Lesson 3
Energy-based docking and scoring
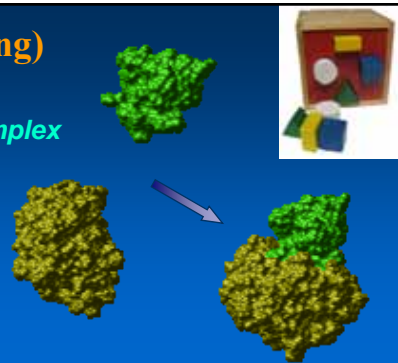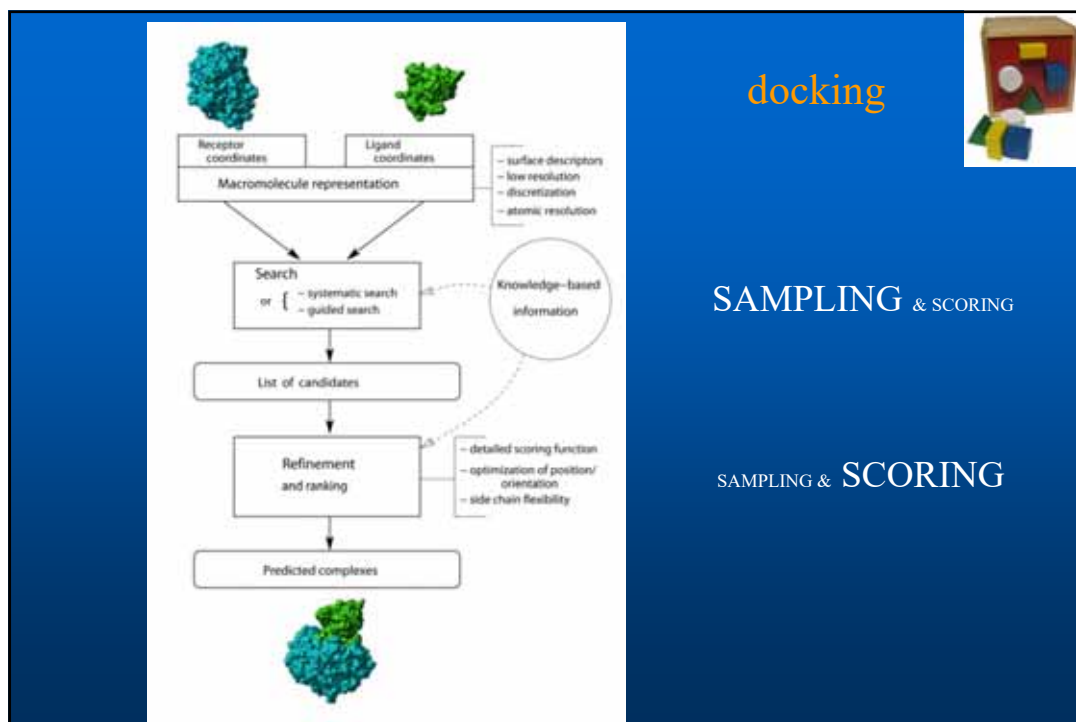
Juan Fernández-Recio
CSIC, BSC

juanf@bsc.es

Complex structure prediction (docking)

Generation of the structure of a protein-protein complex
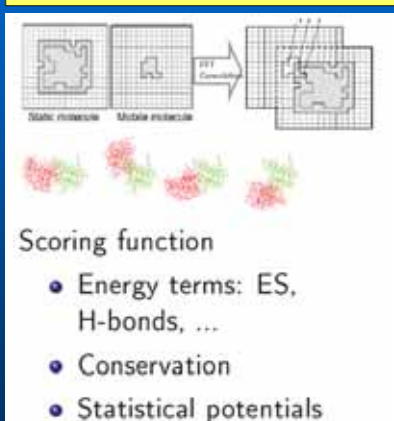*from the individual protein structures*

1

docking

SAMPLING & SCORING

SAMPLING & SCORING

Protein-protein docking methods

**Exhaustive search
(FFT, surface-based)**

Scoring function
- Energy terms: ES, H-bonds, ...
- Conservation
- Statistical potentials

Geometry-based docking

**Stochastic sampling
(Monte-Carlo, minimization)**

Energy-based docking
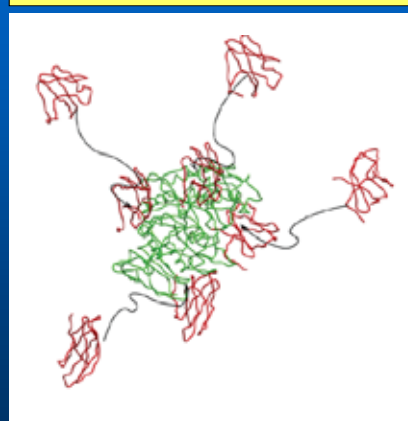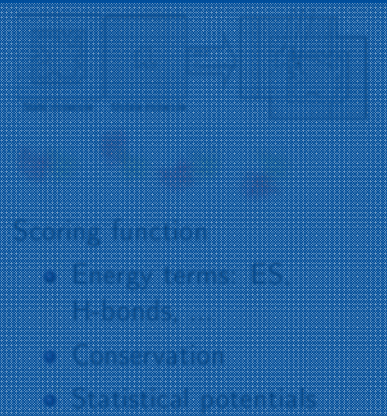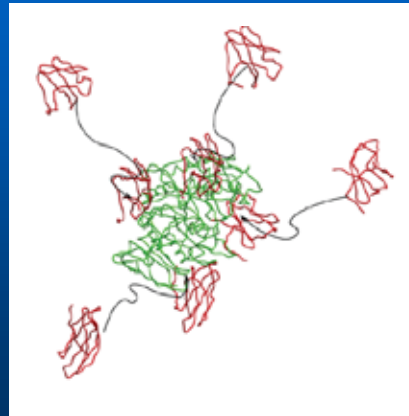
Protein-protein docking methods

Exhaustive search
(FFT, surface-based)

Stochastic sampling
(Monte-Carlo, minimization)

Scoring function
- Energy terms: ES, H-bonds,
- Conservation
- Statistical potentials

Geometry-based docking

Energy-based docking

Docking Search Strategies

- **Pseudo Random**
  - Simulated Annealing / Monte Carlo
  - Genetic Algorithms

- **Directed Search**
  - Geometric Hashing
  - Spherical Harmonic Surface Triangles

- **Brute-Force Search**
  - Explicit Grid Correlations
  - Fast Fourier Transform (FFT) Correlations
  - Spherical Polar Fourier Correlations

- **Refinement Phase**
  - Classical or Soft Potentials (+/- Electrostatics)
  - Desolvation, Solvent Dipoles...
  - Visual Inspection!!
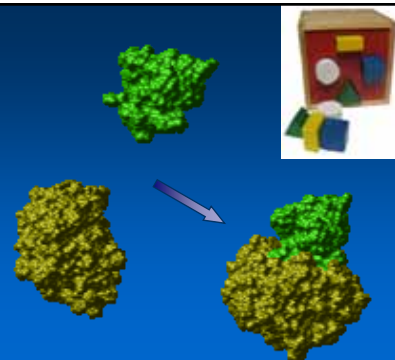
Criteria for Good Docking Orientations

- Low Free Energy (Difficult!)
- Low Pseudo-Energy (Easy) Based On...
- Large Surface Burial: $\sim 1600 \pm 400\text{Å}^2$
- Small van der Waals Overlaps
- No Large Cavities in Interface
- Good H-Bonding: ~ 1 HB/100 Å^2
- Good Charge Complementarity
- Polar/Polar Contacts Favoured
- Polar/Non-Polar Contacts Disfavoured
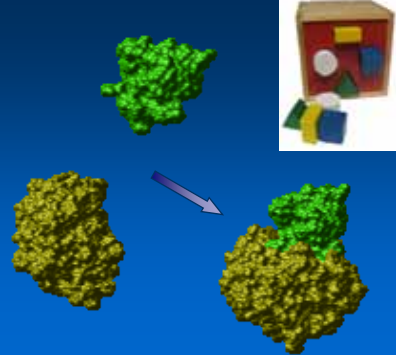
- **Energy-based docking**

- Docking and restraints

- Scoring by energy

- Scoring by statistical energy

- Scoring and restraints

- Applications of energy-based docking
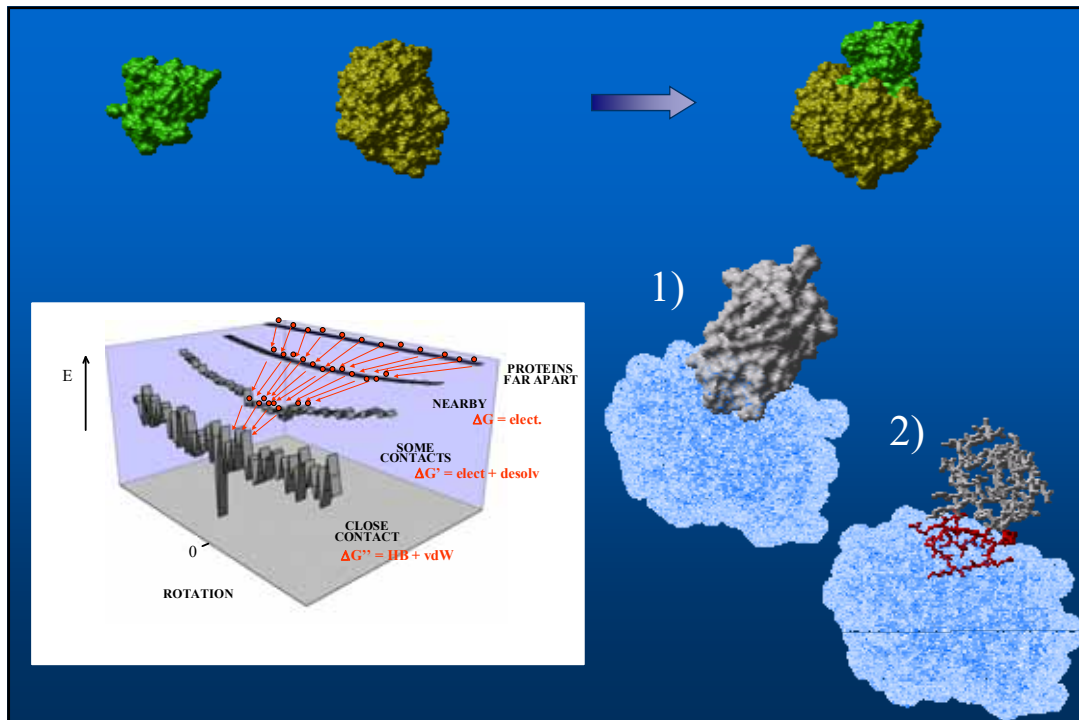
Protein-Protein Docking Energy

