air}: ambiguous interaction restraint energy
- E_{desolv}: desolvation energy using Atomic Solvation Parameters (Fernandez-Recio et al *JMB* 335, 843 (2004))
- BSA: buried surface area

The Not4 – UbcH5B complex

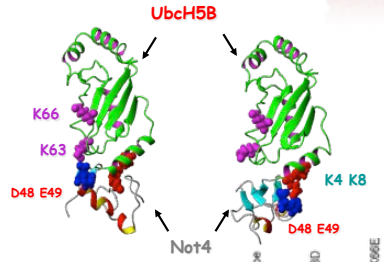
- **Not4: involved in the RNA polymerase II regulation. Contains a N-terminal Ring finger domain** (Hanzawa *et al.*, 2000)



- **UbcH5B: involved in the ubiquitination pathway**

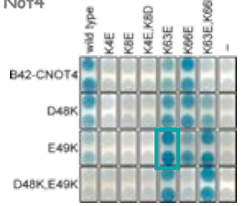


Best Haddock solutions



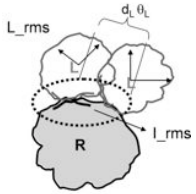
Haddock directed mutagenesis

==> Altered specificity mutants!

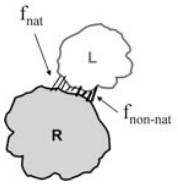


Assessment terminology

	F_{nat}	l-RMSD (Å)	i-RMSD (Å)
High (***)	≥0.5	≤1	≤1
Medium (**)	≥0.3	≤5	≤2
Acceptable (*)	≥0. 1	≤10	≤4
Incorrect	<0. 1	>10	>4



- i-RMSD: Interface RMSD
- l-RMSD: Ligand RMSD
- F_{nat}: Fraction of native contacts



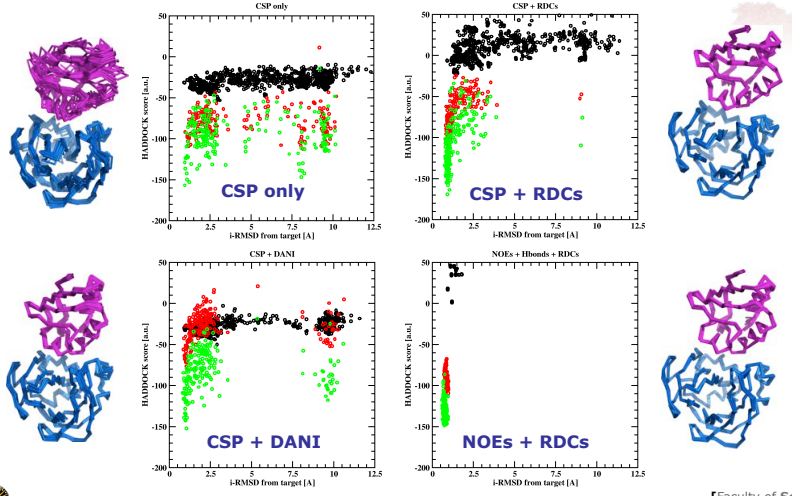
Lensink *et al. Proteins 2007*

Accuracy <-> Data

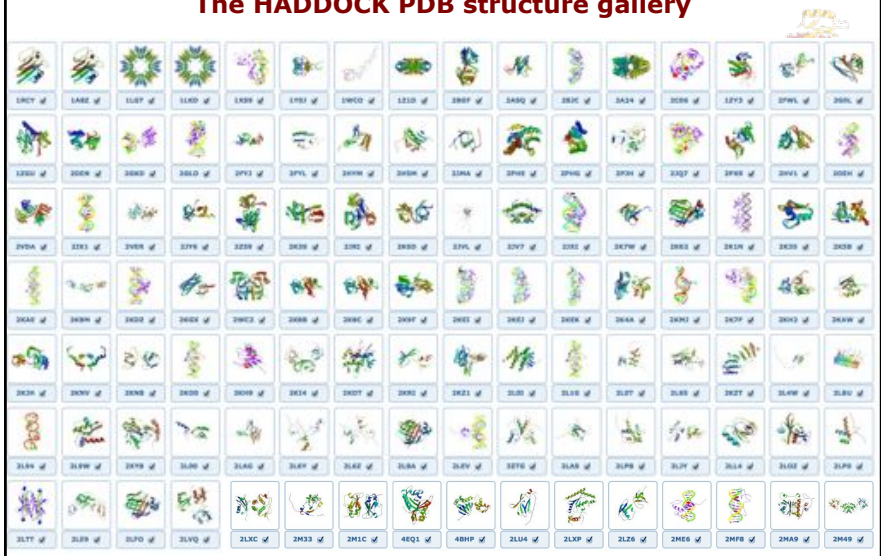
When does the model stop

and the structure start?

Accuracy <-> Data: E2A-HPR



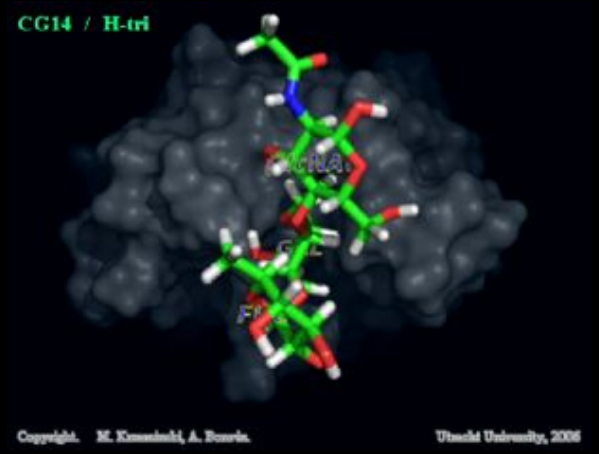
The HADDOCK PDB structure gallery



>110 entries – Jan. 2014

Image collage from http://www.pdb.org

Fully flexible protein-ligand docking

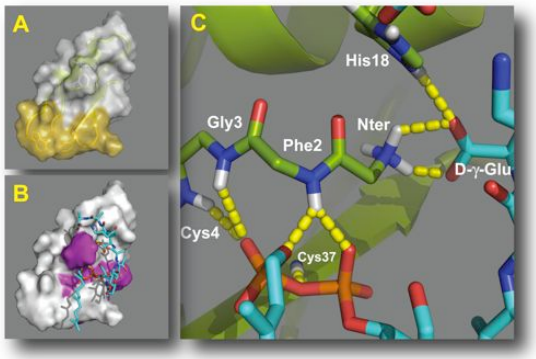


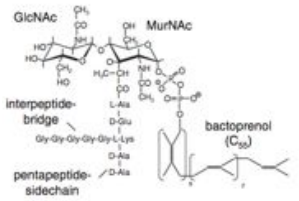
Wu *et al. Glycobiology* 2007

Can deal with complex molecules

Science
AAAS

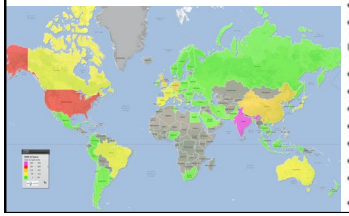
Plectasin, a Fungal Defensin, Targets the Bacterial Cell Wall Precursor Lipid II
Tanja Schneider, *et al.*
Science **328**, 1168 (2010);
DOI: 10.1126/science.1185723





Haddock web portal

- > 3600 r…ered users
- > 60000 se… runs since Jun… 2008
- ~ 12% on the GRID



Visit haddock.science.uu.nl

De Vries *et al.*
Nature Prot. 2010

Overview

- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- **Incorporating biophysical data into docking**
- Multiple choices…
- Challenges
- Conclusions & perspectives

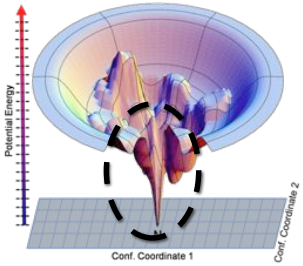
Universiteit Utrecht

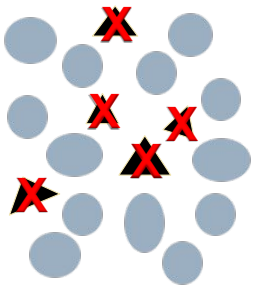
[Faculty of Science Chemistry]

Data Incorporation

a priori: as Restraint

a posteriori: as Filter



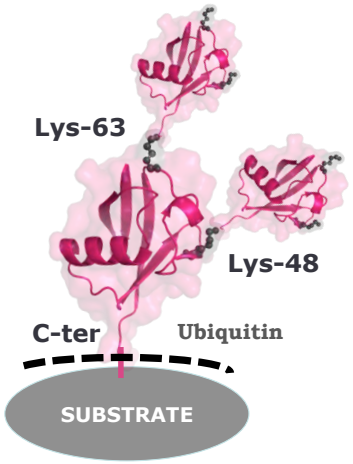


http://www.cs.gmu.edu/~ashehu/?q=ProjectionGuidedExploration

Universiteit Utrecht

[Faculty of Science Chemistry]

NMR Example: CSP-driven docking



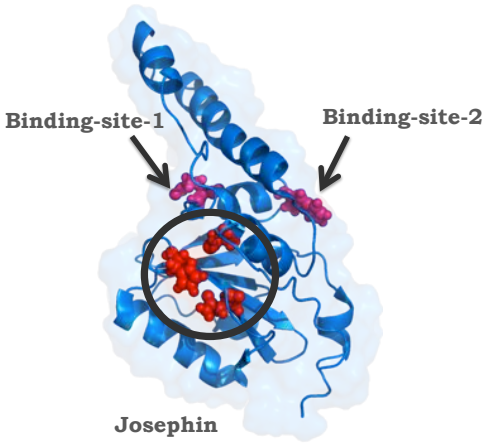
- Ub-cleaving enzyme
  - Josephin
- Which di-Ub linkage type is cleaved, K48 and/or K63 linkage?
- Collaboration with Annalisa Pastore (London, MRC)

Nicastro et al., *Plos One*, 2010

Universiteit Utrecht

[Faculty of Science Chemistry]

NMR Example: CSP-driven docking

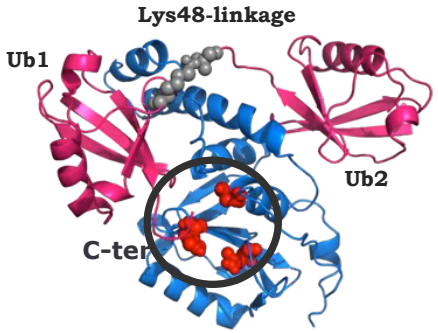


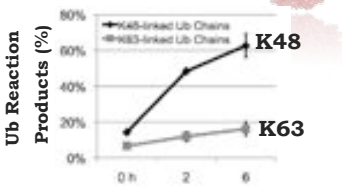
Input for docking:

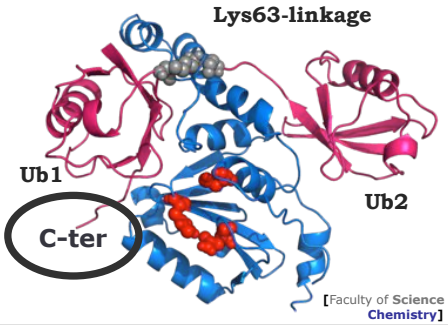
- Catalytic Triad
- 2 Binding-sites
  - CSP + Mutation
- FMD Protocol

Nicastro et al., *Plos One*, 2010

NMR Example: CSP-driven docking







Nicastro et al., *Plos One*, 2010

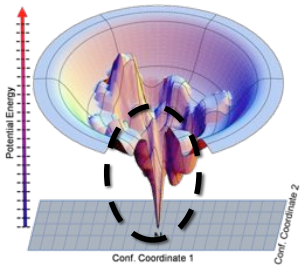
Overview

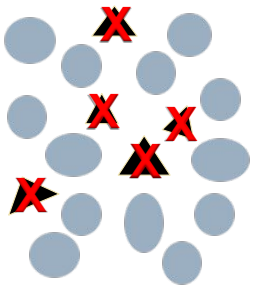
- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- **Incorporating biophysical data into docking**
  - Chemical shift perturbation data
  - Pseudo-contact shifts
  - **SAXS and CCS as filters in docking**
- Multiple choices…
- Challenges
- Conclusions & perspectives

Data Incorporation

a priori: as **Restraint**

a posteriori: as **Filter**

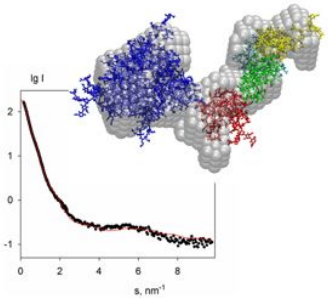


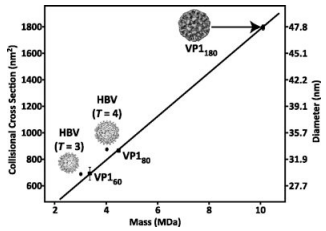


http://www.cs.gmu.edu/~ashehu/?q=ProjectionGuidedExploration

Integration of shape information

Small Angle X-ray Scattering
- SAXS Curve





Ion Mobility Mass Spectrometry
- Collision Cross Section (CCS)

Benchmarking the performance of shape filters

Protein-Protein Docking Benchmark Version 3.0

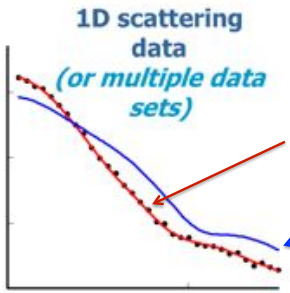
Howook Hwang[1], Brian Pierce[1], Julian Mintseris[2], Joël Janin[3], and Zhiping Weng*[1,4]
[1]Boston University Bioinformatics Program, Boston, USA
[2]Department of Cell Biology, Harvard Medical School, Boston, MA.
[3]Yeast Structural Genomics, IBBMC Université Paris-Sud, CNRS UMR 8619, 91405-Orsay, France
[4]Boston University Biomedical Engineering Department, Boston, USA

124 protein-protein complexes (bound & unbound):
88 Rigid Body
19 Medium
17 Difficult

⇒

***ab-initio* HADDOCK:**
Center of mass restraints
it0: 10.000 models

Measuring the quality of fit



Experimental SAXS Curve (or simulated from Xtal complex

vs.

