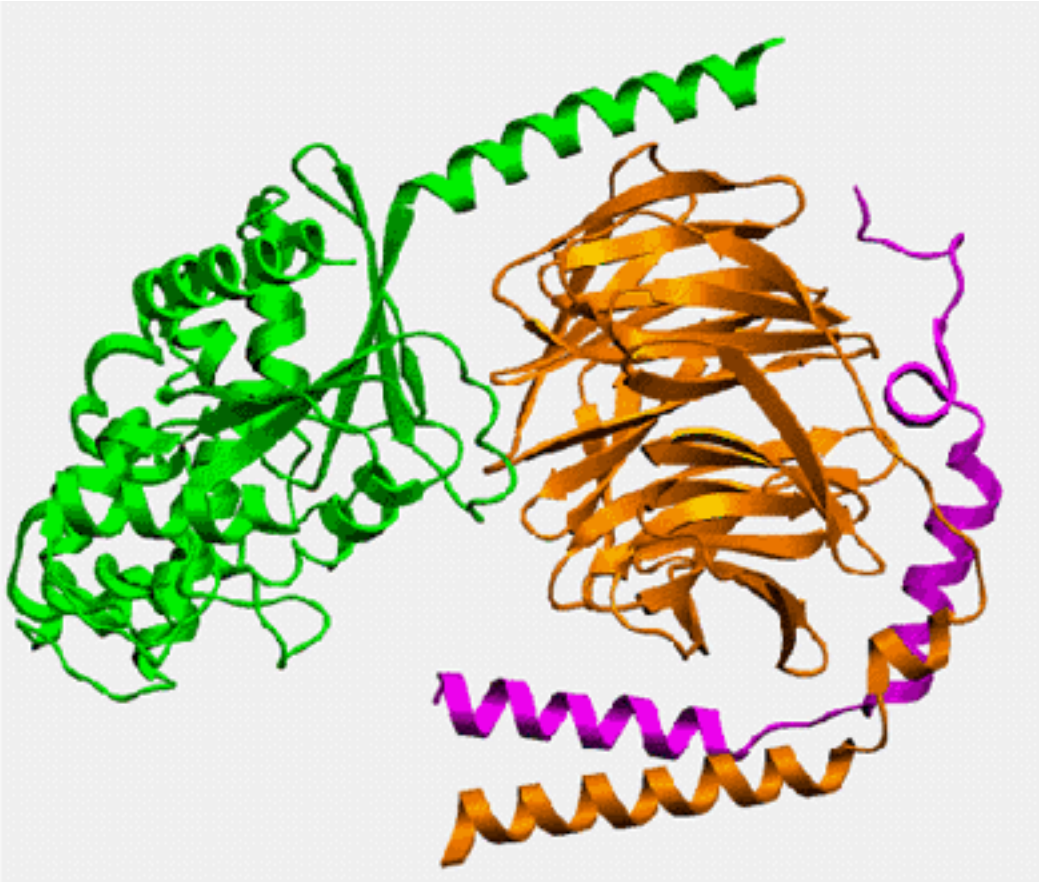
Protein-Protein Docking

Jianlin Cheng

2013

Slides Adapted from Prof. Ora Schueler-Furman at The Hebrew University of Jerusalem

Protein Complex



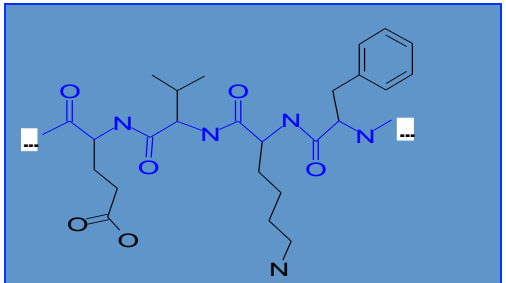
Prediction of protein-protein interactions

1. How do proteins interact?
2. Can we **predict** and **manipulate** those interactions?
3. Prediction of protein quaternary structure

Docking *vs. ab initio* modeling

de novo Structure Prediction

ADEFFGKLSTKK……. *Sequence*



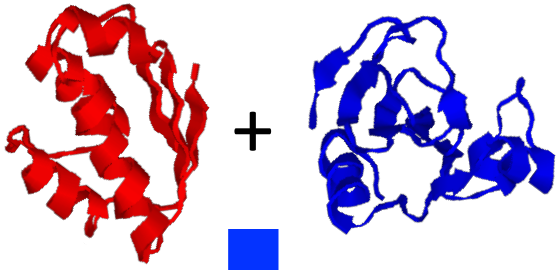
Building Blocks:
backbone & side chains

CASP

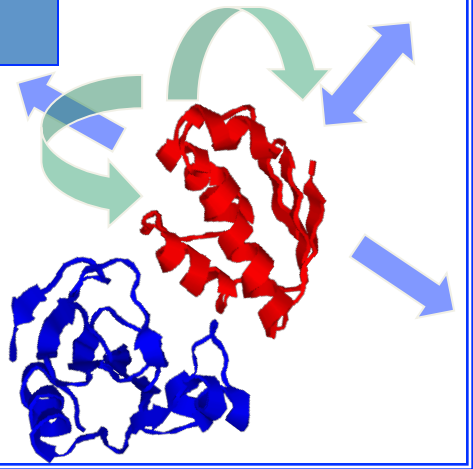
Structure

Docking

Monomers



Rigid body degrees of freedom
3 translation
3 rotation

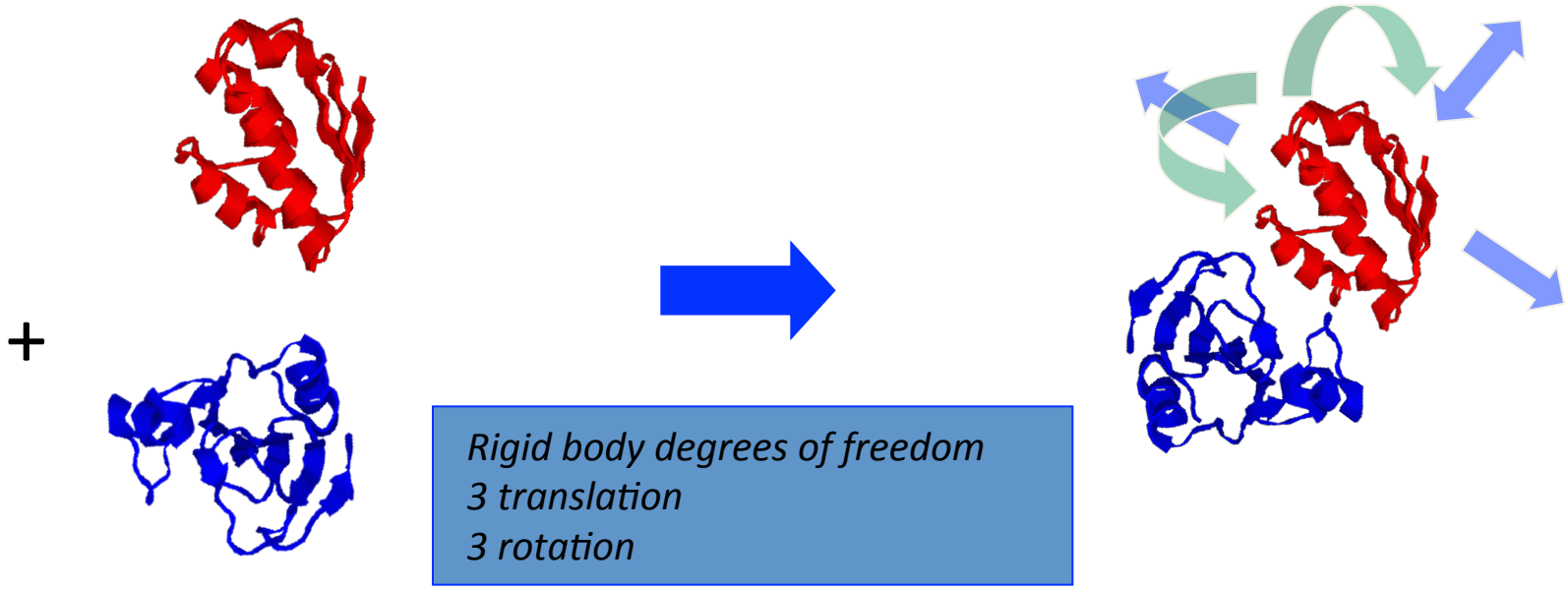


CAPRI

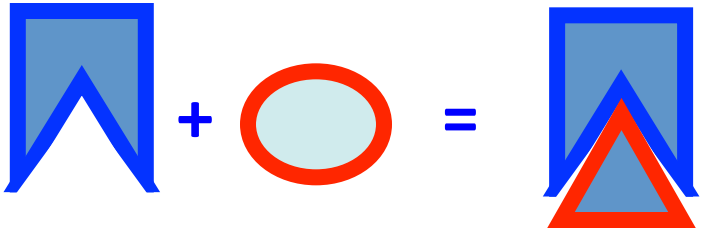
Complex

Protein-protein docking

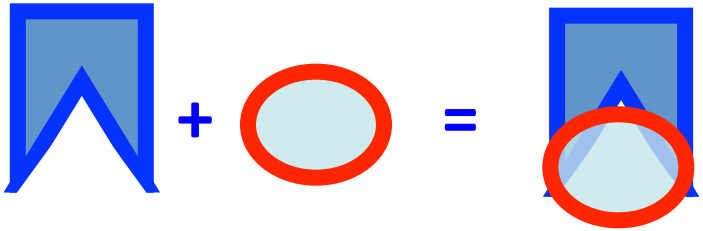
- *Aim*: predict the structure of a protein complex from its partners



Monomers change structure upon binding to partner



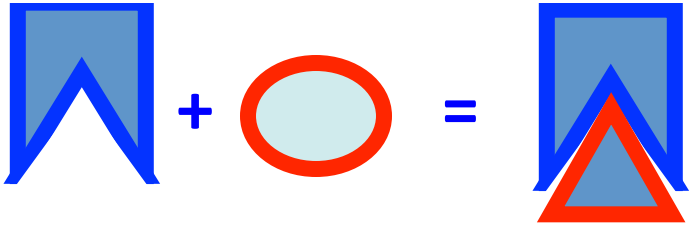
Solution 1: Tolerate clashes



✓ **Fast**

↓ Weak discrimination of correct solution

Solution 2: Model changes



↓ Slow

✓ **Precise**

Protein-protein docking

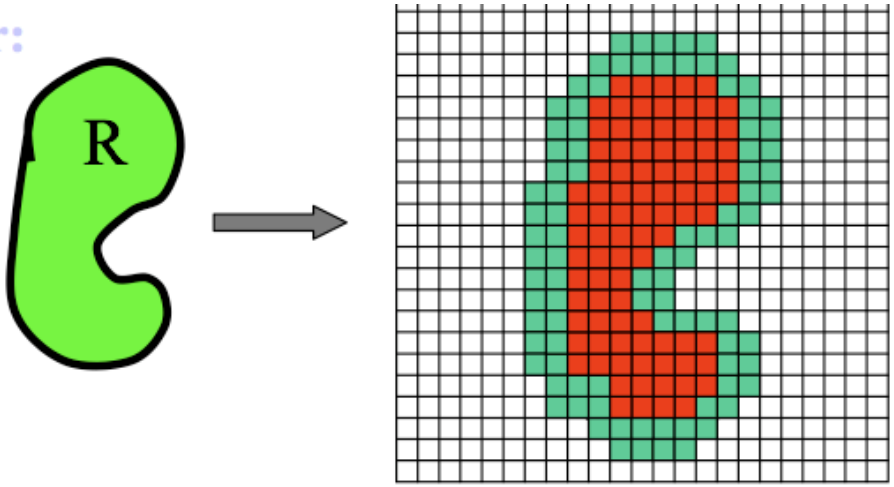
Sampling strategies

- Initial approaches: Techniques for fast detection of shape complementarity
  1. Fast Fourier Transform (FFT)
  2. Geometric hashing
- Advanced high-resolution approaches: model changes explicitly
  3. Rosettadock
- Data-driven docking
  4. Haddock

Find shape complementarity: 1. Fast Fourier Transform (FFT)

Ephraim Katzir

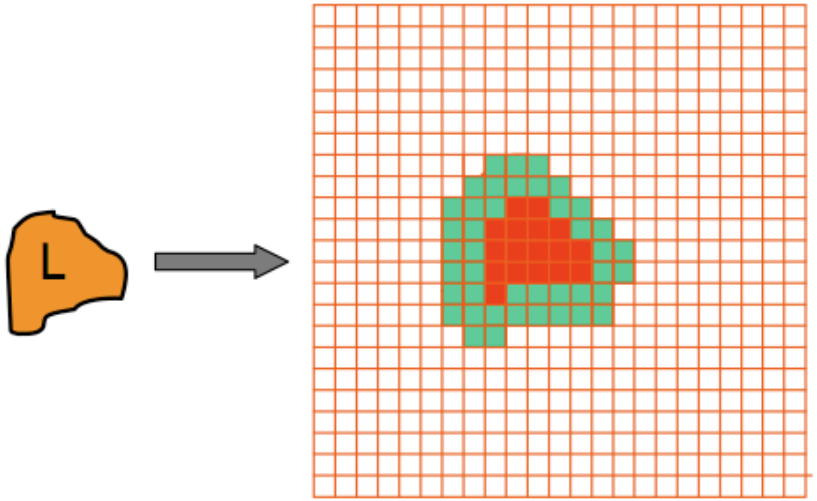
Receptor:



Assign value to each cell:

- Exterior: a(i,j) = 0
- Surface: a(i,j) = +1
- Interior: a(i,j) = -15

Ligand:

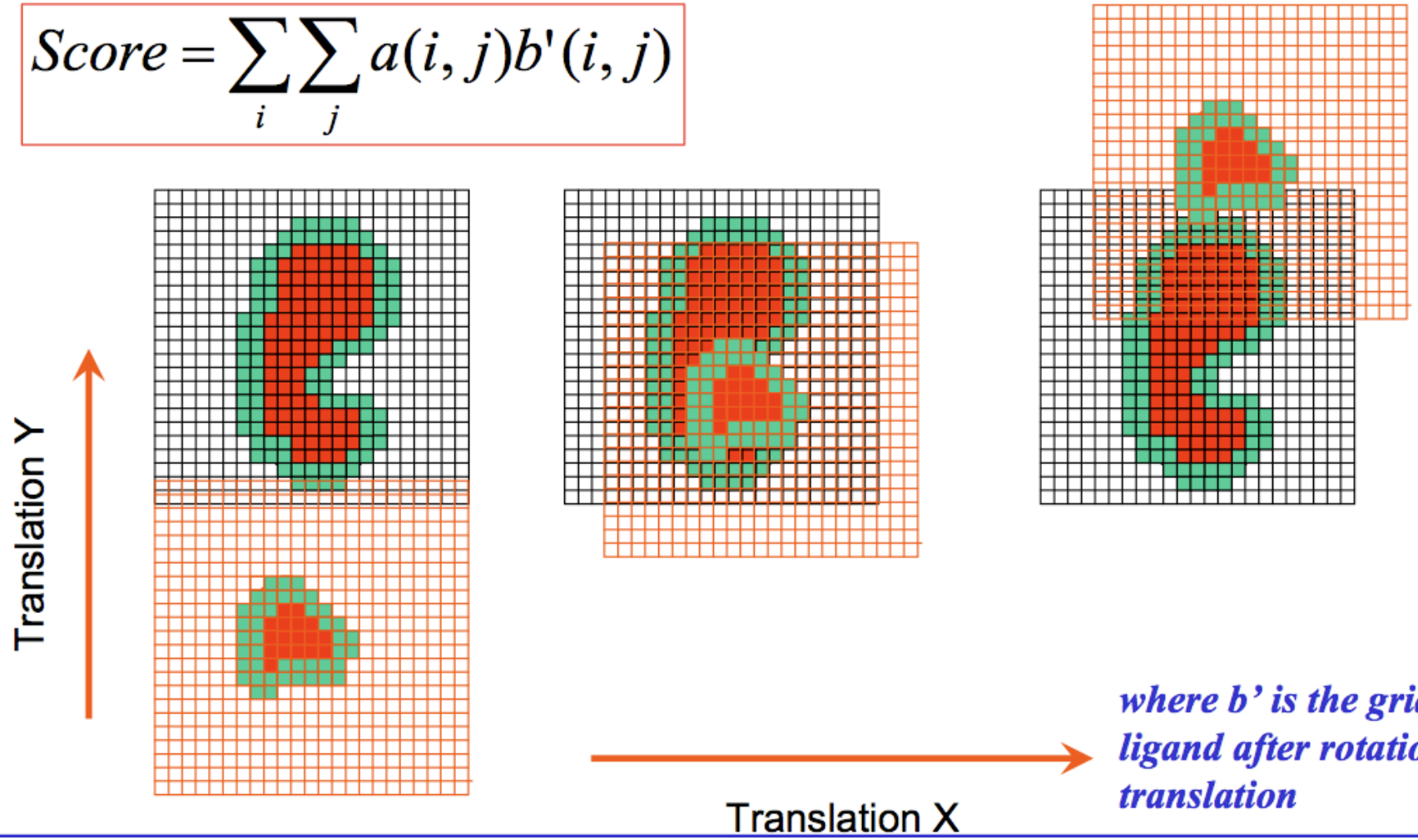


- Exterior: b(i,j) = 0
- Surface: b(i,j) = +1
- Interior: b(i,j) = +15

Find shape complementarity - FFT

Ephraim Katzir

$$Score = \sum_i \sum_j a(i,j)b'(i,j)$$



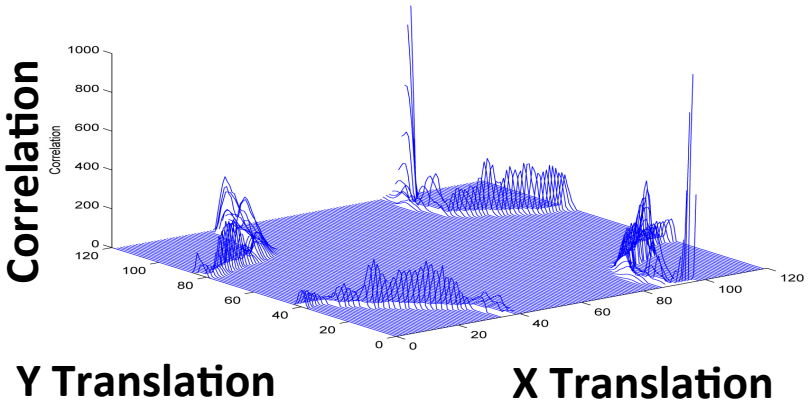
where b' is the grid for the ligand after rotation and translation

Find shape complementarity: Fast Fourier Transform (FFT)

Ephraim Katzir

Test all possible positions of ligand and receptor:

- For each rotation of ligand (R)
  - evaluate all translations (T) of ligand grid over receptor grid



$$S(R,T) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a(i,j,k)\, b'\left(i+T_x, j+T_y, k+T_z\right)$$

= correlation product: can be calculated by FFT

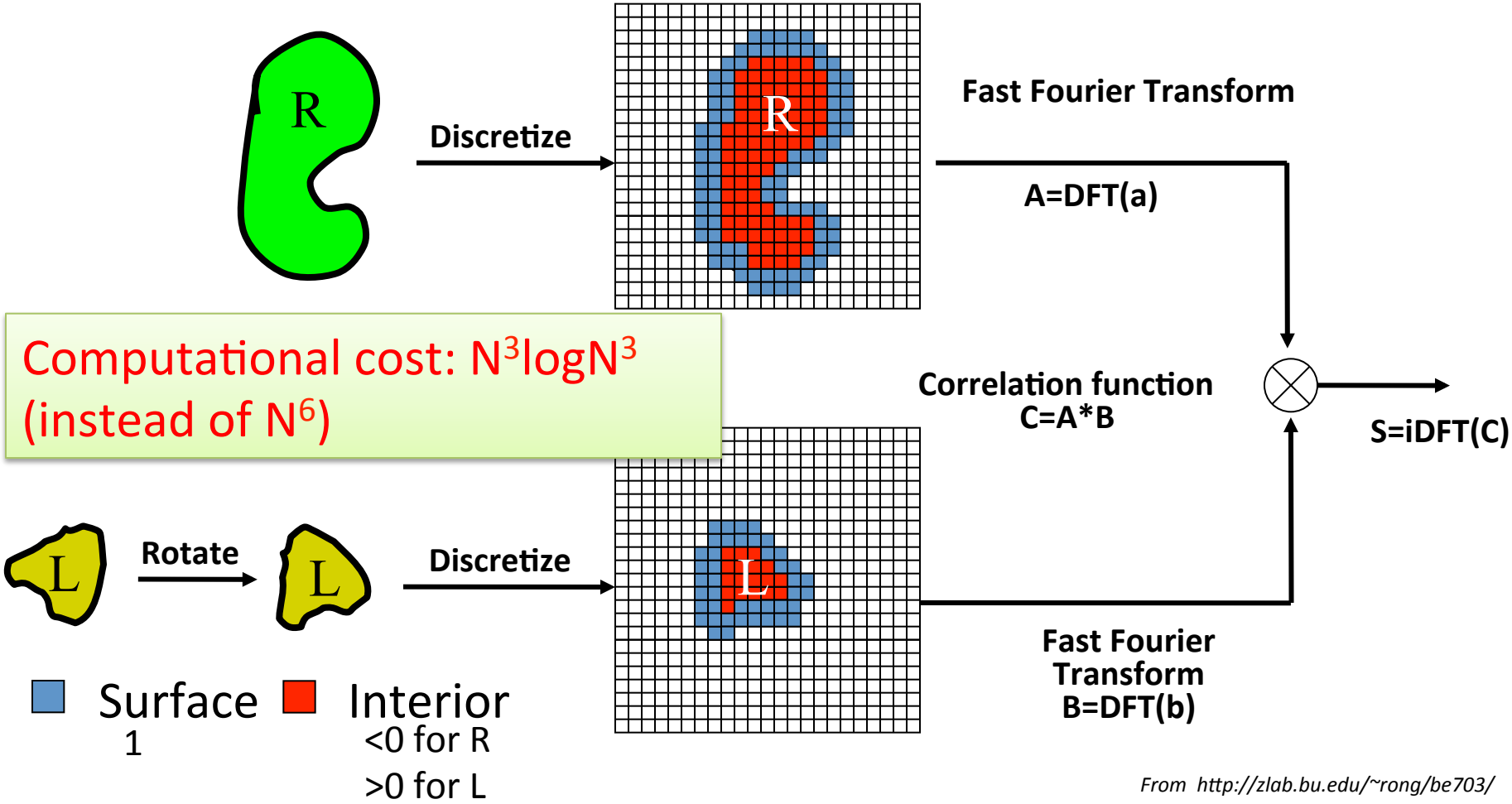
What is the time complexity in terms of N?