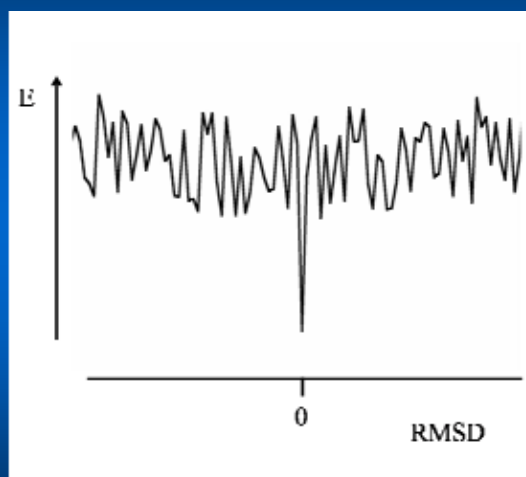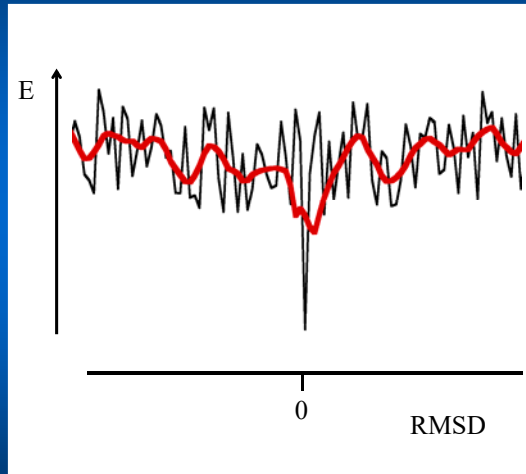# Protein-Protein Docking Energy

E

0   RMSD

Protein-Protein Docking Energy

$$E = E_{vw} + E_{el} + E_{hb} + E_{hp}$$

vdW Energy (kcal/mol)

E_{vw}^0

$$E_{vw}^0 = -\frac{A_{ij}}{R_{ij}^6} + \frac{B_{ij}}{R_{ij}^{12}} \qquad \text{(ECEPP/3)}$$

'soft' vdW $= \begin{cases} E_{vw}^0 & if\ E_{vw}^0 \le 0 \\ \dfrac{E_{vw}^0 E_{max}}{E_{vw}^0 + E_{max}} & if\ E_{vw}^0 > 0 \end{cases}$

'soft' vdW

Distance (Å)
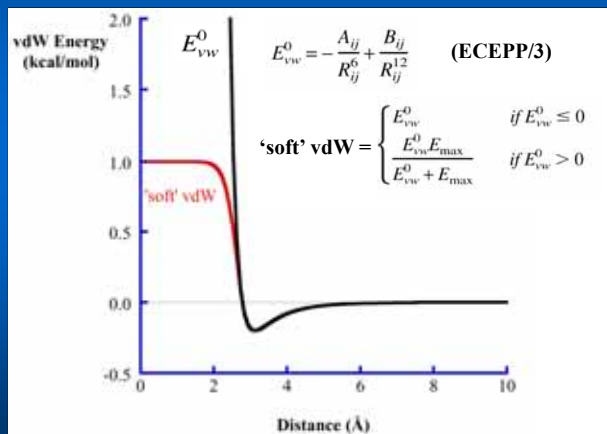
$$E_{el} = 332.0\frac{q_i^s q_j}{4d_{ij}^2}$$

Max E_{el} = 20 kcal/mole
Min E_{el} = -20 kcal/mole

$$E_{hb} = E_{hb}^0 e^{-[(r-r_{ep})^2/d_{hb}^2]}$$

$E^O_{hb} = 2.5$ kcal/mole
$d_{hb} = 1.4$ Å

E_{hp} = 0.03 kcal/mole * ASA(apolar)

ICM-DISCO protein docking

Fernández-Recio et al. 2002 Protein Sci. 11, 280-291

Redocking complexed subunits

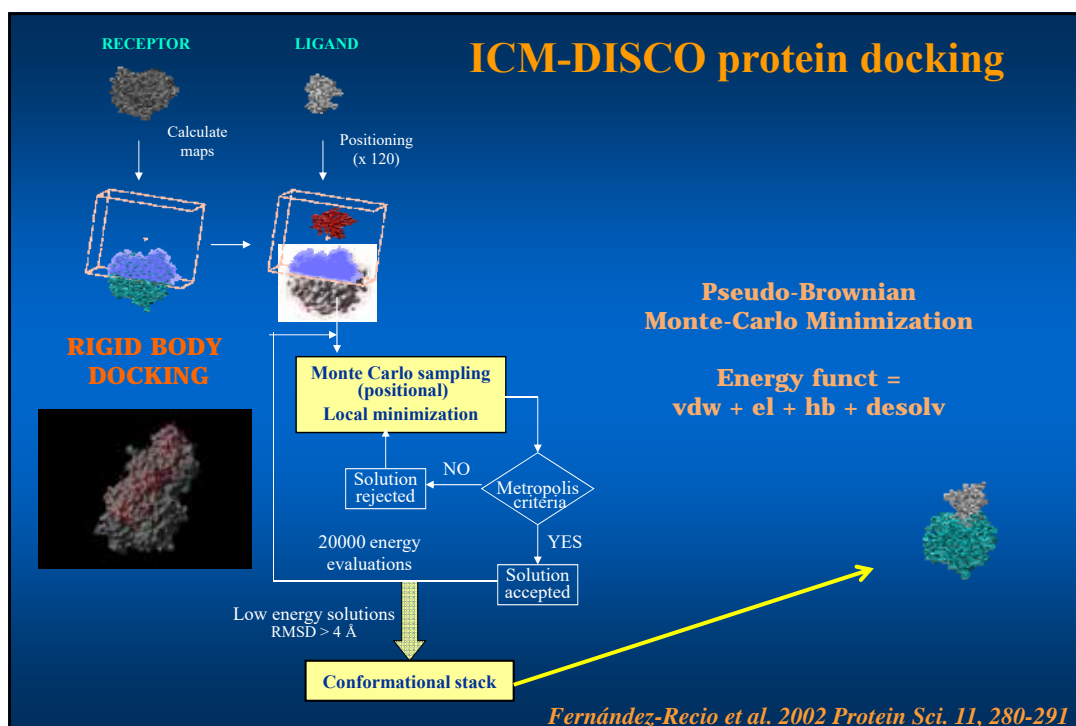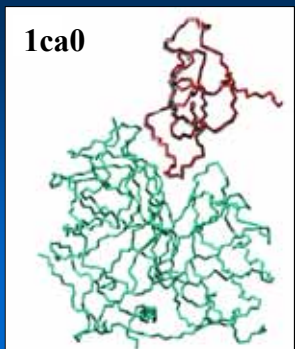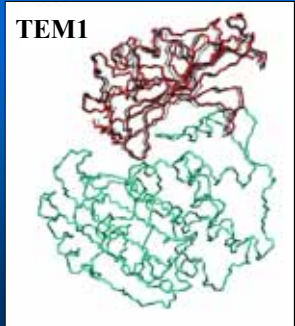| Complex PDB	Res (Å)	Receptor Name	Ligand Name	ICM Docking [a]
Protease-inhibitor				
1ca0	2.10	Chymotrypsin	APPI	**1 of 30** (0.4)
1cbw	2.60	Chymotrypsin	BPTI	**1 of 30** (0.4)
1acb	2.00	Chymotrypsin	Eglin C	**1 of 30** (0.5)
1cho	1.80	Chymotrypsin	Ovomucoid	**1 of 30** (0.3)
1cgi	2.30	Chymotrypsinogen	PTI	**1 of 30** (0.4)
2kai	2.50	Kallikrein A	BPTI	**1 of 30** (0.8)
2sni	2.10	Subtilisin BPN	CI-2	**1 of 30** (0.3)
2sic	1.80	Subtilisin BPN	SSI	**1 of 30** (0.4)
1cse	1.20	Subtilisin Carlsberg	Eglin C	**1 of 30** (0.3)
2tec	1.98	Thermitase	Eglin C	**1 of 30** (0.3)
1taw	1.80	Trypsin	APPI	**1 of 30** (0.7)
2ptc	1.90	Trypsin	BPTI	**1 of 30** (0.4)
3tgi	1.80	Trypsin	BPTI	**1 of 30** (0.3)
1brc	2.50	Trypsin D189G/G226D	APPI	**1 of 30** (0.7)
Enzyme-inhibitor				
1fss	3.00	Acetylcholinesterase	Fasciculin II	**1 of 30** (0.4)
1bvn	2.50	α-amylase	Tendamistat	**1 of 30** (0.4)
1bgs	2.60	Barnase	Barstar	**1 of 30** (0.6)
1ay7	1.70	Ribonuclease sa	Barstar	**1 of 30** (0.7)
TEM1[f]	1.70	TEM-1 β-lactamase	BLIP	**1 of 30** (1.3)
1ugh	1.90	UDG	UGI	**1 of 30** (0.4)
Electron transport				
2pcb	2.80	Cyt. c Peroxidase	Cytochrome c	4 of 30 (1.2)
2pcf	NMR	Cytochrome f	Plastocyanin	**1 of 30** (1.1)
Antibody-antigen				
1mlc	2.10	Fab D44.1	Lysozyme	**1 of 30** (0.4)
1vfb	1.80	Fv D1.3	Lysozyme	**1 of 30** (0.5)

1ca0

TEM1

Fernandez-Recio et al. 2002 Prot. Sci. 11, 280-291

Docking unbound trypsin/BPTI

Prot. Sci., 2002

Distribution of solutions

Real complex

Best docking solution

Energy (kcal / mol)

