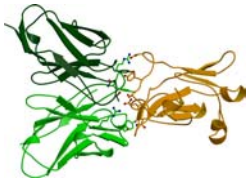
Protein-Protein Docking



Jeffrey J. Gray
Protein Bioinformatics, April 2006

JOHNS HOPKINS UNIVERSITY

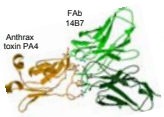
Chemical & Biomolecular Engineering

Institute for Multiscale Modeling of Biological Interactions

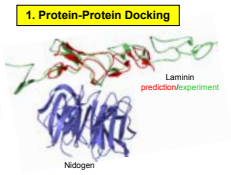
Biomolecular & Nanoscale Modeling Lab

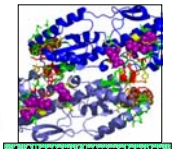
Jeffrey J. Gray, Ph.D. – jgray@jhu.edu – http://graylab.jhu.edu
Chemical & Biomolecular Engineering and Program in Molecular & Computational Biophysics

1. Protein-Protein Docking

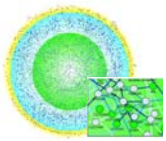


2. Therapeutic Antibodies

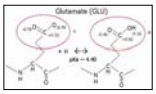


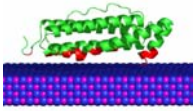


4. Allostery / Intramolecular Signal Transduction



3. Proteome Docking Predictions





5. Protein-Surface Interactions

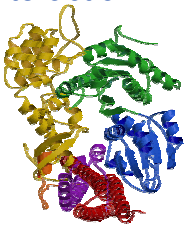
Outline

- Motivation
- Docking Methods
- Results / Evaluation of Method
- Blind Prediction Challenge
- Recent Work: Flexibility & Ensembles

Goal: Demonstrate Current Methodologies & Capabilities in Protein-Protein Docking

Cellular Function Depends on Protein-Protein Interaction

- Signaling
- Regulation
- Recognition
- Enzymes/inhibitors
- Antibodies/antigens

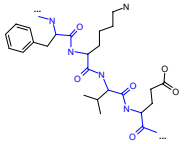


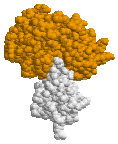
nSec1 + syntaxin1a

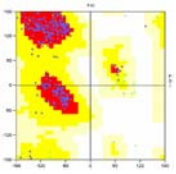
Faulty interactions result in diseases

Protein docking tests our fundamental knowledge of biomolecular physics

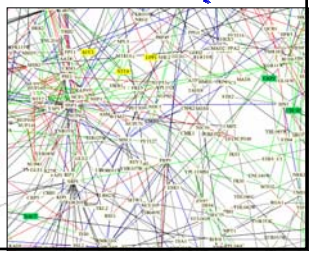
- Conformational space
- Free energy functions
  - Water (solvation)
  - Hydrogen bonding
  - Van der Waals
  - Electrostatics





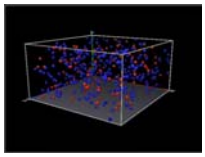
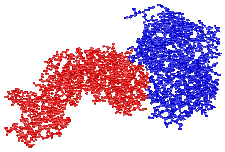
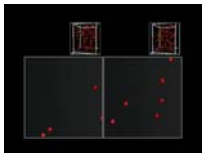
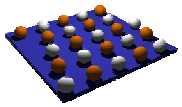


Computational protein docking could help elucidate biological molecular interactions on a genomic scale



Uetz, Fields, Rothberg *et al. Nature* 2000

Protein docking studies may teach us how to design complex devices capable of assembling themselves from nanoscopic (macromolecular) components

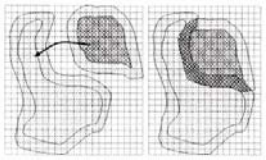
FTDOCK: Fourier-Transform Docking (Rigid Body)

Katchalski-Katzir, Shariv, Eisenstein, Friesem, Afalo & Vakser 1992

- Discretize the protein shape:

$$a_{l,m,n} = \begin{cases} 1, \text{surface of molecule} \\ \rho = -15, \text{core of molecule} \\ 0, \text{outside of molecule} \end{cases}$$

$$b_{l,m,n} = \begin{cases} 1, \text{surface of molecule} \\ \delta = 1, \text{core of molecule} \\ 0, \text{outside of molecule} \end{cases}$$



Ackermann 1998

- And correlate the functions:

$$c_{\alpha,\beta,\gamma} = \sum_{l}\sum_{m}\sum_{n} a_{l,m,n} \cdot b_{l+\alpha,m+\beta,n+\gamma}$$

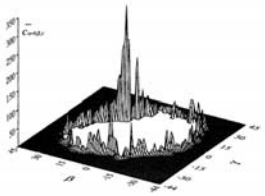
- $l,m,n,\alpha,\beta,\gamma \rightarrow N^6$