Parameters

- Grid interval size (η): 0.7 – 0.8 Angstrom
- Surface thickness: 1.5 – 2.5 Angstrom
- Angular step: 20°
- N*η > the size of the complex
- R used to determine if a point is in interior of a molecule (its distance with the closest atom is less than r) : 1.8 Angstrom

Find shape complementarity: Fast Fourier Transform (FFT)

Ephraim Katzir



Computational cost: $N^3 \log N^3$ (instead of N^6)

From http://zlab.bu.edu/~rong/be703/

Fast Fourier Transformation

- A simple correlation calculation is $O(N^6)$, but …

$$X_{o,p,q} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \exp[-2\pi i(ol + pm + qn)/N] \cdot x_{l,m,n}, \quad [4]$$

where $o, p, q = \{1 \ldots N\}$ and $i = \sqrt{-1}$. The application of this transformation to both sides of Eq. 3 yields (21)

$$C_{o,p,q} = A^*_{o,p,q} \cdot B_{o,p,q}, \quad [5]$$

where C and B are the DFT of the functions $\bar{c}$ and $\bar{b}$, respectively, and A^* is the complex conjugate of the DFT of

In mathematics, complex conjugates are a pair of complex numbers, both having the same real part, but with imaginary parts of equal magnitude and opposite signs

Katchalski-Katzir et al, PNAS, 1991.

Fast Fourier Transformation

$\bar{a}$. Eq. **5** indicates that the transformed correlation function C is obtained by a simple multiplication of the two functions A^* and B. The inverse Fourier transform (20) (IFT), defined as

$$\bar{c}_{\alpha,\beta,\gamma} = \frac{1}{N^3} \sum_{o=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} \exp[2\pi i(o\alpha + p\beta + q\gamma)/N] \cdot C_{o,p,q}, \quad [6]$$

is used to obtain the desired correlation between the two original functions $\bar{a}$ and $\bar{b}$. The Fourier transformations can be performed with the fast Fourier transform algorithm (20), which requires less than the order of $N^3 \ln(N^3)$ steps for transforming a 3D function of $N \times N \times N$ values. Thus, the overall procedure leading to Eq. **6** is significantly faster than the direct calculation of $\bar{c}$ according to Eq. **3**.

Algorithm

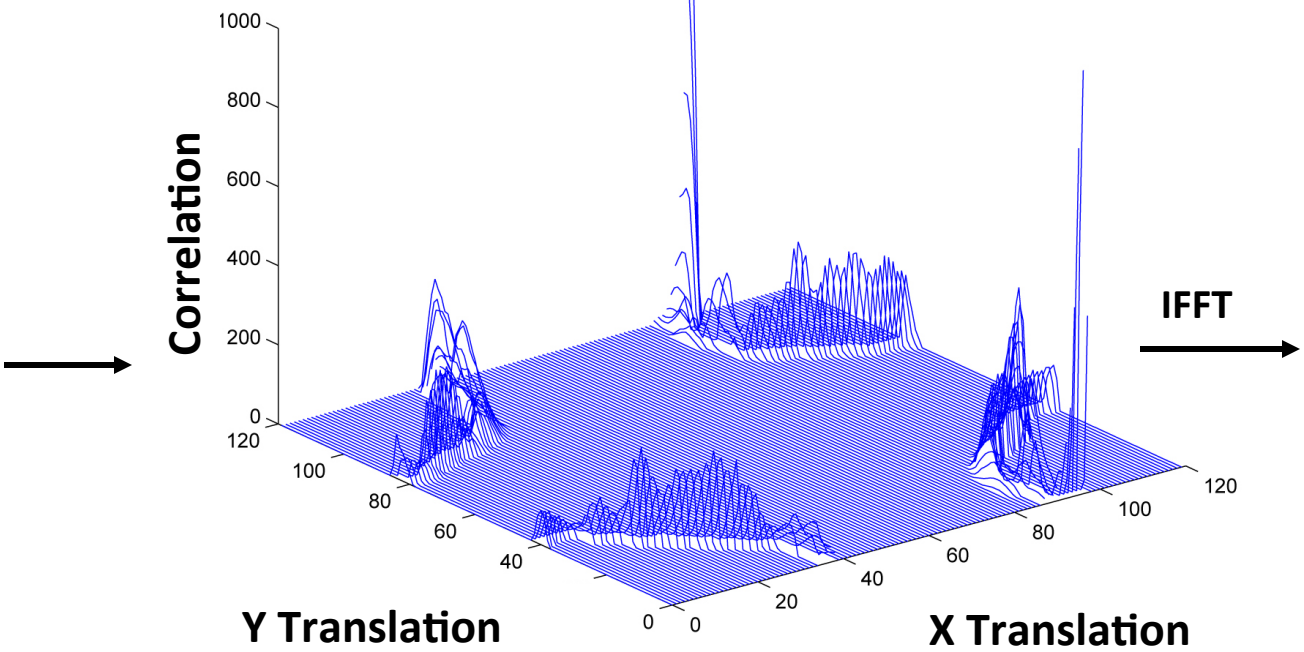
Finally, to complete a general search for a match between the surfaces of molecules **a** and **b**, the correlation function $\bar{c}$ has to be calculated for all relative orientations of the molecules. In practice, molecule **a** is fixed, whereas the three Euler angles defining the orientation of molecule **b** (*xyz* convention in ref. 22) are varied at fixed intervals of Δ degrees. This results in a complete scan of $360 \times 360 \times 180/\Delta^3$ orientations for which the correlation function $\bar{c}$ must be calculated.

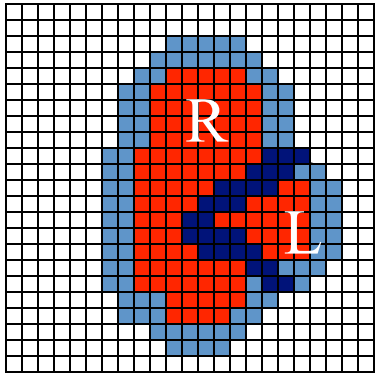
The entire procedure described above can be summarized by the following steps:

(*i*) derive $\bar{a}$ from atomic coordinates of molecule **a** (Eq. **2**),
(*ii*) $A^* = [\mathrm{DFT}(\bar{a})]^*$ (Eq. **4**),
(*iii*) derive $\bar{b}$ from atomic coordinates of molecule **b** (Eq. **2**),
(*iv*) $B = \mathrm{DFT}(\bar{b})$ (Eq. **4**),
(*v*) $C = A^* \cdot B$ (Eq. **5**),
(*vi*) $\bar{c} = \mathrm{IFT}(C)$ (Eq. **6**),
(*vii*) look for a sharp positive peak of $\bar{c}$,
(*viii*) rotate molecule **b** to a new orientation,
(*ix*) repeat steps *iii*–*viii* and end when the orientations scan is completed, and
(*x*) sort all of the peaks by their height.

Find shape complementarity: Fast Fourier Transform (FFT)

Increase the speed by 10^7





Surface Interior Binding Site

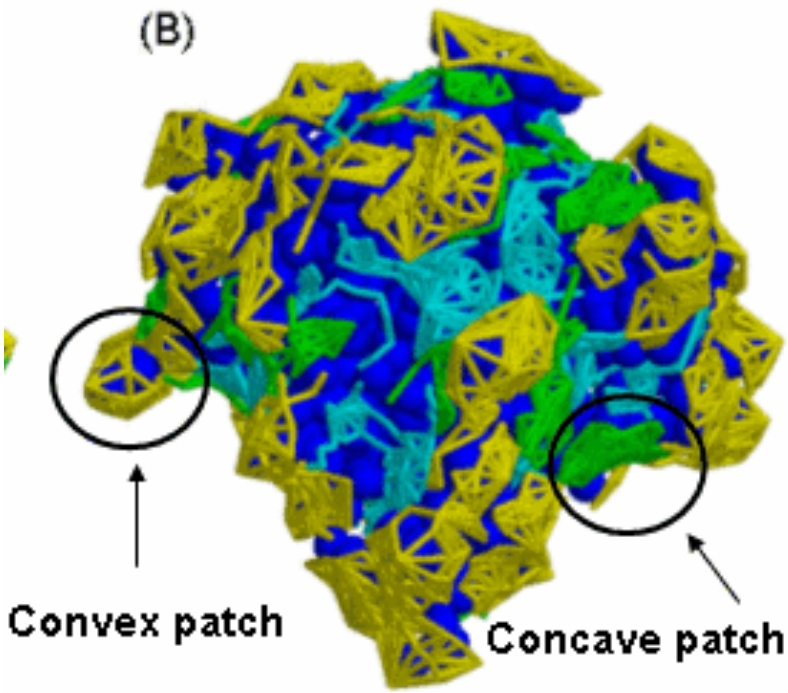
From http://zlab.bu.edu/~rong/be703/

Some FFT-based docking protocols

- Zdock (Weng)
- Cluspro (Vajda, Camacho)
- PIPER (Vajda, Kozakov)
- Molfit (Eisenstein)
- DOT (TenEyck)
- HEX (Ritchie) – FFT in rotation space

Shape complementarity: 2. Geometric hashing *(patchdock, Wolfson & Nussinov)*

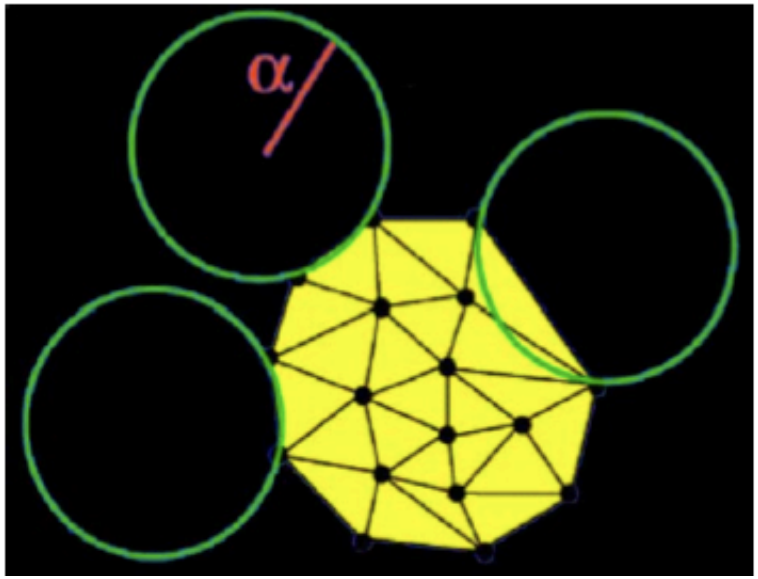
- Matching of puzzle pieces
  1. Define geometric patches (concave, convex, flat)
  2. Surface patch matching
  3. Filtering and scoring

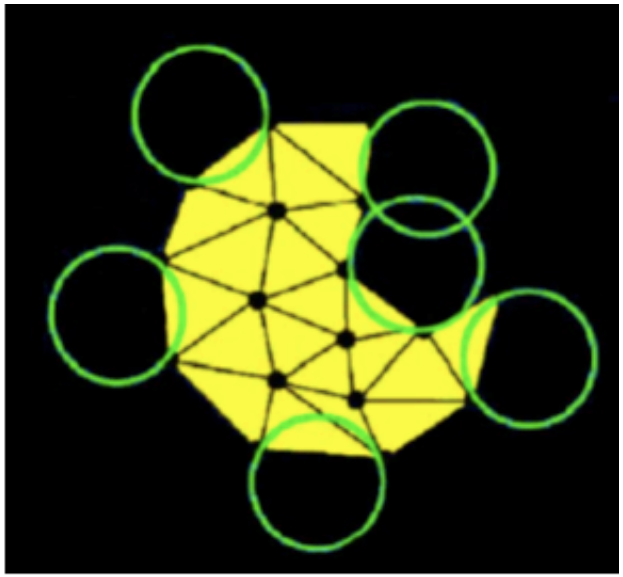


From http://bioinfo3d.cs.tau.ac.il/PatchDock/patchdock.html

Hashing: alpha shapes

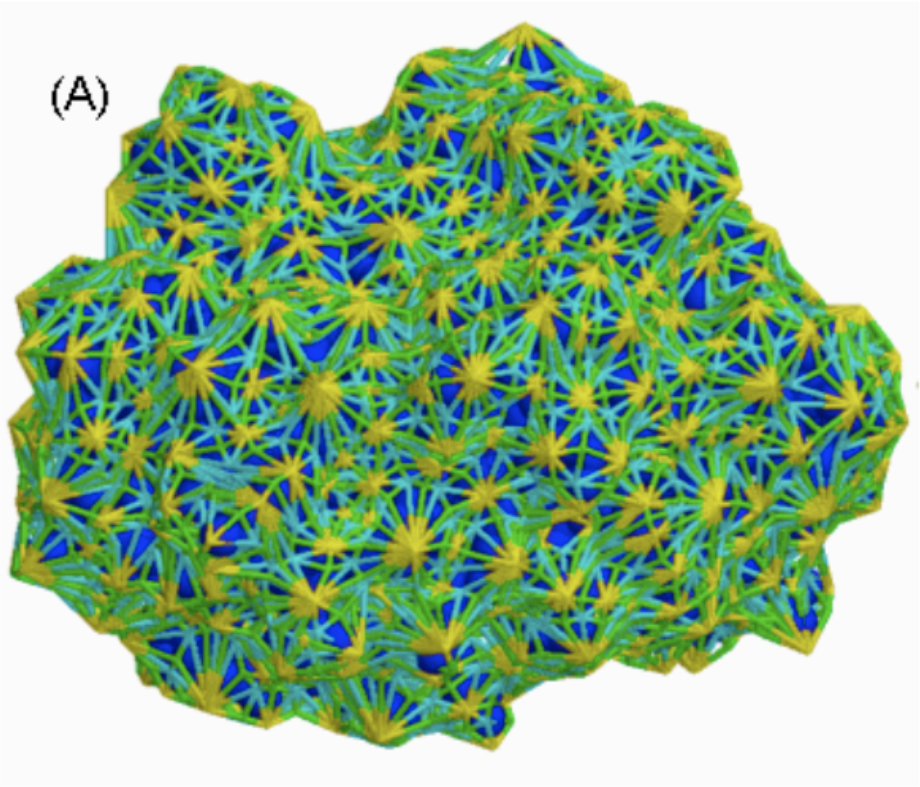
- Formalizes the idea of "shape"
- In 2D an "edge" between two points is "alpha-exposed" if there exists a circle of radius alpha such that the two points lie on the surface of the circle and the circle contains no other points from the point set

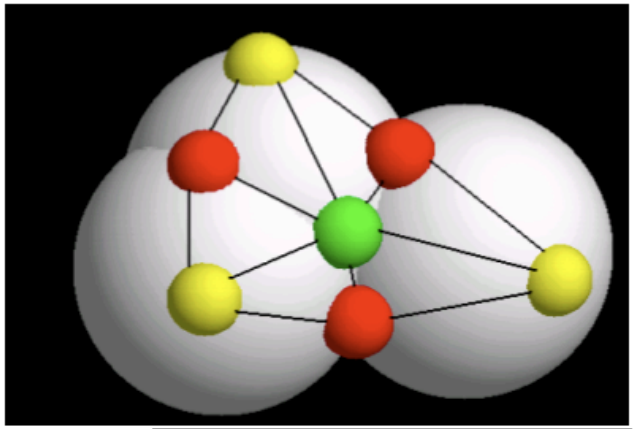


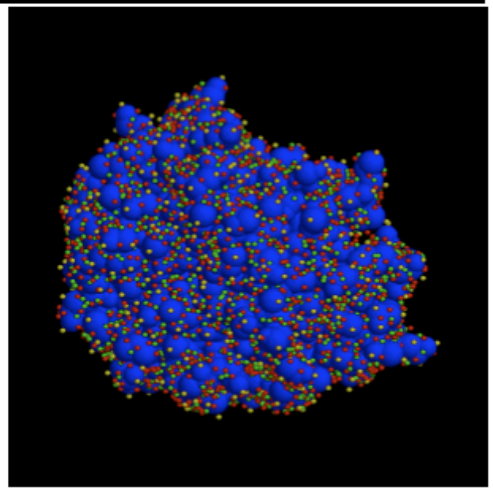


Hashing – sparse surface representation

- ***Caps, pits, belts:***







Slide from Jens Meiler

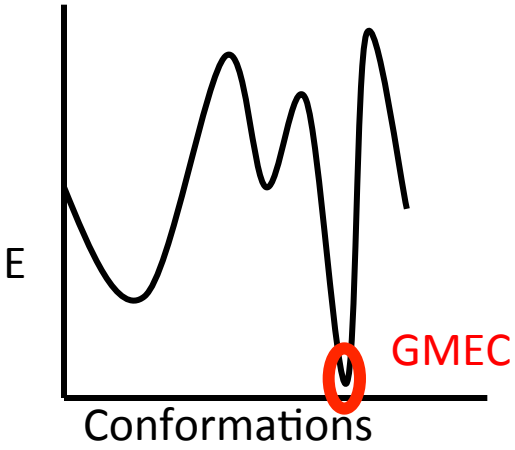
Docking with geometric hashing

PATCHDOCK

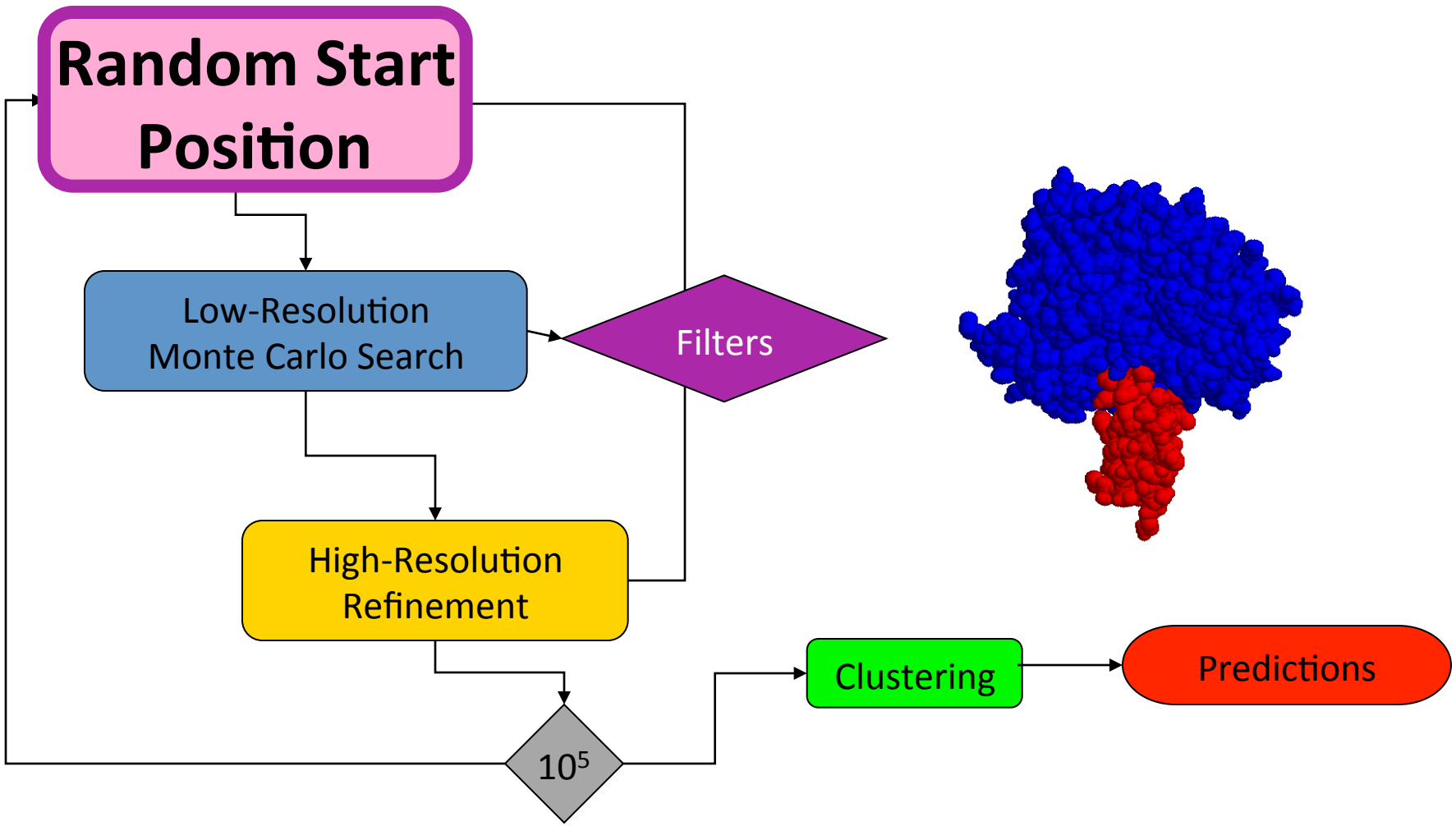
- Fast and versatile approach
- Speed allows easy extension to multiple protein docking, flexible hinge docking, etc
- A extension of this protocol, FIREDOCK, includes side chain optimization (RosettaDock-like) – very flexible, fast and accurate protocol