-40
-50
-60
-70
-80
-90
-100

Docking complexed subunits

Docking unbound

Correct solution

RosettaDock

Random Start Position

Low-Resolution Monte Carlo Search

High-Resolution Refinement

10^5

Clustering

Predictions
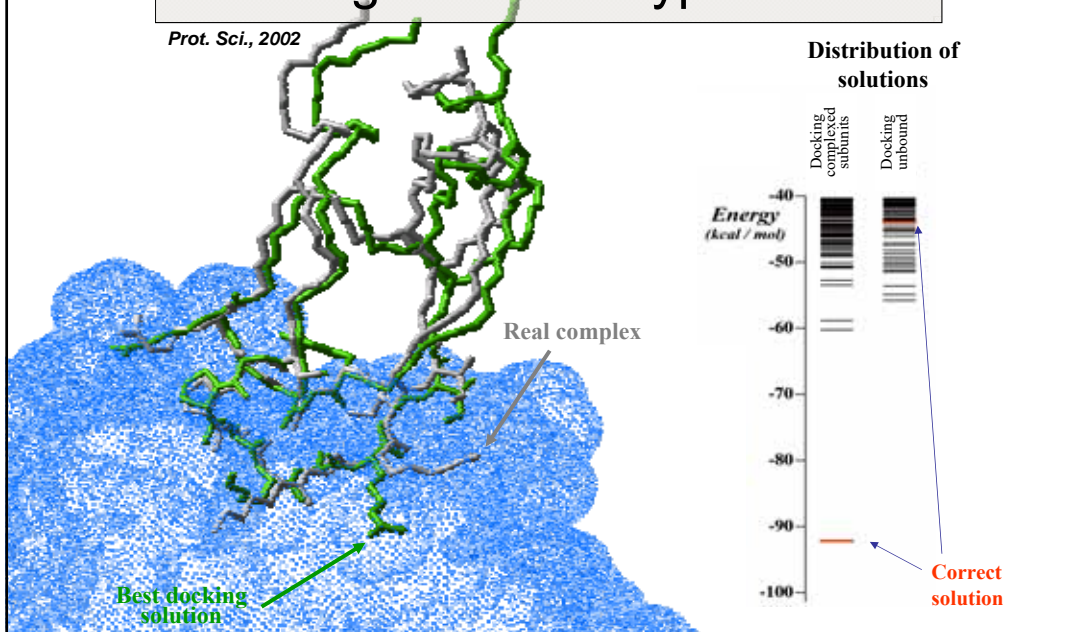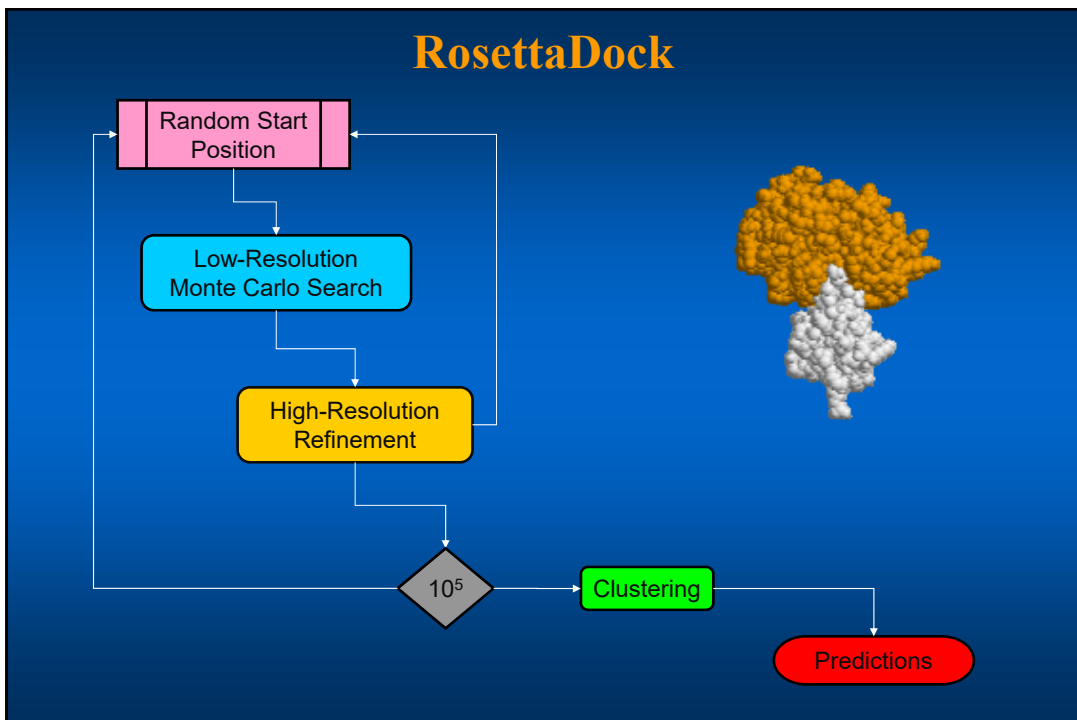
RosettaDock

RANDOM START POSITION

- Creation of a decoy begins with a random orientation of each partner and a translation of one partner along the line of protein centers to create a glancing contact between the proteins
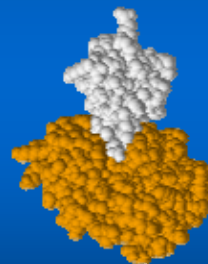

RosettaDock

Monte Carlo Search

Rigid body translations and rotations

Residue-scale interaction potentials

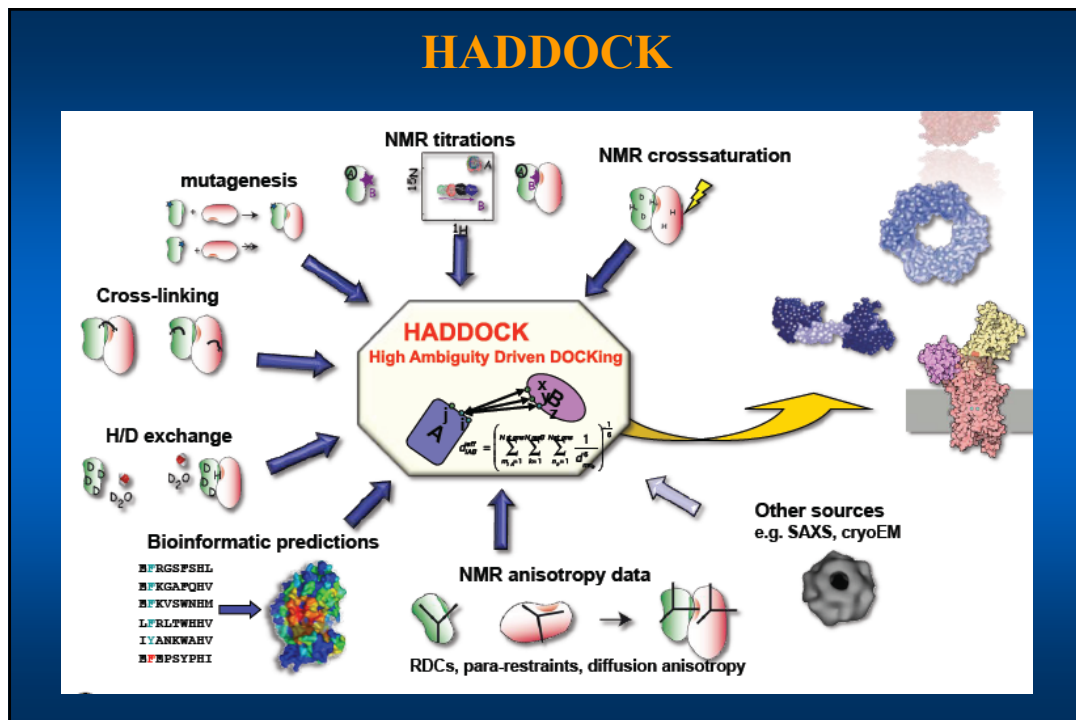Protein representation: backbone atoms + *average centroids*
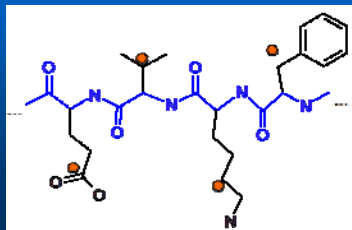
Low-Resolution Search

- Mimics physical diffusion process

- Energy-based docking

- **Docking and restraints**

- Scoring by energy

- Scoring by statistical energy

- Scoring and restraints

- Applications of energy-based docking

HADDOCK

HADDOCK

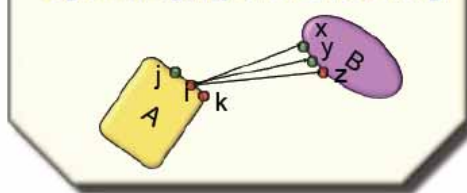List of interface residues
for protein A

List of interface residues
for protein B

(i,j,k)

(x,y,z)

HADDOCK

High Ambiguity Driven DOCKing

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

$$d_{iAB}^{eff} = \left(\sum_{m_{i A}=1}^{N\,at\,oms} \sum_{k=1}^{N\,resB} \sum_{n_k=1}^{Nat\,oms} \frac{1}{d_{mn_k}^6} \right)^{-\frac{1}{6}}$$

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

---

HADDOCK
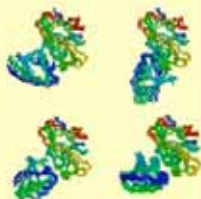
Separate proteins and apply random rotations

Rigid body energy minimization:
- first only rotations
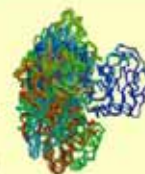- then rotations + translations

Final refinement in explicit water

Clustering and analysis Sorting of clusters according to the HADDOCK score
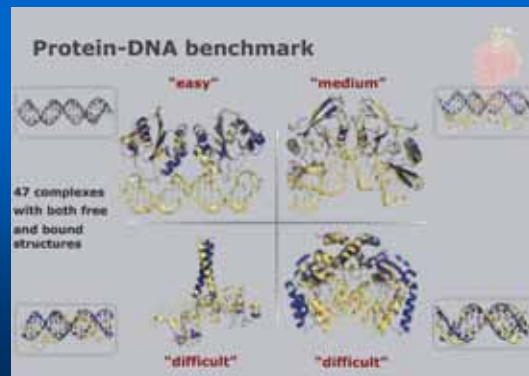
Semi-flexible simulated annealing in torsion angle space:
1) Rigid body search
2) SA with flexible side-chains at interface
3) SA with flexible side-chains and backbone at interface