FTDOCK

Katchalski-Katzir *et al.*, 1992

- Use a Discrete Fourier Transform

$$X_{o,p,q} = \sum_{l}\sum_{m}\sum_{n} \exp\left[\frac{-2\pi i}{N}(ol+pm+qn)\right] \cdot x_{l,m,n}$$



- Mulitply in Fourier Space:

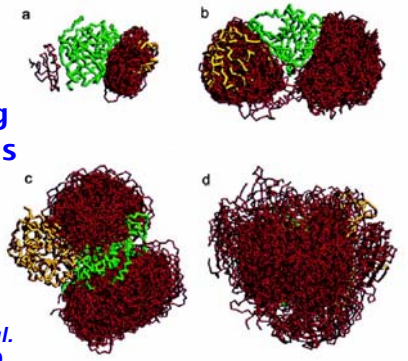
$$C_{o,p,q} = A^{*}_{o,p,q} \cdot B_{o,p,q}$$

- Invert:

$$c_{\alpha,\beta,\gamma} = \frac{1}{N^3}\sum_{o}\sum_{p}\sum_{q} \exp\left[\frac{2\pi i}{N}(o\alpha+p\beta+q\gamma)\right] \cdot C_{o,p,q}$$

- Then, search over rotation space: $\{\theta,\phi,\psi\}$
- DFT $\rightarrow N^3 \ln N^3$

FT Docking Solutions



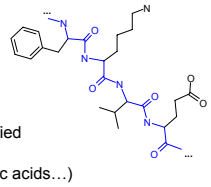
Vakser *et al.*
PNAS 1999

A wide variety of methods have been developed since Katchalski-Katzir

- FFT/Grid (Eisenstein, Sternberg, Weng, Ten Eyck)
- Computer vision / matching knobs & holes / geometric hashing (Wolfson, Nussinov, Norel)
- Electrostatic and VdW filters (Weng, Camacho, Sternberg, Ten Eyck, many others)
- Spherical harmonic shape representations (Ritchie)
- Genetic Algorithm (Gardiner)
- MD (Mustard, Bates) and Minimization (many)
- NMR + docking (Bonvin)
- Residue conservation and co-variance/hotspots (Valencia, Kaznessis)
- Biological information (Sternberg, many others)
- Monte Carlo with physical potentials (Abagyan, US!)

Protein Docking is Difficult!

- Proteins can be large (50-1000+ residues = 500-10,000+ atoms)
- Interactions mediated by *water*
- Proteins are flexible
  - Backbone
  - Side chains
- Ions can be present
- Proteins can be post-translationally modified
- Environment is crowded (other proteins, lipids, membranes, nucleic acids…)
- Multi-protein interactions (chaperones) could be important



Need to simplify!!

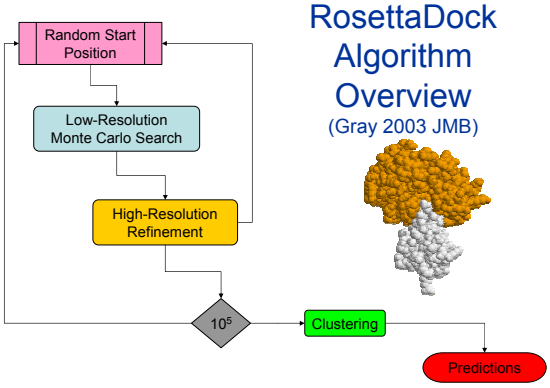
Our Approach to Modeling Proteins

- Model physical forces when possible:
  van der Waals, solvation, hydrogen bonding, electrostatics, …
- Use statistics from the *Protein Data Bank* to compensate for poor physical models
- Generate large numbers of plausible decoys
- Model only necessary degrees of freedom
- Employ multi-scale models for both breadth of search and accuracy of discrimination

Although the problem is tremendously complex, we believe that simple fundamental principles will emerge…

RosettaDock Algorithm Overview

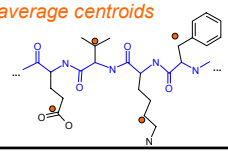
(Gray 2003 JMB)

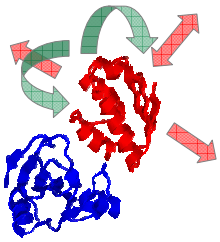


Low-Resolution Search

- Monte Carlo Search
- Rigid body translations and rotations
- Residue-scale interaction potentials

Protein representation:
backbone atoms +
average centroids

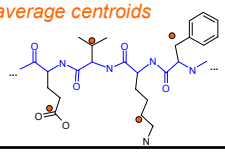


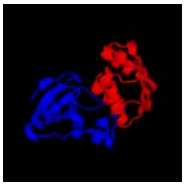


Low-Resolution Search

- Monte Carlo Search
- Rigid body translations and rotations
- Residue-scale interaction potentials

