Protein-Protein Docking: Prediction of Protein Association

Pralay Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Short term course on Bioinformatics, Oct 31,2013, IIT KGP

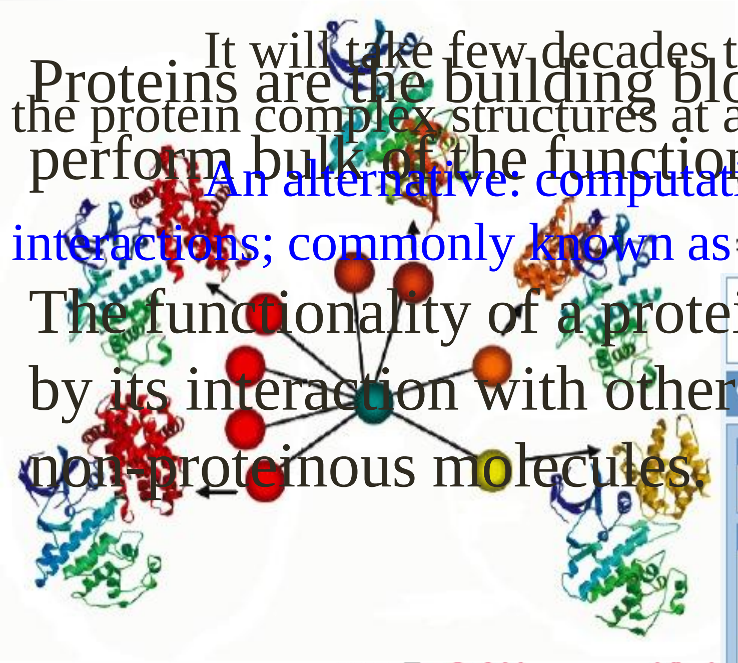
Background

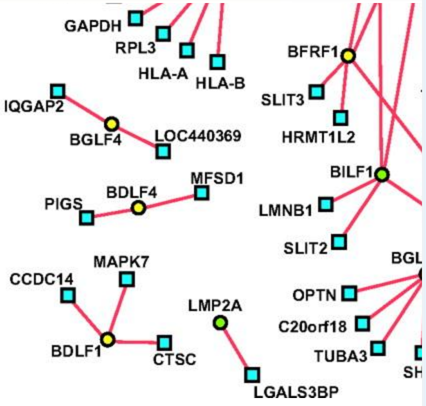
Proteins are the building blocks of the cells; perform bulk of the functions of the cell.

The functionality of a protein is determined by its interaction with other proteinous or non-proteinous molecules.

It will take few decades to experimentally determine all the protein complex structures at atomic level resolution.

An alternative: computational modeling of protein-protein interactions; commonly known as protein-protein docking.





Epstein–Barr virus(

PDB PROTEIN DATA BANK

A MEMBER OF THE wwPDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Aug 16, 2011 at 5 PM PDT there are 75246 Structures | PDB Statistics

Contact Us | Print

PDB ID or Text ▾ PDB ID lookup or Text search of the complete structure file [Search] ? [Advanced Search]

[Customize This Page]

MyPDB Hide
- Login to your Account
- Register a New Account

Home Hide
- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

Deposition Hide
- All Deposit Services
- Electron Microscopy
- X-ray | NMR
- Validation Server
- BioSync Beamlines/Facilities
- Related Tools

Search Hide
- Advanced Search
- Latest Release
- New Structure Papers
- Sequence Search
- Chemical Components
- Unreleased Entries
- Browse Database
- Histograms

Tools Hide
- Download: Entries | Ligands
- Compare Structures
- FTP Services
- File Formats

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Hide Welcome Message

Featured Molecules Hide

Structural View of Biology — List View of Archive By: Title | Date | Category

Enzymes

Molecule of the Month:
Rhomboid Protease GlpG
Proteases, enzymes that cut protein chains, come in many shapes and sizes. The most familiar proteases, like trypsin and pepsin, are machines of destruction used to digest proteins in our diet. However, most of the proteases in our cells are used in a more delicate task. They regulate the action of other proteins by making specific cuts in their protein targets.
Full Article...

Protein Structure Initiative Featured System:
Exploring the Secretome of Gut Bacteria
JCSG researchers are studying the secretome of the bacteria that inhabit the human gut, to reveal how we live in harmony with these symbiotic organisms.
Full Article | Archive | PSI Structural Biology Knowledgebase

Latest Structures Hide

3skr - Crystal structure of the 2'-Deoxyguanosine riboswitch bound to 2'-

New Features Hide
Advanced Search: Protein Modifications
Latest features released:
Website Release Archive: ▾

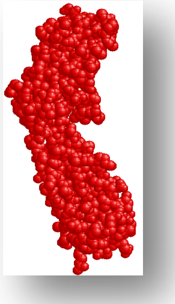
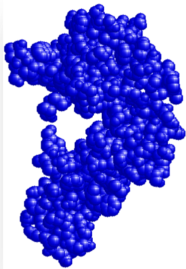
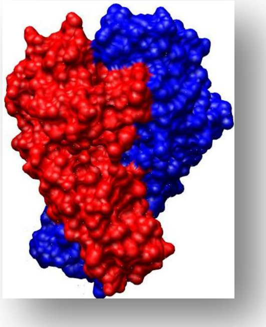
wwPDB News Hide
PDB40 Symposium October 28-30, 2011 Cold Spring Harbor Laboratory
2011-08-11
IUCr XXII: wwPDB Q&A session, exhibit booth, presentations, posters and more
- Due Aug 1: PDB40 Travel Award Applications for Early Career Scientists and Students
- Full wwPDB News
- Statement on Retraction of PDB Entries

RCSB PDB News Hide
Weekly | Quarterly | Yearly
2011-08-16
wwPDB Events at IUCr (August 22 -30, 2011)