HADDOCK
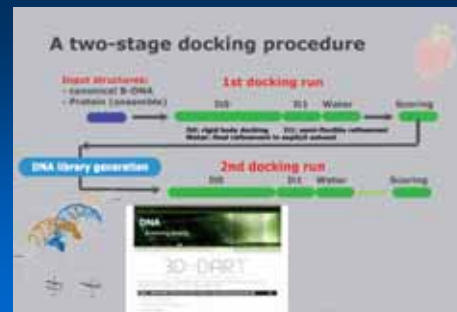
The HADDOCK web portal

A two-stage docking procedure

Protein-DNA benchmark

The HADDOCK PDB structure gallery

- **Energy-based docking**

- **Docking and restraints**

- **Scoring by energy**

- **Scoring by statistical energy**

- **Scoring and restraints**

- **Applications of energy-based docking**

docking

SAMPLING & SCORING

SAMPLING & SCORING

Optimized desolvation for rigid-body protein-protein docking

ICM Global Rigid-body Docking

$$E = 0.46 * (E_{\mathrm{Hvw(max1.0)}} + E_{\mathrm{Cvw(max1.0)}}) + E_{\mathrm{el}}^{\mathrm{solv}} + 1.17 E_{\mathrm{hb}} + 2.01 E_{\mathrm{hp}} + 0.09 SOLV^{\mathrm{w/v}}$$

chymotrypsin / APPI *cytochrome f / plastocyanin* *Fab / lysozyme* *EPOR / EPO*

Optimized desolvation for rigid-body protein-protein docking

ICM Global Rigid-body Docking

$$E = 0.46*(E_{\text{Hvw(max1.0)}} + E_{\text{Cvw(max1.0)}}) + E_{\text{el}}^{\text{solv}} + 1.17E_{\text{hb}} + 2.01E_{\text{hp}} + 0.09SOLV^{\text{w/v}}$$
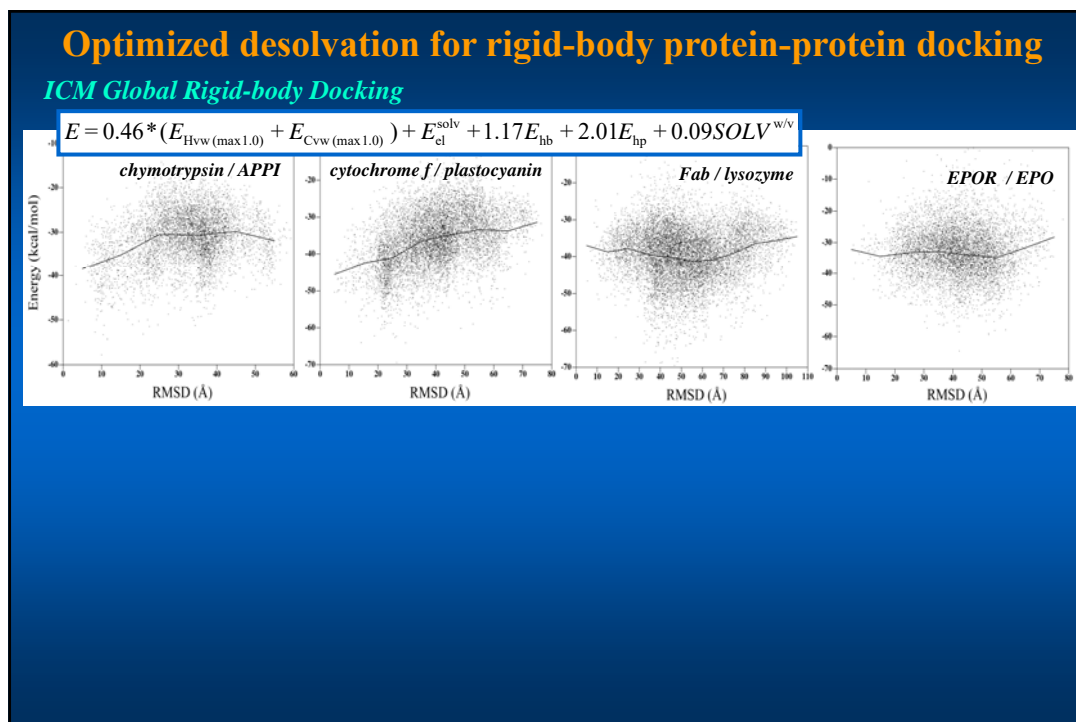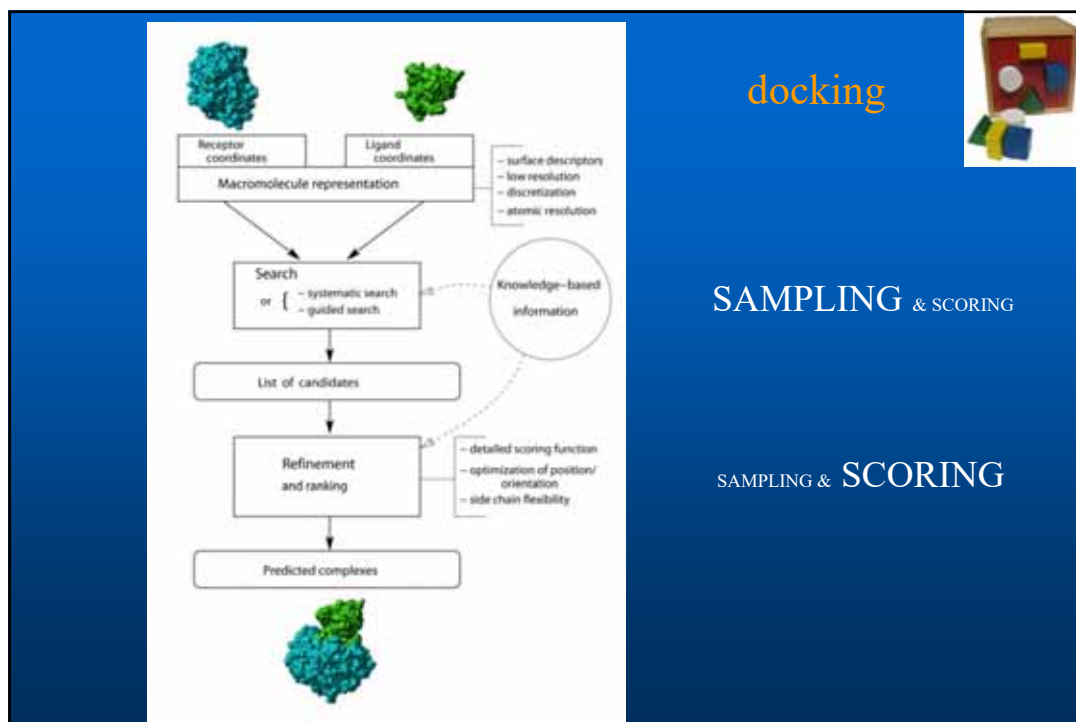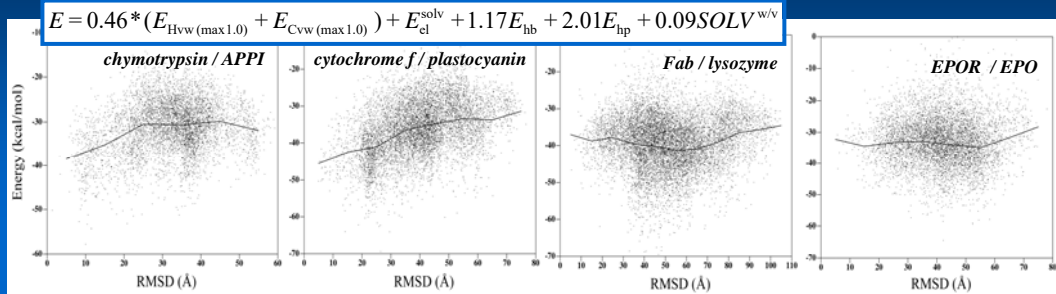
Optimized Desolvation Energy

$$\Delta G_{\text{sol}} = \sum_{k=1}^{N} \sigma_k A_k$$

ASP are optimized for protein docking

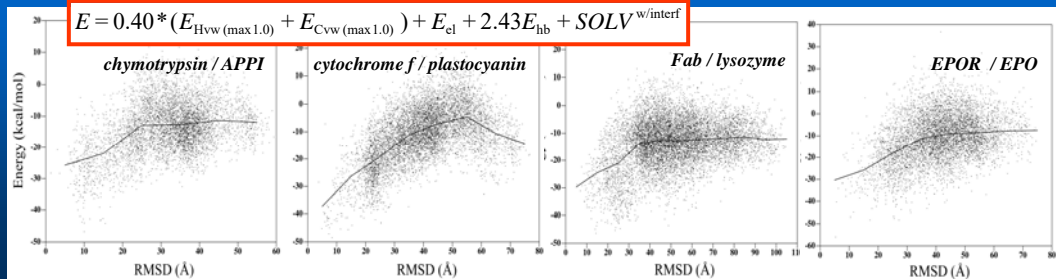| ASPs from vacuum/water	→	ASPs from octanol/water	→	ASPs for interface/water

---

Optimized desolvation for rigid-body protein-protein docking

ICM Global Rigid-body Docking

$$E = 0.46*(E_{\text{Hvw(max1.0)}} + E_{\text{Cvw(max1.0)}}) + E_{\text{el}}^{\text{solv}} + 1.17E_{\text{hb}} + 2.01E_{\text{hp}} + 0.09SOLV^{\text{w/v}}$$
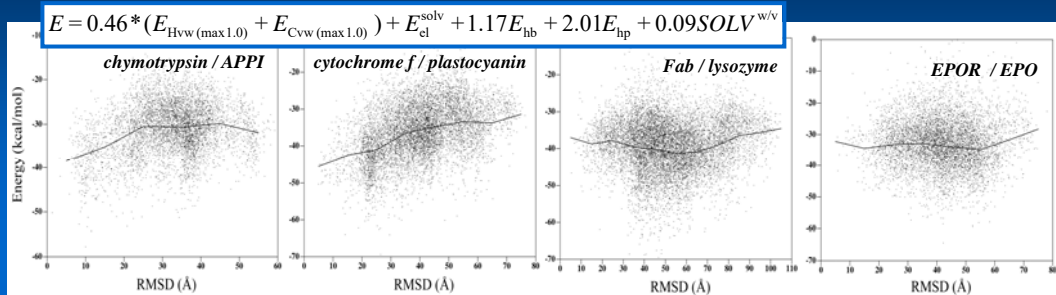
Optimized Desolvation Energy

$$E = 0.40*(E_{\text{Hvw(max1.0)}} + E_{\text{Cvw(max1.0)}}) + E_{\text{el}} + 2.43E_{\text{hb}} + SOLV^{\text{w/interf}}$$

pyDock: electrostatics + desolvation scoring
(evaluation on ICM-DISCO rigid-body docking sets)