SAXS Curve – Docked Model

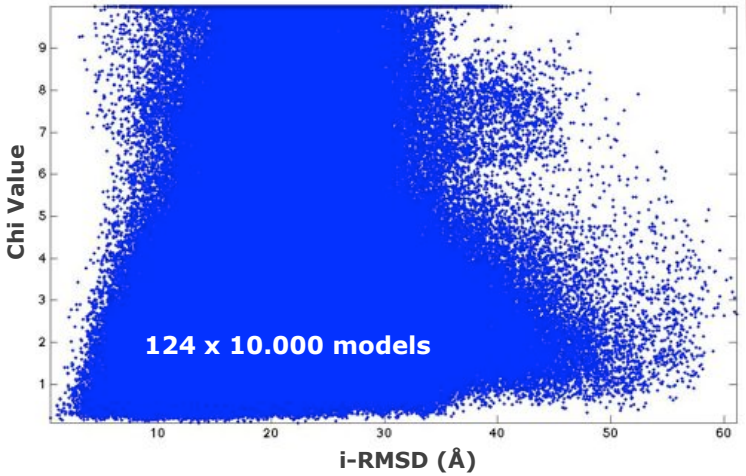
} Chi

CRYSOL (Svergun et al)

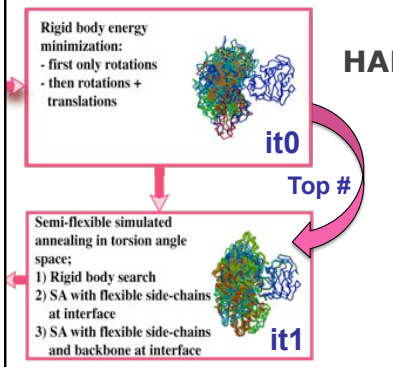
$$\text{discrepancy: } \chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I_{exp}(s_j) - cI(s_j)}{\sigma(s_j)}\right]^2$$

Svergun et al., *J. Appl. Cryst*, 1995

Is SAXS alone enough?



Protein Docking with a SAXS-filter

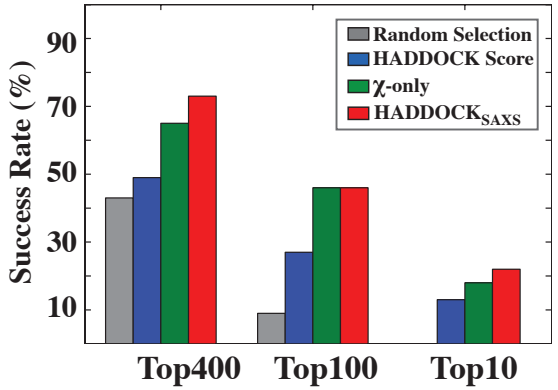


$HADDOCK_{SAXS} = HADDOCK_{Score} + w \cdot SAXS_{Chi}$

w is optimized such to maximize the number of a "acceptable" models in the top 400

"acceptable": i-RMSD ≤ 4Å

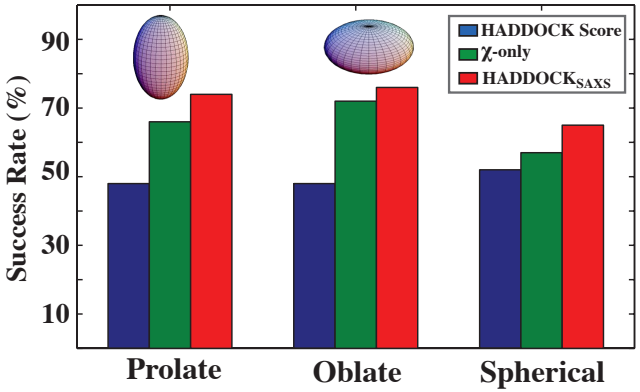
HADDOCK-Scoring can be improved with SAXS



among the cases having at least one acceptable (138/176)

Shape dependency of SAXS-scoring

Top 400 – among acceptable's



Test against experimental SAXS data

PDB id	Complex Type	#aa per monomer	Shape /Anisotropy value
3K3K[b] (Nishimura et al., 2009)	Homodimer	211	Prolate / 0.8
2R15[b] (Pinotsis et al., 2008)	Homodimer	212	Prolate / 26.8
1O6S[u] (Schubert et al., 2002)	Dimer	466/105	Prolate / 11.3

Test against experimental SAXS data

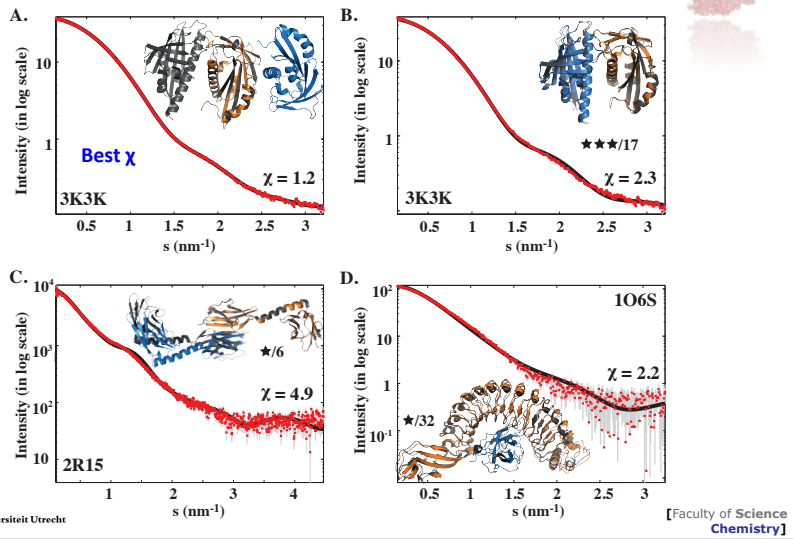
Rigid-body docking

	#hits out of 10000 models	#hits ranked by HADDOCK in the top 400 (%)	#hits ranked by HADDOCK$_{SAXS}$ in the top 400 (%)
3K3K	117	74 (63%)	52 (44%)
2R15	2	0 (%)	2 (100%)
1O6S	25	0 (%)	20 (80%)

After flexible refinement

	Single structure scoring (Quality/Rank)	Cluster-based scoring (Quality/Rank)
3K3K	***/17	***/1
2R15	*/6 &**/12	-
1O6S	*/32	*/1

Test against experimental SAXS data



Overview

- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- **Multiple choices...**
- Challenges
- Conclusions & perspectives

Multiple choice...

Modelling multibody (N>2) assemblies and dealing with large conformational changes

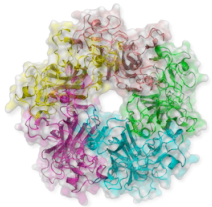
Protein-peptide docking

Protein-DNA modelling

Small molecule docking

HADDOCK's adventures in CAPRI

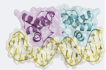
Building large macromolecular assemblies by multi-body docking



- ✓ Simultaneous docking (N≤6)
- ✓ Hetero- or homo-oligomers
- ✓ Symmetry between and within each molecule

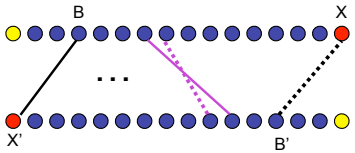
Karaca *et al. Mol. Cell. Proteomics 2010*

Benchmark set

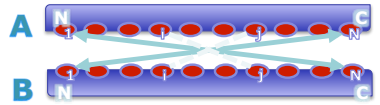
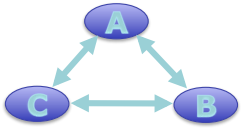
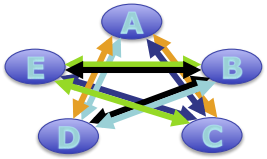
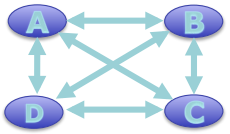
PDB ID	CATH Classification	Complex Type	Docking Type	Symmetry Type	# of residues per Monomer	
1QU9	Mainly Beta	Homotrimer	Bound	C3	128	
1URZ	Alpha Beta	Homotrimer	Unbound	C3	400	
1OUS	Alpha Beta	Homotetramer	Bound	D2	114	
1VIM	Alpha Beta	Homotetramer	Bound	D2	200	
1VPN	Mainly Beta	Homopentamer	Bound	C5	289	
3CRO	Mainly Alpha	Homodimer-ds DNA	Unbound	C2	71 (Protein) 20 (DNA)	

Dealing with symmetry

- Two kinds of symmetry restraints can be used:
  - NCS: non-crystallographic symmetry
    - enforces that monomer A be identical to A' without imposing a symmetry relationship between them
  - C_2 symmetry: enforced by defining pair of distances that must be equal (symmetry potential in CNS (Nilges *et al.*))

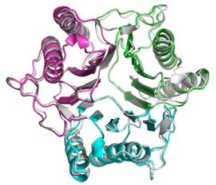


C3, C5 and D2 can be defined by combinations of C2 pairs

C2		$d(A_iB_j) = d(B_iA_j)$
C3		$d(AB) = d(BC)$ $d(BC) = d(CA)$ $d(CA) = d(AB)$
C5		$d(AC) = d(AD)$ $d(BD) = d(BE)$ $d(CE) = d(CA)$ $d(DA) = d(DB)$ $d(EB) = d(EC)$
D2		$d(AB) = d(BA)$ $d(AC) = d(CA)$ $d(AD) = d(DA)$ $d(BC) = d(CB)$ $d(BD) = d(DB)$ $d(CD) = d(DC)$

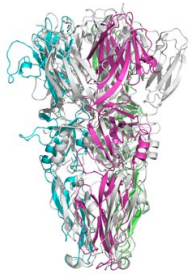
Based on the use of symmetrical distance restraint (Nilges)

Results (quality/rank)

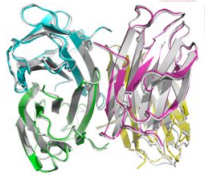


1QU9[b] - CPORT ★★★ / 1

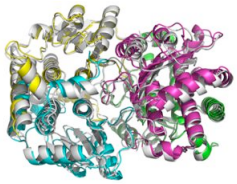
1URZ[u] – CPORT and experimental data



★★/ 1



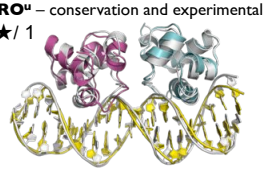
1OUS[b] - CPORT ★★★/ 1

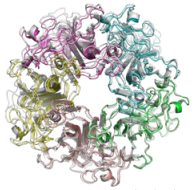


1VIM[b]- CPORT ★★★/ 1

3CRO[u] – conservation and experimental data

★★/ 1





1VPN[b]- CPORT ★★★ / 1

Gray: Xray Structure

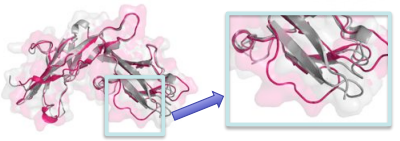
Universiteit Utrecht

Karaca *et al. Mol. Cell. Proteomics 2010*

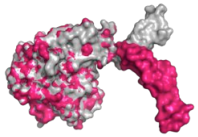
[Faculty of Science Chemistry]

Dealing with conformational changes in docking

- Local changes: (small) loop reorientations and structure changes
  - ✓ Ensemble Docking
  - ✓ Soft Docking



- Global changes: large scale domain motions (hinge, shear)



  - ✓ Divide and Conquer
  - ✓ Multi-Body Docking
- Binding-induced folding events.... ???