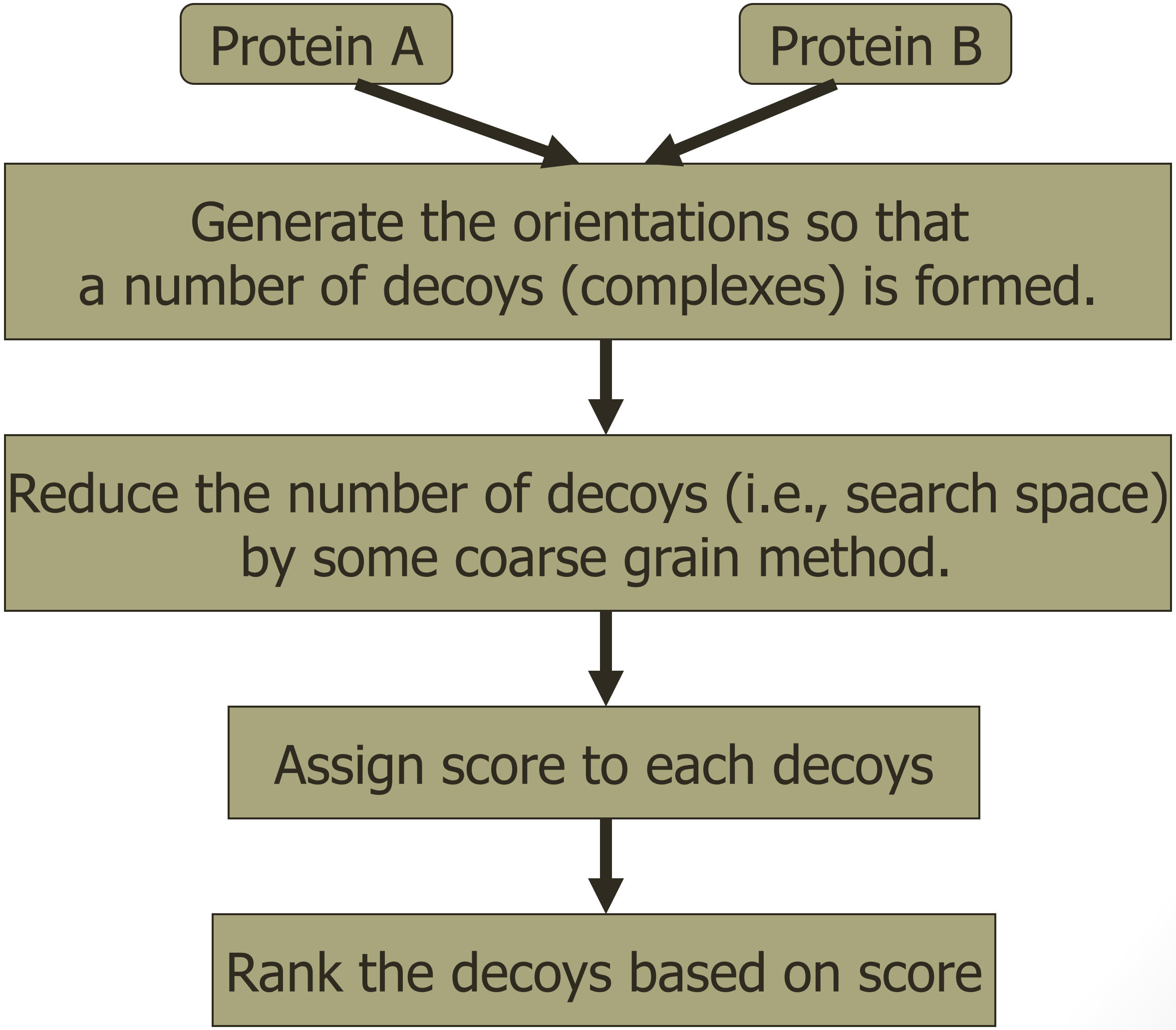
Aloy et al. (2004). *Nat. Biotechnology*

Calderwood M A et al. (2007) *PNAS* **104**, 7606-7611.

Docking Types

- Based on crystallization information
  - Bound docking
  - Unbound docking
  - (A)   (B) (C) 
- Based on protein flexibility
  - Rigid Body
  - Flexible Body

Docking Strategy



Protein A

Protein B

Generating

Generate the orientations so that a number of decoys (complexes) is formed.

Pruning

Reduce the number of decoys (i.e., search space) by some coarse grain method.