ICM-DISCO's docking benchmark (45 unbound cases)
Fernandez-Recio et al. (2004) JMB 335, 843-865
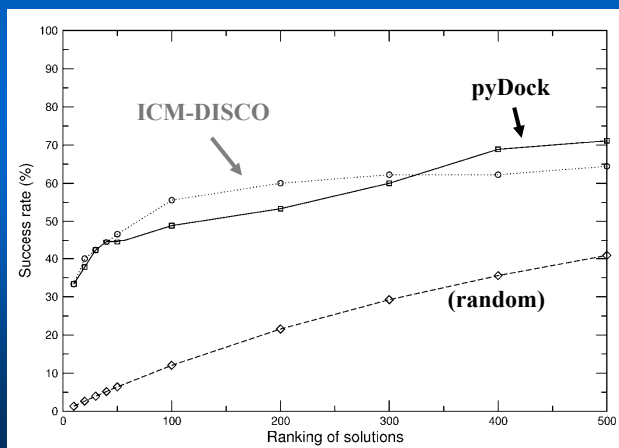
pyDock rescoring

$$E = E_{el} + E_{solv}$$

$$E_{el} = 332.0 \frac{q_i^s q_j}{4 d_{ij}^2}$$

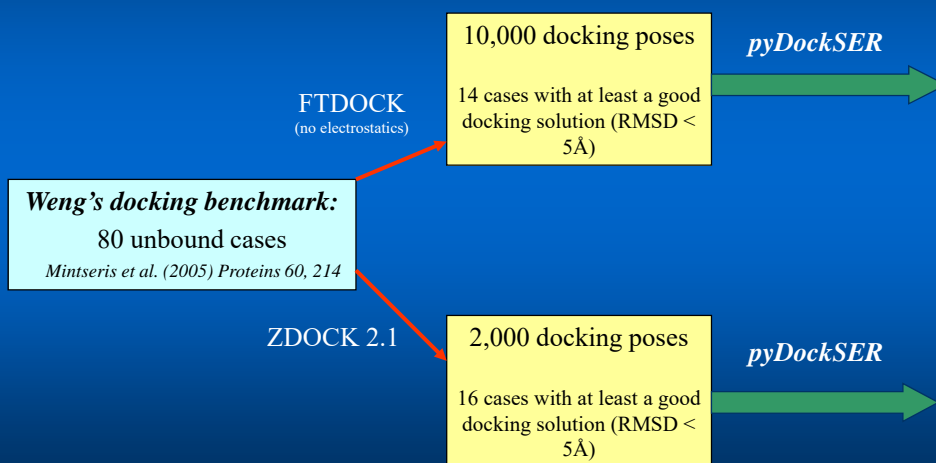$$\text{Max } E_{el} = +1 \text{ kcal/mole}$$
$$\text{Min } E_{el} = -1 \text{ kcal/mole}$$

$$\Delta G_{sol} = \sum_{k=1}^{N} \sigma_k A_k$$

water/interface

$(C_{ali}, C_{aro}, N, N^+_{lys}, N^+_{arg}, O_{-oh}, O_{-co}, O^-, S_{-sh}, S_{-s-})$

Fernandez-Recio et al., 2004

---

pyDock evaluation in other
rigid-body docking sets

FTDOCK
(no electrostatics)

10,000 docking poses

14 cases with at least a good docking solution (RMSD < 5Å)

pyDockSER

Weng's docking benchmark:
80 unbound cases
Mintseris et al. (2005) Proteins 60, 214

ZDOCK 2.1

2,000 docking poses

16 cases with at least a good docking solution (RMSD < 5Å)
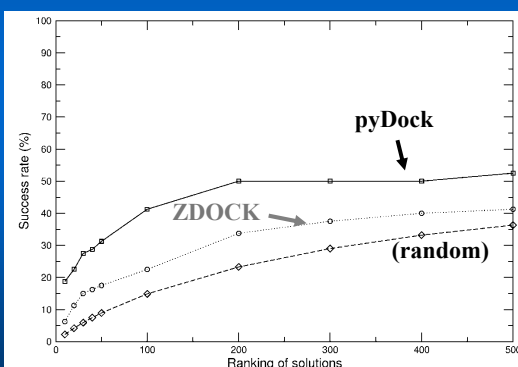
pyDockSER

pyDock evaluation in other rigid-body docking sets

FTDOCK's docking sets:
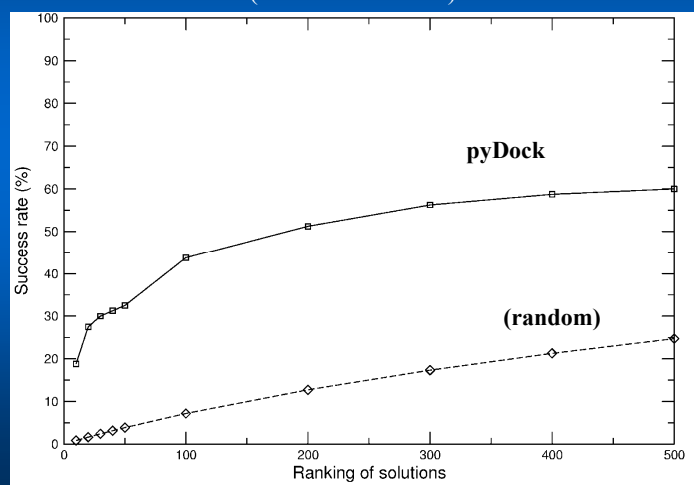(80 unbound cases)
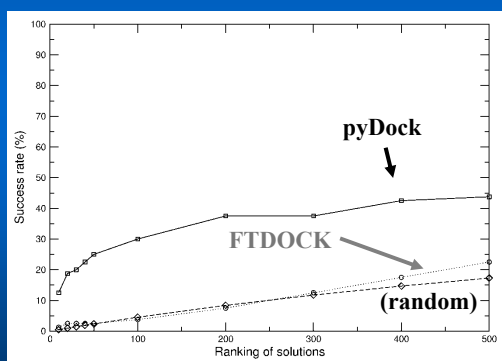
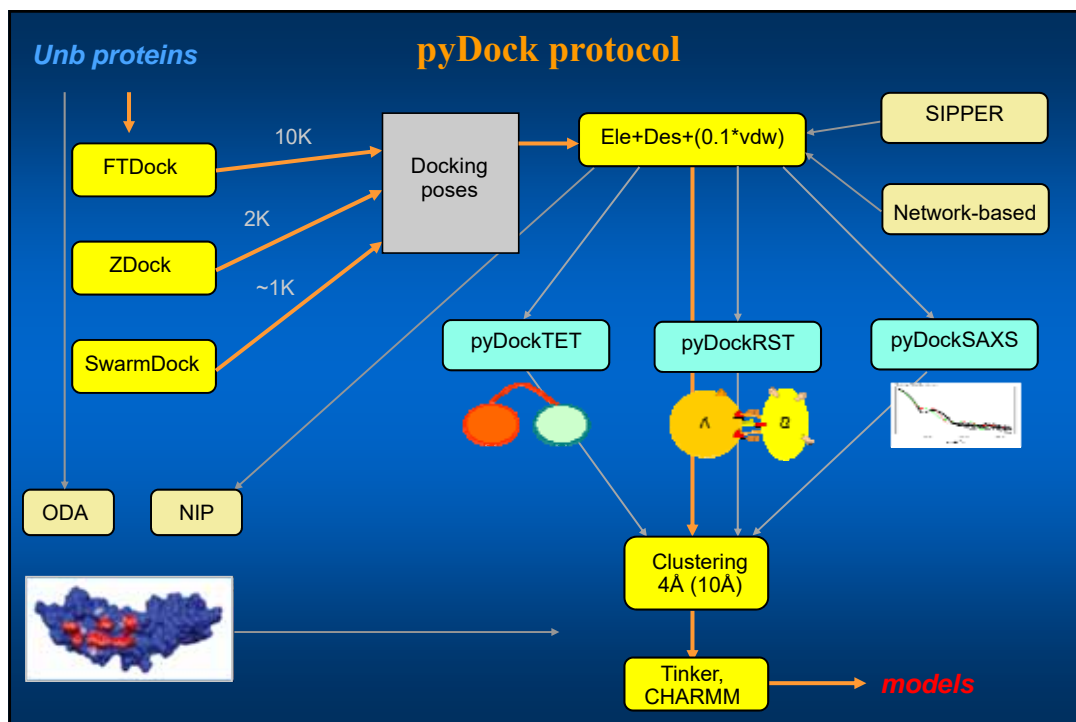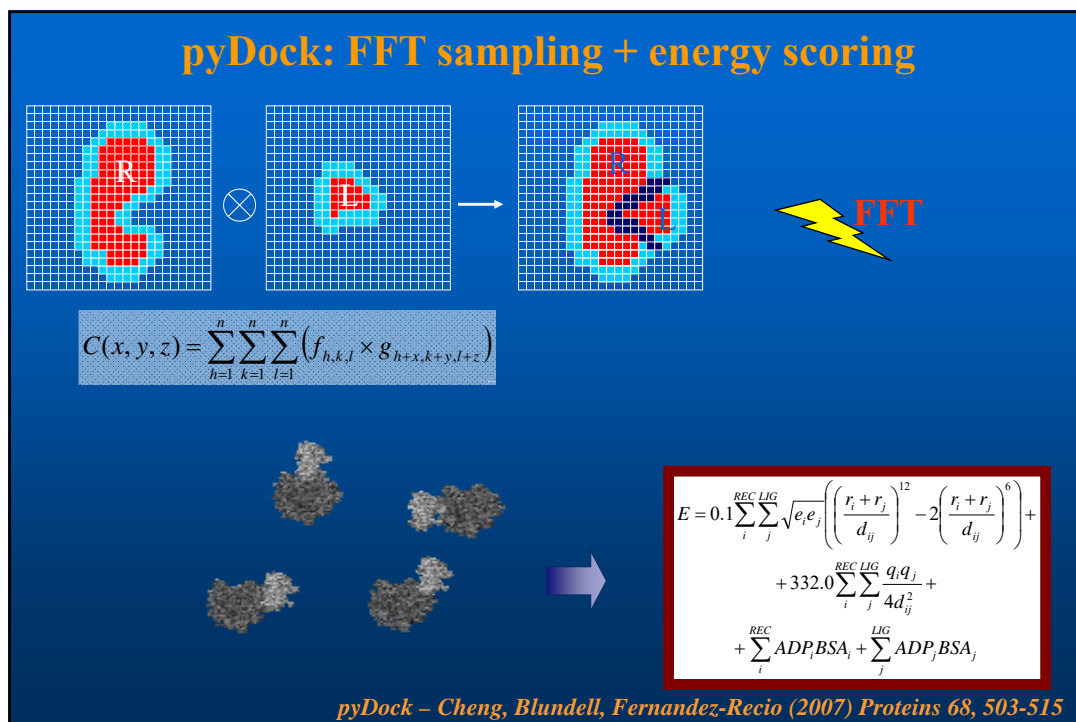**ZDOCK's docking sets:**
(80 unbound cases)

pyDock – Cheng, Blundell, Fernandez-Recio (2007) Proteins 68, 503-515

pyDock evaluation in other rigid-body docking sets

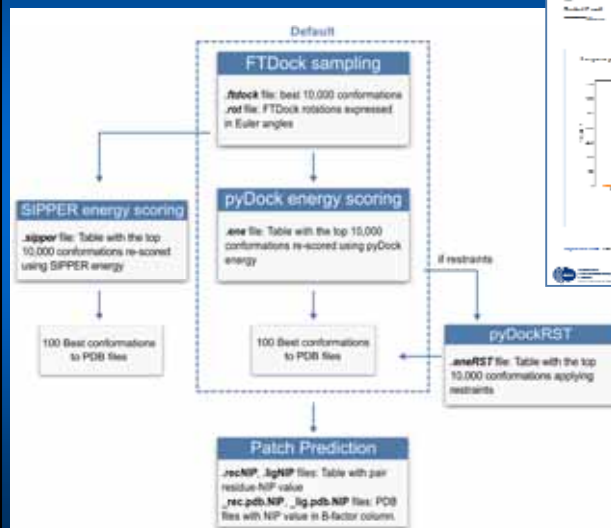**FTDOCK+ZDOCK combined**
(80 unbound cases)

pyDock: FFT sampling + energy scoring

$$C(x, y, z) = \sum_{h=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \left(f_{h,k,l} \times g_{h+x,k+y,l+z} \right)$$