3. High-resolution docking: Explicit modeling of conformational changes

- Parameters:
  - **energy function** (Native structure should be near global energy minimum conformation, GMEC)
  - **sampling strategy** (Locate energy minimum efficiently)
  - energy function and sampling strategy are coupled

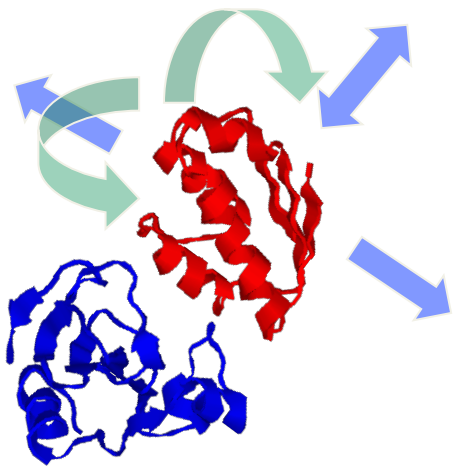


Rosettadock algorithm



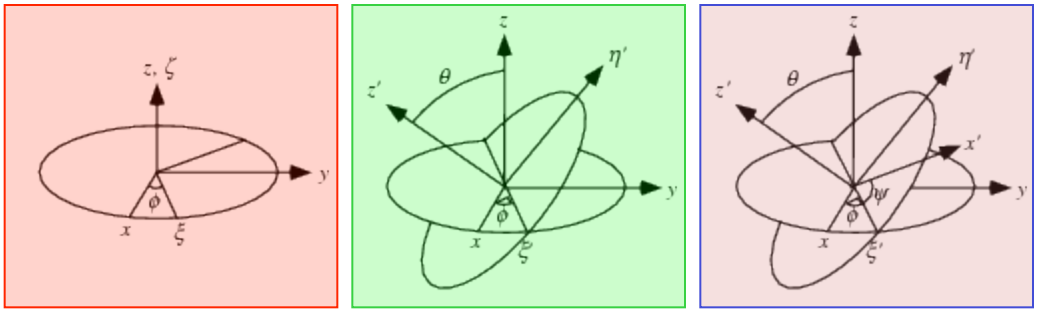
Choosing starting orientations

1. Global search
   - Random Translation
   - Random Rotation (Euler Angles)



1. Tilt direction [0..360°]
2. Tilt angle [0:90°]
3. Spin angle [0..360°]

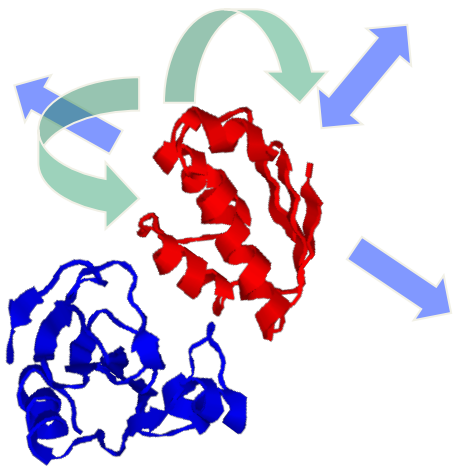
- Euler angles are independent and guarantee non-biased search



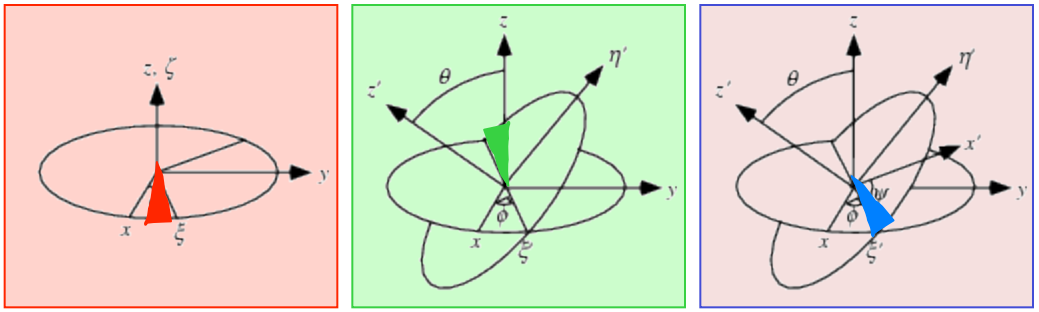
Choosing starting orientations

2. Local Refinement

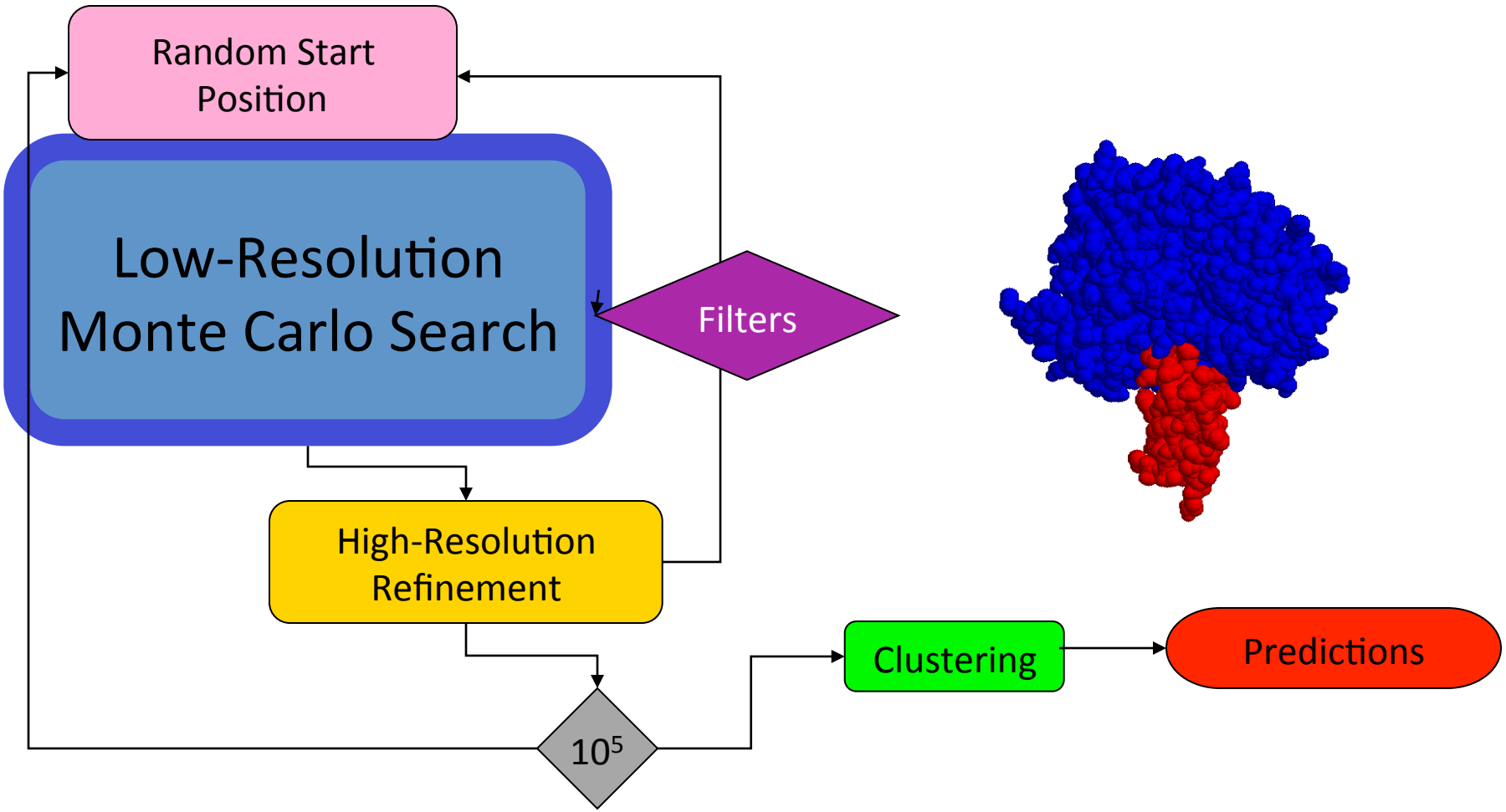
- Translation 3Å normal, 8Å parallel
- Rotation 8^0



1. Tilt direction [0±8°]
2. Tilt angle
3. Spin angle



Overview of docking algorithm

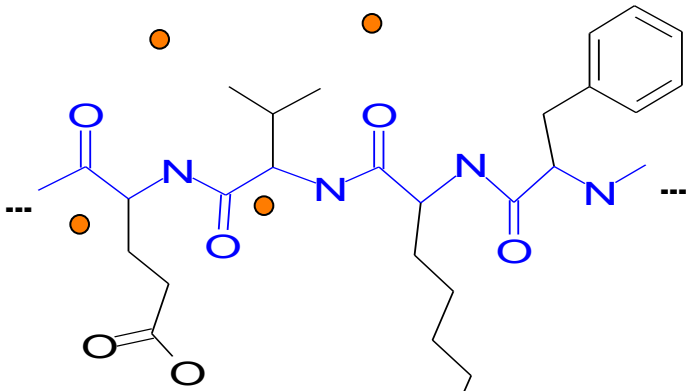


Low-resolution search

1. Perturbation
2. Monte Carlo search
3. Rigid body translations and rotations
4. Residue-scale interaction potentials

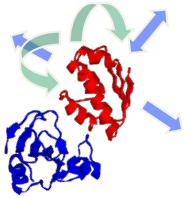
Protein representation:

backbone atoms + *average centroids*



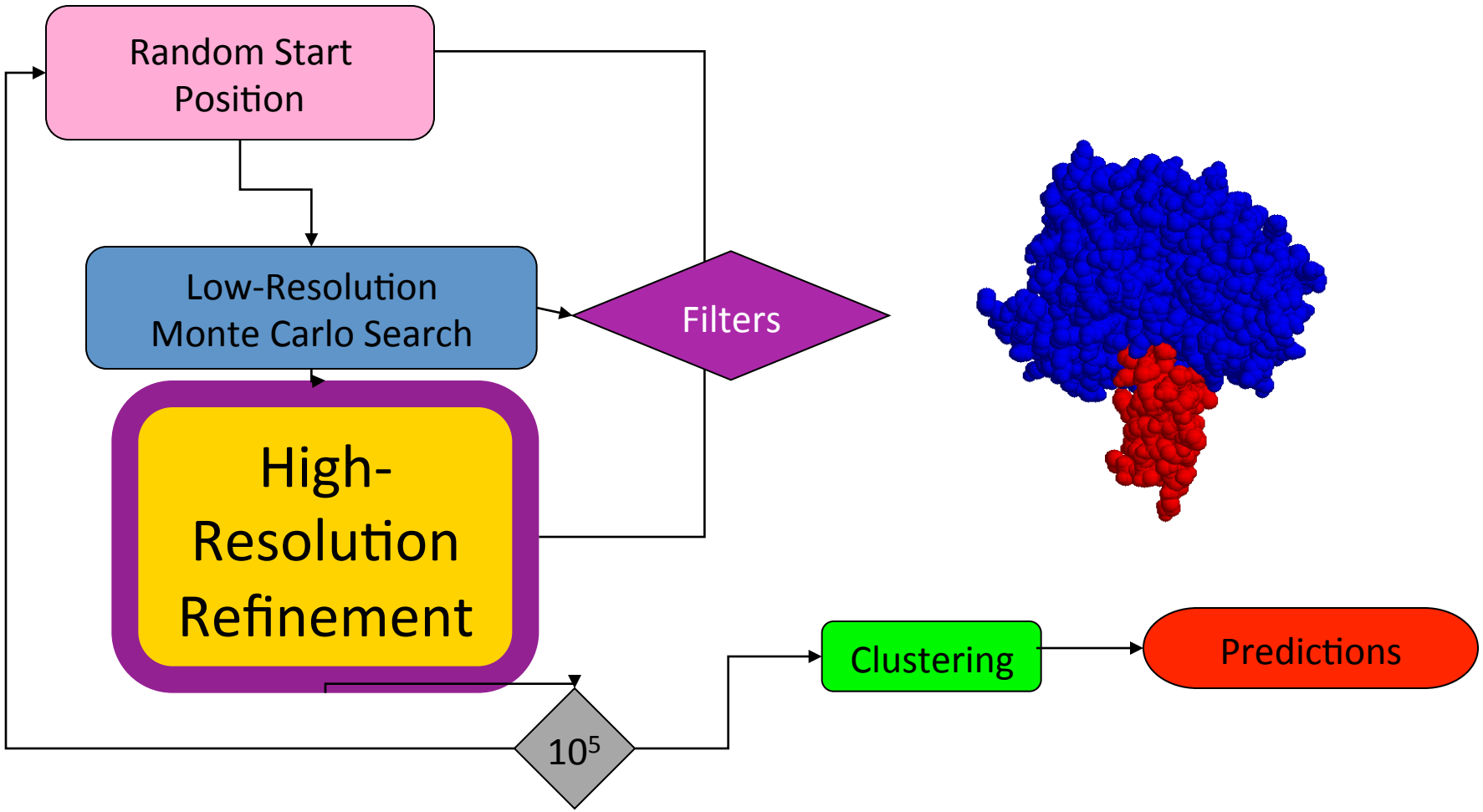
- Mimics physical diffusion process

Residue-scale scoring



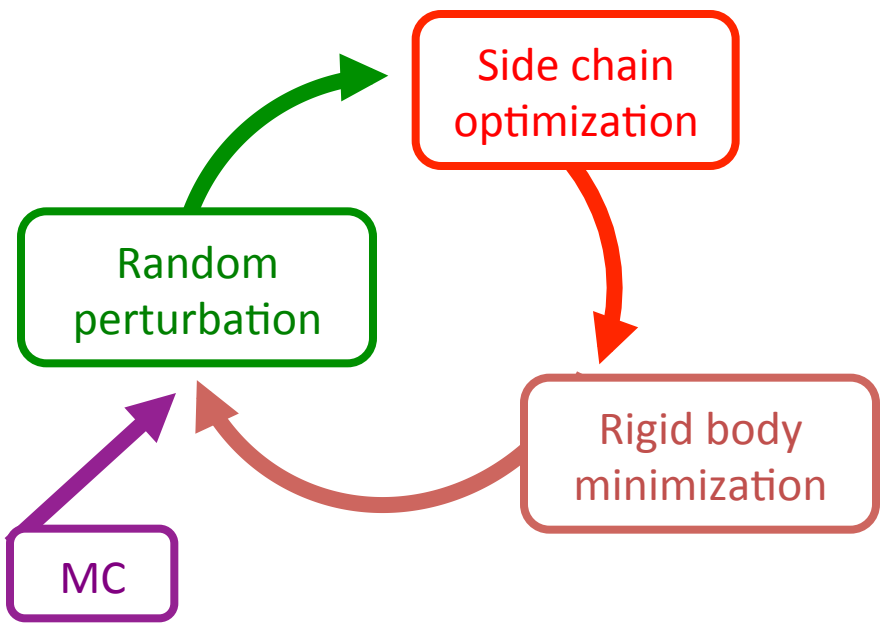
Score	Representation	Physical Force
Contacts	$r_{\text{centroid-centroid}} < 6$ Å	Attractive van der Waals
Bumps	$(r - R_{ij})^2$	Repulsive van der Waals
Residue environment	$-\ln(P_{\text{env}})$	Solvation
Residue pair	$-\ln(P_{ij})$	Hydrogen bonding electrostatics, solvation
Alignment	-1 for interface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

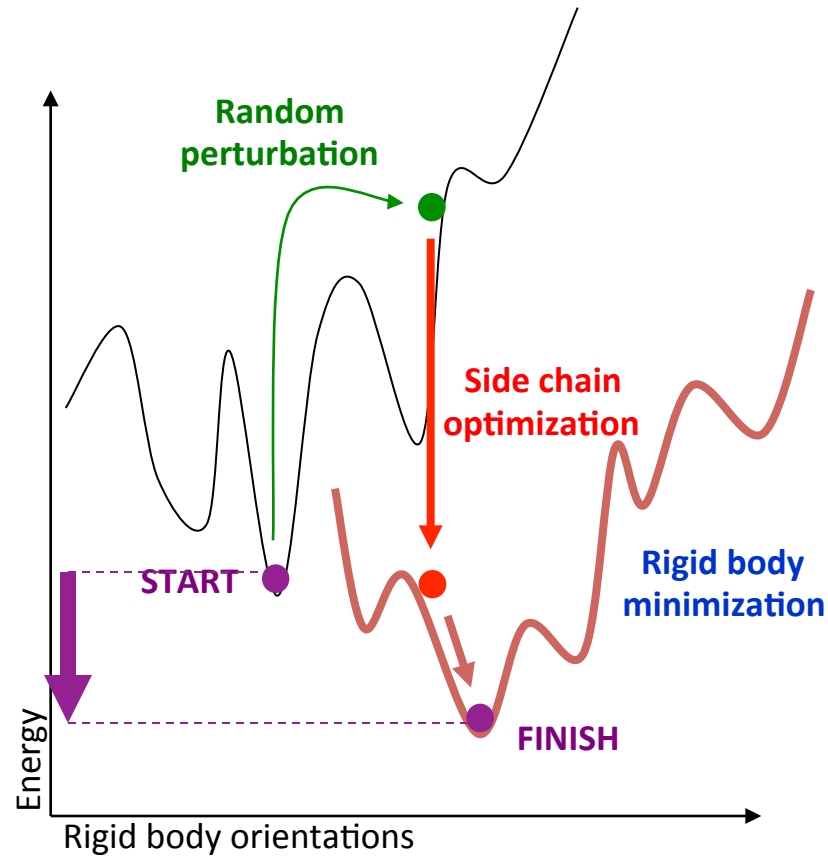
Overview of docking algorithm



High resolution optimization: Monte Carlo with Minimization *(MCM)*

Cycles of iterative optimization





Energy-based model selection

Low-energy models are all accurate

Protocol depends on:

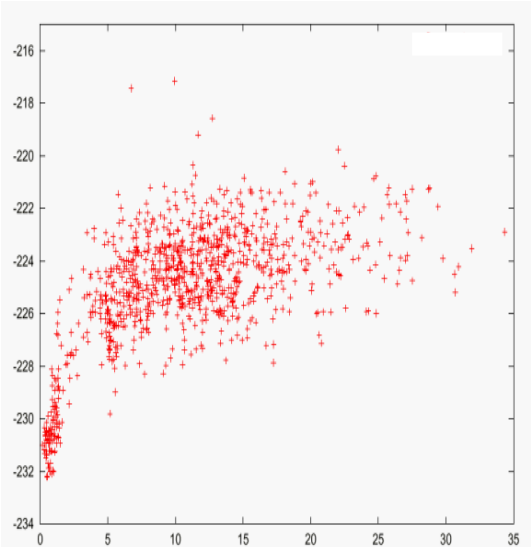
1. Sampling Strategy

   Sample near-native conformation

2. Energy Function

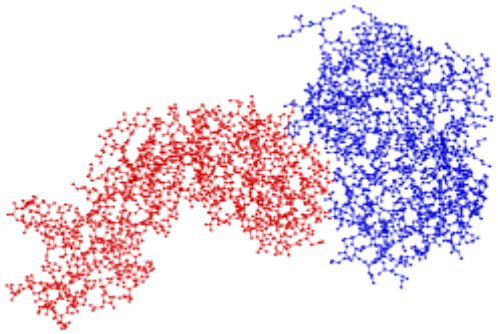
   Energy Function and Sampling are coupled