Scoring

Assign score to each decoys

Ranking

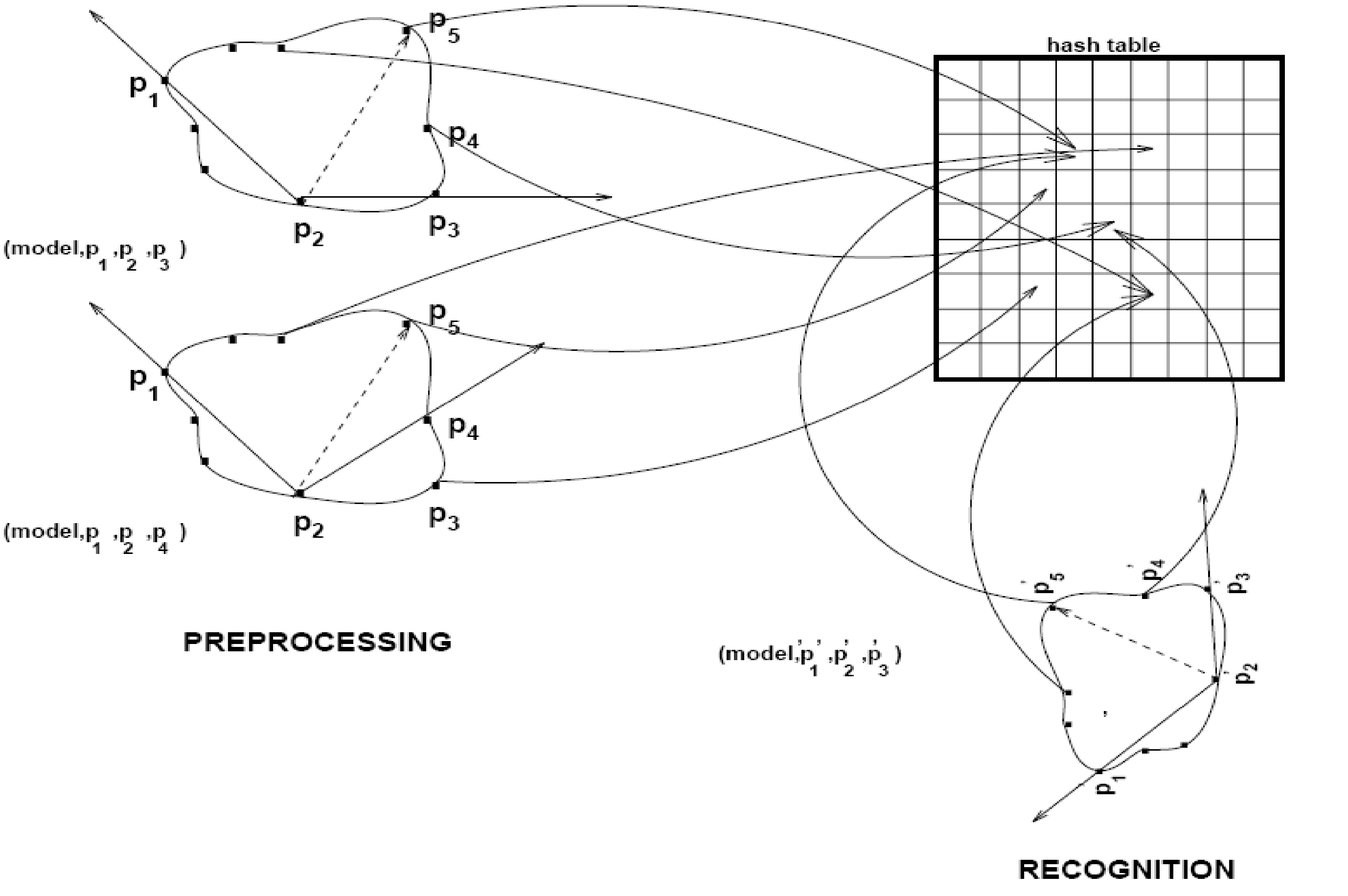
Rank the decoys based on score

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

Geometric Hashing

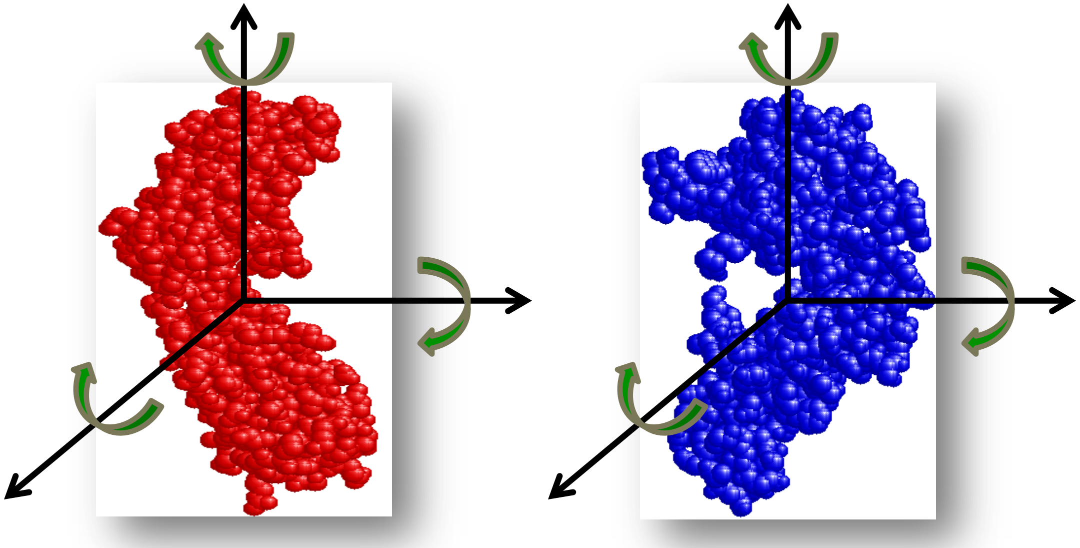
- ❖ Models are represented in a redundant affine invariant way and stored in a table (off-line).
- ❖ Hashing is used for organizing and searching the table.



Geometric Hashing

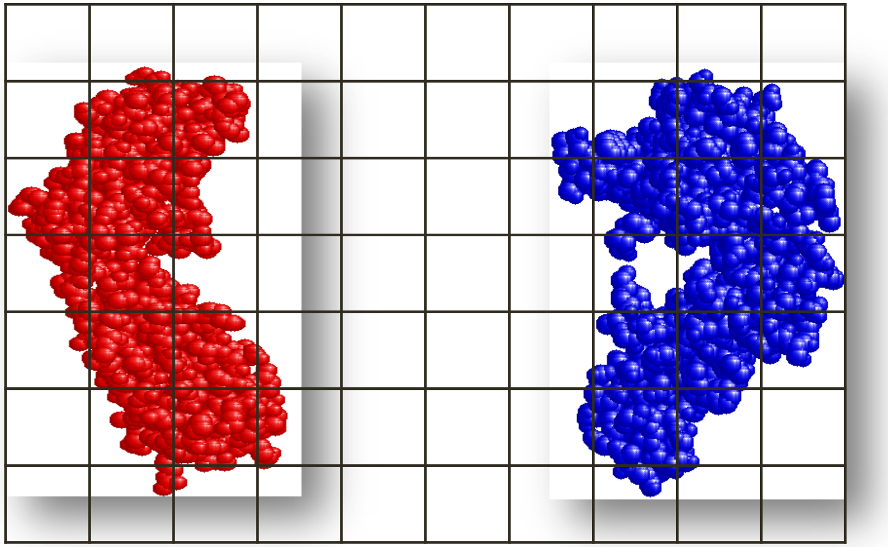
- Pro:
  - Faster
- Con:
  - Storage requirement is very high and increases with the increase in object points.
  - Proper identification of object points are crucial for the success.