Universiteit Utrecht

Karaca & Bonvin. *Structure* 2011

[Faculty of Science Chemistry]

Benchmark

PDB	CATH	Molecular Classification	Backbone RMSD(Å) Receptor	Backbone RMSD(Å) Ligand
1IRA	Mainly Beta	Cytokine Receptor/ Antagonist	19.5	0.7
1H1V	Alpha - Beta	Actin Binding	13.9	1.6
1Y64	Alpha - Beta	Structural Protein	10.3	1.1
1F6M	Alpha - Beta	Oxidoreductase	7.3	0.92
1FAK	Mainly Beta	Blood Clotting	6.0	1.0
1ZLI	Alpha - Beta	Hydrolase/Inhibitor	3.8	0.6
1E4K	Mainly Beta	Immune System	2.9	1.7
1IBR	Alpha - Beta / Mainly Beta	Cell Cycle	2.9	1.1
1KKL	Alpha - Beta	Hydrolase/Transferase	2.6	0.5
1NPE	Mainly Beta	Structural Protein	1.8	-
1DFJ	Alpha - Beta	Endonuclease/Inhibitor	1.5	0.7

RMSD

Challenging

Universiteit Utrecht

[Faculty of Science Chemistry]

Docking Protocol

Treat the molecule as a collection of sub-domains with conn rest betwe

Define the Hinge Regions

Hinge Predictor Based on Elastic Network model
http://www.prc.boun.edu.tr/appserv/prc/hingeprot/

HINGEprot

Define interactions and distance restraints for removed peptide bonds

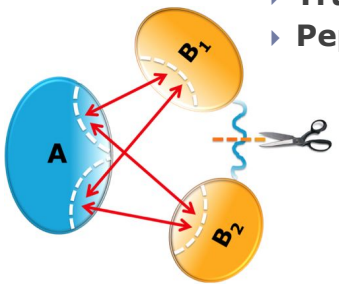
HADDOCKing

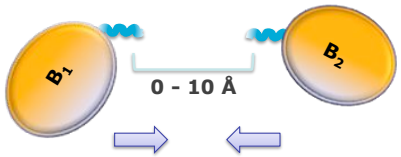
Universiteit Utrecht

[Faculty of Science Chemistry]

Docking setup

- True interface
- Peptide connectivity restraints

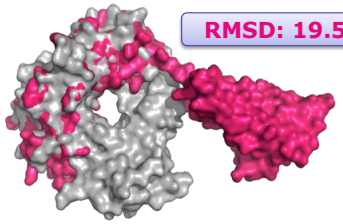




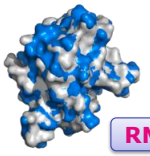
Reduce in final refinement stage
To real peptide bond distance

- Fully flexible hinge regions
- Center of Mass Restraints
- Simultaneous 3-body docking

Case Study: 1IRA



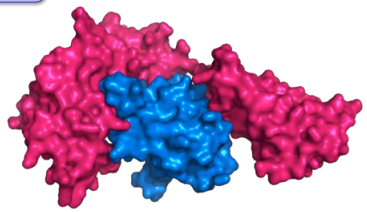
RMSD: 19.5 Å



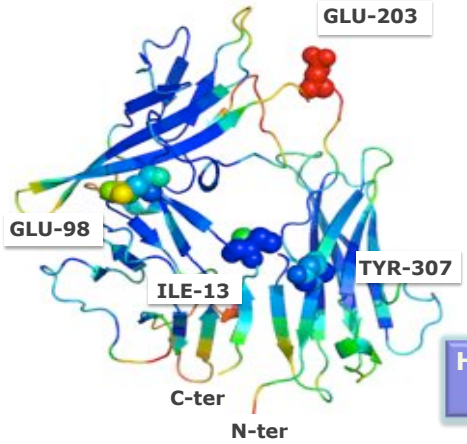
RMSD: 0.7 Å

Receptor:
bound vs. unbound

Ligand:
bound vs. unbound



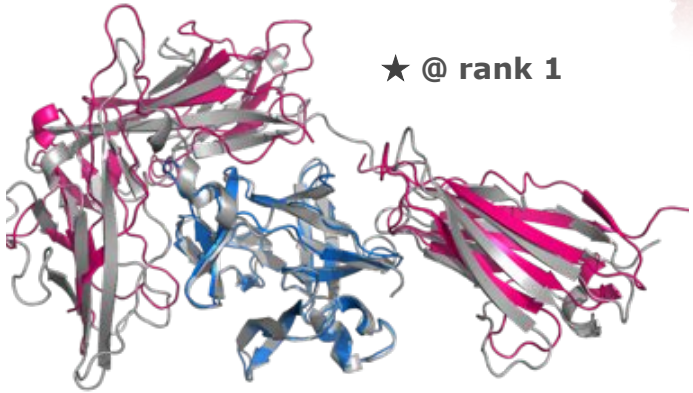
Case Study: 1IRAu



Hinge Prot Predictions:
13, 93, 203, 307

Structure colored according to b-factors.

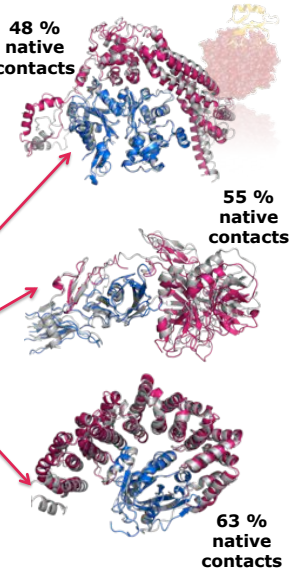
Case Study: 1IRAu



★ @ rank 1

Xray Structure

	i-RMSD(Å) / Rank		
PDB ID	FMD	2-body Docking	Receptor RMSD (Å)
1IRA	**3.9** / 1	17.5 / 1	19.5
1H1V	**4.6** / 11	11.9 / 1	13.9
1Y64	**3.9** / 5	10.3 / 1	10.3
1F6M	**3.5** / 1	14.1 / 1	7.3
1FAK	**3.4** / 2	11.4 / 1	6.0
1ZLI	**2.1** / 1	14.8 / 1	3.8
1IBR	**2.3** / 1	9.6 / 1	2.9
1E4K	**2.3** / 1	4.1 / 1	2.9
1KKL	**2.2** / 1	3.1 / 1	2.6
1NPE	**1.2** / 16	1.7 / 1	1.8
1DFJ	**2.0** / 5	1.8 / 116	1.5



Acceptable

Multiple choice...

- Modelling multibody (N>2) assemblies and dealing with large conformational changes
- Protein-peptide docking
- Protein-DNA modelling
- Small molecule docking
- HADDOCK's adventures in CAPRI

Peptides : Why?

MOTIVATION

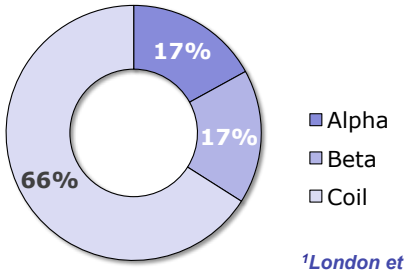
- **Involved in up to 40% of protein interactions**
  - Mainly regulatory processes and protein-protein interaction inhibition – relevant for IDPs and IDRs
- **Promising leads for drug design**

CHARACTERISTICS

- **Short size** - 2-30 amino acids length
- **High flexibility** – often disordered when unbound
- **Small interfaces** – ≈ 500Å² (3-4 times smaller than protein-protein interfaces)
- **Transient interactions** - low binding affinities (μM)

Benchmark:
103 protein-peptide complexes[1]

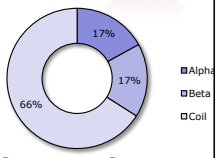
- **Peptide length : From 5 to 15 residues**
- **25 peptides and 78 protein fragments**
- **Different bound conformations**



[1]*London et al., Structure (2010)*

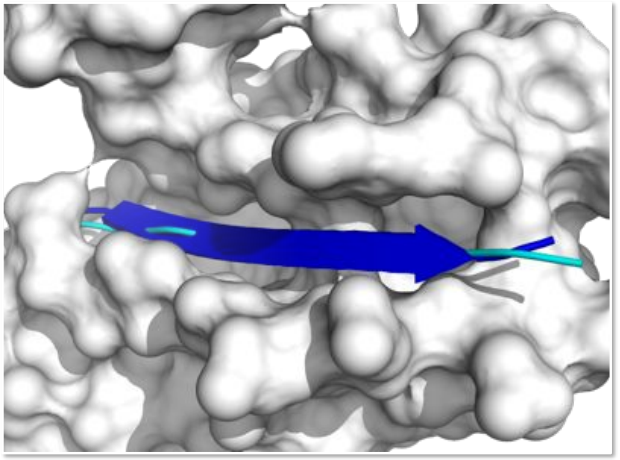
Benchmark:
103 protein-peptide complexes[1]

- Peptide length : From 5 to 15 residues
- 25 peptides and 78 protein fragments
- Different bound conformations
- **No significant conformational changes in protein upon binding**

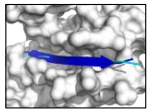


[1]*London et al., Structure (2010)*

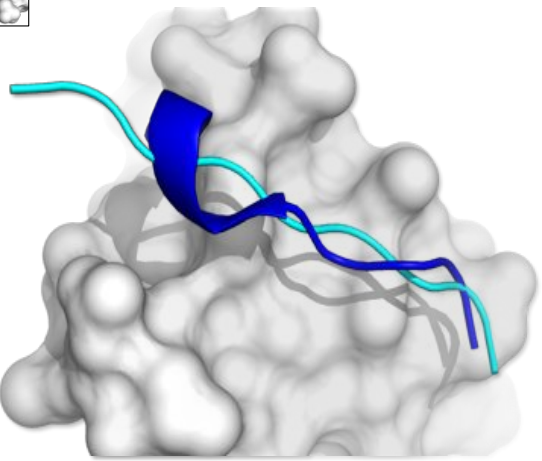
Difficulty range



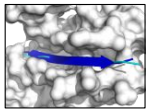
EASY: Peptides RMSD extended vs bound conformation <= 2Å



EASY

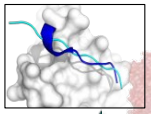


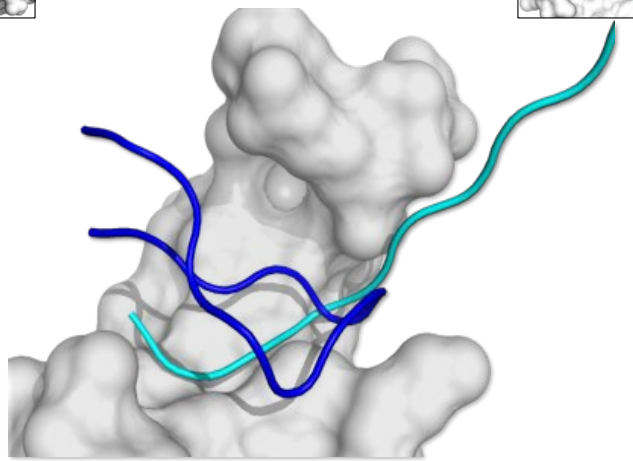
MEDIUM: Peptides RMSD extended vs bound conformation between 2 and 4Å



EASY

MEDIUM



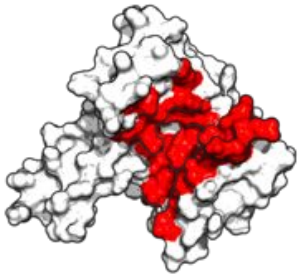


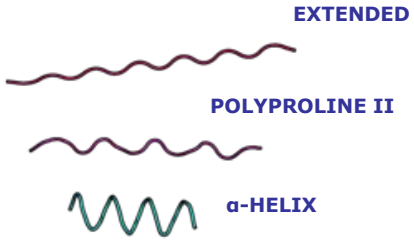
DIFFICULT: Peptides RMSD extended vs bound conformation more than 4Å

Dealing with conformational changes

A combination of:

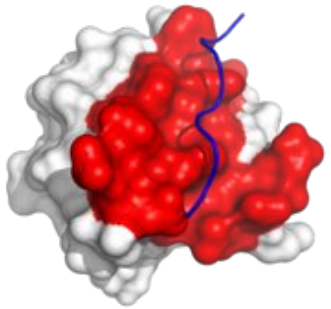
- **conformational selection** (ensemble docking from 3 conformations)
- **and induced fit** (flexible refinement of both protein and peptide – increased flexibility and flexible refinement length)