3. Sampling Intensity

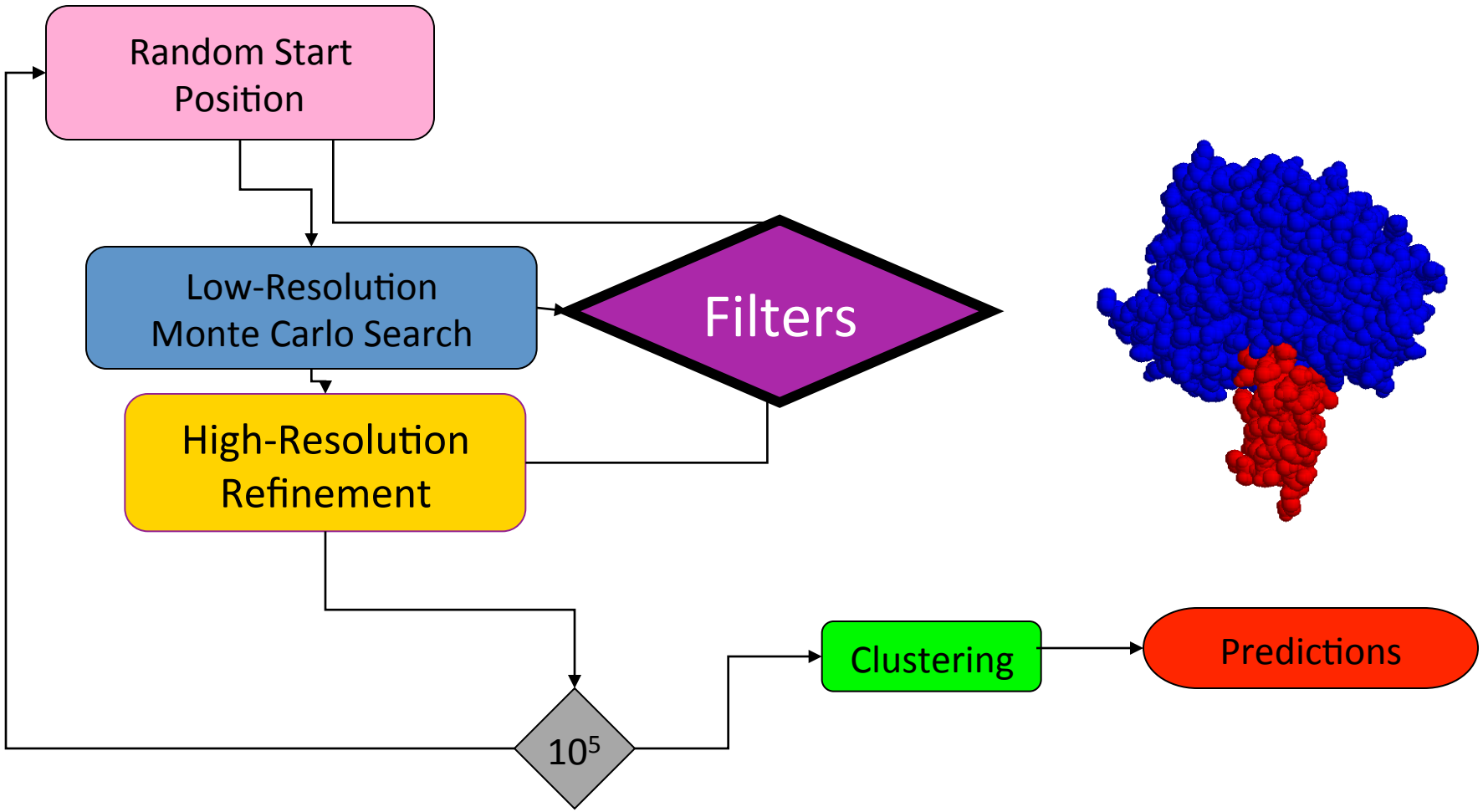


Full-atom scoring

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)

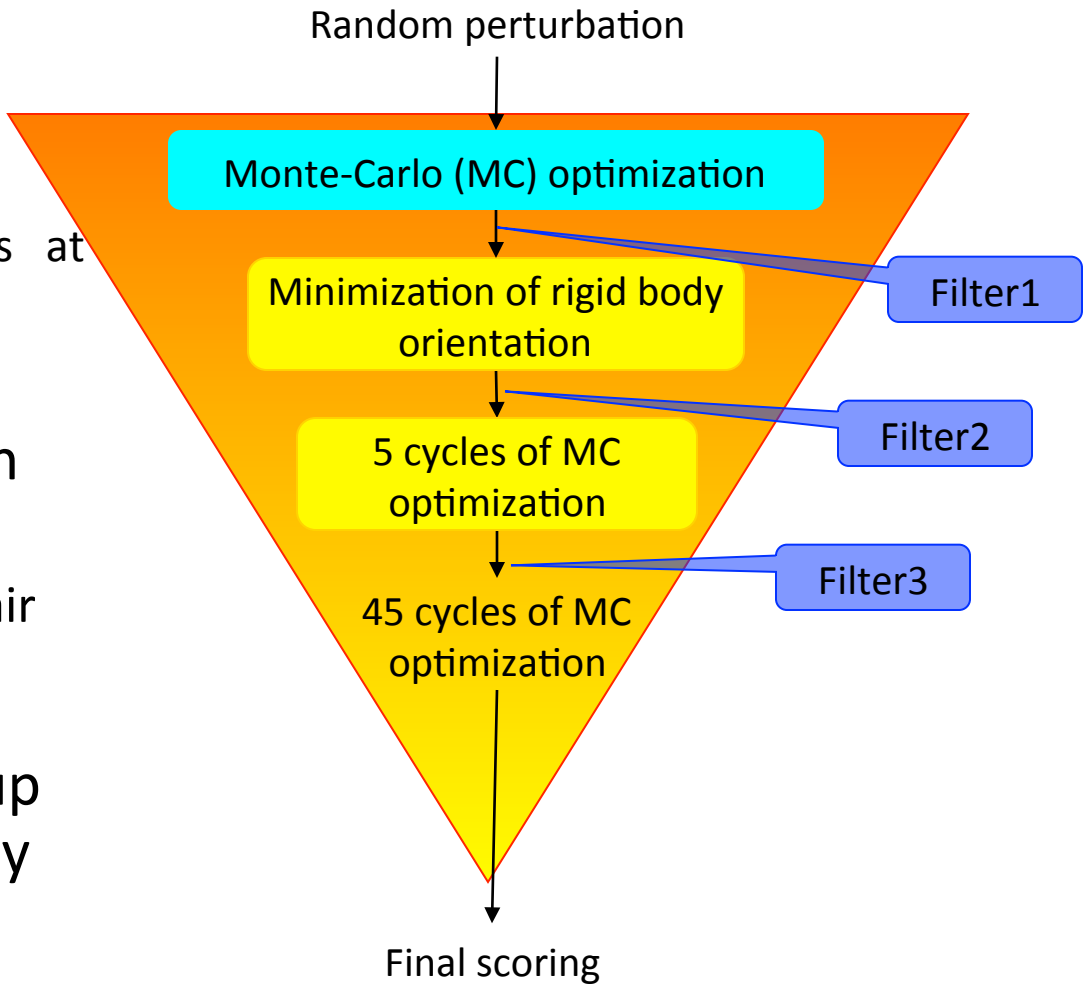


Overview of docking algorithm



Filters

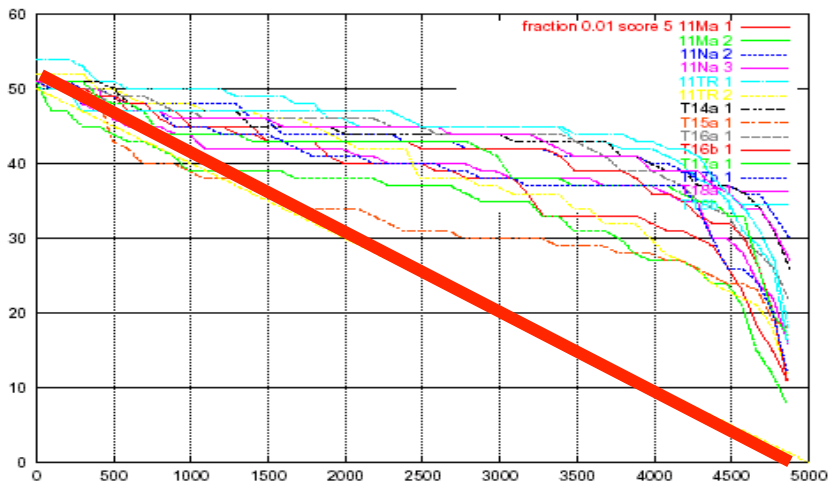
- Low resolution
  - Antibody profiles
    - Antigen binding residues at interface
  - Contact filters
  - Biological information
    - Interface residues
    - Interacting residue pair
- High resolution
  - Energy filters speed up creation of low energy models



Energy filters

Enrichment = $\dfrac{\text{Fraction of “good decoys” after applying filter}}{\text{Fraction of “good decoys” before applying filter}}$

TT

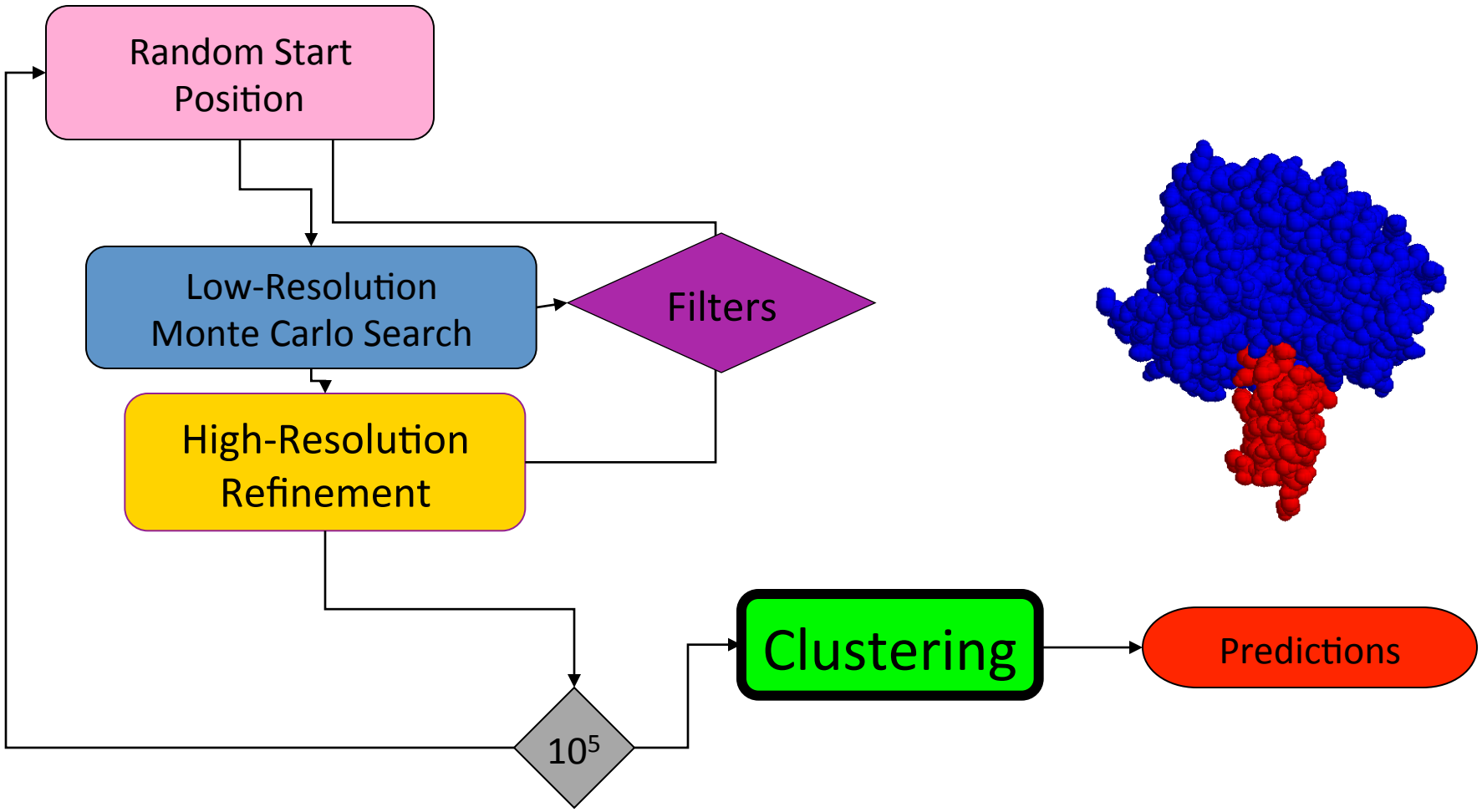


FF

ROC curve

	Pass filter	Do not pass filter
“bad decoys” high final energy	**TF**	**FF**
“good decoys” low final energy	**TT**	**FT**

Overview of docking algorithm

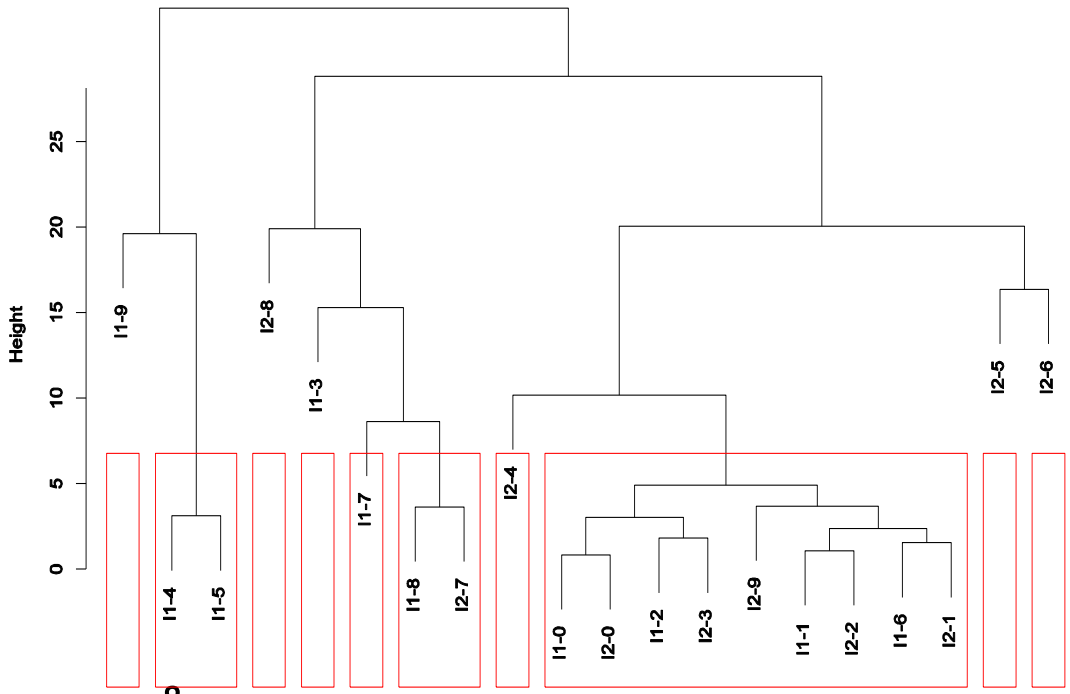


Clustering

- Compare all top-scoring decoys pairwise

$$\text{rmsd} = \sqrt{\sum_i |x_i - y_i|^2}$$

- Cluster decoys hierarchically



- Decoys within e.g. 2.5Å form a cluster

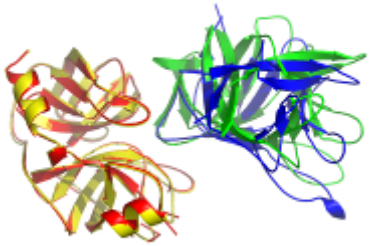
Represents ENTROPY

Assessment 1: Benchmark studies

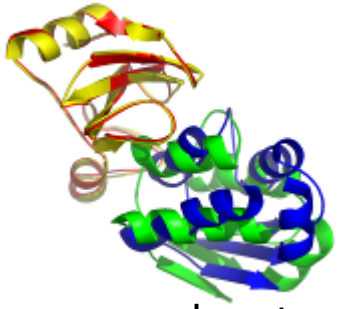
Benchmark set contains 54 targets for which *bound* and *unbound* structures are known

http://zlab.bu.edu/zdock/benchmark.shtml

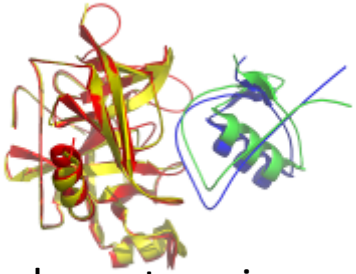
- Bound-Bound
  - Start with bound complex structure, but remove the side chain configurations so they must be predicted
- Unbound-Unbound
  - Start with the individually-crystallized component proteins in their unbound conformation
- Bound-Unbound (Semibound)



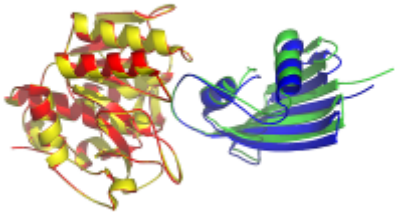
trypsin + inhibitor



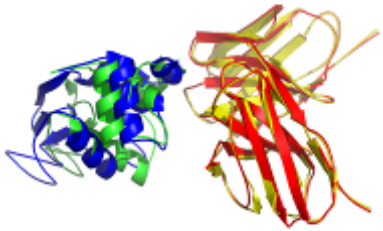
barnase + barstar

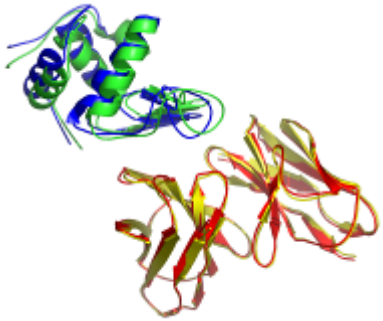


α-chymotrypsin
+ inhibitor

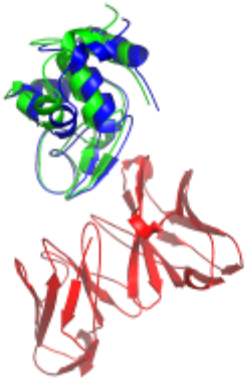


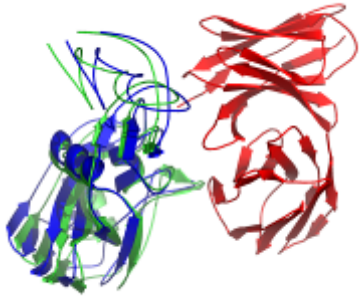
subtilisin + inhibitor



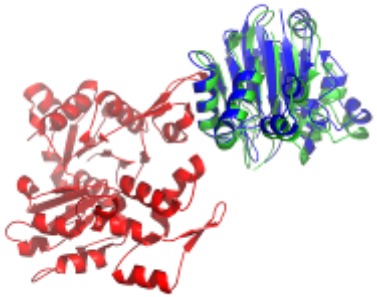


lysozyme + antibodies

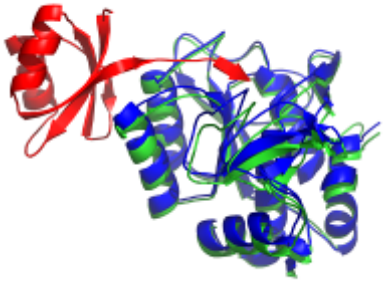




hemagglutinin
+ antibody



actin + deoxyribonuclease I

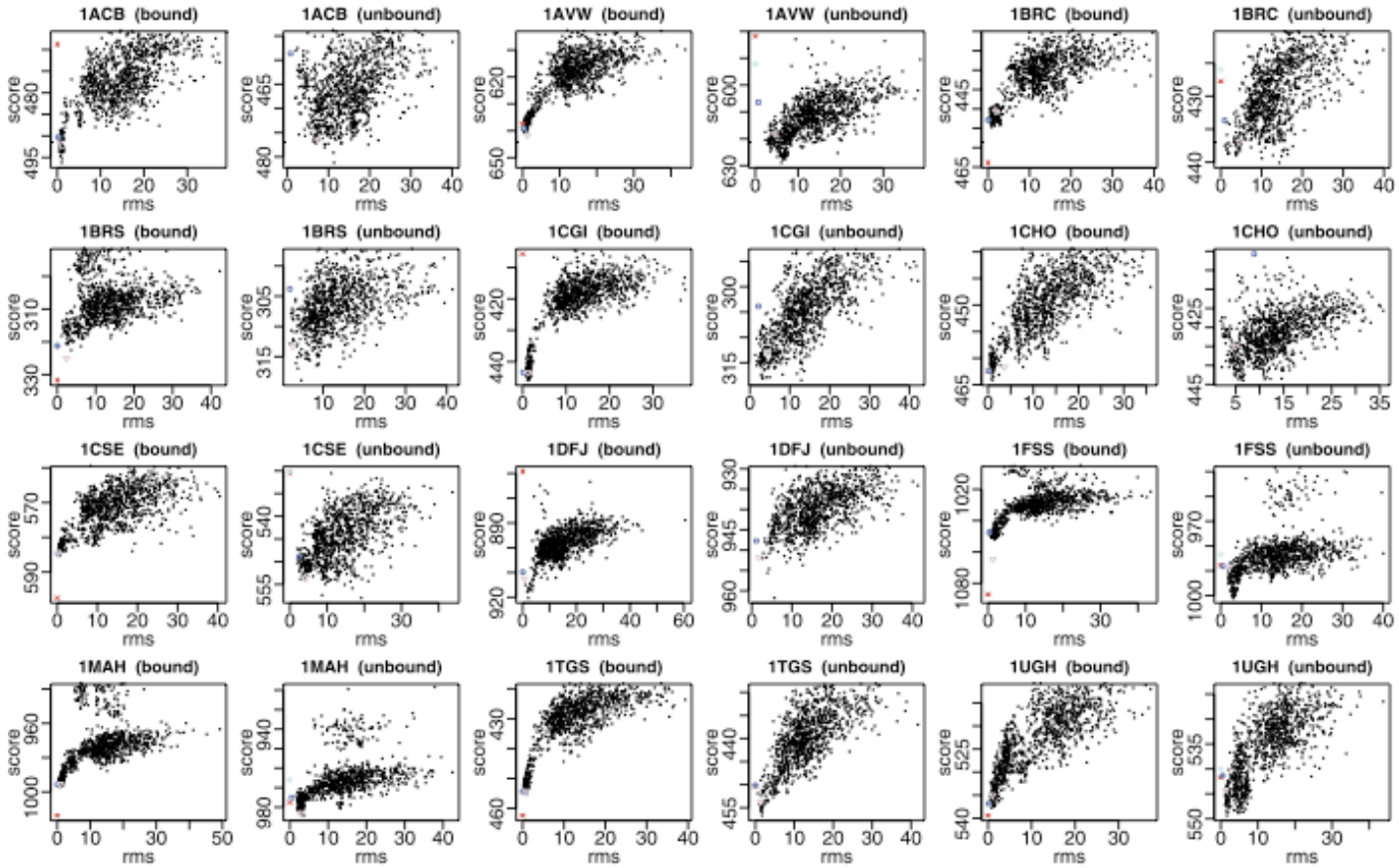


subtilisin + prosegment

Assessment of method on benchmark

(54 proteins, Gray et al., 2003)

- funnel - 3/5 top-scoring models within 5A rmsd



.....