Generation methods



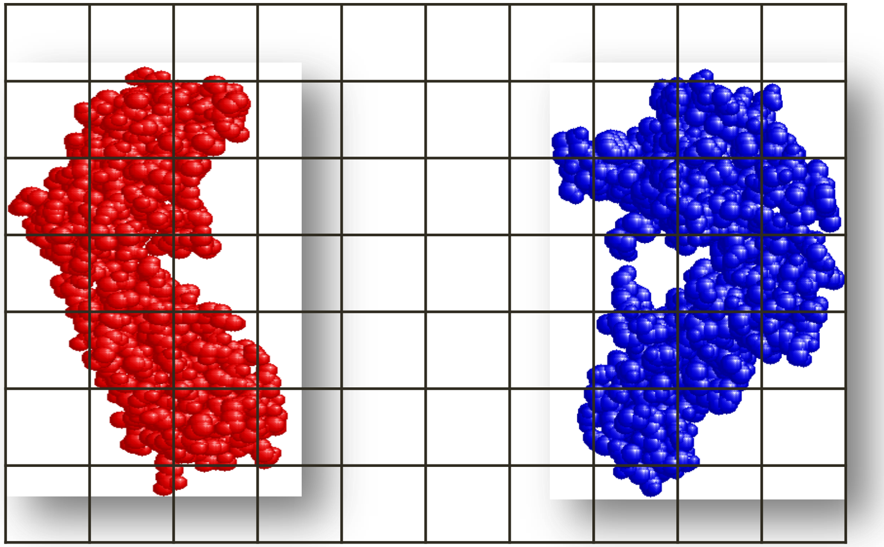
- Tagline – “Higher the decoys; better the possibility of having a hit”
- How many is good?
- Move to discrete space

Generation methods



On an average some brute force method can generate ~10^7 decoys.

Fast Fourier Technique

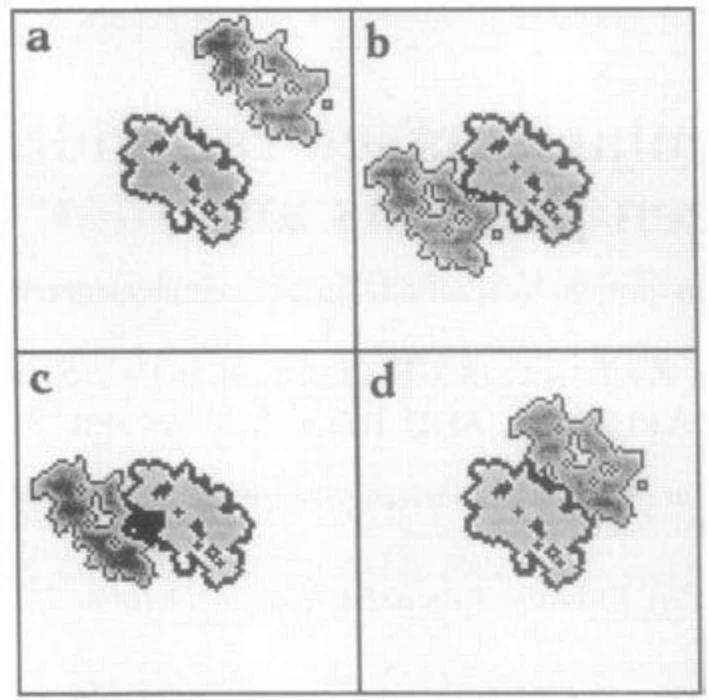


$$\bar{a}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the molecule} \\ \rho & \text{inside the molecule} \\ 0 & \text{outside the molecule,} \end{cases}$$

and

$$\bar{b}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the molecule} \\ \delta & \text{inside the molecule} \\ 0 & \text{outside the molecule,} \end{cases}$$