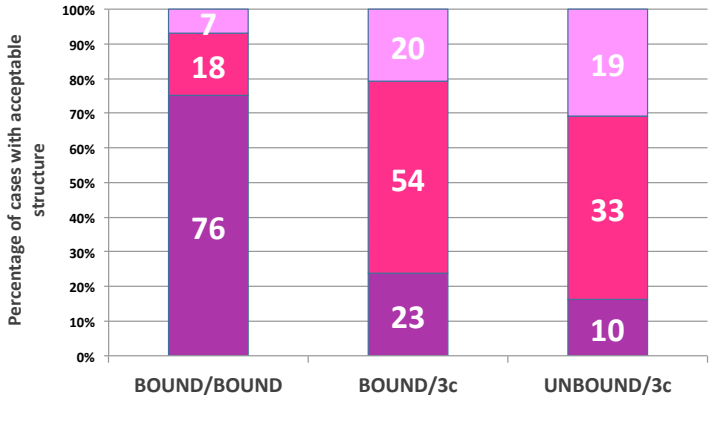


HADDOCK parameters

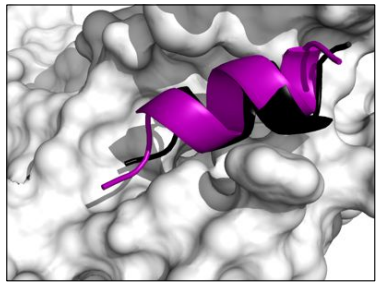
- **Interface:**
  - **Calculated from experimental complexes -> TRUE interface**
  - **Protein -> Residues within 5Å of the peptide chain**
  - **Peptide -> All passive**



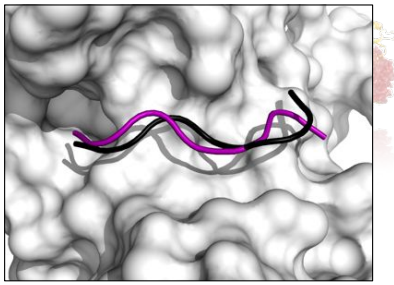
HADDOCK results



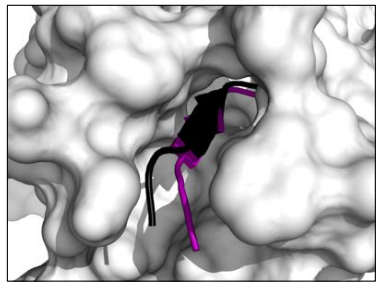
- **NOT ACCEPTABLE**: Interface RMSD > **2Å** from the native complex
- **NEAR NATIVE**: Interface RMSD < **2Å** from the native complex
- **SUB-ANGSTROM**: Interface RMSD < **1Å** from the native complex



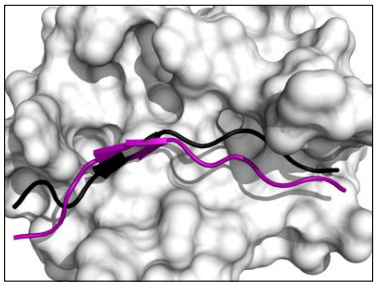
1NX1 (hard) 1st cluster – 1.4Å



1CZY (medium) 1st cluster – 1.2Å

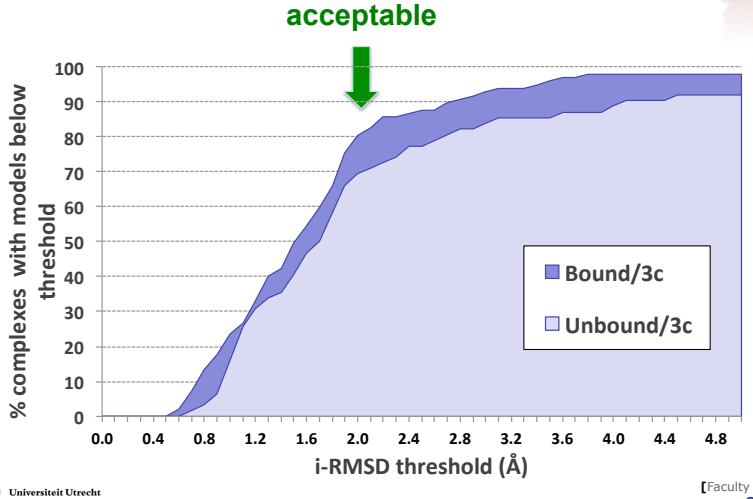


1LVM (easy) 1st cluster – 1.1Å

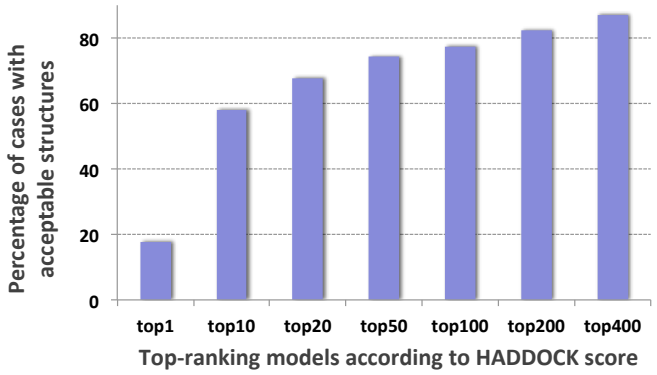


1D4T (medium) 2nd cluster – 1.6Å

Analysis – *models resolution*

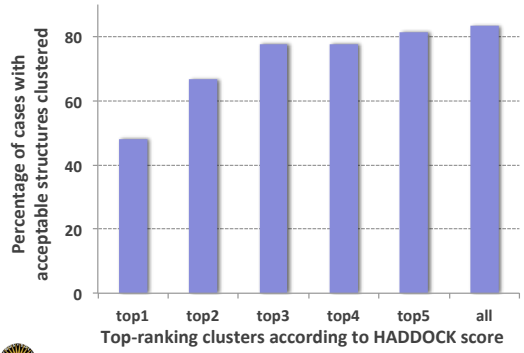


Analysis
ranking of individual models
(unbound – C3)



Analysis - *clustering*
(unbound – C3)

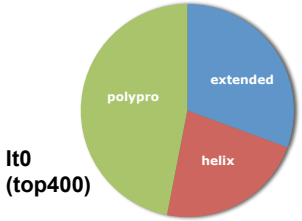
- **Clustering assessment:**
  - At least one **acceptable** model in **top 4** of at least **one** cluster
  - RMSD clustering at **5Å**



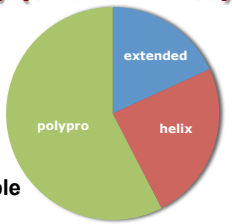
Acceptable cluster in 32 out 40 cases with at least one acceptable model among the clusters representatives.

The clustering is successful for 80% of the cases

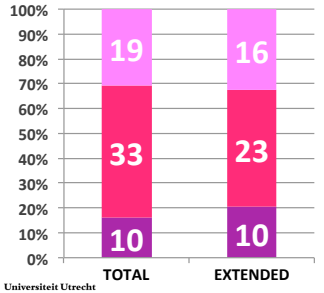
Conformational selection performance
extended conformation recovery (unbound – C3)



It0 (top400)



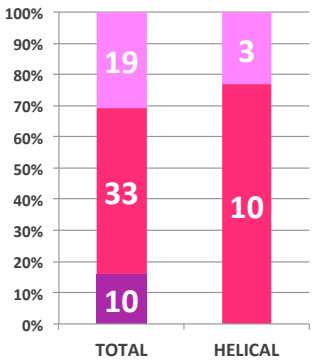
It0 all acceptable



33% of cases without **acceptable** models
- **56%** from it0 **extended** selection

67% of cases with **acceptable** models
- **18%** from it0 **extended** selection

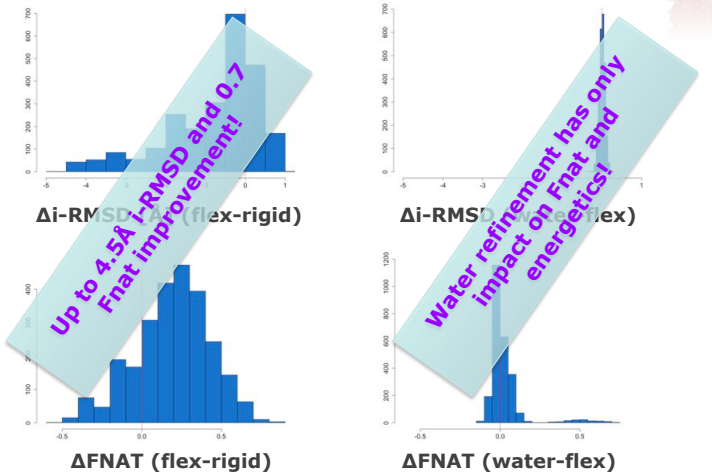
Conformational selection performance
helix recovery (unbound – C3)



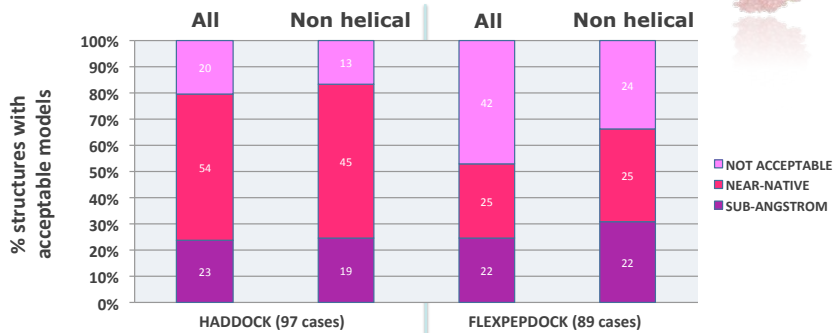
23% of cases **without acceptable** models
- **2** from it0 polypro selection
- **1** from it0 extended selection

77% of cases **with acceptable** models
- **All** from it0 helix/polypro selection

Induced fit performance
(improvements of flexible refinement?)



Comparison with FlexPepDock



- **Large** putative interacting **surface**
- bb-i-RMSD on **protein and peptide**
- **Bound** protein + **3 peptide conf.**
- Conformational selection + induced fit **docking**

- Peptide **anchor residue**
- bb-i-RMSD on **peptide only**
- **Bound** protein + **extended** peptide
- Flexible **refinement**

Raveh *et al.* Proteins 2010

Conclusions

- Our protein-peptide docking protocol combines **conformational selection** and **induced fit** models of biomolecular recognition.
- **Good performance** on benchmark with 69% success rate* (<2Å) on unbound cases.
- **Sub-ångstrom** resolution in 17% of unbound cases.
- Large conformational changes remain challenging.
- CAPRI (blind docking experiment) is now accepting peptide targets. So please consider contributing!

** 87% success rate (31% sub-ångstrom) if using a metric that is independent from conformational changes on the protein receptor*

Trellet *et al.* PloS One 2013

Multiple choice...

Modelling multibody (N>2) assemblies and dealing with large conformational changes

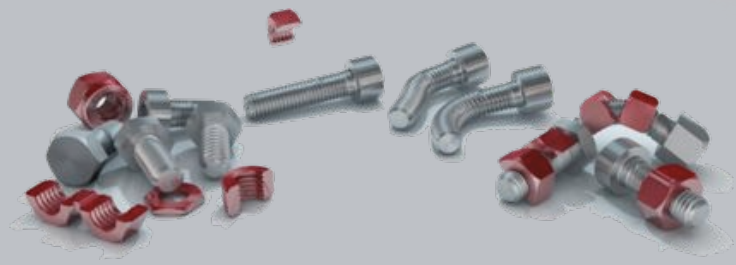
Protein-peptide docking

Protein-DNA modelling

Small molecule docking

HADDOCK's adventures in CAPRI

Modeling protein-DNA interactions: Bend and Twist it to make it fit



Modelling of Protein-DNA complexes: a two-stage protocol

Input structures:
- canonical B-DNA
- Protein (ensemble)

1st docking run

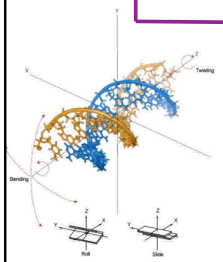
It0 → It1 → Water → Scoring

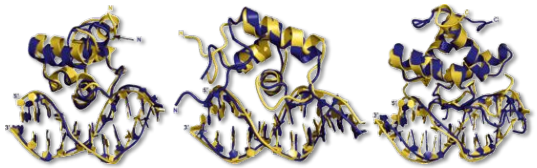
It0: rigid body docking It1: semi-flexible refinement
Water: final refinement in explicit solvent

DNA library generation

2nd docking run

It0 → It1 → Water → Scoring





Cro - O_1R iRMSD = 1.62 Å

Lac - O1 iRMSD = 2.02 Å

Arc - operator iRMSD = 1.90 Å

Van Dijk *et al. Nucl. Acid. Res.* 2006

Generating (custom) nucleic acids structures

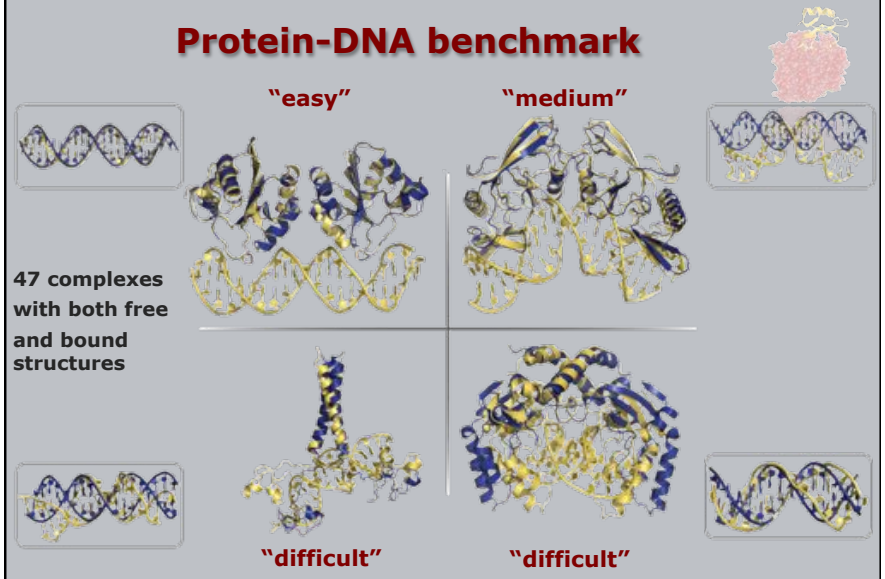
Control over global conformation (bend & twist)
Uses 3DNA (Lu & Olson, NAR 2003)