FFT

$$E = 0.1 \sum_{i}^{REC} \sum_{j}^{LIG} \sqrt{e_i e_j} \left(\left(\frac{r_i + r_j}{d_{ij}} \right)^{12} - 2 \left(\frac{r_i + r_j}{d_{ij}} \right)^{6} \right) +$$
$$+ 332.0 \sum_{i}^{REC} \sum_{j}^{LIG} \frac{q_i q_j}{4 d_{ij}^2} +$$
$$+ \sum_{i}^{REC} ADP_i BSA_i + \sum_{j}^{LIG} ADP_j BSA_j$$

pyDock – Cheng, Blundell, Fernandez-Recio (2007) Proteins 68, 503-515

pyDock protocol

Unb proteins

FTDock → 10K → Docking poses → Ele+Des+(0.1*vdw) ← SIPPER

ZDock → 2K → Docking poses

SwarmDock → ~1K → Docking poses

Network-based

pyDockTET   pyDockRST   pyDockSAXS

ODA   NIP

Clustering 4Å (10Å)

Tinker, CHARMM → *models*

pyDock Web Server

http://life.bsc.es/servlet/pydock

- **Energy-based docking**

- **Docking and data-based restraints**

- **Scoring by binding energy**

- **Scoring by statistical energy**

- **Scoring and data-based restraints**

- **Applications of energy-based docking**

Faster scoring by residue-based SIPPER

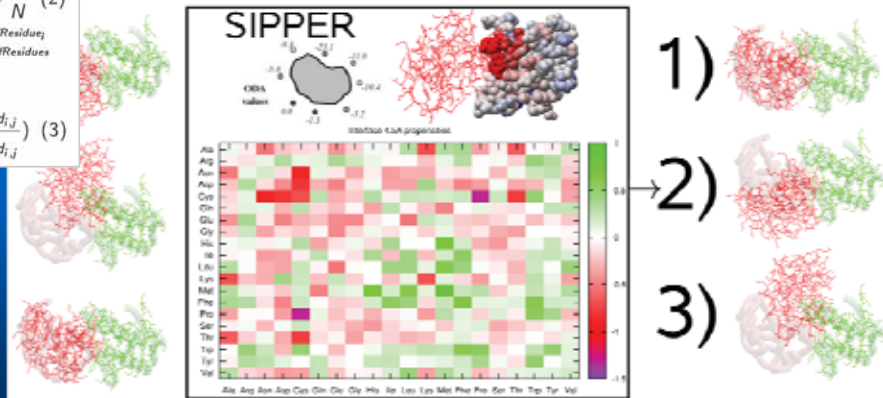$$f_{observed_{i,j}} = \frac{o_{i,j}}{O} \quad (1)$$

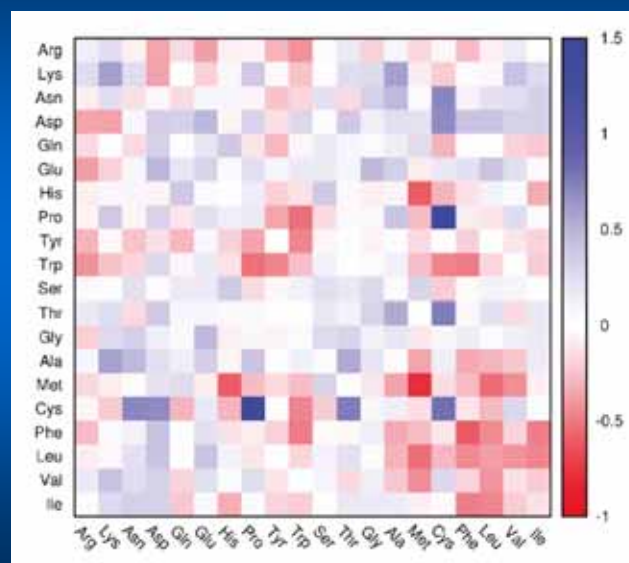$$o_{i,j} = InterproteinContacts_{i,j}$$
$$O = totalInterproteinContacts$$

$$f_{expected_{i,j}} = \frac{n_i}{N} * \frac{n_j}{N} \quad (2)$$

$$n_i = OccurrencesOfResidue_i$$
$$N = totalNumberOfResidues$$

$$p_{i,j} = ln\left(\frac{f_{observed_{i,j}}}{f_{expected_{i,j}}}\right) \quad (3)$$

70 PDB structures
of hetero-dímers

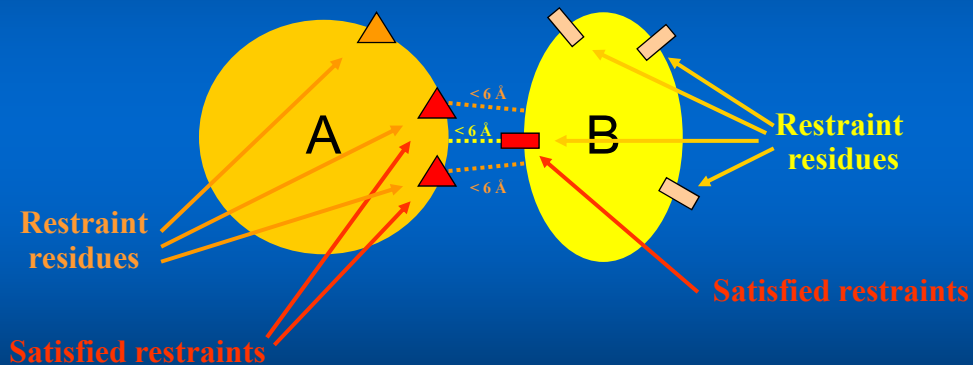# Faster scoring by residue-based SIPPER

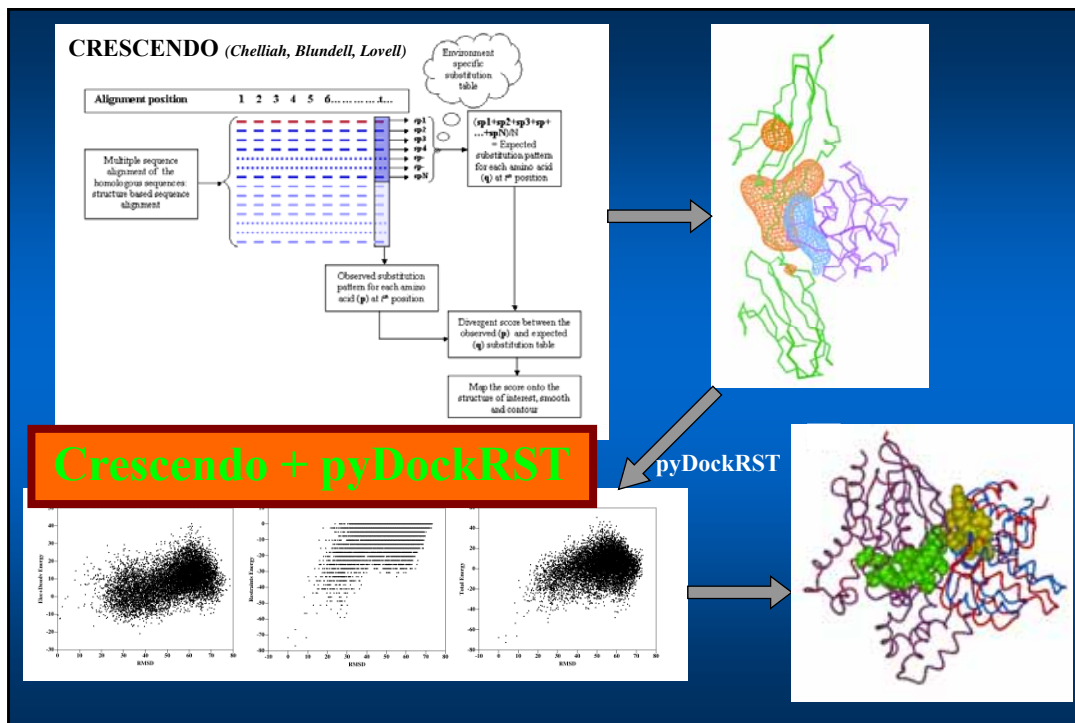## Faster scoring by residue-based SIPPER
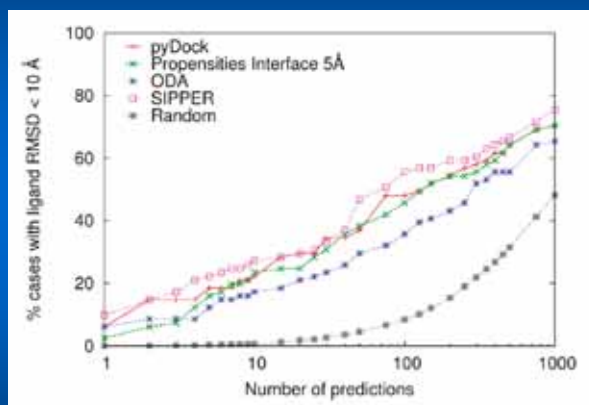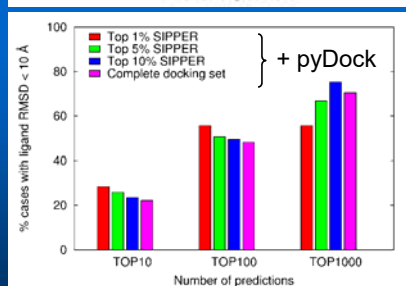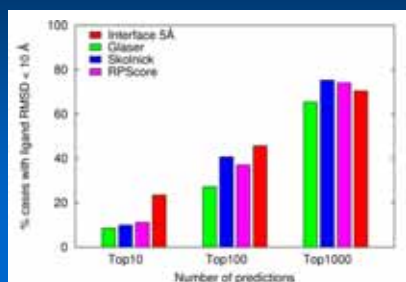
- **Energy-based docking**

- **Docking and restraints**

- **Scoring by energy**

- **Scoring by statistical energy**

- **Scoring and restraints**

- **Applications of energy-based docking**

pyDockRST: use of restraints to filter docking solutions

Docking solution *i*

Restraint residues

Restraint residues

Satisfied restraints

Satisfied restraints

< 6 Å
< 6 Å
< 6 Å

A

B

Restraint pseudo-energy = -1.0*(% satisfied restraints)