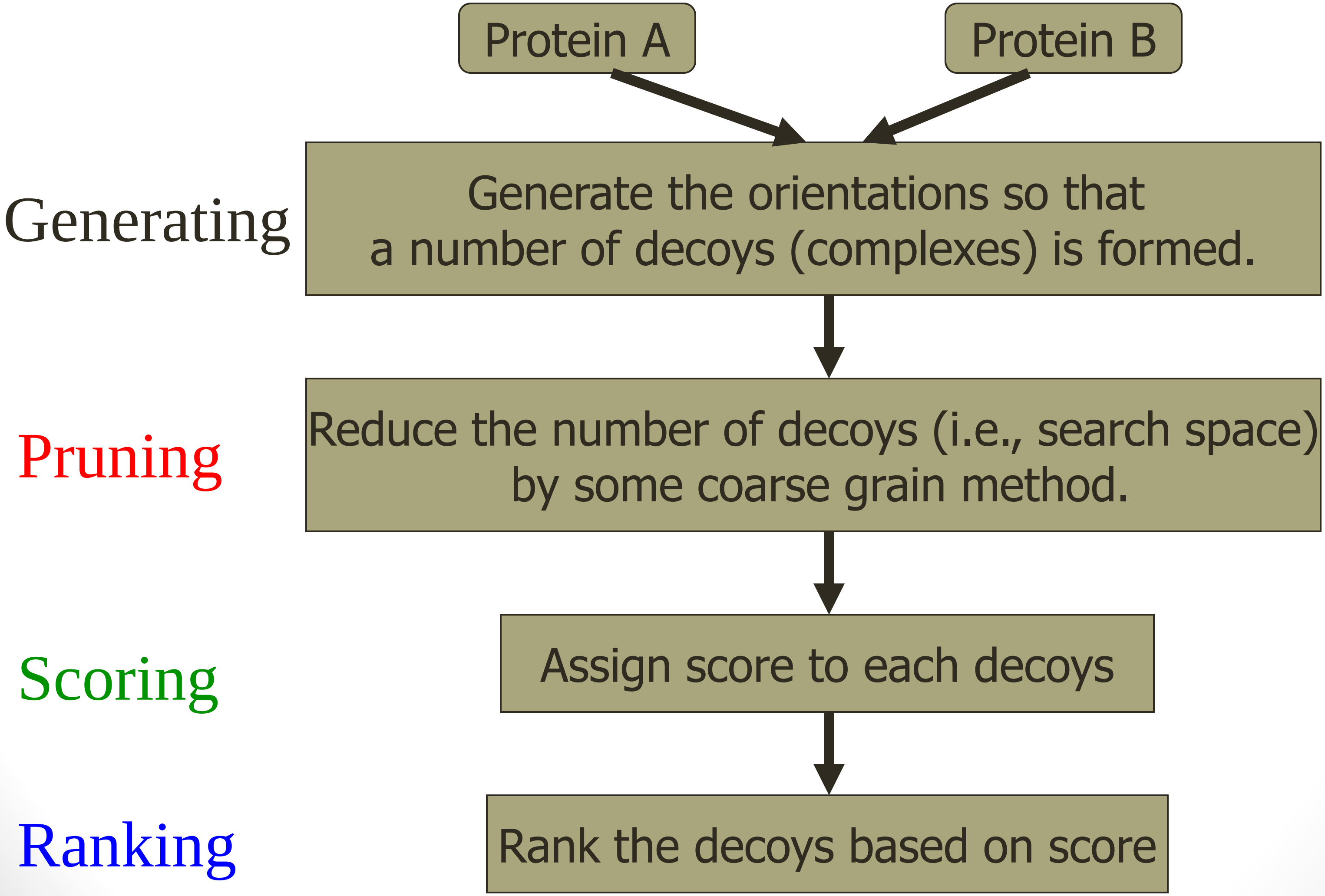
Fast Fourier Technique*



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

*Katchalski-Katzir *et al*, (1992) *PNAS*

Docking Strategy



Protein A

Protein B

Generating

Generate the orientations so that a number of decoys (complexes) is formed.

Pruning

Reduce the number of decoys (i.e., search space) by some coarse grain method.

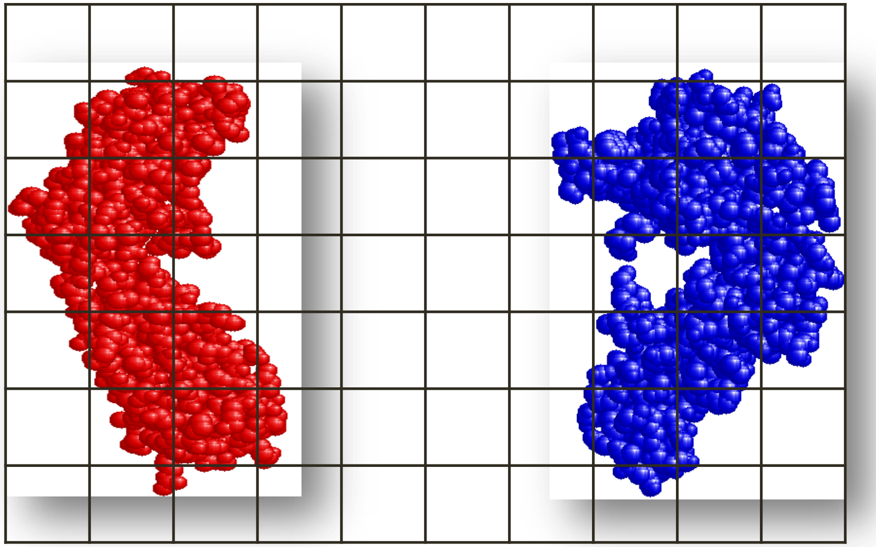
Scoring

Assign score to each decoys

Ranking

Rank the decoys based on score

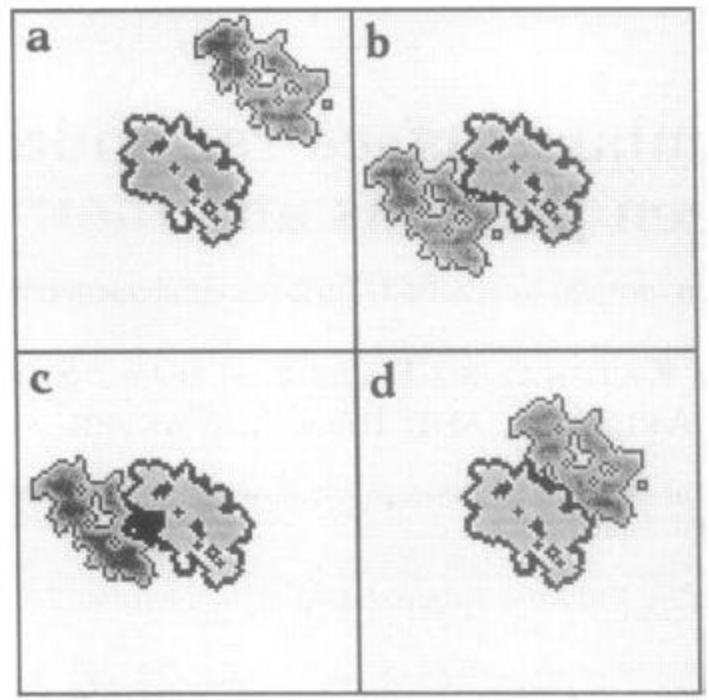
Generation methods



On an average some brute force method can generate ~10^7 decoys.

Assuming processing of each decoy takes 1 sec; total processing time ~115 days.