Protein representation:
backbone atoms +
average centroids



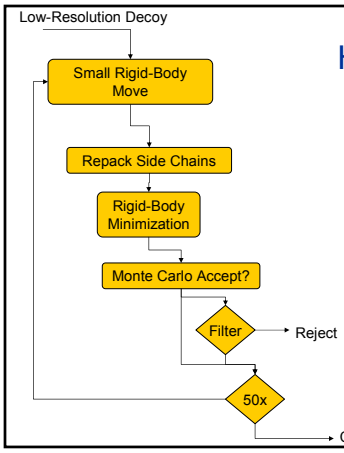


- Mimics physical diffusion process

Residue-scale scoring

Score	Representation	Physical Force
Contacts	$r_{\text{centroid-centroid}} < 6$ Å	Attractive van der Waals
Bumps	$(r - R_{ij})^2$	Repulsive van der Waals
Residue environment	$-\ln(P_{\text{env}})$	Solvation
Residue pair	$-\ln(P_{ij})$	Hydrogen bonding electrostatics, solvation
Alignment	-1 for interface residues in Antibody CDR	(bioinformatic)
Constraints	varies	(biochemical)

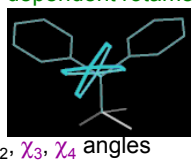
High-Resolution Refinement



- Simultaneous rigid-body and side-chain refinement

Side Chain Packing

- Build amino acid side chains
  - Choose side chains from Dunbrack's backbone-dependent rotamer library



Phenylalanine rotamers
(Richardson, 2000)

  - Vary χ_1, χ_2, χ_3, χ_4 angles
  - Minimize a full-atom energy function w.r.t. all rotamer combinations
  - With strict VdW parameters, extra angles are necessary (Chu Wang)

(Brian Kuhlman & David Baker, Nature Struct. Biol. 2001)

Rotamer Statistics

(Roland Dunbrack, *Curr. Op. Struct. Biol.* 2002)

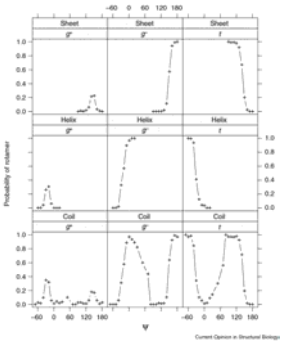
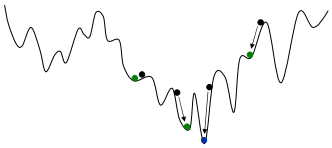
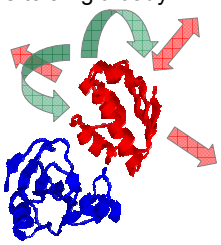


Fig. 1. Observed frequency of the *gauche*$^+$ (χ_1 +60°), *gauche*$^-$ (χ_1 -60°), and *trans* (χ_1 180°), rotamers of valine (horizontally, respectively) in sheet, helix, and coil regions (vertically, respectively) of proteins as a function of the backbone dihedral angle ψ. Data were taken from a list of 850 proteins, 1.7 Å resolution or better, and mutual sequence identity less than 50% (http://www.fccc.edu/research/labs/dunbrack/culledpdb.html). A B-factor cutoff of 40 was used as recommended by Lovell et al. [15].

Minimization

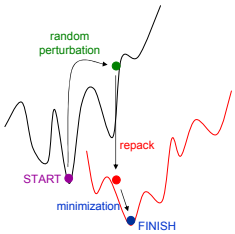
- Full atom rigid-body minimization
  - Use a conjugate-gradient search to find the local score minimum relative to a rigid body translation and rotation





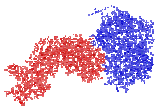
Refinement Cycle

- Simultaneous rigid-body displacement and side chain minimization



Full-Atom scoring

Score	Form / Source	Discriminatory z-value
Repulsive van der Waals	Modified Lennard-Jones 6-12	73.0
Attractive van der Waals	Lennard-Jones 6-12	45.0
Surface area solvation	Surface area (see Tsai 2003)	28.5
Gaussian solvent-exclusion	Lazaridis & Karplus, 1999	27.2
Rotamer probability	Dunbrack & Cohen, 1997	19.6
Hydrogen bonding	Empirical, Kortemme *et al.* 2003	14.9 & 6.8 (BB/BB)
Residue pair probability	Empirical, Kuhlman & Baker 2000	6.9
Electrostatics	Coulomb model with simple charges	0.4-15.1 (LR rep)