Van Dijk & Bonvin NAR 2009

Generate A-DNA or B-DNA from sequence
Full control over base-pair(step) parameters

haddock.science.uu.nl/dna

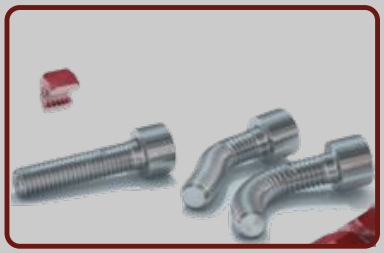
Protein-DNA benchmark



47 complexes with both free and bound structures

Van Dijk *et al*. NAR 2008

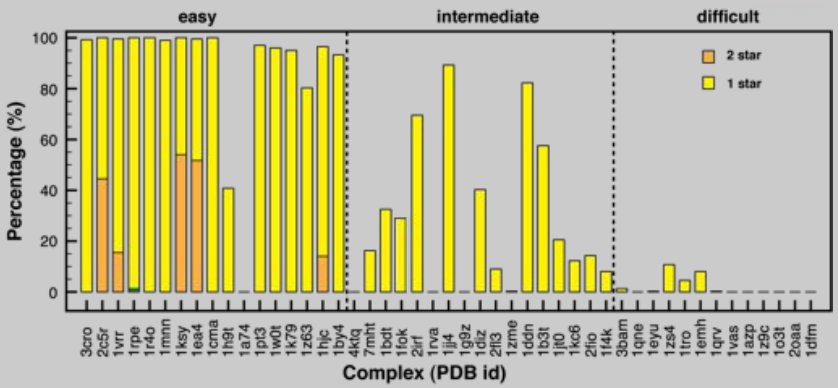
Unbound-Unbound using canonical B-DNA and true interface restraints



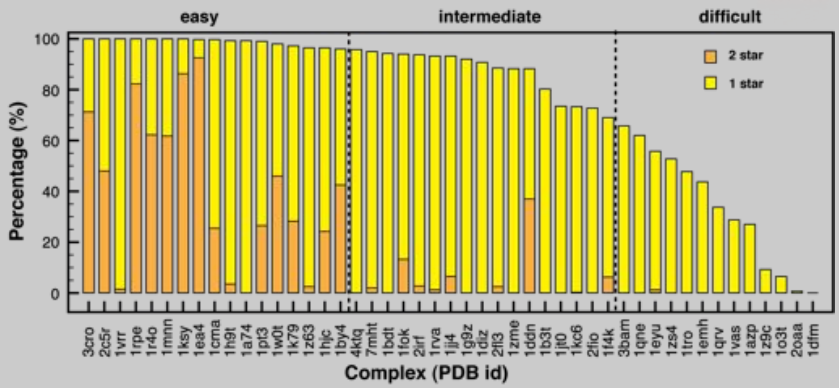
Is the protein-DNA docking procedure able to account for conformation changes, and to what extend?

Van Dijk & Bonvin. NAR 2010

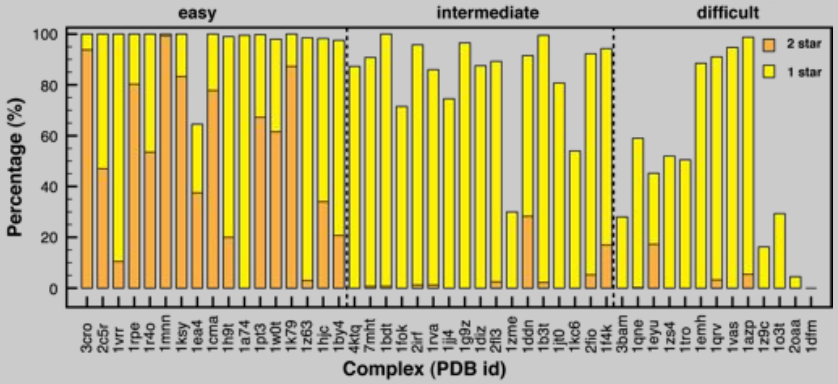
Performance of rigid-body docking only



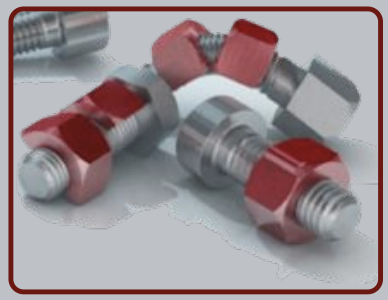
Performance after flexible refinement (1 cycle)



Performance after the 2 steps protocol with custom DNA library



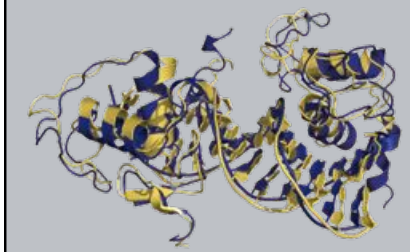
Unbound-Unbound using canonical B-DNA with experimental information



How well does the procedure perform when knowledge-based restraints are used?

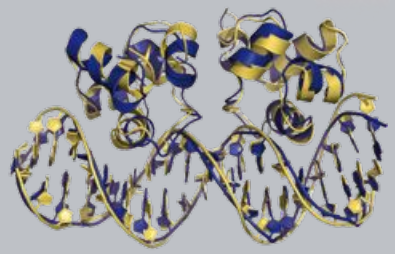
"easy" cases

Retinoic acid receptor



1by4 **
fnat = 0.40
iRMSD = 3.55 Å
dRMSD = 1.50 Å

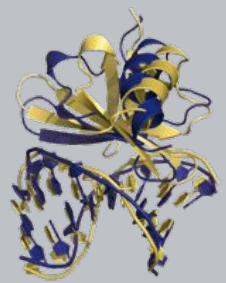
434 Cro protein



3cro **
fnat = 0.50
iRMSD = 2.23 Å
dRMSD = 1.93 Å

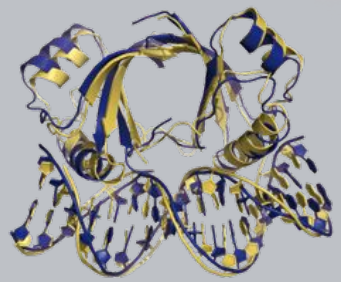
"medium" cases

Hyperthermophile chromosomal protein SAC7D



1azp *
fnat = 0.11
iRMSD = 3.44 Å
dRMSD = 1.58 Å

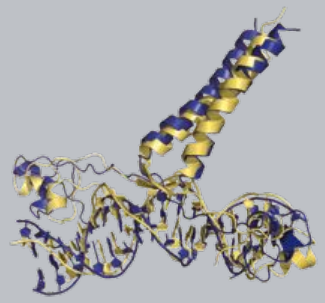
papillomavirus type 18 E2



1jj4 **
fnat = 0.44
iRMSD = 2.63 Å
dRMSD = 2.26 Å

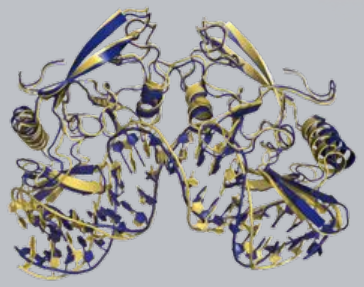
"difficult" cases

PUT3



1zme *
fnat = 0.15
iRMSD = 3.75 Å
dRMSD = 3.23 Å

1-PPOL homing endonuclease



1a74 **
fnat = 0.31
iRMSD = 3.24 Å
dRMSD = 3.70 Å

Multiple choice...

Modelling multibody (N>2) assemblies and dealing with large conformational changes

Protein-peptide docking

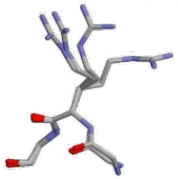
Protein-DNA modelling

Small molecule docking

HADDOCK's adventures in CAPRI

Small molecules docking with HADDOCK

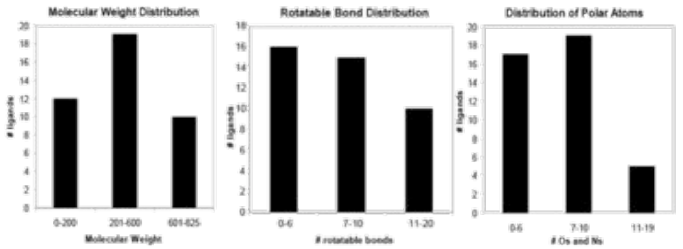
- Docking protocol issues:
  - Pre-sample ligand conformations
  - use ensemble for docking
  - same for protein



  - If flexibility is expected to play an important role (e.g. docking of an unstructured peptide onto a protein), perform a **fully flexible docking during the simulated annealing phase**

Validation of Haddock for small ligands

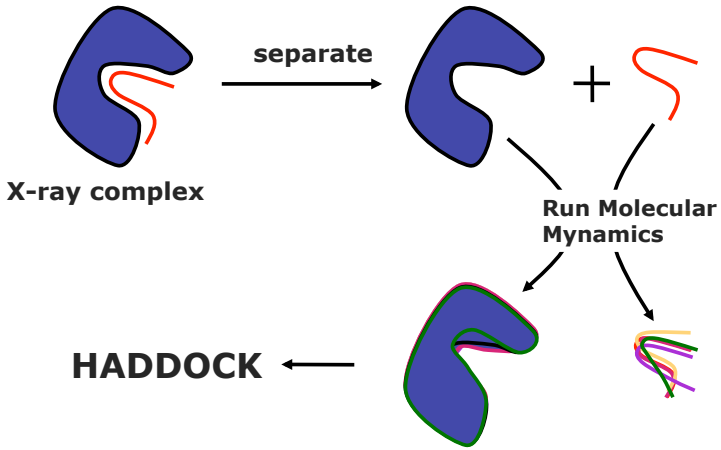
- **Benchmark of 41 complexes protein-ligand**
  - Diversity of ligand in term of: Size, Flexibility, Polarity, Charge, Position of the active site, *etc*



  - Good resolution of the complexes by X-ray (<2Å)
  - Already tested with other protein-ligand software (DOCK,FlexX and GOLD)

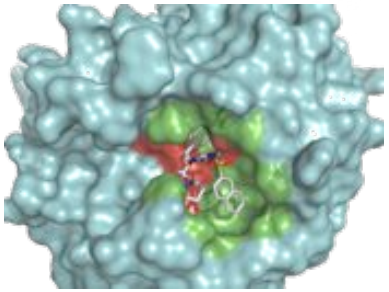
Erickson et al. *J.Med.Chem* (2004)

Input data:
Eliminating the bias of the bound form



Input data:
Simulated experimental NMR data (CSP)

- Simulation of NMR data
  - Residues of the protein found as important in literature are **actives**.
  - All residues of the protein close to 5Å of the ligand are **passives**.
  - Ligand is one residue and is active.