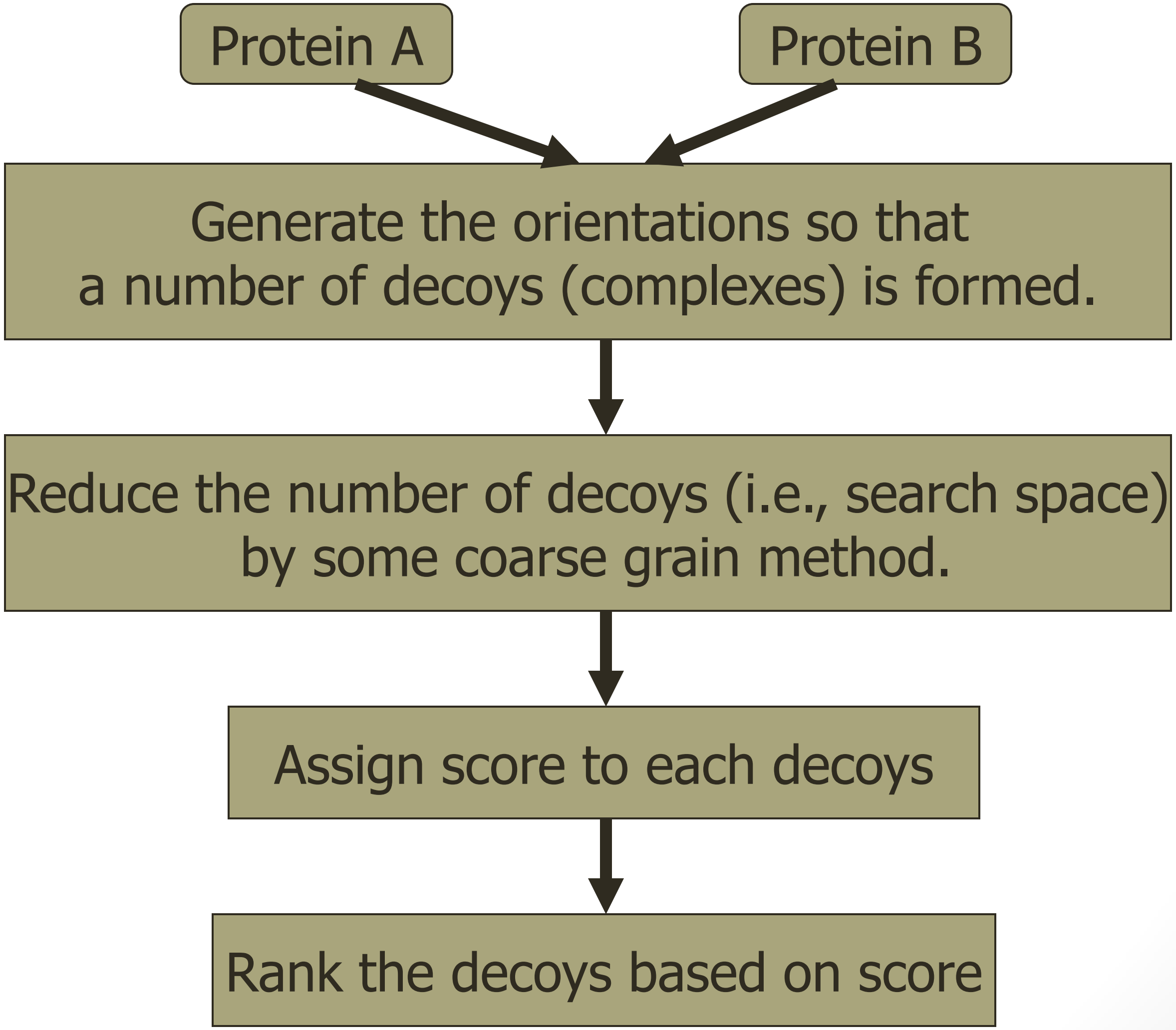
Fast Fourier Technique*



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

*Katchalski-Katzir *et al*, (1992) *PNAS*

Docking Strategy



Protein A

Protein B

Generating

Generate the orientations so that a number of decoys (complexes) is formed.

Pruning

Reduce the number of decoys (i.e., search space) by some coarse grain method.

Scoring

Assign score to each decoys

Ranking

Rank the decoys based on score

Scoring methods

Ab initio scoring (Physics based)

- Contact Area
- Contact Packing
- Non-bonded interactions
- Solvation Energy
- Etc.

Evolutionary scoring (Template based)

Ab initio method

- Interface area (IA)
- Normalized interface packing (NIP)
- Normalized surface complementarity (NSc)
- Non-bonded energy (NE):

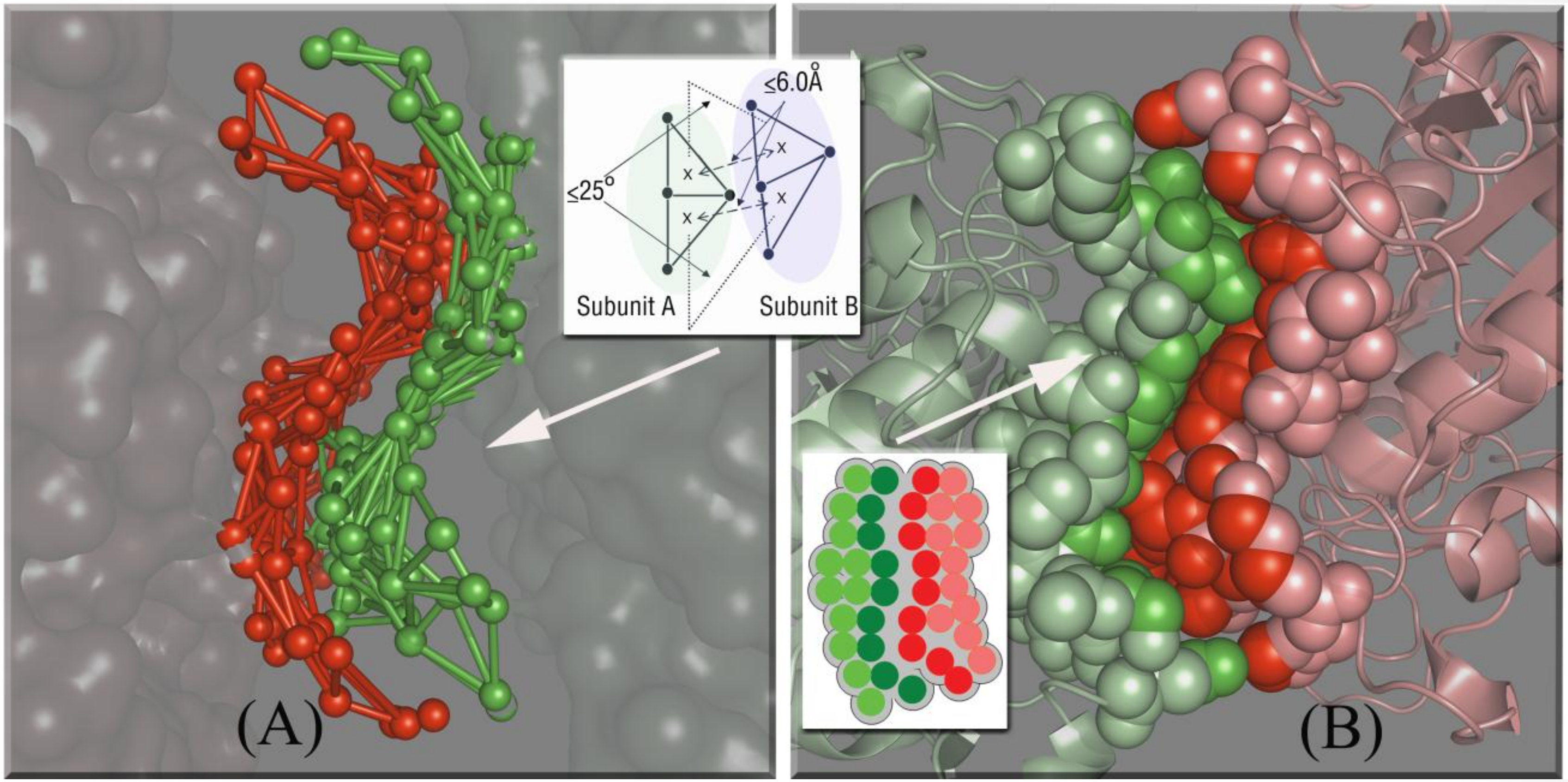
$$NE = \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{4\Pi \varepsilon R_{ij}} \right)$$