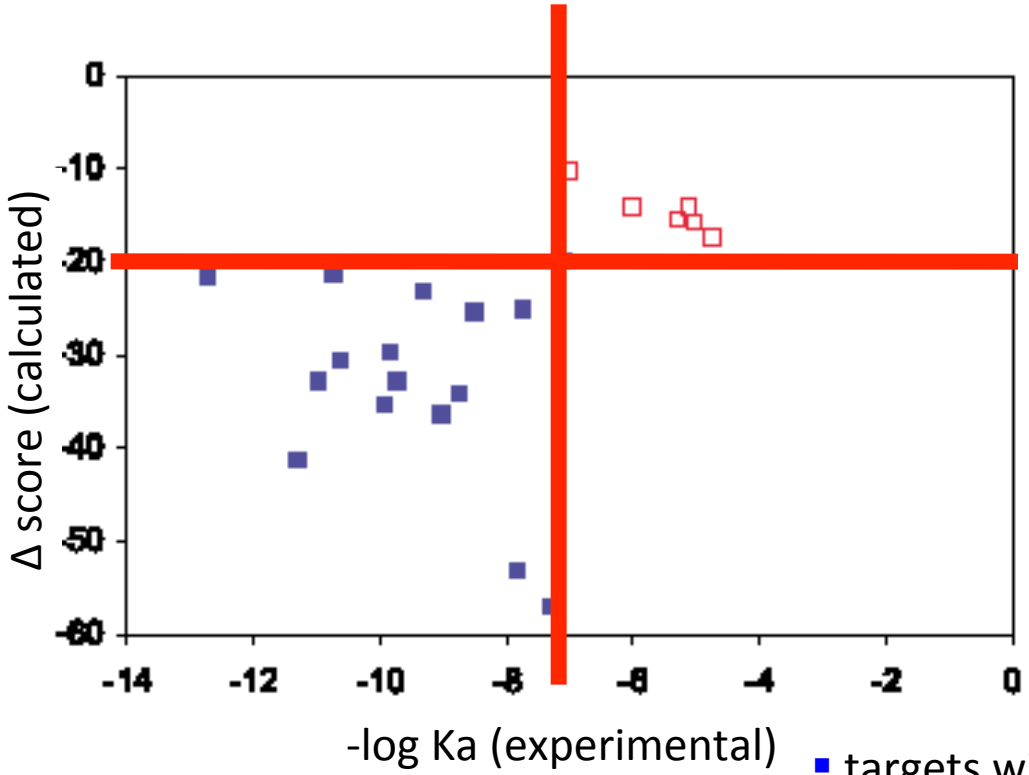
RosettaDock benchmark performance

Docking Benchmark	Bound Docking Perturbation[1]	Unbound Docking Perturbation[2]	Unbound Docking Global[3]
Enzyme/Inhibitor	21/22	18/22	17/18
Antigen/Antibody	10/16	9/16	8/9
Others	5/10	5/10	3/5
Difficult	6/6	0/6	N/A
Total	42/54	32/54	28/32

1. More than **three** of top **five** decoys (by score) that have rmsd less than **5** Å
2. More than three of top five decoys (by score) that predict more than 25% native residue contacts
3. The rank of the first cluster with >**25%** native residue contacts

Benchmark: R. Chen *et al*, 2003 ; RosettaDock: Gray *et al* , 2003

Score and performance are correlated with binding affinity



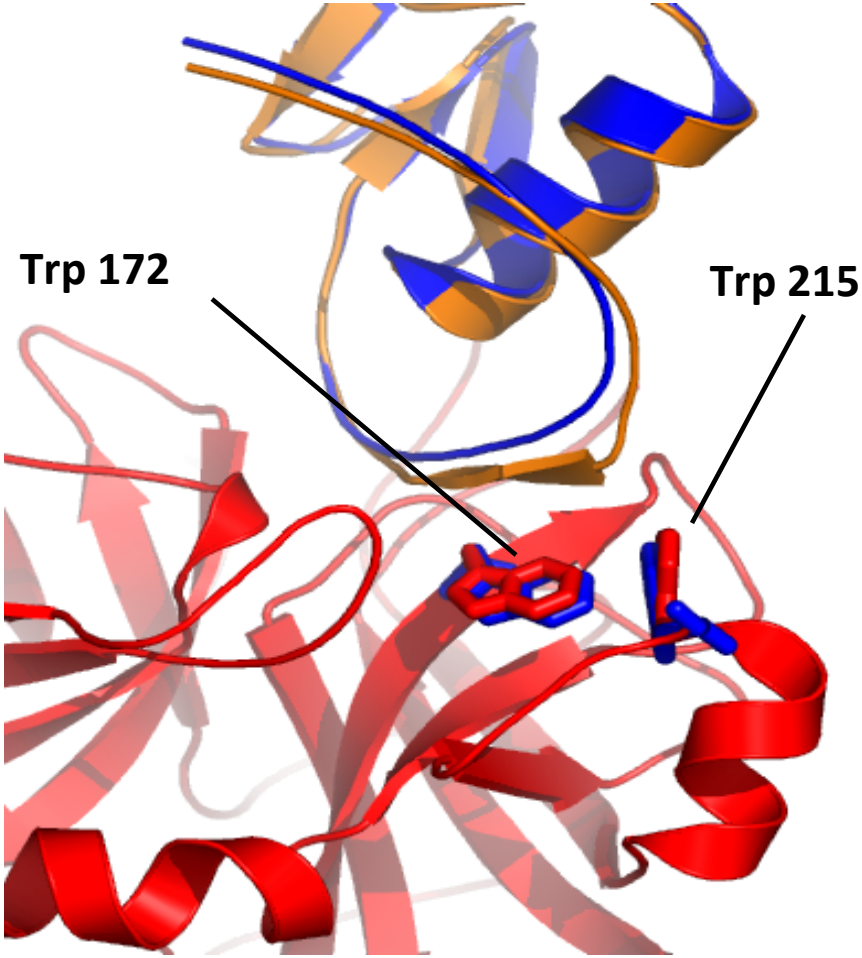
▪ targets with funnels

▫ targets without funnels

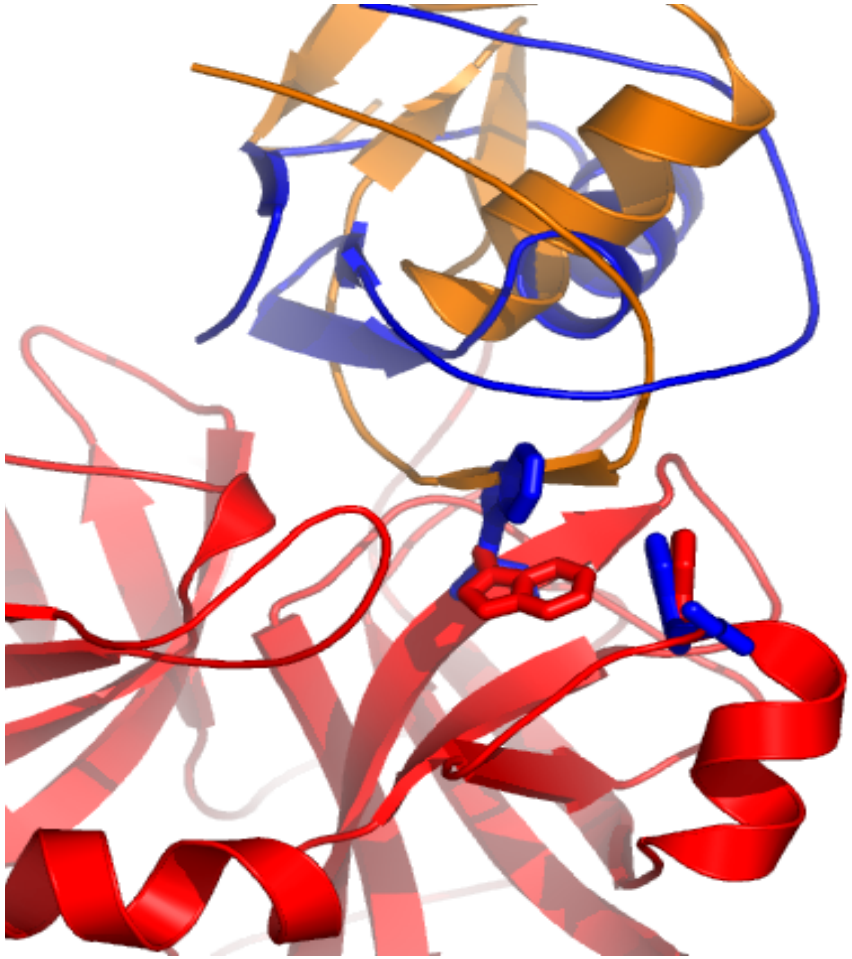
Δ score for bound backbone docking

Limitation of "rotamer-based" modeling

Near-native model with clash



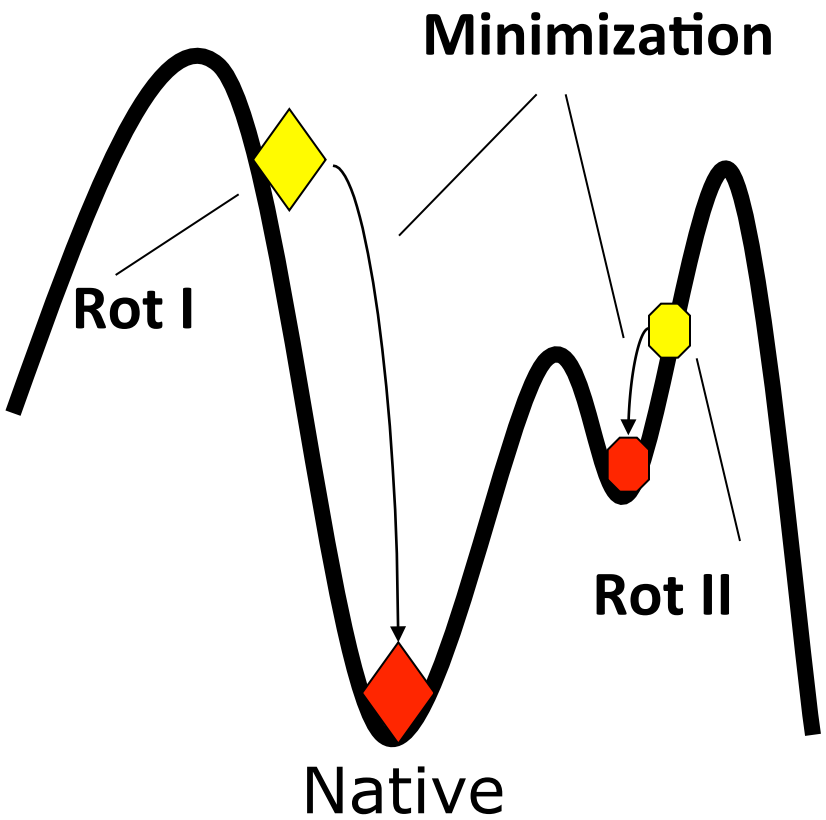
Non-native model without clash



Orange and red: native complex; Blue: docking model.

PDB code: 1CHO

Improved side chain modeling at interface



Rtmin: rotamer trial with minimization

- Randomly pick one residue.
- Screen a list of rotamers.
- Minimize each of these rotamers.
- Accept the one that yields the lowest energy.

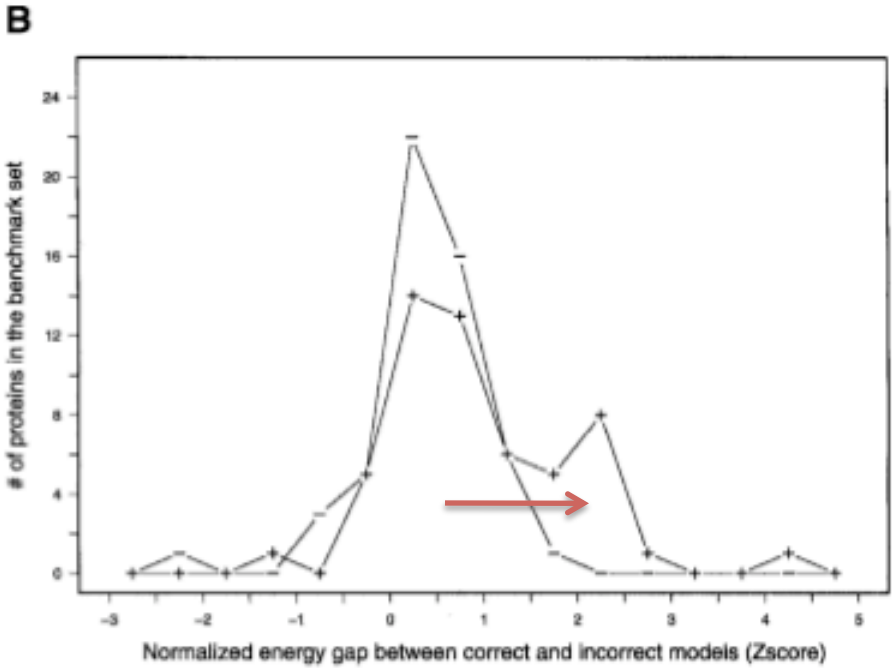
Additional rotamers

- Include free side chain conformation in rotamer library

Wang, OSF & Baker, 2005

More accurate side chain modeling improves predictions

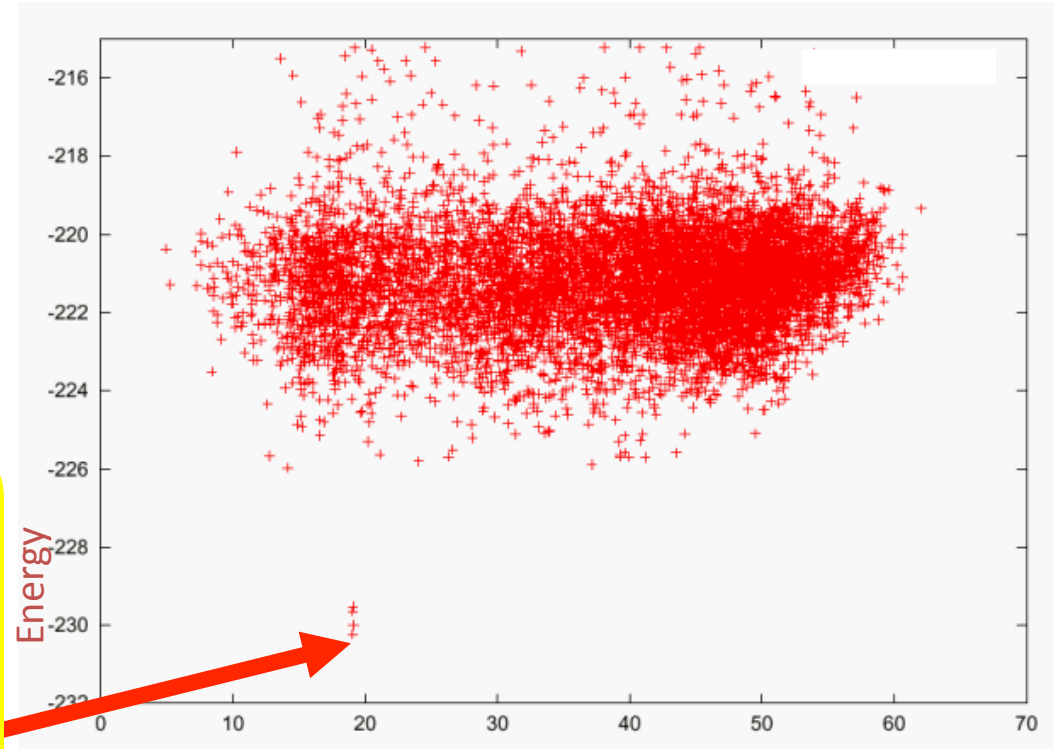
- Rotamer trial minimization and inclusion of free side chain conformations increases normalized energy gap between correct and incorrect models (Z-score)



RosettaDock simulation

- 1 model/simulation: *energy vs RMSD* *(structural similarity to starting model)*

- Final model selected based on *energy* (and/ or *sample density*)

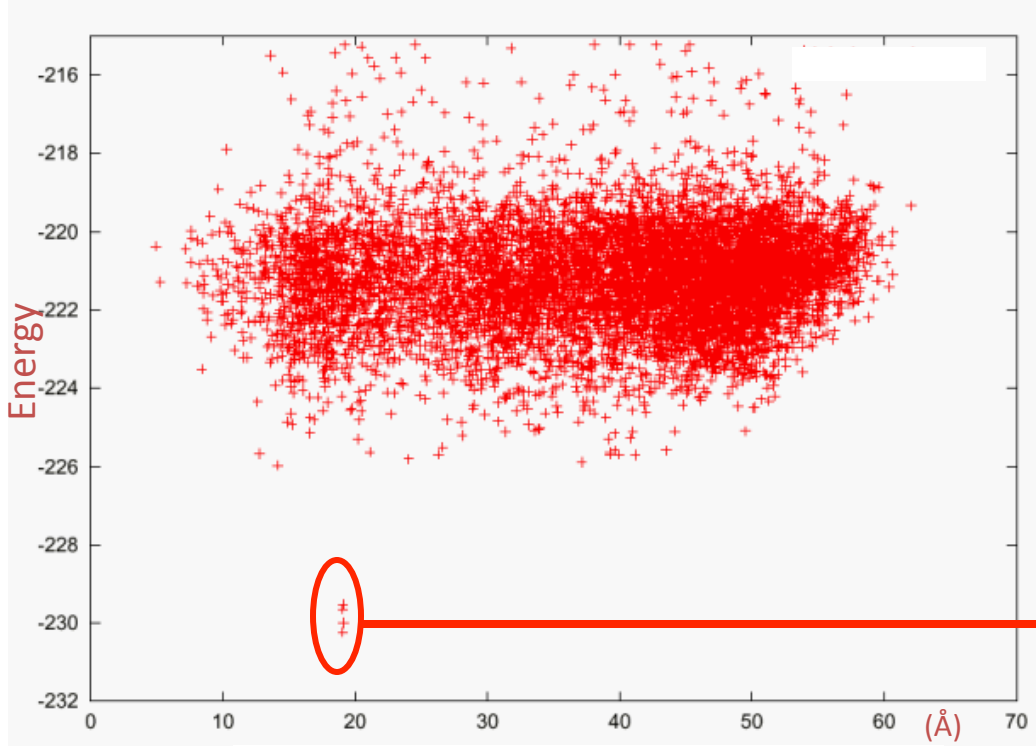


Rigid body orientations:
RMSD to arbitrary starting structure (Å)