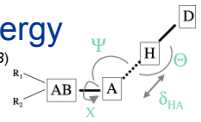


Scoring Weights

Score	Weight (P)	Weight (M)	Weight (D)	z-value
Repulsive van der Waals	0.80	0.338	0.08	73.0
Attractive van der Waals	0.80	0.338	0.338	45.0
Surface area solvation	-	-	0.344	28.5
Gaussian solvent-exclusion	0.80	0.279	0.279	27.2
Rotamer probability	0.79	0.069	0.069	19.6
Hydrogen bonding	2.1	0.441	0.441	
SC/SC + SC/BB				14.9
BB/BB				6.8
Residue pair probability	0.66	0.164	0.164	6.9
Simple electrostatics				
Short-range repulsive	-	0.025	0.025	3.2
Short-range attractive	-	0.025	0.025	8.3
Long-range repulsive	-	0.098	0.098	15.1
Long-range attractive	-	0.0020	0.0020	0.4

Hydrogen Bonding Energy

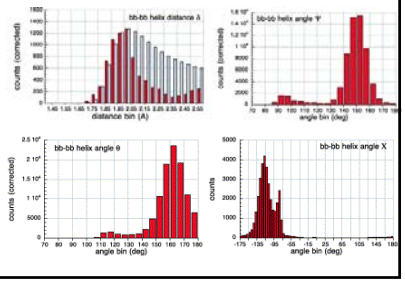
(Kortemme, Morozov & Baker 2003 *JMB*)



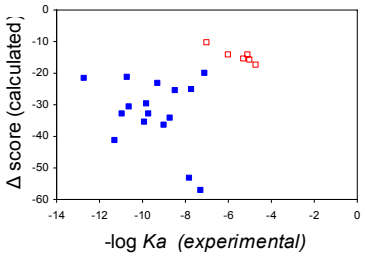
Based on statistics from high-resolution structures in the Protein Data Bank (rcsb.org)

$$\Delta G = -kT \ln P$$

$$E_{HB} = W_{HB}[E(\delta_{HA}) + E(\Theta) + E(\Psi) + E(X)$$



Score correlates with Binding Energy



Filled symbols – targets with funnels
Open symbols – targets without funnels

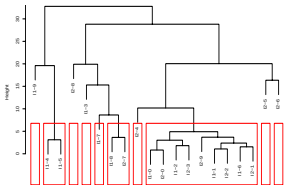
Δ score for bound backbone docking

Clustering

- Compare all top-scoring decoys pairwise

$$RMSD = \left(\sum_{\text{ligand}} (x_i - y_i)^2 \right)^{1/2}$$

- Cluster decoys hierarchically



- Decoys within 2.5Å form a cluster

Represents ENTROPY

IBM BladeCenter
Supercomputing Facility

60 CPUs
0.5 TB storage
1.5 GB RAM/node
1 GB network

Capable of producing
~10^5 protein structures/day

Chemical & Biomolecular

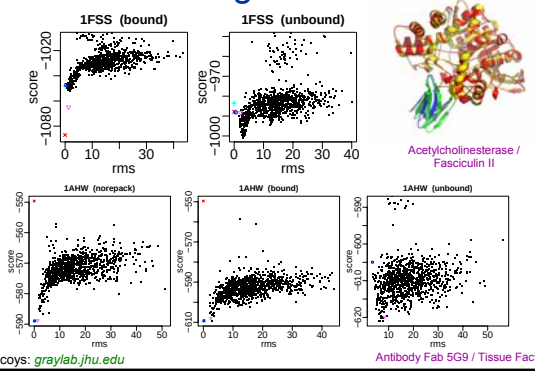
Benchmark Studies

Benchmark set contains 54 targets for which *bound* and *unbound* structures are known

http://zlab.bu.edu/~rong/dock/benchmark.shtml

- Bound-Bound
  - Start with bound complex structure, but remove the side chain configurations so they must be predicted
- Unbound-Unbound
  - Start with the individually-crystallized component proteins in their unbound conformation
- Bound-Unbound (Semibound)

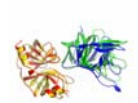
Binding Funnels



Acetylcholinesterase / Fasciculin II

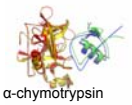
Antibody Fab 5G9 / Tissue Factor

Decoys: *graylab.jhu.edu*

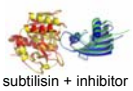


trypsin + inhibitor

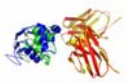
barnase + barstar



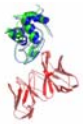
α-chymotrypsin + inhibitor



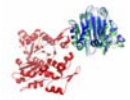
subtilisin + inhibitor



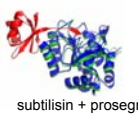
lysozyme + antibodies



hemagglutinin + antibody



actin + deoxyribonuclease I



subtilisin + prosegment

Benchmark Results

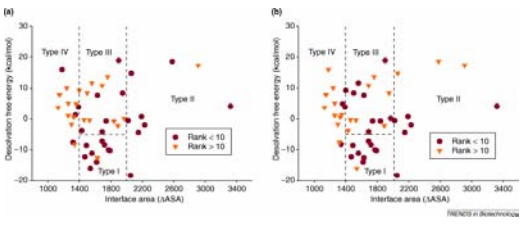
	Bound Perturbations	Unbound Perturbations	Global Searches
Enzyme/Inhibitor	21/22	18/22	17/22
Antibody/Antigen	10/16	9/16	8/16
Other	5/10	5/10	3/10
Difficult	6/6	0/6	0/6
TOTAL	42/54	32/54	**28/54**

Number of successful dockings, starting from either bound or unbound protein backbones and searching either near the native structure or globally.