- Solvation energy (SE):*

$$SE = \sum_{\text{interface atoms}} \Delta\sigma(\text{Atom Type}) \times \Delta\text{ASA}$$

*Eisenberg and McLachlan (1986) *Nature*

NSc and NIP at protein interface*



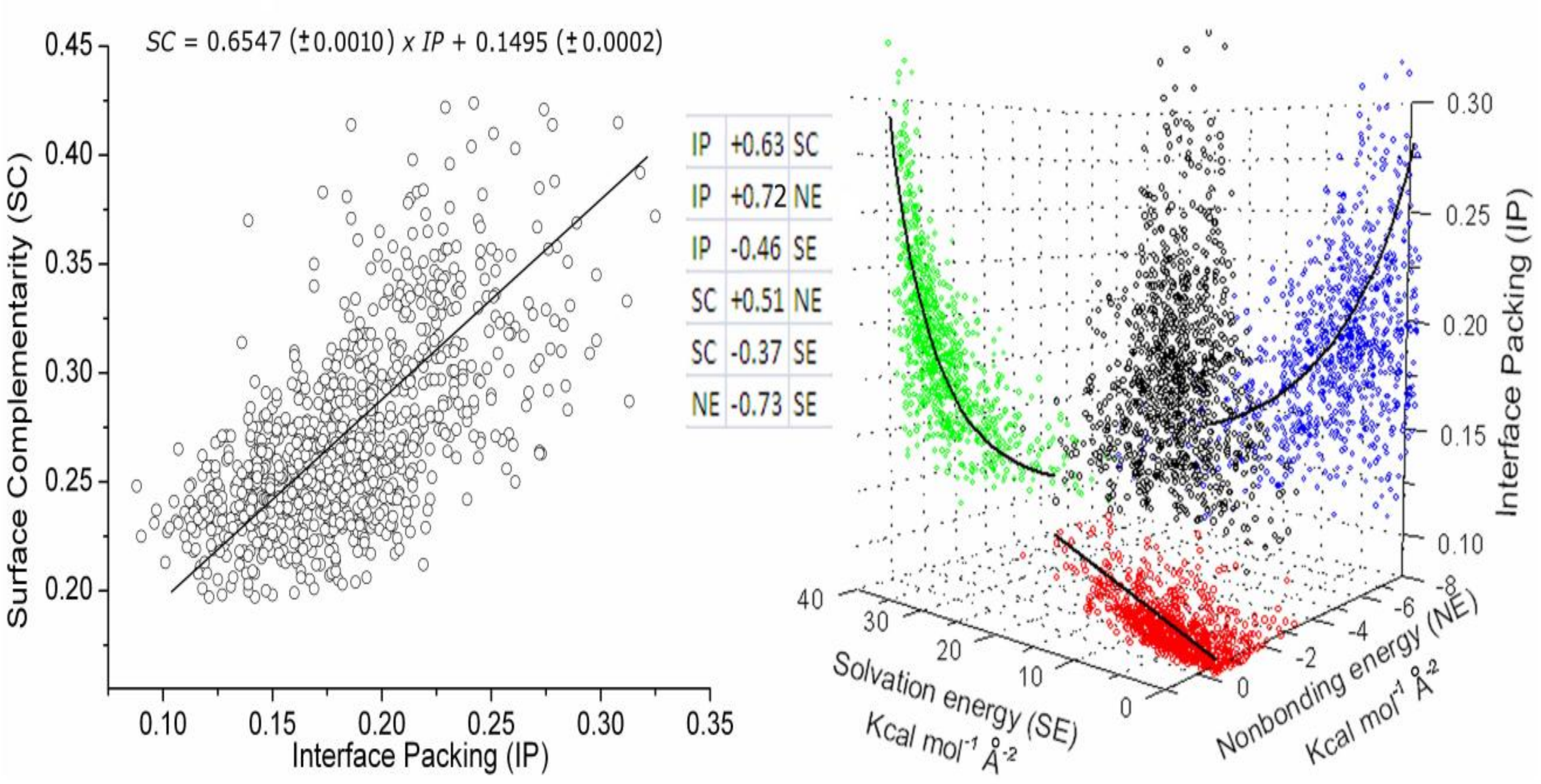
Correlation coefficient of NIP and NSc is **+0.95**

*Mitra and Pal (2010) *FEBS Lett*

*Mitra and Pal (2011) *J. Comput. Chem.*

Scoring methods

Correlation among the four physico-chemical properties at the protein interfaces



IP	+0.63	SC
IP	+0.72	NE
IP	-0.46	SE
SC	+0.51	NE
SC	-0.37	SE
NE	-0.73	SE

Scoring and Ranking*

Compute *IP*, *SC*, *NE* and *SE* at the decoy interface

↓

Group the decoys such that all decoys with RMSD<1.0Å and difference in SP<0.04 is in a group G, where SP = |SC-IP×0.6547-0.1495|