CRESCENDO *(Chelliah, Blundell, Lovell)*

Crescendo + pyDockRST

pyDockRST

Crescendo + pyDockRST
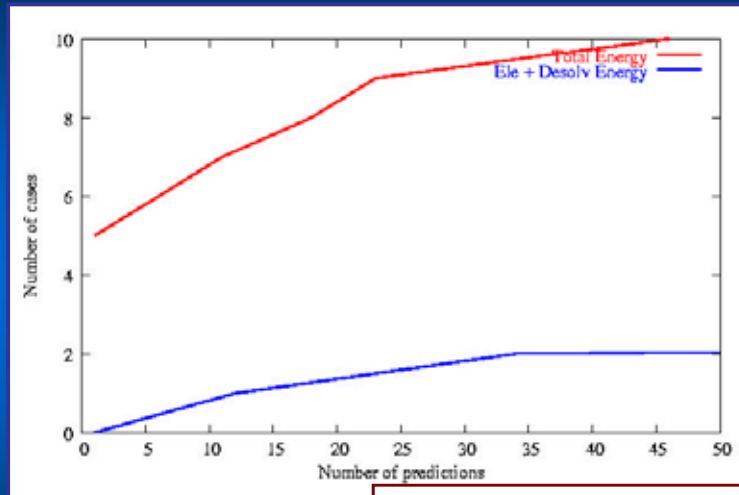
Introduction of evolutionary restraints dramatically improves the docking results

Integrating docking and low-resolution data: SAXS

22

Integrating docking and low-resolution data: SAXS

70 cases

1) Docking with FTDock

2) Scoring of docking poses:

- FTDock default
- pyDock (E)
- CRYSOL (χ^2)
- pyDock (E) + CRYSOL (χ^2)
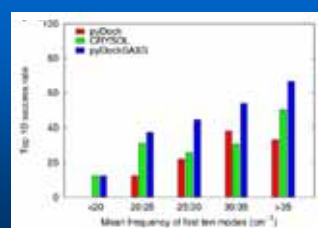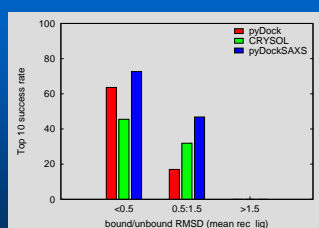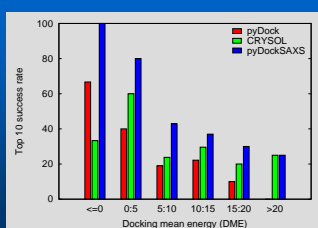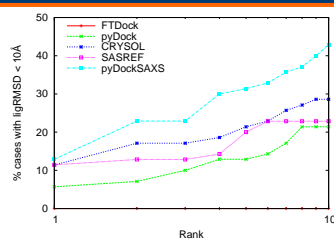
Pons et al. 2010 JMB

• **Energy-based docking**

• **Docking and restraints**

• **Scoring by energy**

• **Scoring by statistical energy**

• **Scoring and restraints**

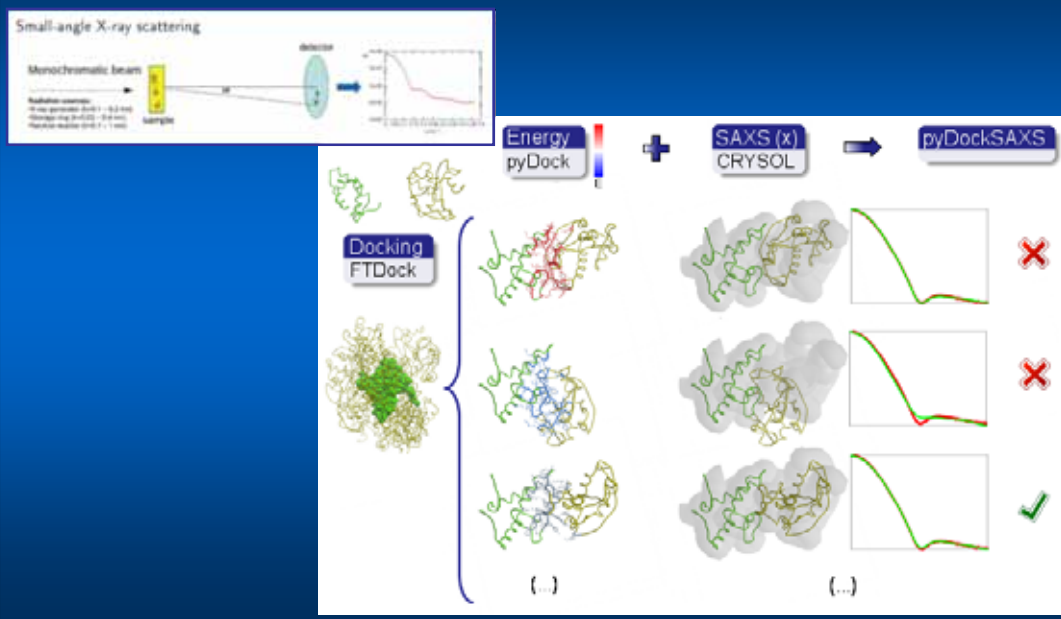• **Applications of energy-based docking**

Docking energy landscapes

1ca0

2pcf

1mlc

Docking energy landscapes

1ca0

2pcf

1mlc
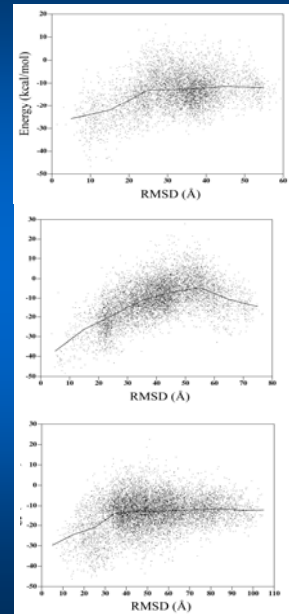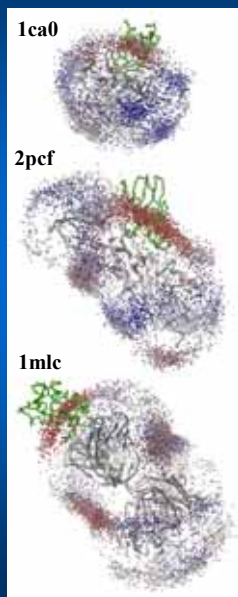
Docking energy landscapes
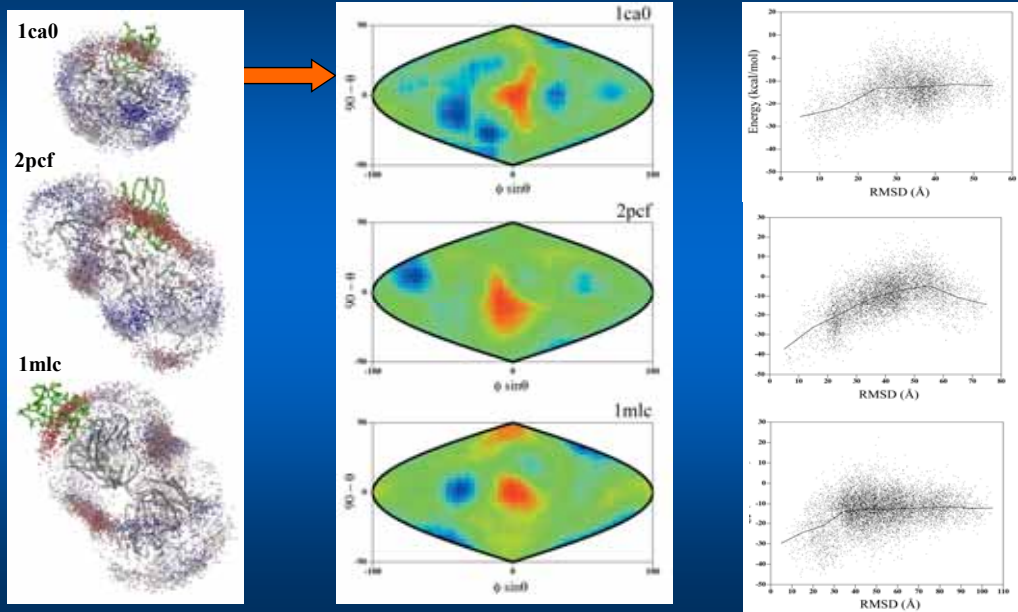
Docking helps to describe binding mechanism

**HPr / EIN
rigid-body docking
(only electrostatics)**

**HPr / EIN
PRE data**
*Tang et al. (2006) Nature 444,
383*
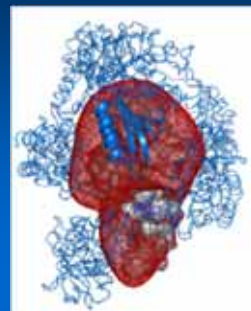
Energy-based docking: conclusions

- *Fewer available methods*

- *Slower than geometry-based*

- *Better success rates*

- *Better geometries*

- *Better understanding of binding mechanism*

- *Desolvation is very important in docking*

Assessment of docking performance
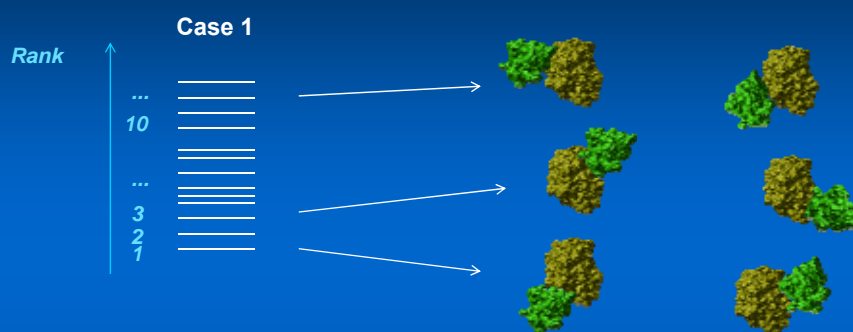
Case 1