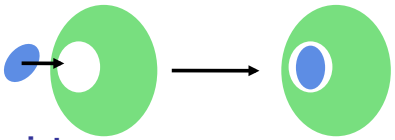
AIRs: The ligand has to be close (2 Å) to at least one passive residue and all active residues

Adapted protocol for ligand docking

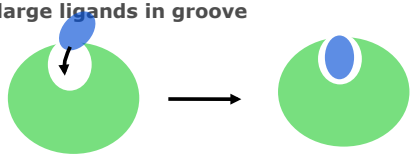
- **Ligand parameter and topologies from PRODGR** (van Aalten *et al.*)
- **Rigid body energy minimization**
  - Passives residues are "active" to attract the ligand in the binding pocket
- **Semi-flexible refinement (it1)**
  - No rigid-body simulating annealing (often lead to ejection from the active site
  - Ligand fully flexible
  - 500 steps: 500K to 50K (protein side-chains flexible)
  - 500 steps: 300K to 50K (protein side-chains+backbone flexible)

Additional changes

- **inter_rigid:**
  - Scaling down (1%) of intermolecular interaction during it0
  - Useful for buried ligand with protein in closed conformation

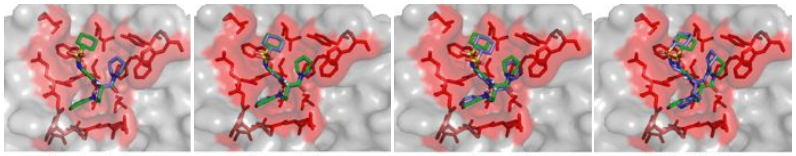


- **center_of_mass restraints:**
  - Force ligand to be / to stay 'inside' the protein (it0 & it1)
  - Useful for large ligands in groove

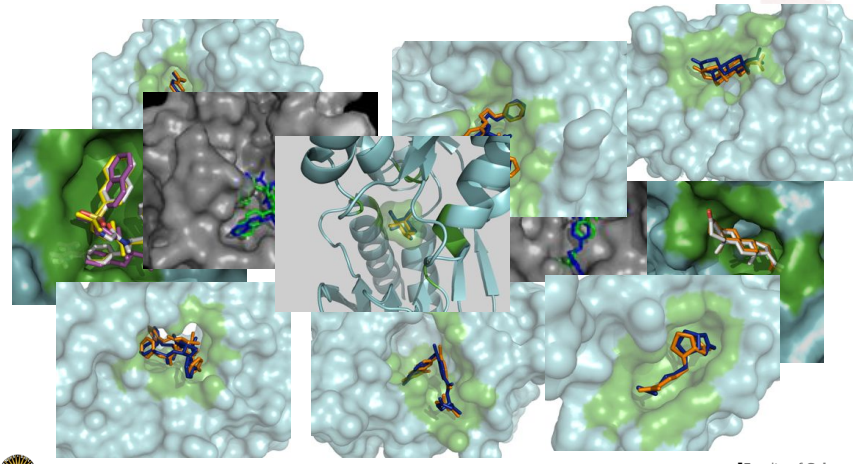


Analysis

- Fit on the reference protein complex
- Look at the ligand position
- Calculate RMSD on heavy atom of the ligand
- RMSD < 2Å considered as an acceptable solution

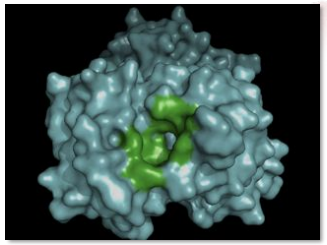


Let's have a look at some examples ...

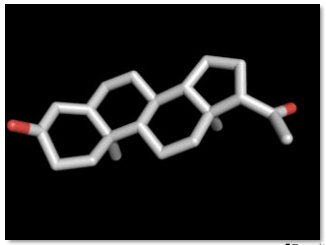


Example: 1DBB complex

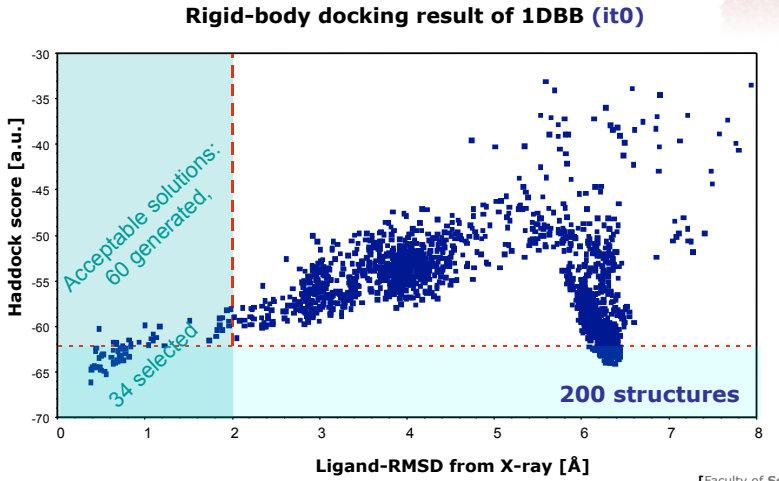
- **Protein:**
  - Antibody/Antigen complex
  - Active site not buried



- **Ligand:**
  - Medium size (306 KDa)
  - No flexibility
  - Hydrophobic
  - No charge

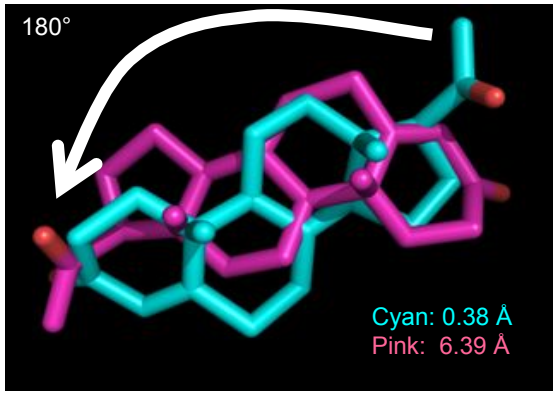


Example: 1DBB complex



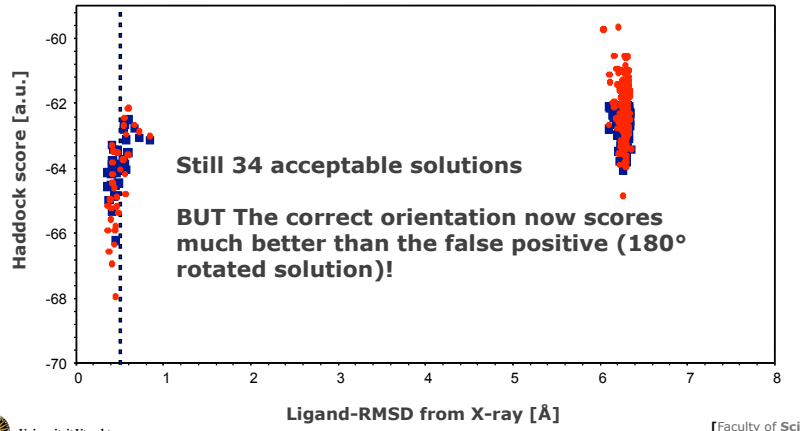
Example: 1DBB complex

Two families of structures selected corresponding to a 180° rotation



Example: 1DBB complex

Semi-flexible docking results (it0 + it1)

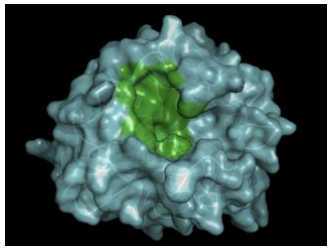


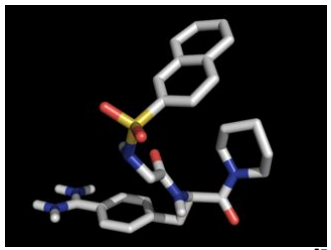
Still 34 acceptable solutions

BUT The correct orientation now scores much better than the false positive (180° rotated solution)!

Example: 1ETS complex

- **Protein:**
  - Serine protease
  - Active site not buried
- **Ligand:**
  - Medium size (523 kDa)
  - High flexibility
  - Medium polarity
  - Positive charge



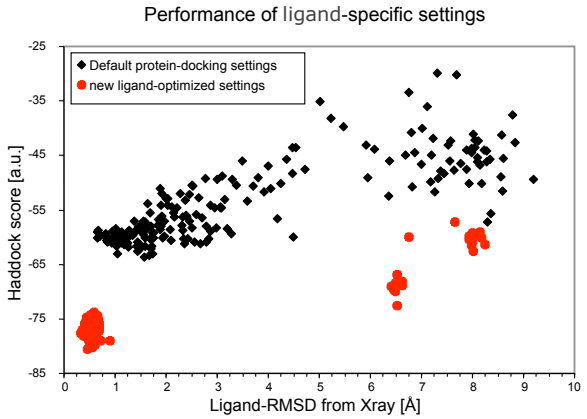


Example: 1ETS complex

- **Improvement in term of:**

Quality of the solution:
lower Rmsd with ligand-specific settings

Quantity of good solutions:
90% of the solutions have RMSD < 1Å



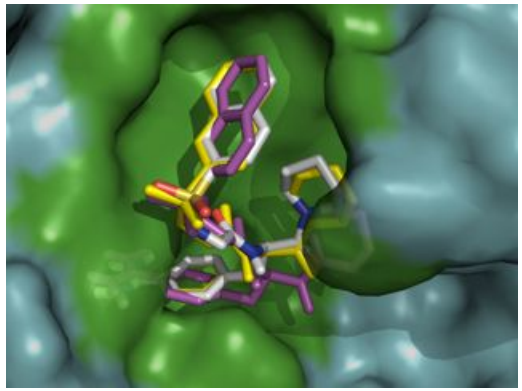
Example: 1ETS complex

- **Improvement in term of:**

Quality of the solution:
lower Rmsd with ligand-specific settings

Quantity of good solutions:
90% of the solutions have RMSD < 1Å

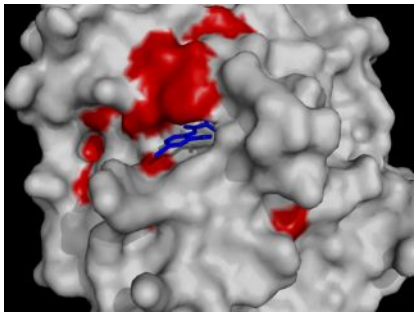
Old settings: 1.5Å
New setting: 0.5Å



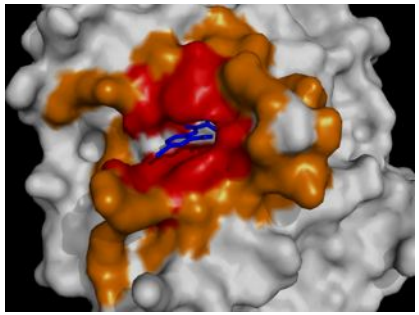
HADDOCKing with real NMR data

- **Test on PTP1B**
  - **With a apo form of the protein (1PTY)**

AIRs from Schwalbe's group (Frankfurt)



Simulated AIRs (5Å or 10Å)

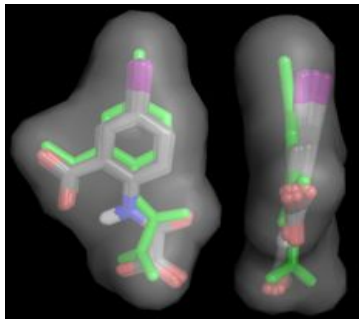


Schieborr *et al. ChemBioChem* (2005)

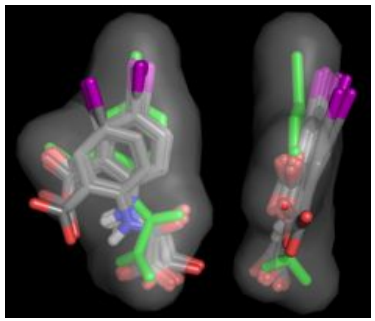
HADDOCKing with real NMR data

- **Comparison with reference structure (1ECV)**
  - **RMSD on ligand (10 best ranked structures):**

0.8 ± 0.1 Å (NMR CSP data)



1.0 ± 0.4 Å (10Å simulated)



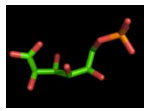
0.9 ± 0.2 Å (5Å simulated)

Summary of the docking result : 1 x 1 structure (bound form)

- **For the 41 complexes:**

For 83 % of the complexes, HADDOCK finds an acceptable solution:
- **44 % are perfect solutions (<0.5Å)**
- **29 % are nice solutions (<1.0Å)**
- **10 % are good solutions (<2.0Å)**

17 % of bad solutions:
- **4 large and flexible ligands**
- **2 ligands with symmetry**
- **1 small ligand with co-factors**



Summary of the docking result:
10 x 10 structures (dynamics forms)

- For the 41 complexes:

For 78 % of the complexes, HADDOCK finds an acceptable solution:

- 49 % are perfect solution (<0.5Å)
- 22 % are nice solutions (<1.0Å)
- 7 % are good solutions (<2.0Å)

22 % of bad solutions:

- 7 large and flexible ligands
- 1 ligand with symmetry
- 1 not selected (bad scoring)

Summary of the scoring results
1x1 structure (bound docking)

- For the 34 complexes (83% of the docking)

For 91 % of the complexes, HADDOCK scores an acceptable solution at the first position:

- 30 % are perfect solution (<0.5Å)
- 35 % are nice solutions (<1.0Å)
- 26 % are good solutions (<2.0Å)

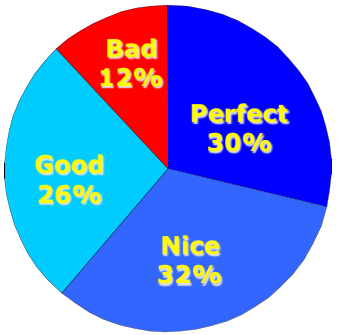
9 % of bad solutions at top rank (3 solutions):

- all small ligands
- 2 are monosaccharide
- 1 "linear ligand" with negative charge at each side

In the top 10 solutions:

- 100 % acceptable: 32% perfect + 41% good + 26% acceptable

Scoring on a cluster basis



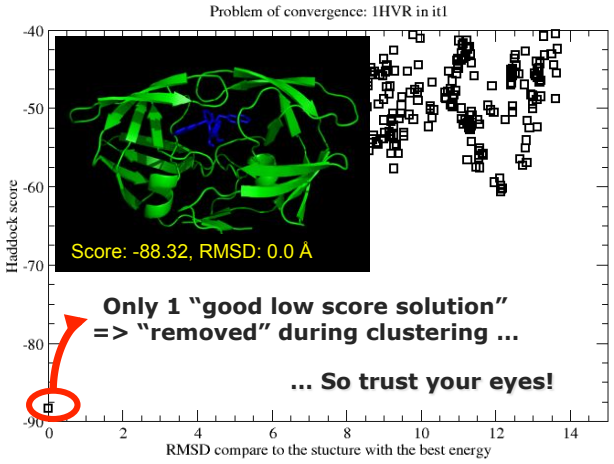
For 88 % of the complexes, HADDOCK ranks an acceptable cluster at the first position.