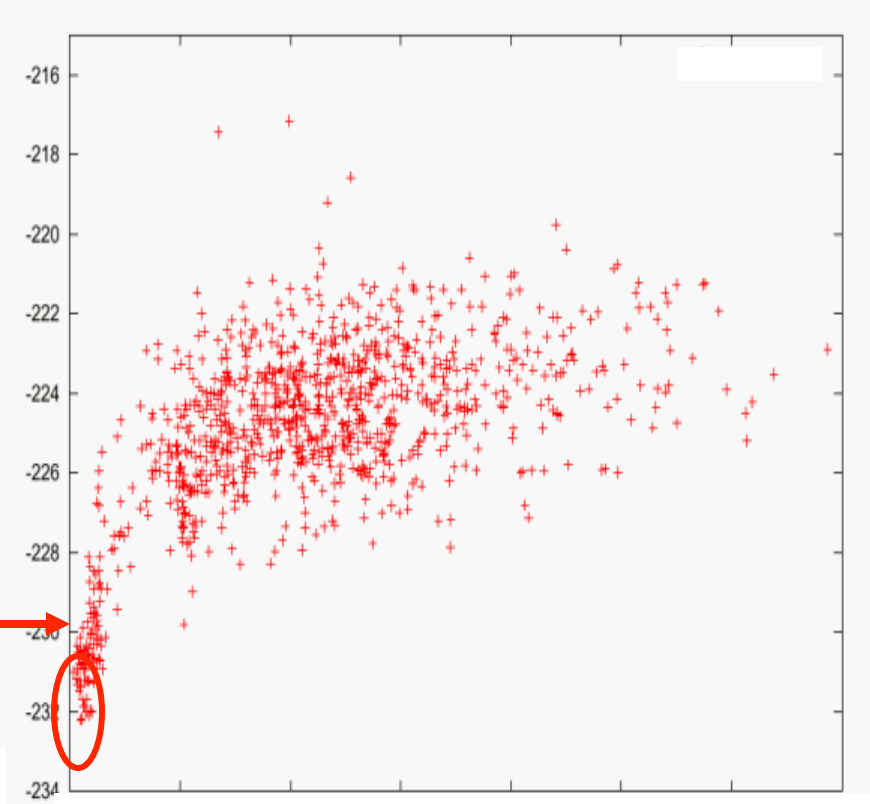
RosettaDock simulation

1. Initial Search

2. Refinement



RMSD to arbitrary starting structure



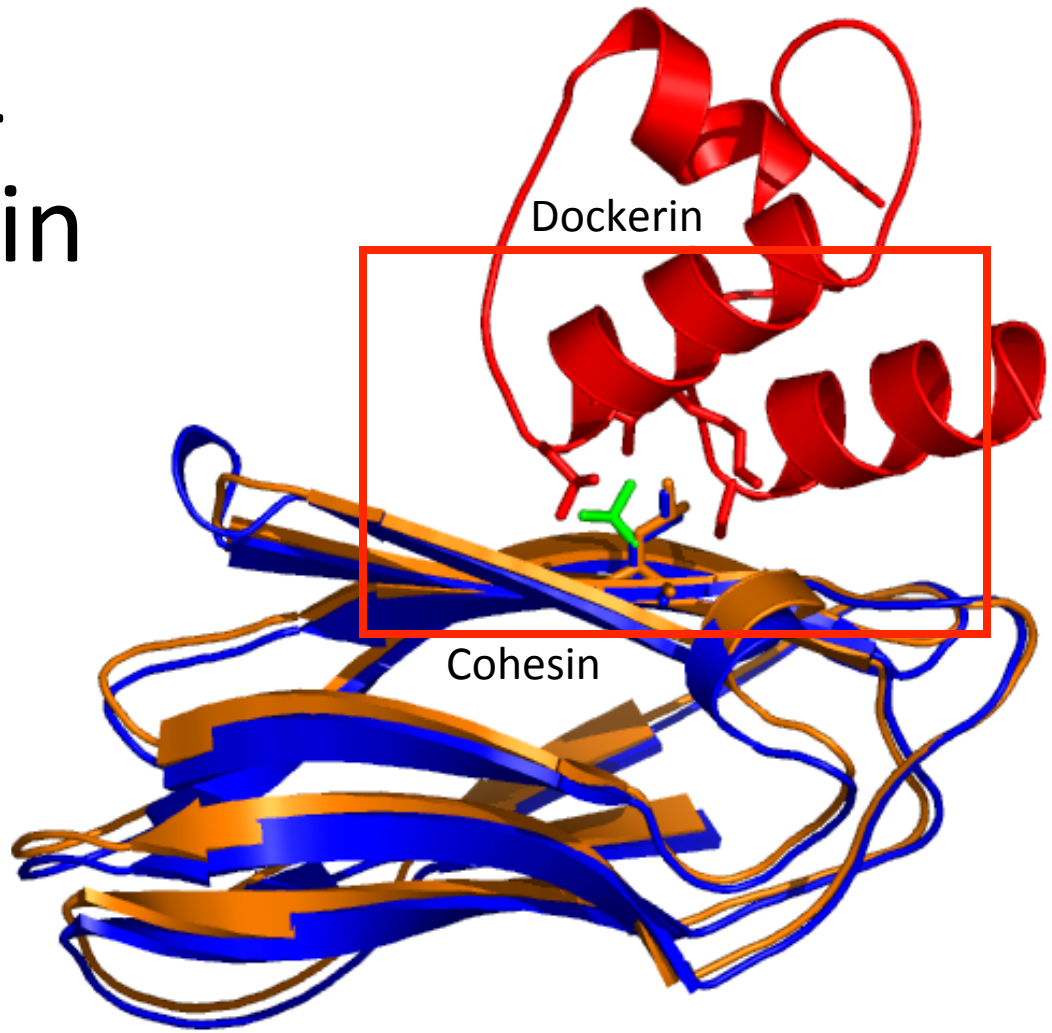
RMSD to starting structure of refinement

Side chain flexibility is important

CAPRI Target 12
Cohesin-Dockerin

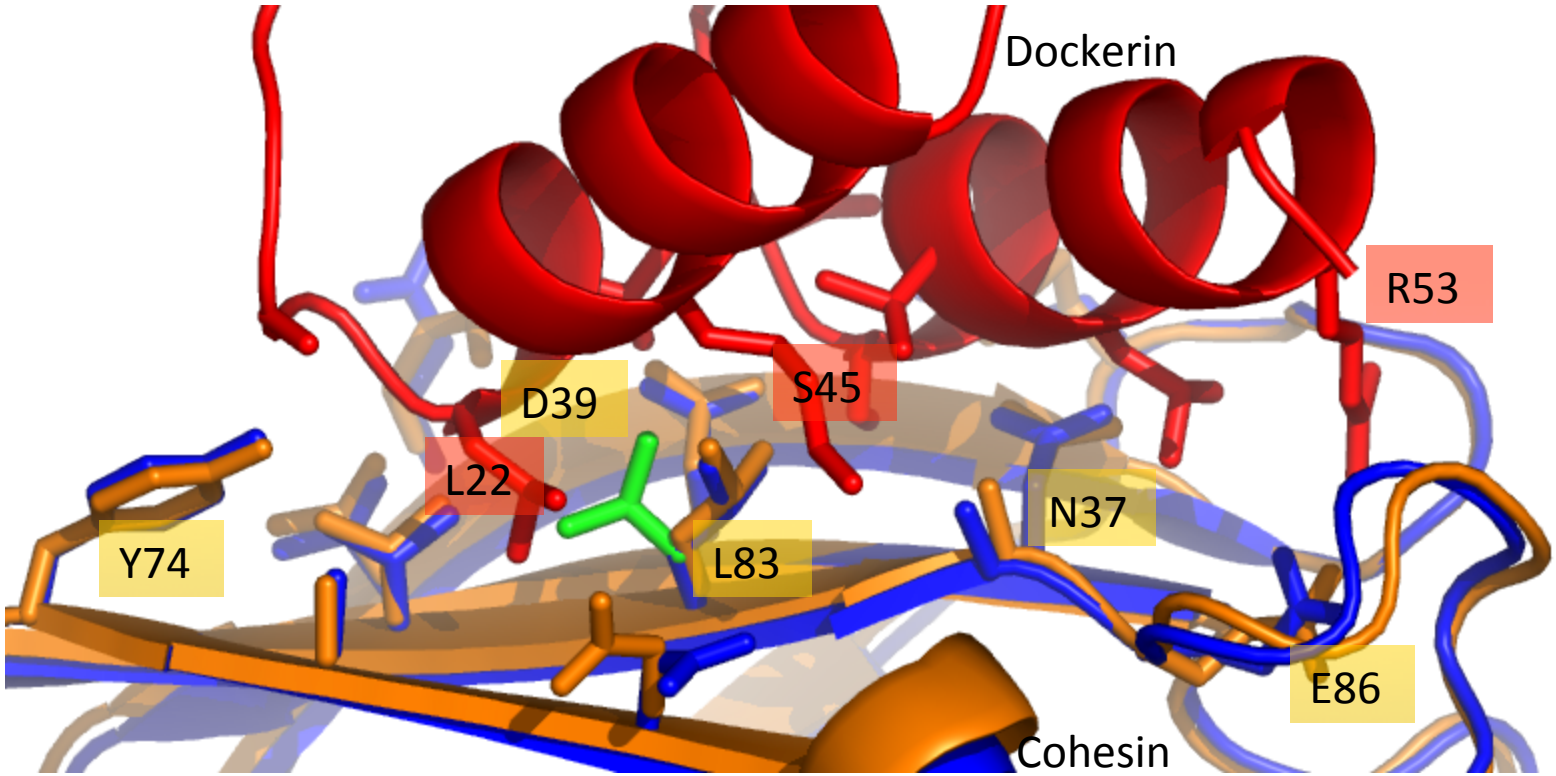
- 0.27Å interface rmsd
- 87% native contacts
- 6% wrong contacts
- Overall rank 1

red,orange– xray
blue – model;
green – unbound



Carvalho et. al (2003) PNAS

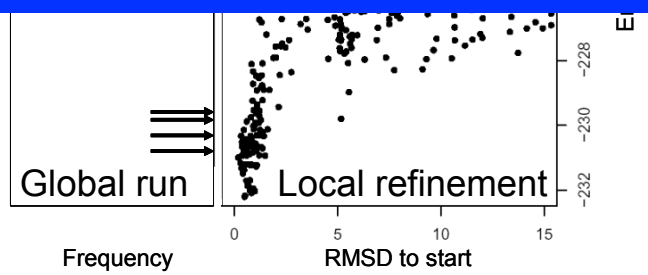
Details of T12 interface

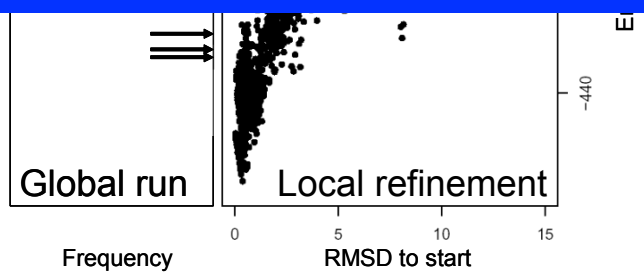


red,orange– xray
blue - model

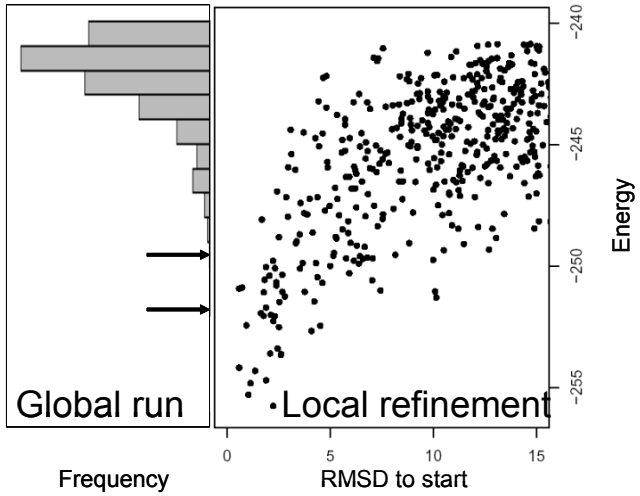
Energy landscapes with funnels

Correct model can be selected based on energy criteria only

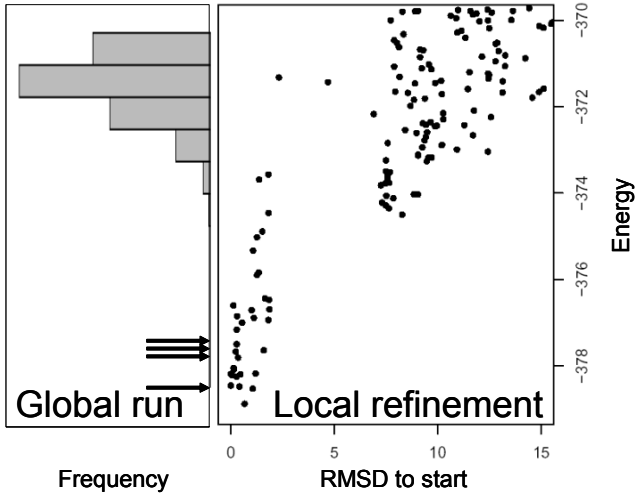




Target 15

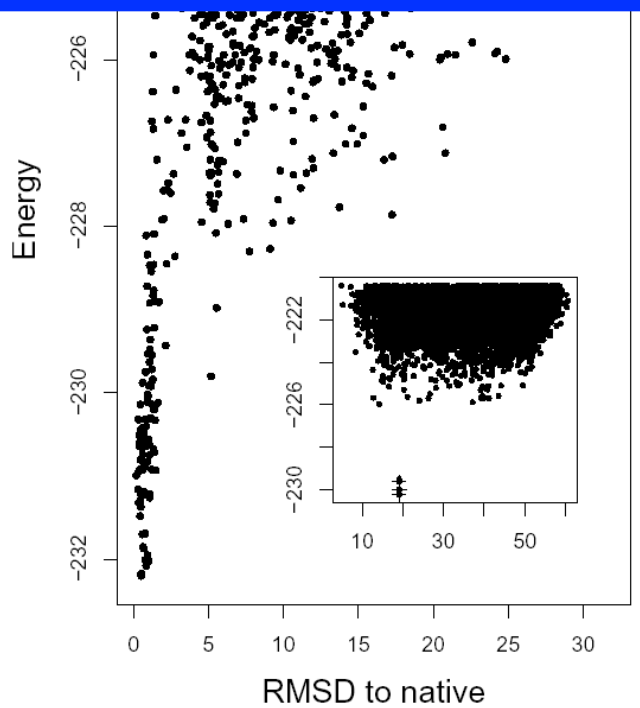


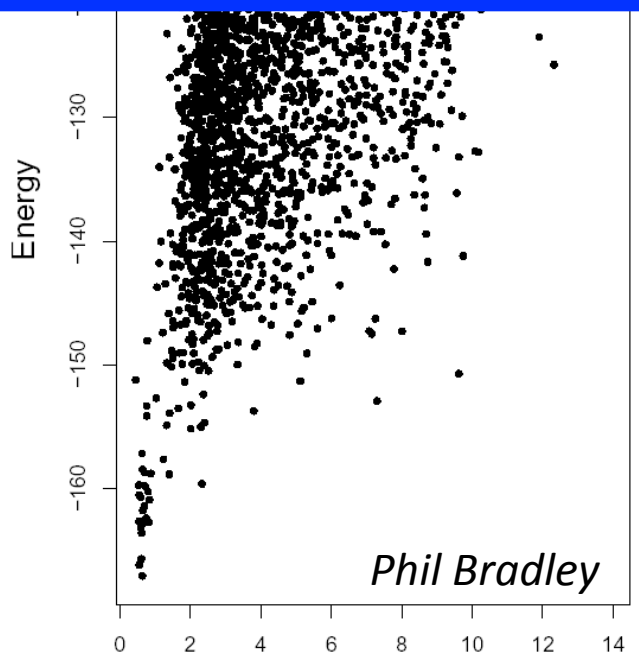
Target 19



Similar landscapes for different Rosetta predictions

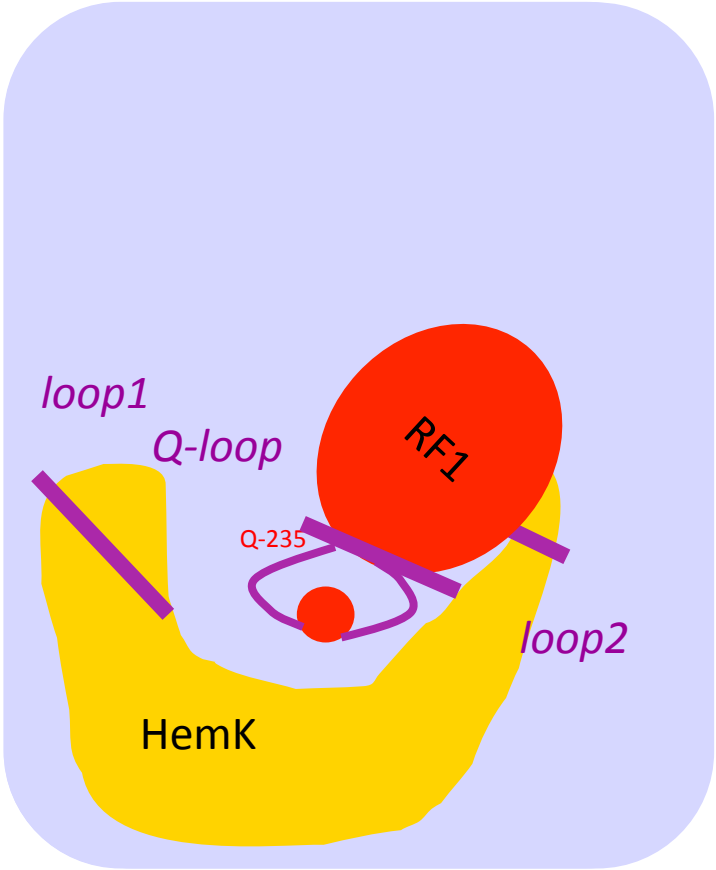
Energy function describes well principles underlying the correct structure of monomers and complexes





Schueler-Furman et. al (2005) Science

A Challenging Target RF1-HEMK (T20)



Challenge:

- Large complex
- RF1 to be modeled from RF2
- Disordered Q-loop

Hope:

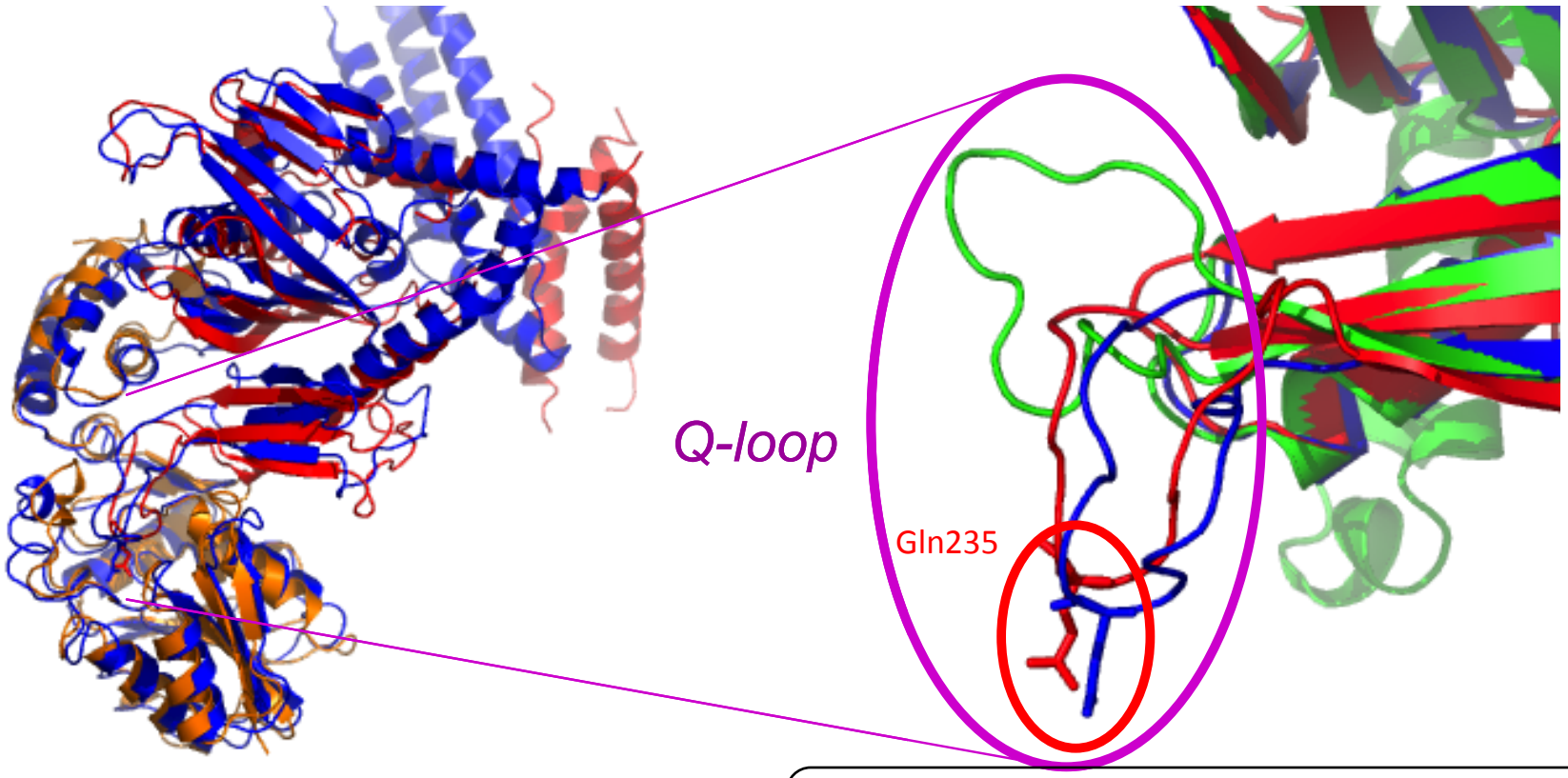
- Q235 methylated
- A Gln analog in HemK crystal