Rank

...
10

...
3
2
1

27
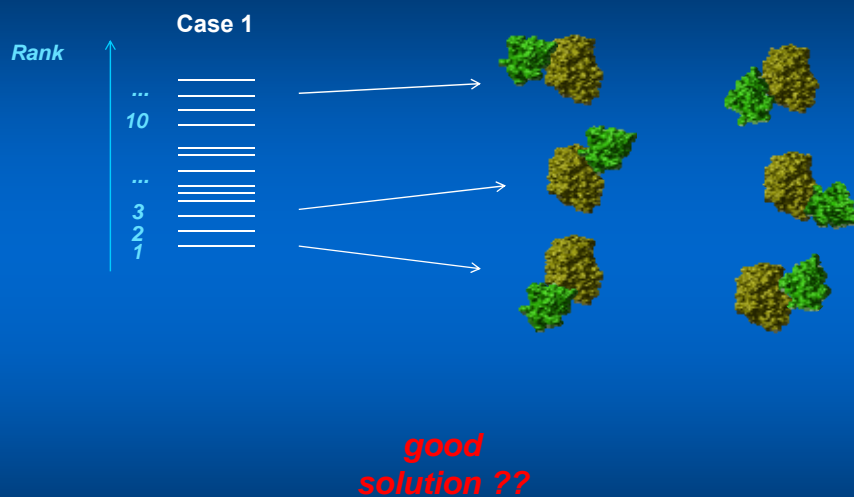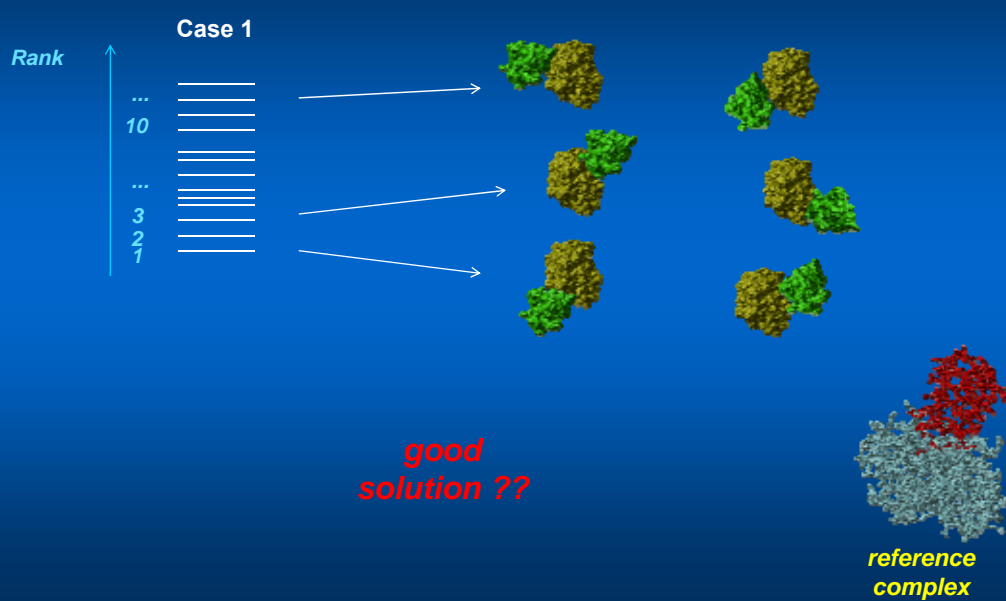
Assessment of docking performance

- F_{nat}: fraction of native contacts (within 5Å)

$$F_{nat} = \frac{\text{Correctly predicted contacts}}{\text{Total number of contacts in the target}}$$

- l-RMSD: RMSD on second protein after superposition on first

near-native:
L-RMSD < 10Å
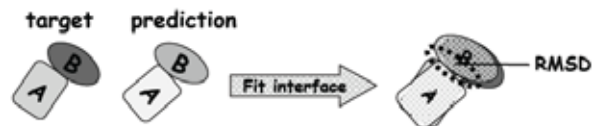
- i-RMSD: RMSD on interface residues (within 10 Å)

Assessment of docking performance

Case 1

Rank | L-RMSD

Best rank NN   3

Rank	Scoring	L-RMSD
1	2569	16.3
2	2456	25.8
3	2019	5.4
...	...	...
...	...	...
10	1567	23.5

DOCKING VALIDATION
CAPRI: A Critical Assessment of PRedicted Interactions

T01
Hpr (*unbound*)
HPr kinase (*unbound*)

T02
VP6 (*unbound*)
Fab (*bound*)

T03
Hemagglutinin (*unbound*)
Fab (*bound*)

T04, T05, T06
α-amylase (*unbound*)
VHH (*bound*)

T07
TCRβ (*unbound*)
speA (*unbound*)

http://www.ebi.ac.uk/msd-srv/capri/
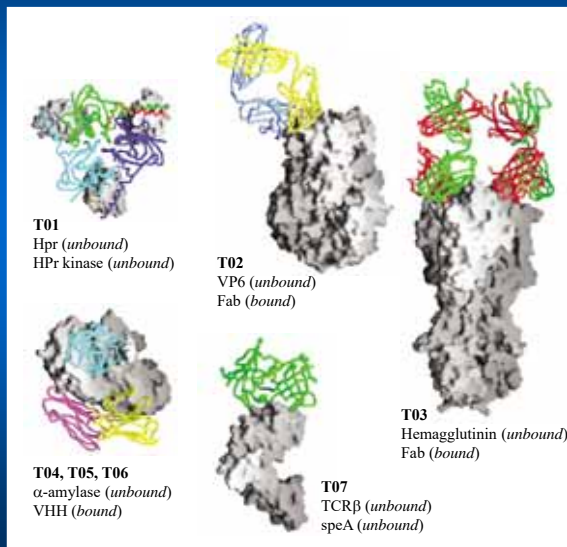
1st CAPRI – Sep02 La Londe (France)
PROTEINS: Structure, Function, and Genetics 52
(July 2003)

2nd CAPRI – Dec04 Gaeta (Italy)
PROTEINS: Structure, Function, and Bioinformatics 60
(July 2005)

3rd CAPRI – Apr07 Toronto (Canada)
PROTEINS: Structure, Function, and Bioinformatics 69
(December 2007)

4th CAPRI – Dec10 Barcelona (Spain)
PROTEINS: Structure, Function, and Bioinformatics 78
(November 2010)

5th CAPRI – Apr13 Utrecht (Netherlands)
PROTEINS: Structure, Function, and Bioinformatics 81
(December 2013)

DOCKING VALIDATION
CAPRI: A Critical Assessment of PRedicted Interactions

Rank	F_{native}	l-RMSD [Å]	i-RMSD [Å]
High ***	> 0.5	x < 1.0	or x < 1.0
Good **	> 0.3	1 < x < 5	or 1 < x < 2
Acceptable *	> 0.1	5 < x < 10	or 2 < x < 4
Incorrect	< 0.1		

- F_{native}: fraction of native contacts (within 5Å)
- l-RMSD: rmsd on second protein after superposition on first
- i-RMSD: rmsd on interface residues (within 10Å)

DOCKING VALIDATION
CAPRI: A Critical Assessment of PRedicted Interactions

**ROUND 9 T25
Results**

Predictor groups	37
Evaluated predictions	336
High accuracy (***)	1

Model	Predictor	f_{nat}	f_{n-nat}	f_{IR-R}	f_{IR-L}	#CI	L-$_{rmsd}$	I-$_{rmsd}$	Θ	D_L
10 **H**	Eisenstein	0.827	0.246	0.913	0.840	4	2.203	0.904	2.70	1.482
01 **M**	Schomburg	0.808	0.276	0.913	0.920	5	1.829	1.062	3.69	0.407
10	PATCHDOCK	0.692	0.357	0.913	0.960	19	2.334	1.166	6.28	1.139
01	GRAMM-X	0.827	0.295	0.913	0.960	10	2.824	1.246	8.94	1.294
03	Vajda	0.635	0.400	0.913	0.960	2	2.832	1.297	9.10	1.293
03	Fernandez-Recio	0.788	0.281	0.913	0.880	35	2.844	1.234	9.36	1.563
02	Totrov	0.692	0.122	0.783	0.960	2	3.018	1.246	10.70	1.643
01	Facemyer	0.788	0.211	0.957	0.920	21	3.153	1.415	14.03	1.151
07	SKE-DOCK	0.788	0.281	0.913	0.920	9	3.307	1.330	11.07	2.029
07	Takeda-Shitaka	0.808	0.323	0.913	1.000	8	3.324	1.335	11.08	2.059
07	Weng	0.808	0.364	0.913	0.960	6	3.804	1.506	13.01	2.450
02	Smith	0.673	0.426	0.913	0.960	23	3.929	1.510	9.27	3.085
01	SMOOTHDOCK	0.385	0.592	0.739	0.760	15	4.849	2.192	23.98	1.966
05	Negi	0.673	0.470	0.913	0.920	55	5.745	1.862	17.36	4.029
01	Bonvin	0.365	0.708	0.913	0.960	10	6.644	3.522	21.05	4.809
04 **A**	Camacho	0.308	0.765	0.826	0.840	7	7.533	4.088	34.68	4.202
06	CLUSPRO	0.654	0.477	0.739	0.840	12	7.564	2.845	34.45	4.634
05	Bates	0.442	0.681	0.957	0.880	17	9.825	3.178	24.42	8.412
	Unbound	0.923	0.094	0.957	0.960	0	0.282	1.134	0.65	0.109

5th CAPRI

	Predictor Group – Per Target Data	Predictor Summary
HADDOCK	Bonvin	9/1***/3**
SwarmDock	Bates	8/2**
GRAMM/homol	Vakser	7/1***
PIPER	Vajda	6/2***/3**
pyDock, SDU	Fernandez-Recio, Shen	6/1***/3**
ITScore+entropy	Zou	6/1***/2**
ATTRACT	Zacharias	6/1***
CLUSPRO	CLUSPRO	6/4**
MolFit	Eisenstein, Grudinin	5/1***/2**
HADDOCK server, ZDOCK, RosettaDock	HADDOCK, Weng, Seok, Gray	4/1***/1**
MLSBF	Nakamura	4/1***
SwarmDock server	SWARMDOCK	4/1**
PatchDock/FireDock	Wolfson	3/2**
...	DOCK_PIE, Zhou	3/1**
	Cui	2/1***/1**
	HEXSERVER, Luethy	2/1***
	Mitchell, Elber	2/1**
	FIBERDOCK, Ritchie, Wang	1***
	Ten Eyck, Camacho, LZERD	1**
	Pal, Zhang, SURFIT, Poupon, Kihara, Dokholyan	1*
	26 other groups	0

Total 62 groups

2ⁿᵈ CAPRI – _Target 14_

Template:
protein phosphatase 1α

Modelled Ligand:
protein phosphatase 1β

ID 92.9%

Bound receptor:
MYPT1

Fernandez-Recio et al. (2005) Proteins 60, 308-313

2ⁿᵈ CAPRI – _Target 14_

Template:
protein phosphatase 1α

Modelled Ligand:
protein phosphatase 1β

ID 92.9%

Docking

Rank 1
Ligand
RMSD:
0.6Å

Bound receptor:
MYPT1

Fernandez-Recio et al. (2005) Proteins 60, 308-313

2nd CAPRI – *Target 14*

ID 92.9%

Template: protein phosphatase 1α

Modelled Ligand: protein phosphatase 1β

Bound receptor: MYPT1

Docking

Rank 1 Ligand RMSD: 0.6Å

Fernandez-Recio et al. (2005) Proteins 60, 308-313