Benchmark set assembled by R. Chen *et al.*, see *Proteins* 2003

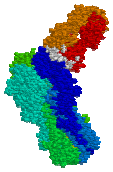
ZDOCK / RosettaDock

(Vajda & Camacho 2004)



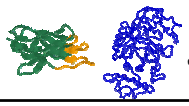
CAPRI: Critical Assessment of PRotein Interactions

- International Blind Prediction Challenge
- 25-35 Participating Research Groups
- Organized by Janin, Wodak, Sternberg
  - Rounds 1-2: 2001-2002 (T1-7)
  - Rounds 3-5: 2003-2004 (T8-19)
  - Round 6-8: 2005 (T20-23)
  - Round 9: *NOW!* (T24-25)

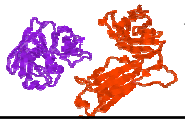


Hemagglutinin + Antibody

- See Janin *et al.* 2003 *Proteins* **52**:2 and Mendez *et al.* 2003 *Proteins* **52**:51.



α-amylase + Camelid Antibody

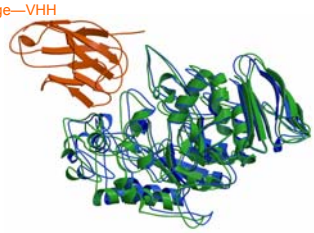


T-Cell Receptor + Strep. pyrogenic exotoxin A (superantigen)

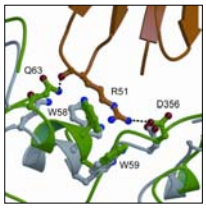
Target 6 (Round 2, Mar 2002)

- α-amylase + VHH, model #1:
  - 48/65 contacts, distance 1.33Å, rotation 3º, rmsd 1.5Å

Orange—VHH



Blue—Native αA
Green—Predicted αA



Xtal by C. Cambillau, CNRS

Target 8 (Round 3, M. Daily, Jan 2003)

- Laminin + Nidogen, model #2:
  - 53% contacts, rmsd 4.6 Å, interface rmsd 0.66 Å

Red—Native laminin
Green—Predicted laminin

Blue—Nidogen

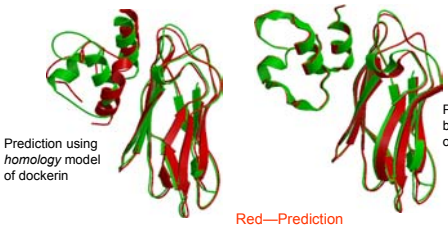
D800, N802, V804 constrained near interface

Xtal by T. Springer, Harvard

Docking a Homology Model (Round 4, Sep 2003)

CAPRI T11/12: Cohesin + Dockerin
Model #6 (T11): 42% contacts, 6.1 Å rmsd, 1.9 Å interface rmsd

- Dockerin coordinates *modeled by homology* via the *Robetta* server
- *RosettaDock* produced the *best model* by correct contacts



Prediction using *homology* model of dockerin

Prediction using bound coordinates of dockerin

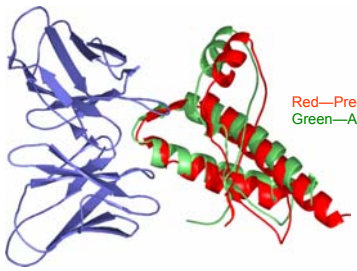
Red—Prediction
Green—Experimental structure

Xtal by Romao, Carvalho, Fontes et al., Lisbon

Prediction by *Mike Daily*
Methods in Gray *et al.* 2003 JMB

Target 19: prion + Fab, model #2

64% contacts, rmsd 3.64 Å, interface rmsd 1.27 Å



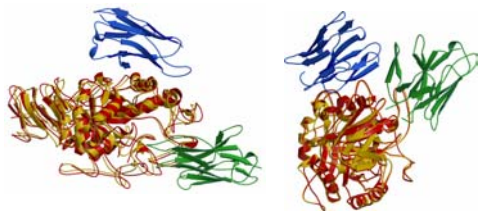
Blue—Fab

Red—Predicted prion
Green—Actual prion

Prion constructed manually from a 95% identical homologue

Targets 4 and 5 (Round 2)