Strategy:

- Trimming – Docking – Loop Modeling - Refining

Keys to success: Location of interface with truncated protein
Separate modeling of large conformational change in key loop

Prediction of large conformational change

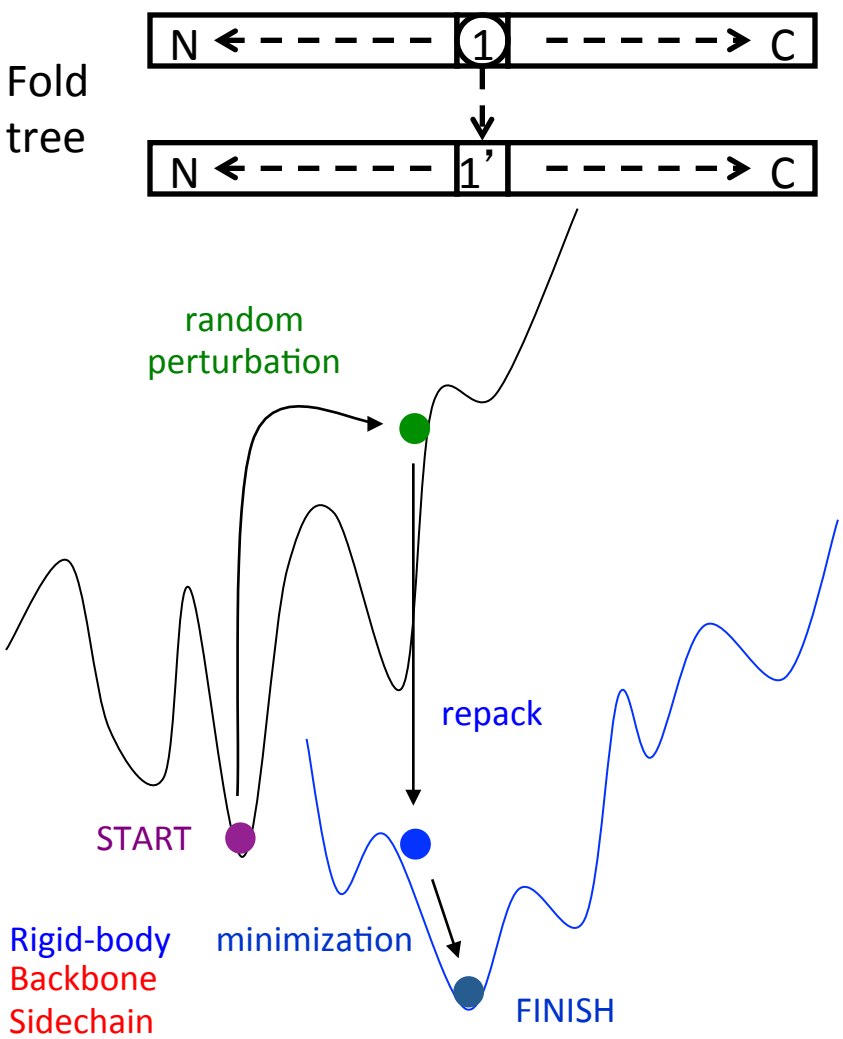


GLN235 Cα atom shift:14.13Å to 3.91 Å
Q-loop global Cα rmsd: 11.8 Å to 4.8 Å

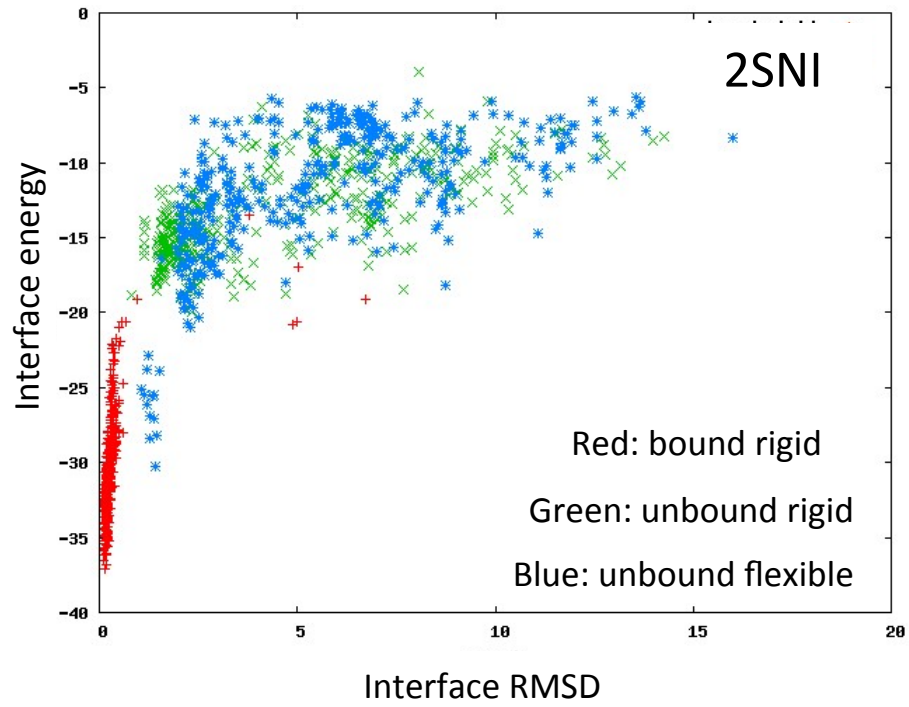
I_rmsd 2.34 Å
F_nat 34.2%

Red, orange – bound; Green,– unbound; Blue -- model

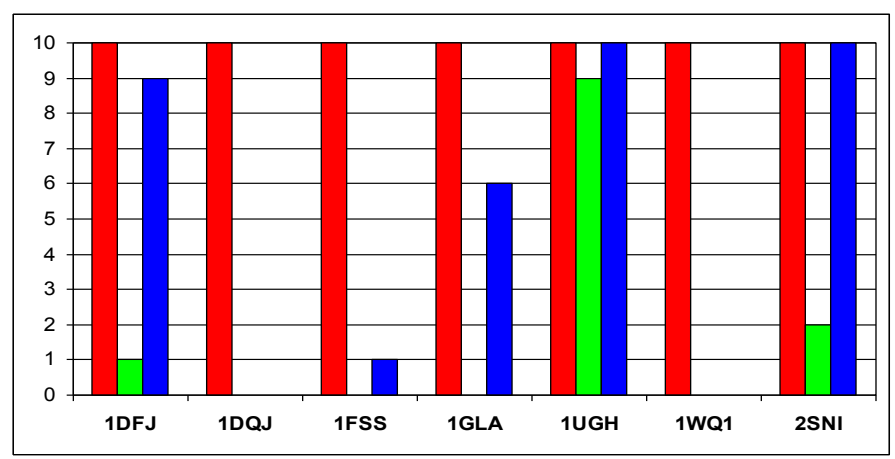
Docking with backbone minimization



Docking Monte Carlo Minimization (MCM)

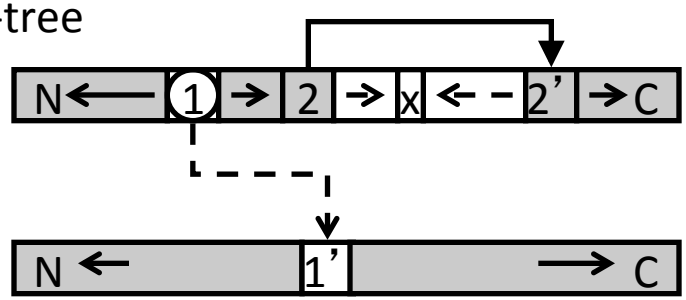


of "hits" in top 10 models



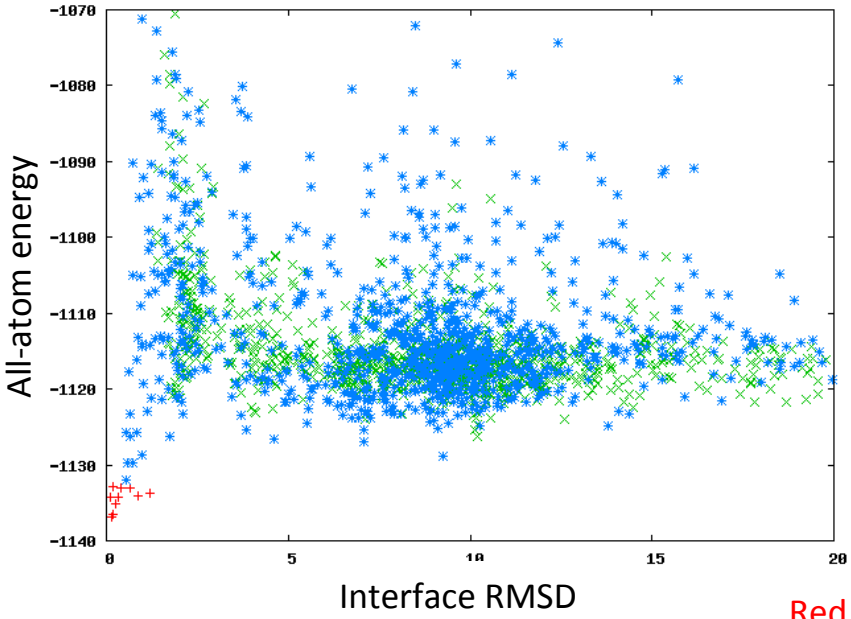
Docking with loop minimization

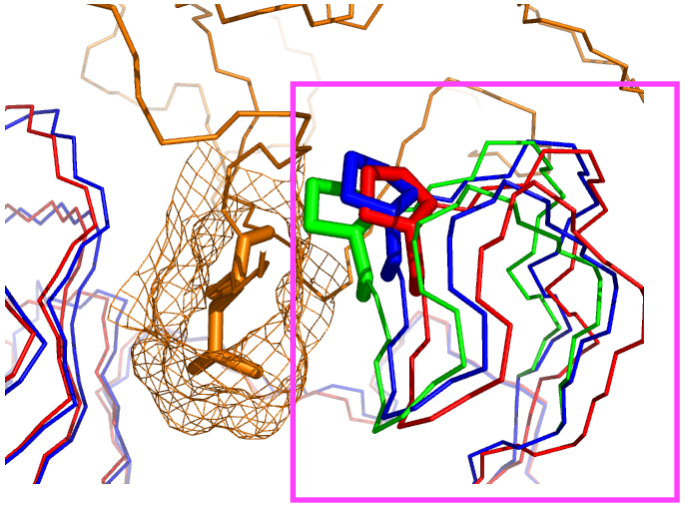
Fold-tree



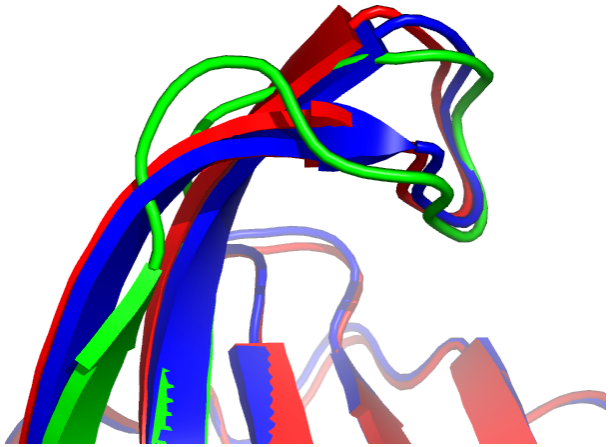
Minimize rigid-body and loop simultaneously

Flexible Docking



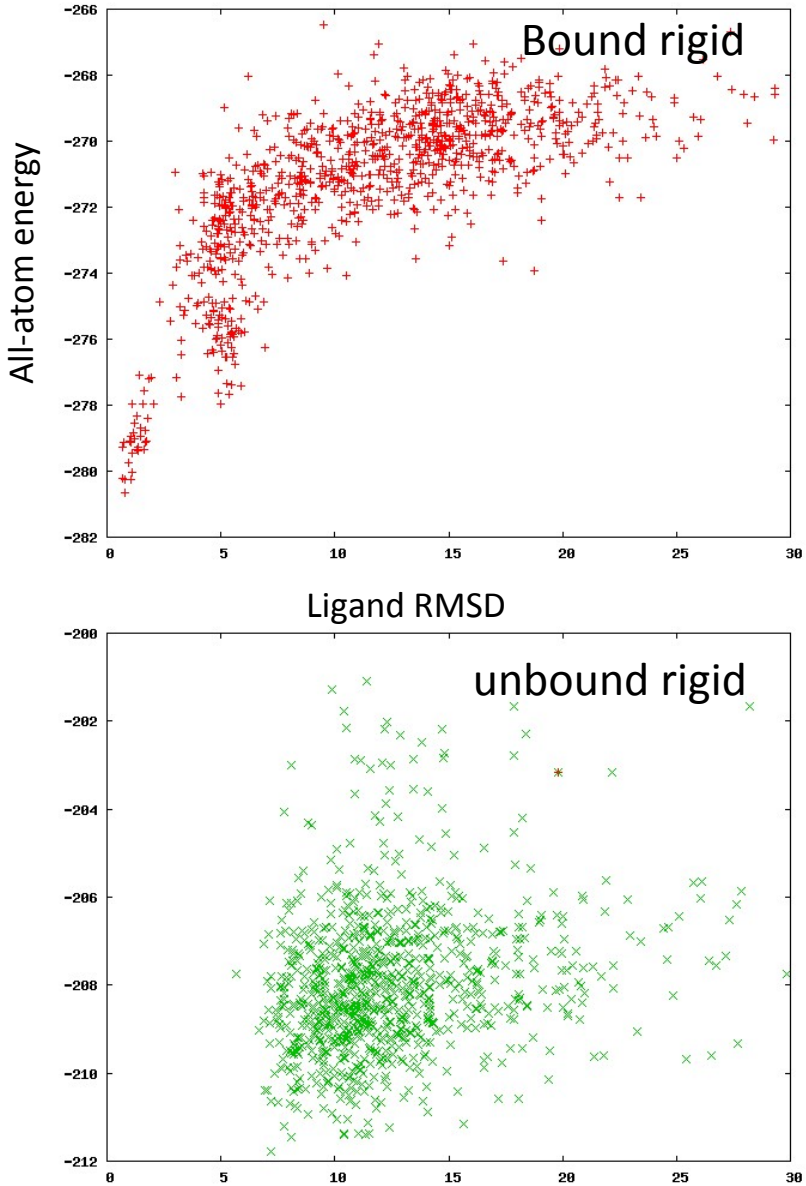


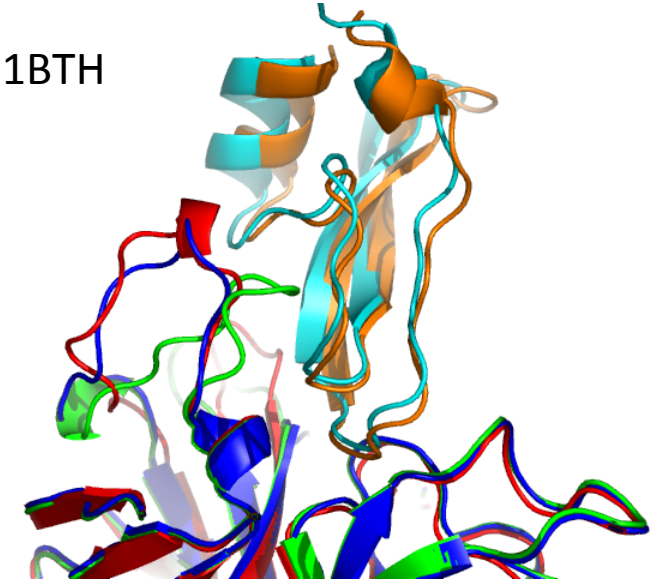
Correctly predicted loop conformation



Red, orange – bound (1T6G, Sansen, S. *et al,* J.B.C.(2004));
Blue – model; Green – unbound (1UKR, Krengel U. *et al,* JMB (1996))

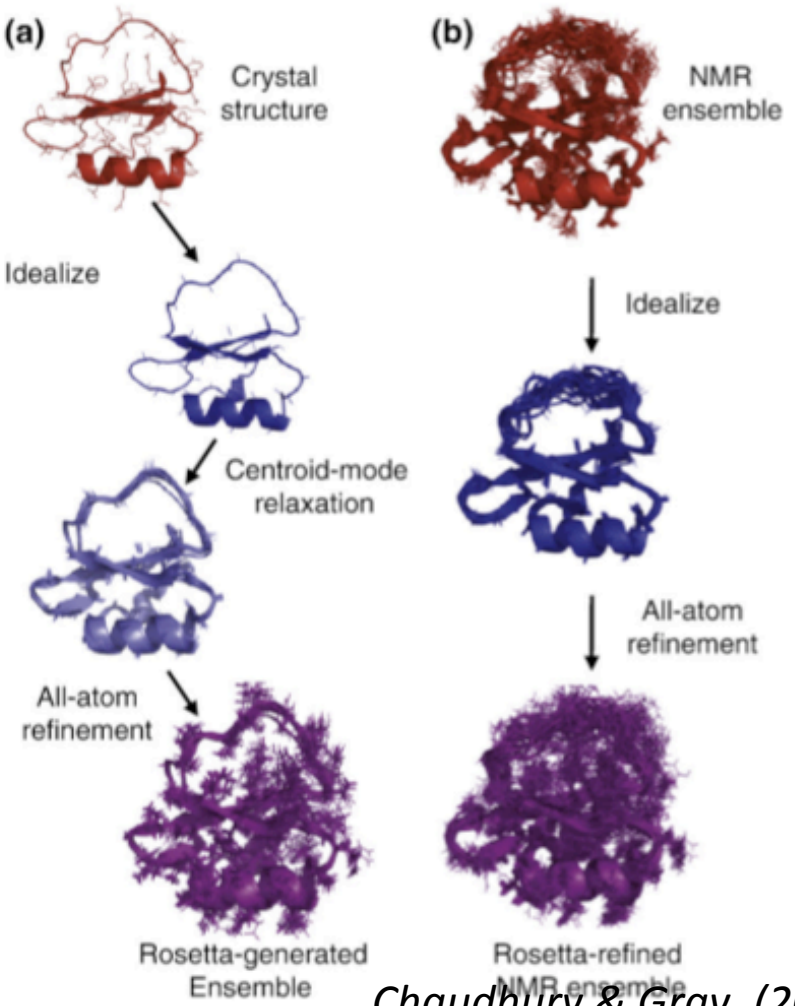
Docking with loop rebuilding





Flexible backbone protein–protein docking using ensembles

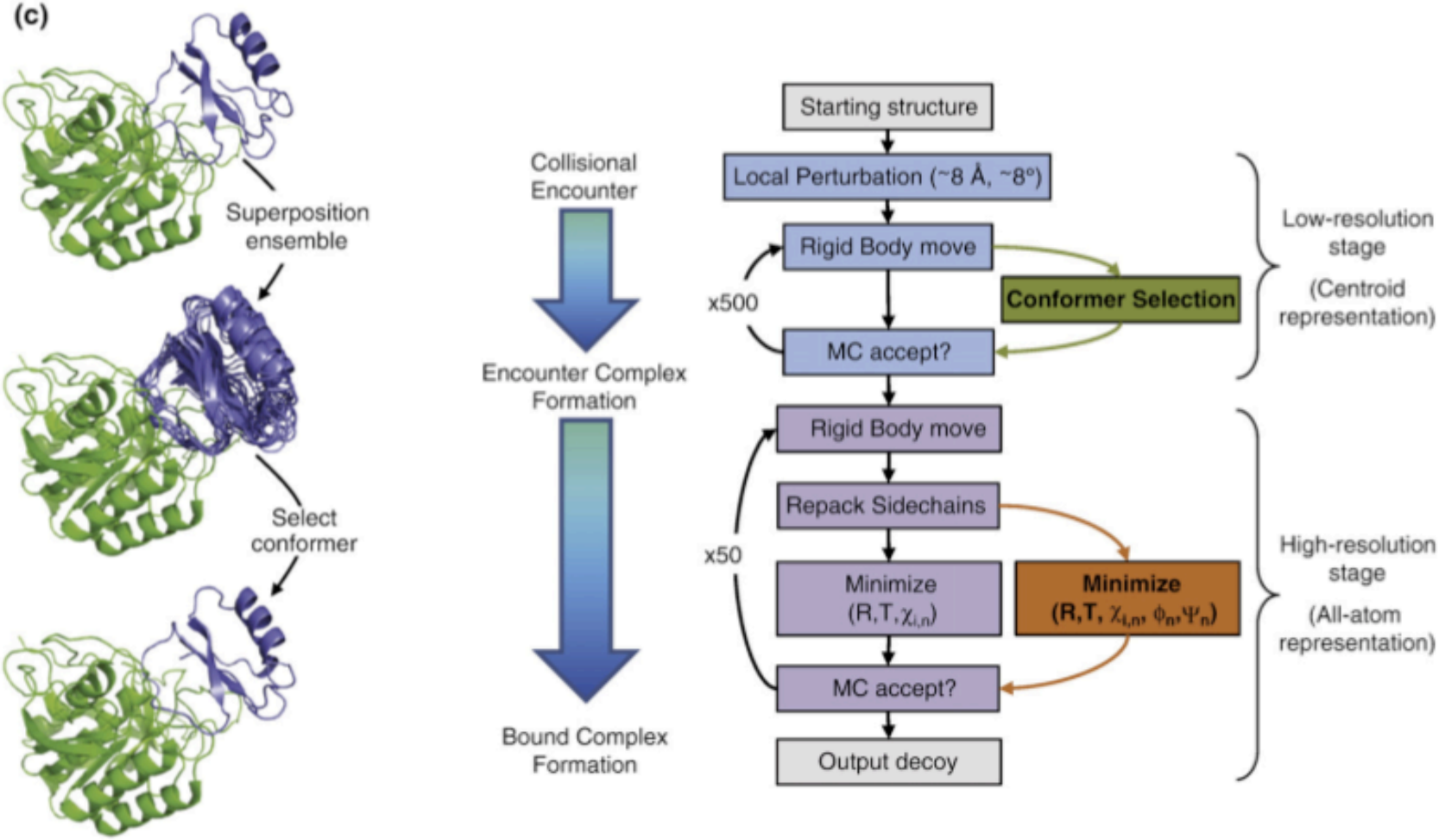
- Incorporate backbone flexibility by using a set of different templates
- Generation of set of ensembles: with Rosetta relax protocol, from NMR ensembles, etc