- For most of the solutions, clustering improve the quality of the first solution.

BUT

- if there is no or poor convergence, clustering fails...

Scoring on a cluster base...



Scoring remains a challenge!

Multiple choice...

Modelling multibody (N>2) assemblies and dealing with large conformational changes

Protein-peptide docking

Protein-DNA modelling

Small molecule docking

HADDOCK's adventures in CAPRI

HADDOCK's adventures in CAPRI

"Critical assessment of predicted interactions"

http://capri.ebi.ac.uk

- CAPRI is a blind test for protein-protein docking
- Usually 3 weeks for a predictions, 10 models can be submitted
- We participated to rounds 4 to 27
- For HADDOCK, we derived information to define AIRs from literature and bioinformatic predictions

Van Dijk et al. *Proteins* 2005; de Vries *et al. Proteins* 2007,2010

Performance of the HADDOCK team in CAPRI rounds 13-19

- 29 [1, 1, 2, 1, 1, 1, 0, 0, 0, 0] BU
- 30 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UU
- 32 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UU
- 33 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UH
- 34 [2, 2, 1, 2, 1, 1, 0, 0, 0, 0] UB
- 35 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] HH
- 36 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] BH

} Two-domain protein – crystal structure incompatible with covalently linked domains!!!

- 37 [0, 0, 2, 2, 0, 0, 0, 0, 0, 0] UH (2 *** uploaded)
- 38 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UH
- 39 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UB
- 40 [3, 3, 3, 3, 3, 3, 3, 3, 3, 3] UB
- 41 [1, 1, 2, 2, 1, 1, 1, 1, 1, 1] UH
- 42 [0, 0, 0, 0, 0, 0, 0, 0, 0, 1] HH(H)

1 ***, 4 **, 1 *, 12 stars

Performance of the HADDOCK server in CAPRI rounds 15-19

- 32 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UU
- 33 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UH
- 34 [1, 1, 1, 1, 1, 1, 0, 0, 0, 1] UB
- 35 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] HH
- 36 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] BH

} Two-domain protein – crystal structure incompatible with covalently linked domains!!!

- 37 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UH
- 38 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UH
- 39 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0] UB
- 40 [0, 0, 3, 0, 0, 0, 0, 0, 0, 0] UB
- 41 [1, 1, 2, 1, 0, 0, 0, 0, 0, 0] UH
- 42 [0, 0, 0, 0, 0, 0, 0, 0, 1, 0] HH(H)

1 ***, 1 **, 2 *, 7 stars

HADDOCK's performance in CAPRI

- **Overall performance**:
  - **3***, 9**, 3*** 15 out of 25 (60%)
- **Unbound only performance**:
  - **6**, 2*** 8 out of 13 (62%)
- As good as it gets… (among the top performing methods)
- "wrong" solutions still often have correctly predicted interfaces, but wrong orientations of the components
- ==> still useful to direct the experimental work

Van Dijk et al. *Proteins* 2005; de Vries *et al. Proteins* 2007,2010

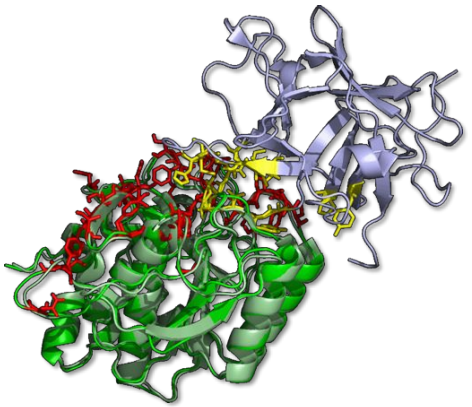
Post-docking interface prediction

Target	Fraction true interface coverage		Fraction overprediciton	
	ligand	receptor	ligand	receptor
T29	0.92	0.88	0.11	0.20
T30	0.84	0.73	0.26	0.39
T32	0.87	0.75	0.25	0.31
T33	0.61	0.42	0.20	0.50
T34	0.61	0.87	0.17	0.10
T37	0.36	0.89	0.66	0.27
T40	0.90	0.96	0.05	0.03
T41	0.89	0.83	0.04	0.15
T42	0.87	0.87	0.14	0.14

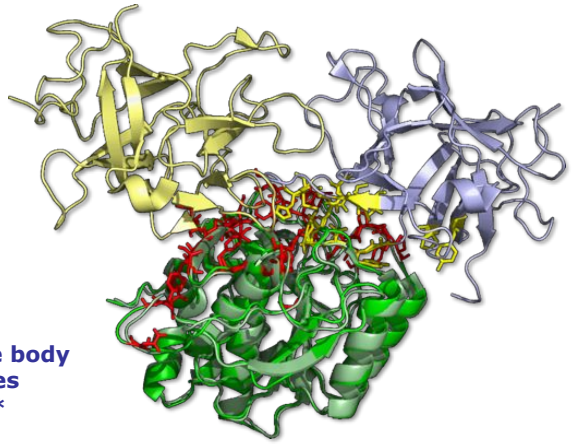
HADDOCK's weakness (one of them)

Information-driven…

Our T32 failure… (the "easy" one)



Our T32 failure… (the "easy" one)

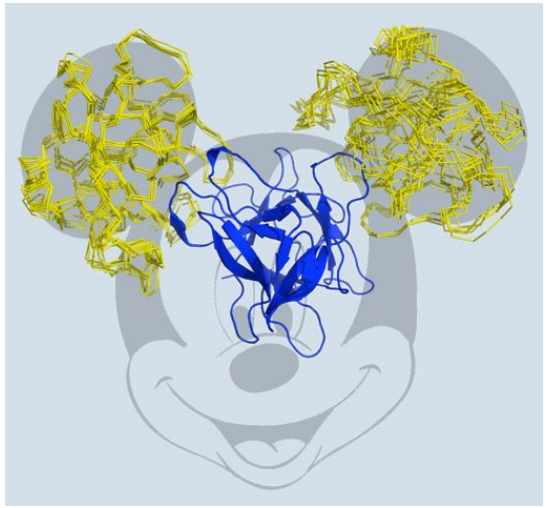


Note: Three body docking does generate ** solutions…

HADDOCK's strength (one of them)

Information-driven…

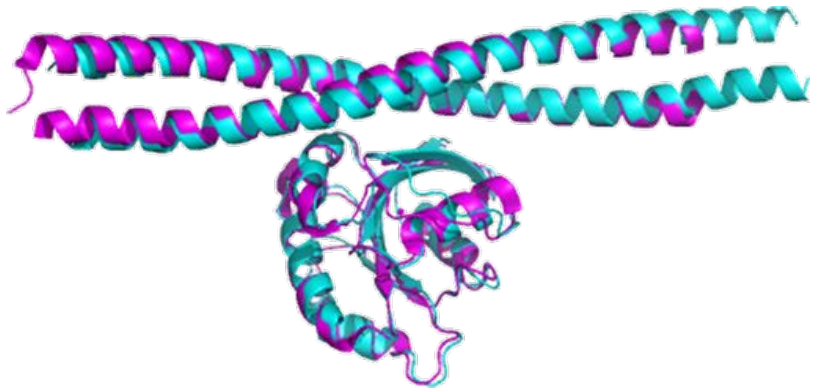
T40



10x ***

T37

** submitted, *** uploaded



Performance of the HADDOCK team in CAPRI rounds 20-27

- T46 [0,1,0,1,0,1,0,0,0,0,0] HU
- T47 [3,2,2,2,2,2,3,2,2,3,3] UU
- T48 [0,0,0,0,0,0,0,0,1,1,0] UU
- T49 [0,0,0,0,0,1,0,0,0,0,0] UU
- T50 [0,0,0,0,0,0,0,2,2,0,0] HU
- T51 [0,0,0,0,0,0,0,1,1,0,0] UU
- T53 [2,2,2,0,0,0,0,0,0,0,1] UU
- T54 [0,0,0,0,0,0,0,0,0,0,0] UU
- T58 [0,0,0,0,0,0,0,0,1,0,0] UU

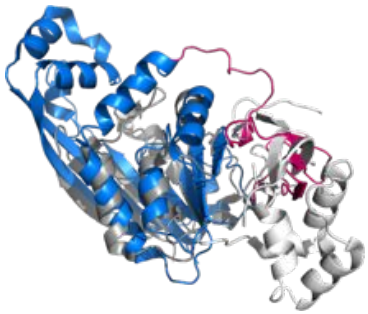
- **Success rate: 8 out of 9 = 88,9% acceptable or better!**

Some examples...

T46: Interface in eukaryotes, core in prokaryotes...

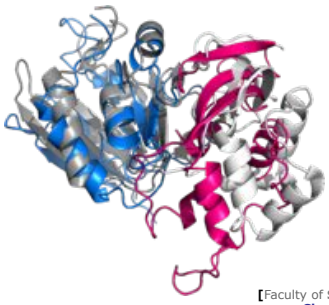
- T46 [0,1,0,1,0,1,0,0,0,0] HH
- Server predicted a single cluster (10/10 *)

chain	SeqID	RMSD
receptor	19%	2.5Å
ligand	12%	4.2Å



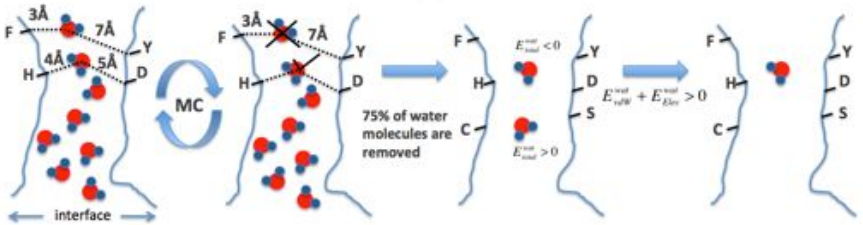
Cα-Cα distance restraints between Rossman-fold methyltransferase and Nter zinc finger domains

iRMSD best HADDOCK model = 3.4Å



T47: predicting interfacial waters A success

- T47 [3,2,2,2,2,3,2,2,3,3] UU / water prediction: [0***/6**/4*]
- Solvated docking based on distance restraints derived from known complex
- Manual submission: **Top 25% of hydration prediction**
- Server submission (only server): **Top 50% of hydration prediction**

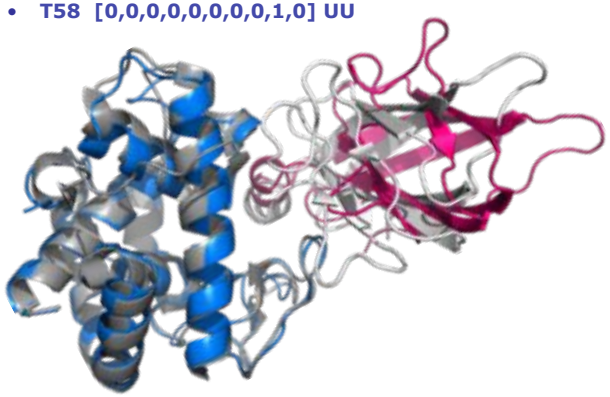


- Server successful and similar hydration prediction [0***/4**/6*]

van Dijk & Bonvin Bioinformatics, 2006
Kastritis et al. Proteins, 2013

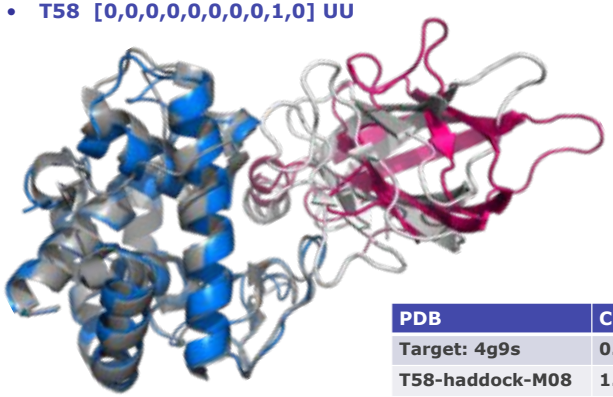
T58: the one with SAXS data

- T58 [0,0,0,0,0,0,0,0,1,0] UU



T58: Not SAXS – solvated docking!