- α-amylase + VHH
  - Incorrectly assumed binding occurs at CDRs



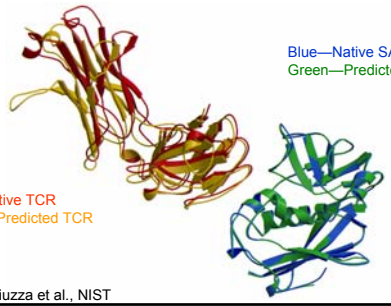
Red—Native α-amylase
Yellow—Predicted α-amylase

Blue—Native VHH
Green—Predicted VHH

Xtal by C. Cambillau et al., CNRS

Target 7 – "Homology Target"

- Streptococcal pyrogenic exotoxin A (superantigen) + T Cell Receptor β chain
  - Predicted by overlaying 1SBB using Mastodon
  - Model #1: 22/37 contacts, distance 3.6Å, rotation 11°
  - Refinement did not improve model



Xtal by Ray Mariuzza et al., NIST

RosettaDock correctly predicts binding sites in 6/10 non-difficult targets

Target	Complex	Type	Nres	Model	Fnat	L_rmsd	I_rmsd	Acc.
15*	Colicin D – immD	BB-BB	194	7	0.88	0.547	0.243	***
12	Cohesin-dockerin	U-B	196	1	0.87	0.99	0.51	***
11	Cohesin-dockerin	U-H	196	5	0.42	6.11	1.93	**
19	Ovine prion – fab	H-B	312	2	0.64	3.64	1.27	***
8	Laminin-nidogen	U-B	427	2	0.53	4.63	0.66	***
17*	GH11 xylanase - XIP	H-U	464	5	0.07	12.91	8.78	-
13	sag1-fab	U-B	474	NP	NP	NP	NP	-
18	GH11 xylanase - TAXI	U-B	552	NP	NP	NP	NP	-
16*	GH 10 xylanase – XIP	H-U	575	7	0.14	8.13	11.64	*
14	mypt1-PP1	U-B	600	NP	NP	NP	NP	-
9	LicT homodimer	U-U	412					
10	TBEV envelope trimer	U-U	1146					

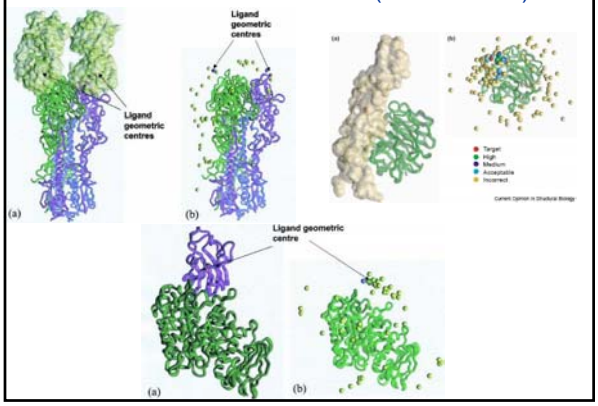
Standard targets; homology targets; not submitted
NP: not predicted

Many Docking Players (Vajda/Camacho 2004)

Table 1. Algorithms of some current protein–protein docking methods[a]

Method (investigator)	Rigid-body search	Re-scoring, ranking, filtering and refinement	Accuracy of CAPRI 1 and 2 submissions[b]
ICM (Abagyan[6])	Pseudo-Brownian Monte Carlo with grid-based energy function	Clustering and selection of 400 conformations. Flexible refinement of interface side-chains and re-scoring with a detailed free-energy function	One high and two medium
ClusPro (Camacho and Vajda [11])	Fast Fourier transform (FFT) correlation approach using the program DOT [16] with a shape complementarity scoring function	Re-scoring with empirical potentials and clustering. Refinement of the 25 largest clusters by the flexible docking method SmoothDock	Two high and one acceptable
MolFit (Eisenstein[4])	FFT with a weighted shape complementarity target function	Clustering of good solutions, filtering using *a priori* information and small, local, rigid rotations around selected conformations	One high and two acceptable
3D-Dock (Sternberg [15])	FFT correlation docking using the program FTDOCK	Complexes re-ranked with a pairwise potential using RPScore. After clustering, side-chains in selected structures are refined using a mean-field approach by Multidock	One high and two acceptable
DOT (TenEyck [16])	FFT correlation approach with shape complementarity and electrostatics	None	One medium and two acceptable
(Gray and Baker [10])	Monte-Carlo search using simplified protein geometry and scoring function	Iterative re-packing of side-chains and rigid-body docking repeated until convergence. Final selection by clustering	One high and one medium
Hex (Ritchie[5])	FFT correlation using polar coordinates and Gaussian density representation of protein shape	None	One high and one medium
ZDOCK (Weng [8])	FFT correlation with shape complementarity, electrostatics and desolvation	Clustering of conformations to avoid redundancies	Two medium
(Nussinov and Wolfson[3])	Geometric matching using knob/hole representations of interacting surfaces	None	One high and one acceptable
GAPDOCK (Gardiner[9])	Genetic algorithm with a shape-based test function	None	Two acceptable
GRAMM (Vakser[7])	FFT correlation with simplified geometry using shape complementarity and hydrophobicity in	Clustering of conformations	One acceptable