Chaudhury & Gray, (2008)

Sampling among conformers during docking

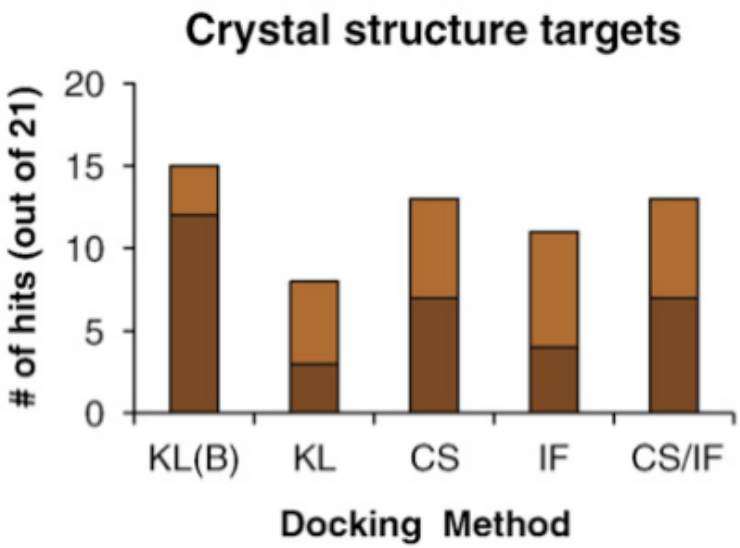
- Exchange between templates during protocol



Evaluation of 4 different protocols

1. key-lock (KL) model
   rigid-backbone docking
2. conformer selection (CS) model
   ensemble docking algorithm
3. induced fit (IF) model
   energy-gradient-based backbone minimization
4. combined conformer selection/induced fit (CS/IF) model

- Can teach us about the possible binding mechanism (e.g. induced fit vs key-lock)



Brown: high-quality decoys
Orange: medium-quality decoys

RosettaDock - summary

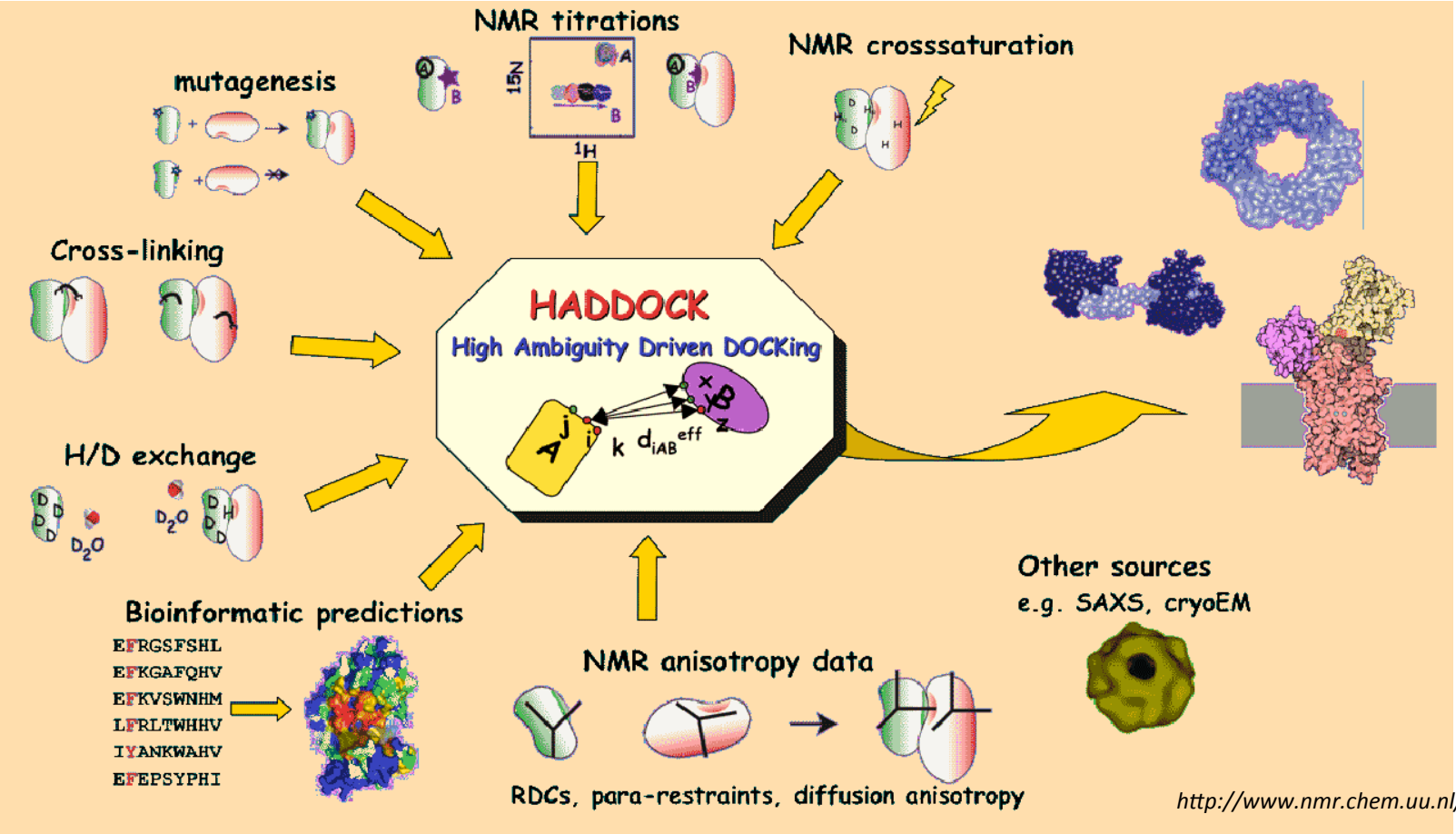
- First program to introduce general (side chain) flexibility during docking
- Advanced the docking field towards unbiased high-resolution modeling
- Many other protocols have since then incorporated RosettaDock as a high-resolution final step
- Targeted introduction of backbone flexibility can improve modeling dramatically

4. Data-driven docking

- Challenges:
  - Large conformational space to sample
  - Conformational changes of proteins upon binding
- Approach: restrict search space by previous information
  - HADDOCK (**H**igh **A**mbiguity **D**riven protein-protein **Dock**ing)

Scheme of Haddock *Bonvin, JACS 2003*

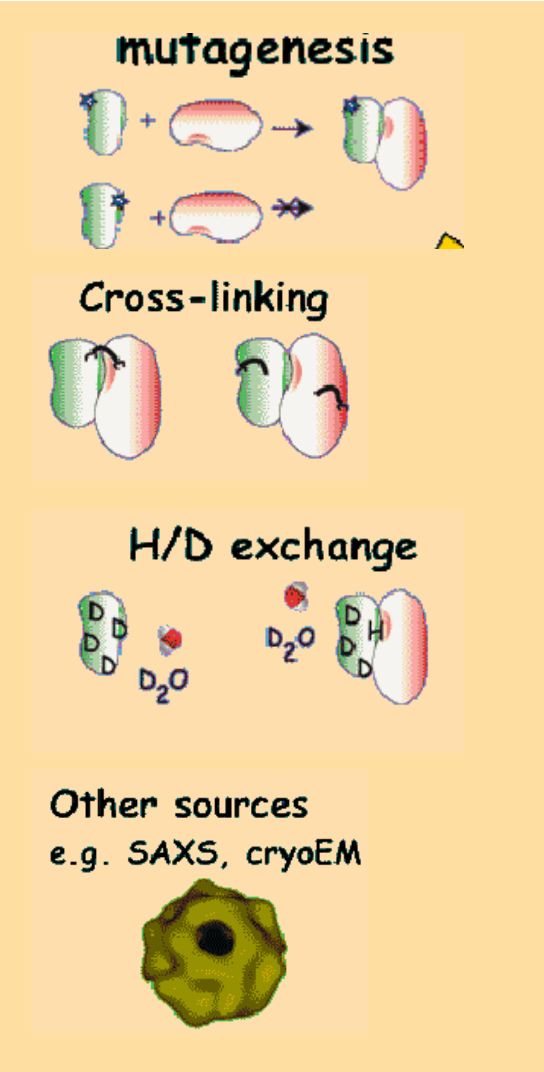
- Information about complex can be retrieved from several sources



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

Sources of information: Experimental

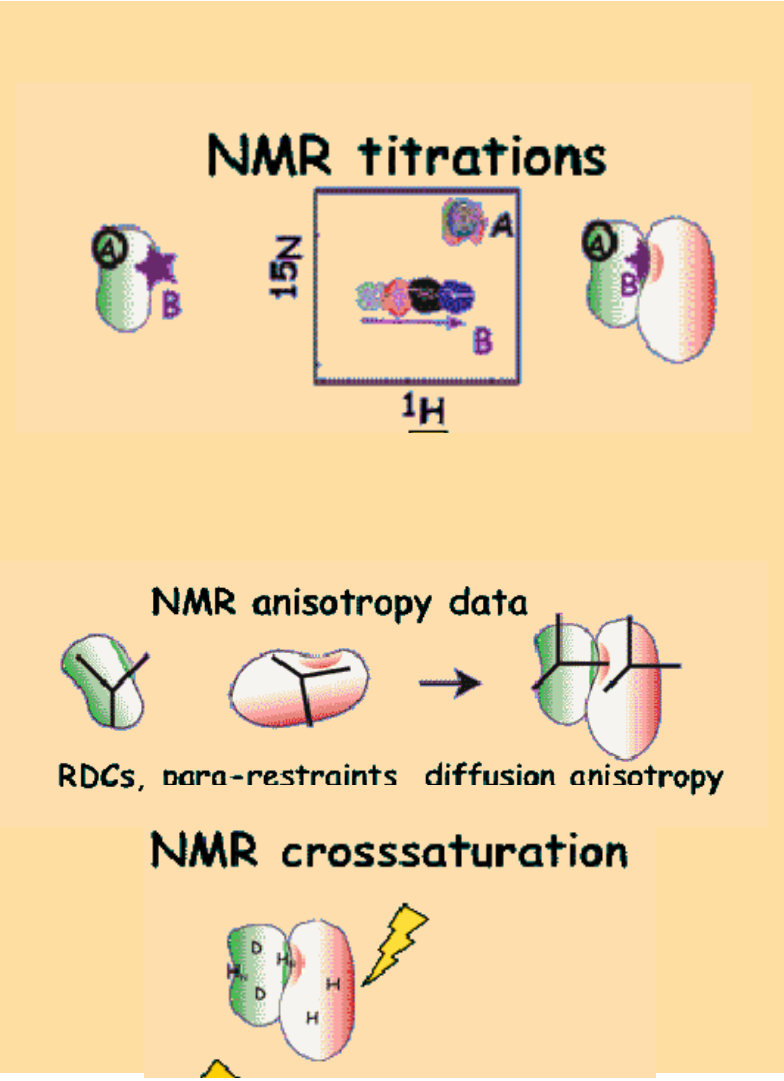
- Mutagenesis
- Cross-linking
- H/D exchange:
  - What protons are protected at the interface from hydrogen-> deuterium exchange?
- Cryo-Electron microscopy pictures
- Angle scattering information



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

Sources of information: Experimental

- NMR-titrations:
  - ^{15}N- HSQC (Heteronuclear Single Quantum Correlation) spectrum detects chemical shifts (changes in environment of amide proton)
  - > finger print of each residue
- NMR anisotropy
- NMR cross-saturation

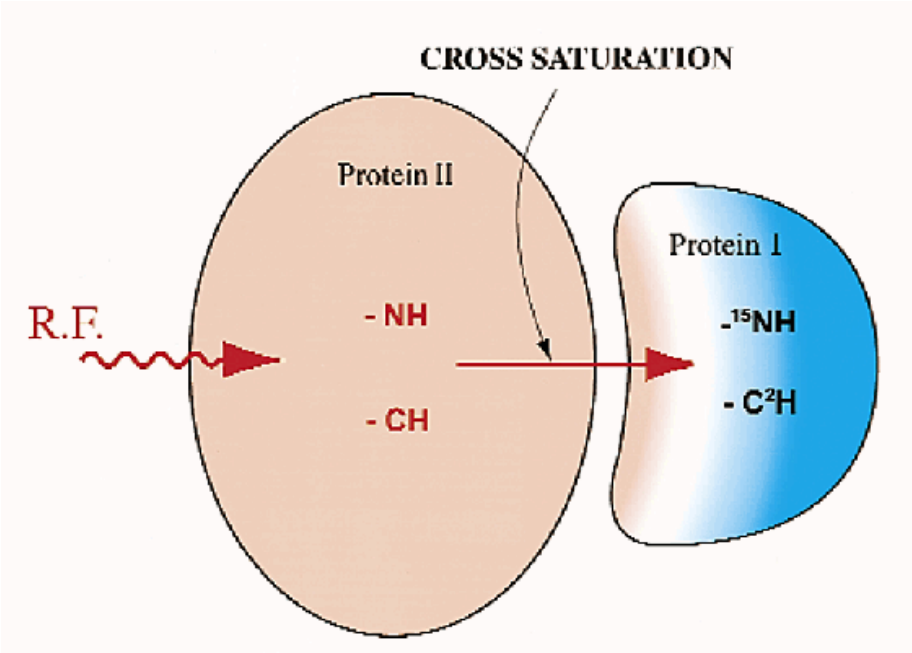


http://www.nmr.chem.uu.nl/haddock/

Cross-saturation

- The protein with residues to be identified in the complex interface (protein I) is labeled with ^{2}H and ^{15}N.
- The saturation caused by irradiation of the nonlabeled protein (protein II) is transferred to the doubly labeled molecule and is limited to the interface.

Spin diffusion effect



A novel NMR method for determining the interfaces of large protein–protein complexes *Takahashi, Nakanishi, Kami, Arata & Shimada. Nature Structural Biology 2000*

Haddock computational scheme

1. Derive Ambiguous Interaction Restraints *(AIRs):*
   - *Active residues:* involved in interaction, and solvent accessible
   - *Passive residues:* neighbors of active residues
2. Create CNS restraints file (Used in NMR structure determination)

Rational:

- Include AIRs in energy function
- find protein complex structure with minimum energy

  Similar to
  - solving a structure by NMR
  - Homology modeling with constraints (e.g. Modeler)

Overview of Haddock

Start Position

Rigid body energy minimization:

1. rotational minimization
2. rotational & translational

- Align molecules if anisotropic data is available
- **Satisfy maximum number of AIC**
- Retain top200

Semi-flexible simulated annealing (SA)

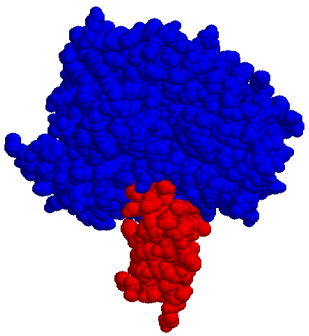
- High temperature rigid body search
- Rigid body SA
- Semi-flexible SA with flexible side-chains at the interface
- Semi-flexible SA with fully flexible interface (both backbone and side-chains)

Flexible explicit solvent refinement

- Improves energy ranking

Clustering

Predictions



Docking – summary & outlook

- Efficient search using
  - fast sampling techniques (e.g. FFT, Geometric hashing), or/and
  - Restraints to relevant region (e.g. biological constraints, etc)
- Challenge: conformational changes in the partners
- Introduction of flexibility has improved modeling to high resolution
  - Full side chain flexibility (Rosetta)
  - Targeted introduction of backbone flexibility
- Larger changes can be incorporated using techniques such as Normal Mode Analysis

Preliminary CAPRI Assessment

Rank	Group	T46	T47 (Water-mediated interactions)	T48	T48 (Trimer)	T49	T49 (Trimer)	T50	T51.1	T51.2	T51.3	T53	T54	T57 (Not assessed yet)	T58 (with SAXS data)	Summary: #Targets / *** + ** + *
1	Bonvin	*	**		*		*	**	*			**			*	8 / 3 ** + 5 *
2	Bates		**	*		*		*		*		*			**	7 / 2 ** + 5 *
3	Fernandez-Recio		*		*		*	**				**			**	6 / 3 ** + 3 *
3	**Shen**		*	**	**	**	**	*				**	*			**6 / 3 ** + 3 ***
5	Vakser		**	*	*	*	*	*					*		*	6 / 1 ** + 5 *
6	Vajda		**		**		*	**				***				5 / 1 *** + 3 ** + 1 *
7	Eisenstein		**		**	*	*	**				*				5 / 3 ** + 2 *
7	Zou		***	**	*	*	*	*							*	5 / 1 *** + 1 ** + 3 *
9	Zacharias		***		*		*	*				*				5 / 1 *** + 4 *
10	ClusPro				**		*	**				**				4 / 3 ** + 1 *
10	Grudinin		**					**				*			**	4 / 3 ** + 1 *
12	Nakamura		***						*			*	*			4 / 1 *** + 3 *
13	Weng		*			*	*	*				**				4 / 1 ** + 3 *
14	Gray		**									*			**	3 / 2 ** + 1 *
14	Seok		**									**			*	3 / 2 ** + 1 *
16	HADDOCK	*	**				*									3 / 1 ** + 2 *
16	PIE/DOCK				*		*	**								3 / 1 ** + 2 *
16	SwarmDock											*	*		**	3 / 1 ** + 2 *
16	Wolfson		*	*	**	*	*									3 / 1 ** + 2 *
20	Zhou		*	*	*		*									3 / 3 *
21	Elber				*			**								2 / 1 ** + 1 *
21	Fernandez-Fuentes							**				*				2 / 1 ** + 1 *
21	Ritchie		**												*	2 / 1 ** + 1 *
24	Camacho							**								1 / 1 **
24	Cui			*	**											1 / 1 **
24	LZerD											**				1 / 1 **
24	Ten Eyck											**				1 / 1 **
24	Wang		**													1 / 1 **
29	Kihara														*	1 / 1 *
29	Luethy							*								1 / 1 *
29	Pal							*								1 / 1 *
29	Poupon											*				1 / 1 *
29	SurFit											*				1 / 1 *
29	Zhang											*				1 / 1 *
35	About 24 Others															0 / 0 *

Notes:

1. All assessments are official results according to the CAPRI website. Tied teams are given the same rank and alphabetically ordered.
2. For all targets but T47, predictions are classified as * (acceptable), ** (medium), and *** (high). Blank space means that no acceptable predictions were submitted.
3. The only, slight exception in classifying predictions was for T47, where the real challenge is the prediction of water-mediated interactions between a given protein sequence and an unbound protein. Here, the classification is * (fair), ** (good), *** (excellent), and **** (outstanding).

Reading Assignment

- http://www.loria.fr/~ritchied/papers/ritchie_cpps_2008.pdf
- D. Ritchie. Recent progress and future directions in protein-protein docking. Current Protein and Peptide Science, 2008.

Project 3

- Apply three docking tools to two CAPRI targets
- Combine tools to improve accuracy if possible
- Assess the performance using a few complementary measures (% true contacts, RMSD)
- Reading: Fourier transformation or CAPRI paper