- T58 [0,0,0,0,0,0,0,0,1,0] UU



PDB	Chi	iRMSD
Target: 4g9s	0.72	0.0
T58-haddock-M08	1.21	2.6
Cluster7_1	0.77	13.3

Beware of SAXS fitting...!

Multiple choice...

Modelling multibody (N>2) assemblies and dealing with large conformational changes

Protein-peptide docking

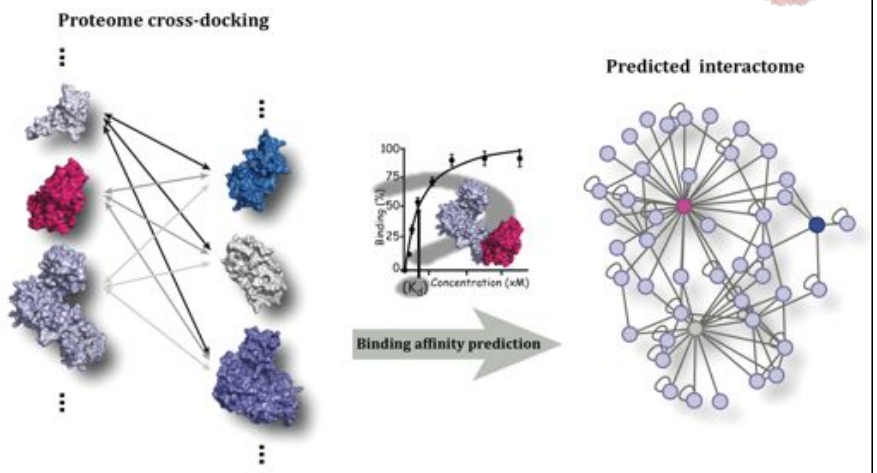
Protein-DNA modelling

Small molecule docking

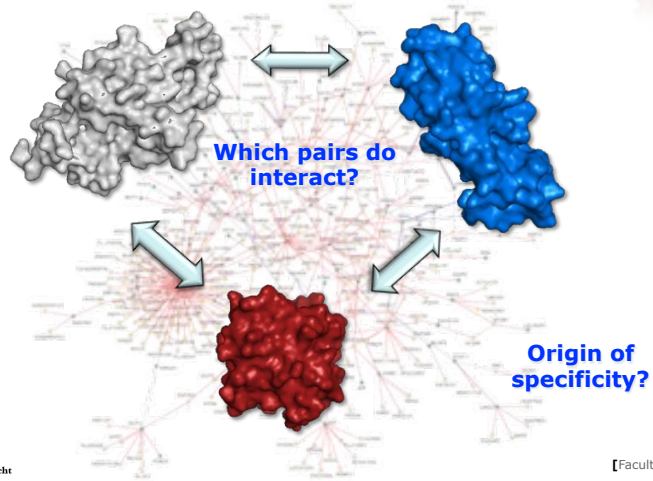
HADDOCK's adventures in CAPRI

Overview

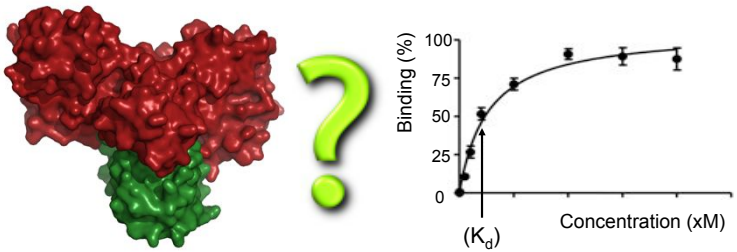
Predicting interactomes by docking… a dream?



Predicting interactomes by docking… a dream?



How good are our scoring functions in predicting binding affinities?



A binding affinity benchmark

Complexes from the Docking Benchmark 3.0 (Hwang et al. Proteins 2008)
From literature search: Kd, pH, T, method, ref.
Classification according to binding affinity (high/medium/low)

124 complexes → **HADDOCK** Software web portal (http://haddock.chem.uu.nl/) → **HADDOCK score**

248 optimized complexes
- Missing side-chains built
- Two sets:
  - Short EM only
  - HADDOCK's water refinement

Kastritis & Bonvin, *J. Proteome Res.* 2010

Further scoring of the complexes…

EMBL-EBI **PISA SERVER** http://www.ebi.ac.uk/msd-srv/prot_int/	→	ΔiG, ΔG_{int}, $T\Delta S_{diss}$, ΔG_{diss}
FireDock http://bioinfo3d.cs.tau.ac.il/FireDock/	→	FIREDOCK score
FastContact 2.0 http://structure.pitt.edu/servers/fastcontact/	→	Electrostatic & Desolvation Free Energy
Score by DFIRE Method http://sparks.informatics.iupui.edu/hzhou/dfire.html	→	DFIRE score
ZRANK	→	ZRANK score
Rosetta	→	Binding energy + I_int

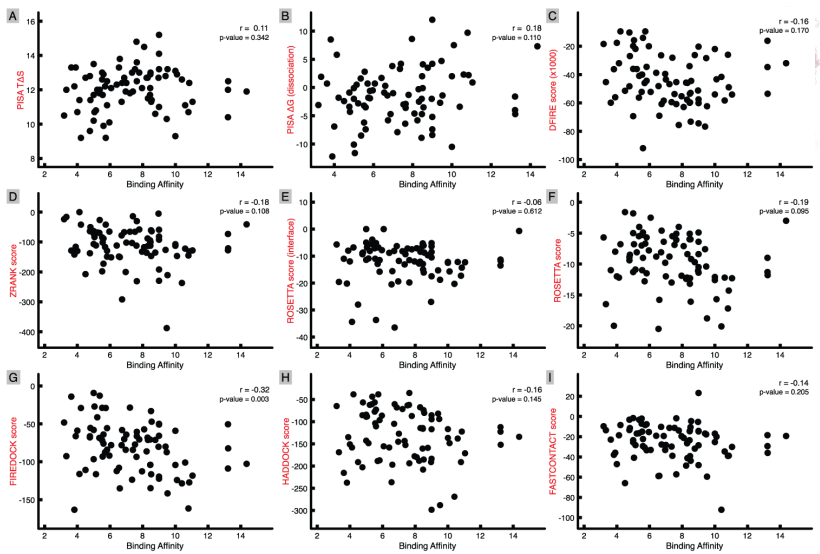
+ all components of the various scores

Is there a linear correlation between binding affinity and scoring?

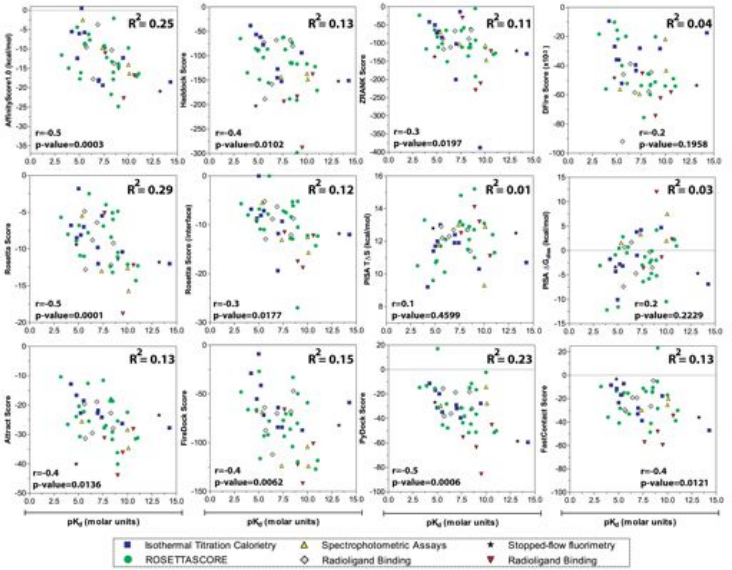
Binding affinity ⟺ $R^2 \sim= 0$

- FASTCONTACT score
- FIREDOCK score
- DFIRE score
- HADDOCK score
- ROSETTAscore
- ZRANKscore
- PISA SERVER score
- ATTRACT score
- PyDock score
- Affinity score

K_d value	Number of Complexes
High (>10^{-9} M)	22
Medium (10^{-6} < C < 10^{-9} M)	35
Low (>10^{-6} M)	26
Unknown	41

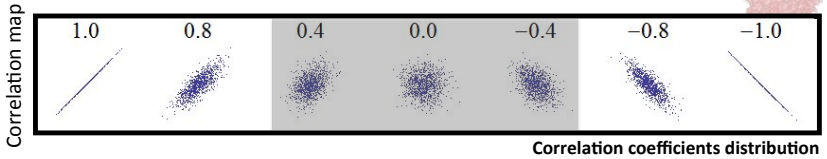


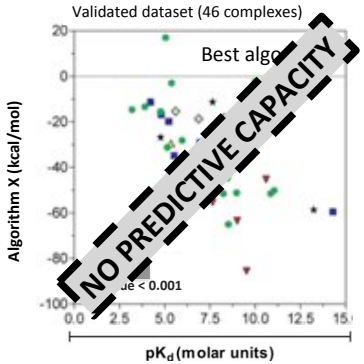
More is needed… Classifiers, neural networks, more indicators…



"Clean subset (48)" (collaborative effort with Janin/Bates/Weng)

Our results show…

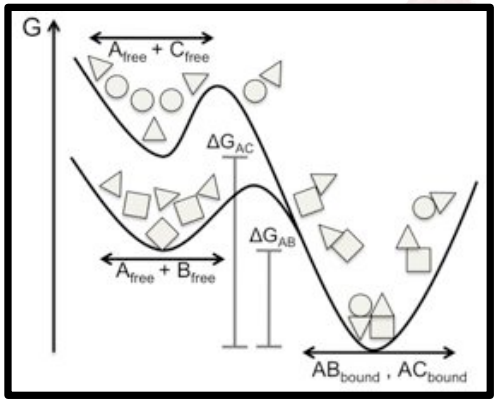




Kastritis and Bonvin, J Proteome Res 2010

Possible Reasons for limitations of current models

- quality of experimental data
- ambiguity of crystal coordinates
- conformational changes
- co-factors
- solvent effects
- free state neglected



Melquiond, Karaca, Kastritis and Bonvin, Comput Mol Sci 2011

A structure-based benchmark for protein-protein binding affinity

Class	Number (All)	Non-cognate	High (Kd < 10^{-10} M)	Medium (K_d 10^{-6} to 10^{-10})	Low (Kd > 10^{-6} M)	ΔG (kcal.mol^{-1}) Mean	S.D.	Conformational changes (iRMSd≥1.5Å)
Antigen-antibody	19	2	2	16	1	12.2	1.3	1
Enzyme/ inhibitor	40	4	17	22	1	13.8	2.3	5
Other enzyme complexes	21	1	0	12	9	9.2	1.9	7
G-proteins	17	-	1	6	10	8.9	2.5	6
Receptors	13	-	1	11	1	11.5	2.1	4
Miscellaneous	34	2	0	22	12	9.3	2.2	10
All	144	9	20	90	34	11.0	2.9	33

Kastritis, P.L. Moal I.H., Hwang H., Weng, Z., Bates P.A., A.M.J.J. Bonvin, Janin J. Protein Science 2011

Overview

Conclusions & Perspectives

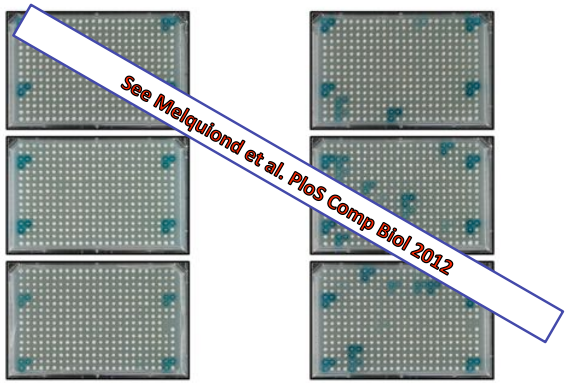
- HADDOCK is highly versatile and can deal with a variety of systems
  - Protein-protein
  - Protein-nucleic acids
  - Protein-small ligand
  - Multi-body assemblies
- Data-driven docking is useful to generate models of biomolecular complexes, even when little information is available
- **While models from docking may not be fully accurate, they provide working hypothesis and can still be sufficient to explain and drive the molecular biology behind the system under study**

Conclusions & Perspectives

- Data-driven docking is complementary to classical structural methods
- Many challenges however remain:
  - Scoring
  - Predicting and dealing with conformational changes
  - Predicting binding affinities
  - …
- **… and, we still don't understand many aspects of biomolecular recognition…**

The butterfly effect in protein-protein interactions

ProteinX ProteinX D120E



ProteinY-like (98% identity)
Y2H interaction profile

HADDOCK Inc.

Gydo van Zundert, Charleen Don, Adrien Melquiond, Ezgi Karaca, Marc van Dijk, Joao Rodrigues, Mikael Trellet, …, Koen Visscher, Manisha Anandbahadoer, Christophe Schmitz, Panos Kastritis, Jeff Grinstead