CAPRI Submissions (Mendez 2003)



RosettaDock Assumptions

- Rigid protein backbones
- Side chains in rotamer conformations
- Native structure is minimum (free) energy
- Entropy captured by clustering *or* convergence compensates for poor energy model
- Energy functions!
  - Linearly separable
  - Choice of contributions
  - Parameters…

What RosettaDock study tells us about Proteins

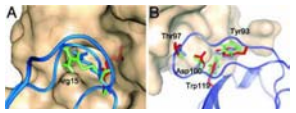
- Packing dominates free energy
- Solvation, hydrogen bonding also important
- Electrostatics not important?
- Energy function is closer to correct than past models
- A short list of probable best docking structures

What it *doesn't* tell you about Proteins

- *THE* energy function
- Unambiguously the "best" conformation
- How specificity is achieved
- Binding affinities

Side chain movement
(Camacho 2004 PNAS)

- Most side chains do not change rotameric conformation upon binding (Weng)
- "Anchor" residue = deeply buried residue at center of interface, usually no conformational change
- "Latch" residue = peripheral interface residue, moves upon binding



co-xtal, unbound, MD

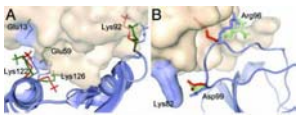
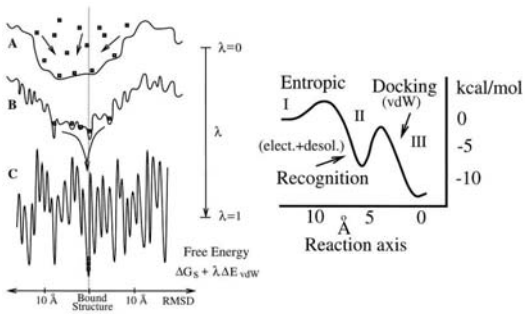


Fig. 1. Shapes of the binding free energy landscape as a function of some arbitrary coordinate measuring the rmsd from the native conformation
Fig. 2. Binding free energy funnel



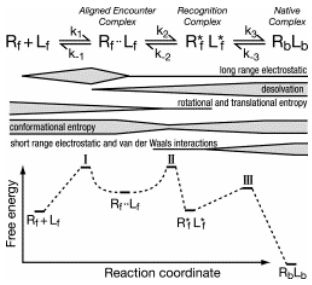
Camacho, Carlos J. and Vajda, Sandor (2001) Proc. Natl. Acad. Sci. USA 98, 10636-10641

PNAS

Three-step mechanism

Grünberg, Leckner & Nilges 2004 *Structure*

I. diffusion
II. free conformer selection (recognition)
III. induced fit (refolding)

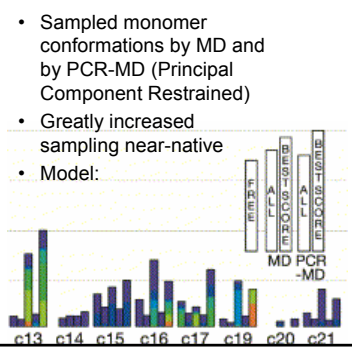


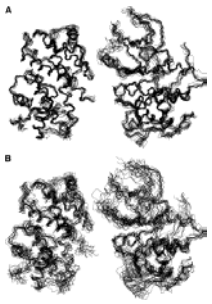
Docking Ensembles

(similar ideas by Smith, Sternberg & Bates 2005)

Grünberg, Leckner & Nilges 2004 *Structure*

- Sampled monomer conformations by MD and by PCR-MD (Principal Component Restrained)
- Greatly increased sampling near-native
- Model:





Loop Flexibility

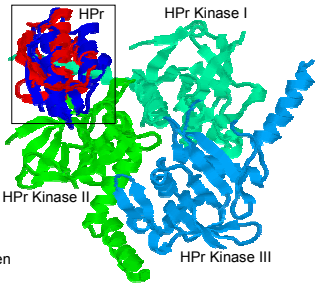
- Currently exploring ways of moving loops during protein-protein docking to simulate an induced fit binding mechanism

Rohl, CA *et al* 2004 to appear

Target 1: HPr + HPr Kinase:
(Round 1,Sep 2001)

- Model #8 among the closest:

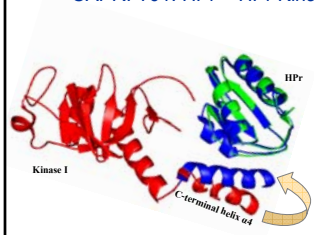
2/52 contacts
distance 2.6Å
rotation 55º
RMSD 8.8Å



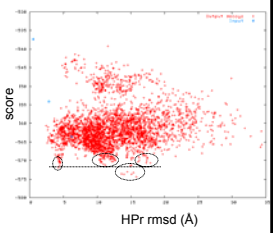
Distance constraint between
Ser157C and Asp46A

Xtal by Fieulaine et al., CNRS

Backbone Conformational Change
CAPRI T01: HPr + HPr Kinase (Round 1, Sep 2001)



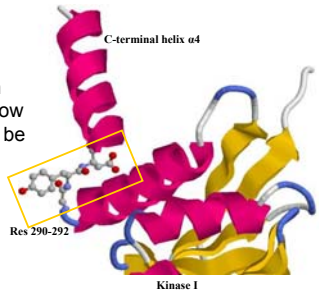
Terminal helix *swings* upon docking,
nuzzling HPr in a pocket

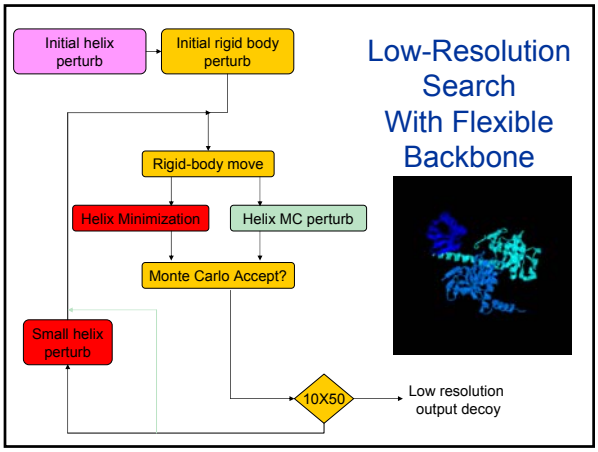


No energy funnel for binding the unbound components

Torsion Angle Perturbation

Torsion angle movement in residues 290-292 would allow the correct conformation to be observed.

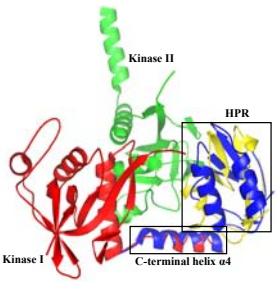




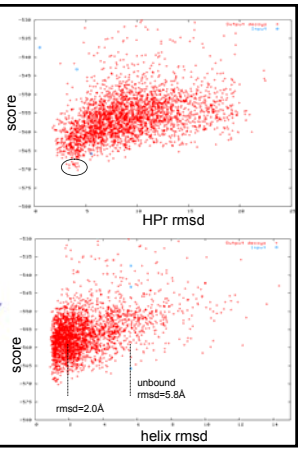
Low-Resolution Search With Flexible Backbone

Flexible Docking Results

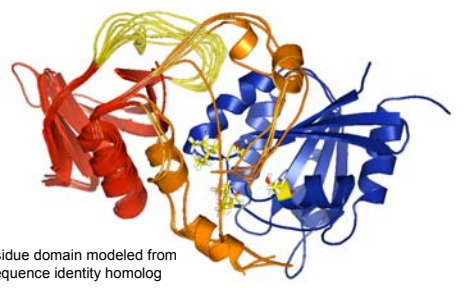
With torsion angle perturbations and explicit minimizations



18/36 contacts, translation 1.8Å, rotation 18°



CAPRI T24: Homology model flexible docking (March 2006)



105 residue domain modeled from 30% sequence identity homolog

Prediction by Monica Berrondo and Aroop Sircar

Summary

- A variety of protein-protein docking techniques have been developed combining advanced techniques in applied mathematics and biophysics
- Benchmark and CAPRI performance is encouraging – but work remains
- Significant challenges persist in sampling (particularly for flexible backbones and large targets) and correction of the energy function

- RosettaDock Software & Decoys:
  - *graylab.jhu.edu*
  - Gray *et al., JMB* **331**:281, 2003
  - Gray *et al., Proteins* **52**:118, 2003

JOHNS HOPKINS UNIVERSITY

Recommended References

1. "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations," Gray, Moughon, Wang, Schueler-Furman, Kuhlman, Rohl & Baker, *J. Mol. Biol.* 2003 **331**, 281-299.
2. "Complementarity of structure ensembles in protein-protein binding," Grunberg, Leckner & Nilges, *Structure* 2004 **12**, 2125-2136.
3. "Prediction of protein-protein interactions by docking methods," Smith & Sternberg, *Curr. Op. Struct. Biol.* 2002 **12**, 36-40.
4. "Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures," Mendez, Leplae, Lensink & Wodak, *Proteins* 2005 **60**, 150-169.

Docking into EM maps

(Aloy, Bork, Serrano, Russell, et al. *Science* 2004)