↓

Nonbonded energy for a group G: $NE^G = \overline{NE} - \sigma(NE)$

Solvation energy for a group G: $SE^G = \overline{SE} + \sigma(SE)$

↓

NE_i^G : *NE*^{G} bin number in all groups' *NE* histogram

SE_i^G : *SE*^{G} bin number in all groups' *SE* histogram

↓

$$Score = \sqrt{((NE_i^G \times NE_i^G) + (SE_i^G \times SE_i^G))} + SP^G \times 10.0$$

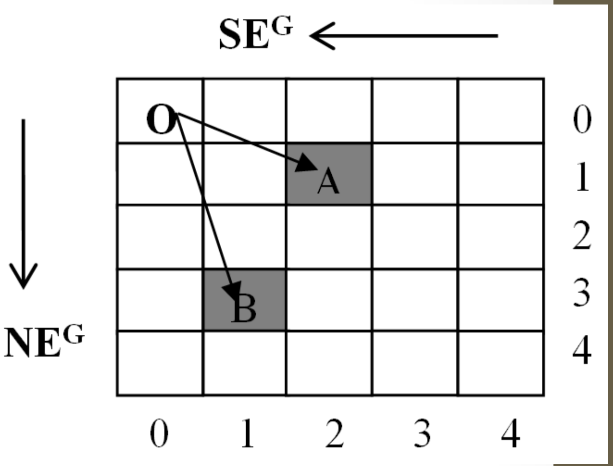
where, *SP*^{G} is minimum SP of the group G.

→

Sort (in ascending order) the group of decoys based upon their scores.

↑

Rank of a decoy is its position in the sorted list

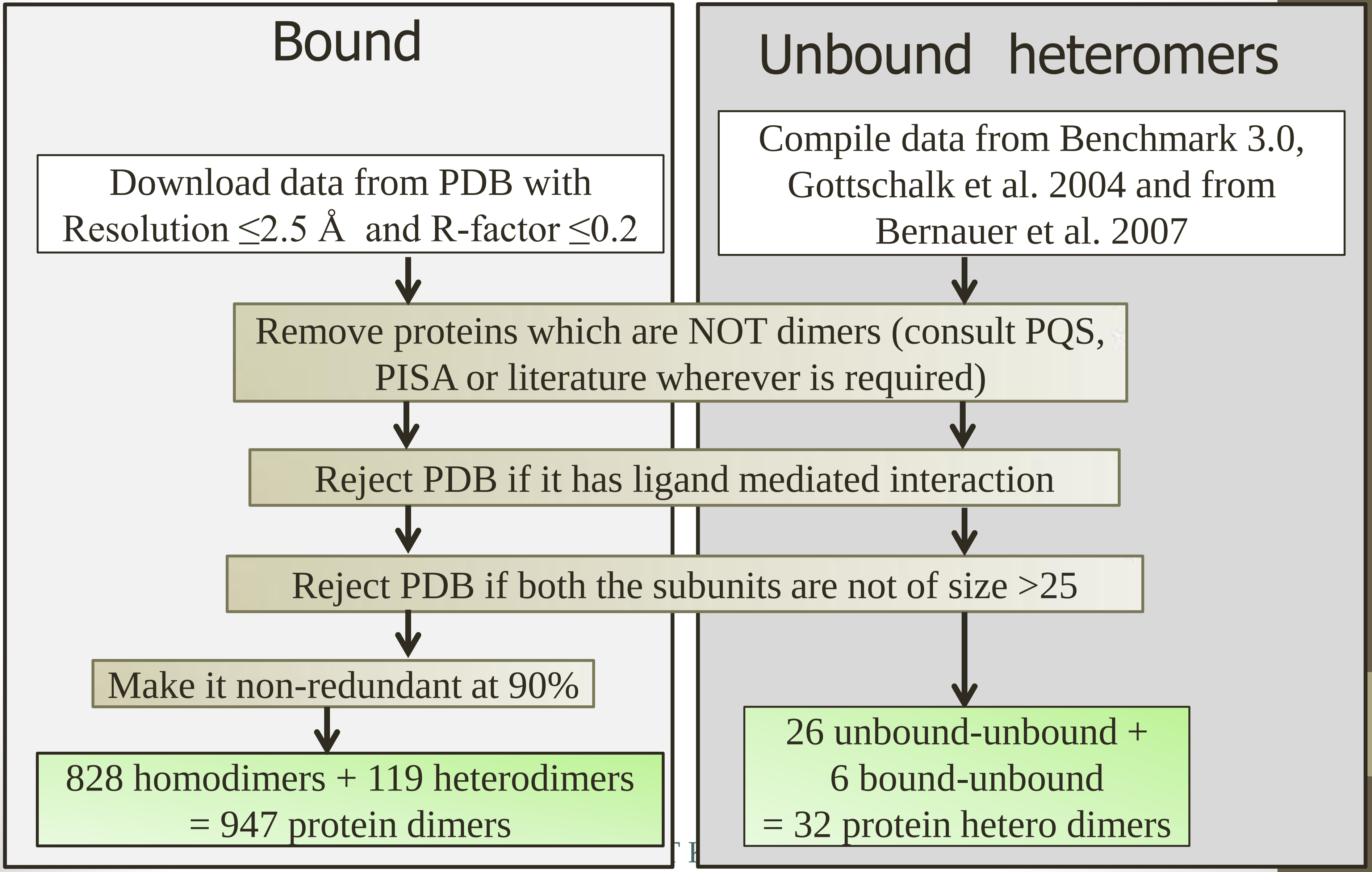


*Mitra and Pal (2011) *J. Comput. Chem.*

Docking types

- Bound docking
  - The crystal structure of complex is available. Interacting/docking partners are taken from that complex structure.
  - Easy to model since the side chain orientation is proper.
- Predictive/Unbound docking
  - The docking partners and complex structure is separately crystallized.
  - Side chain refinement is required

The Dataset



Bound

Download data from PDB with Resolution ≤2.5 Å and R-factor ≤0.2

Unbound heteromers

Compile data from Benchmark 3.0, Gottschalk et al. 2004 and from Bernauer et al. 2007

Remove proteins which are NOT dimers (consult PQS, PISA or literature wherever is required)

Reject PDB if it has ligand mediated interaction

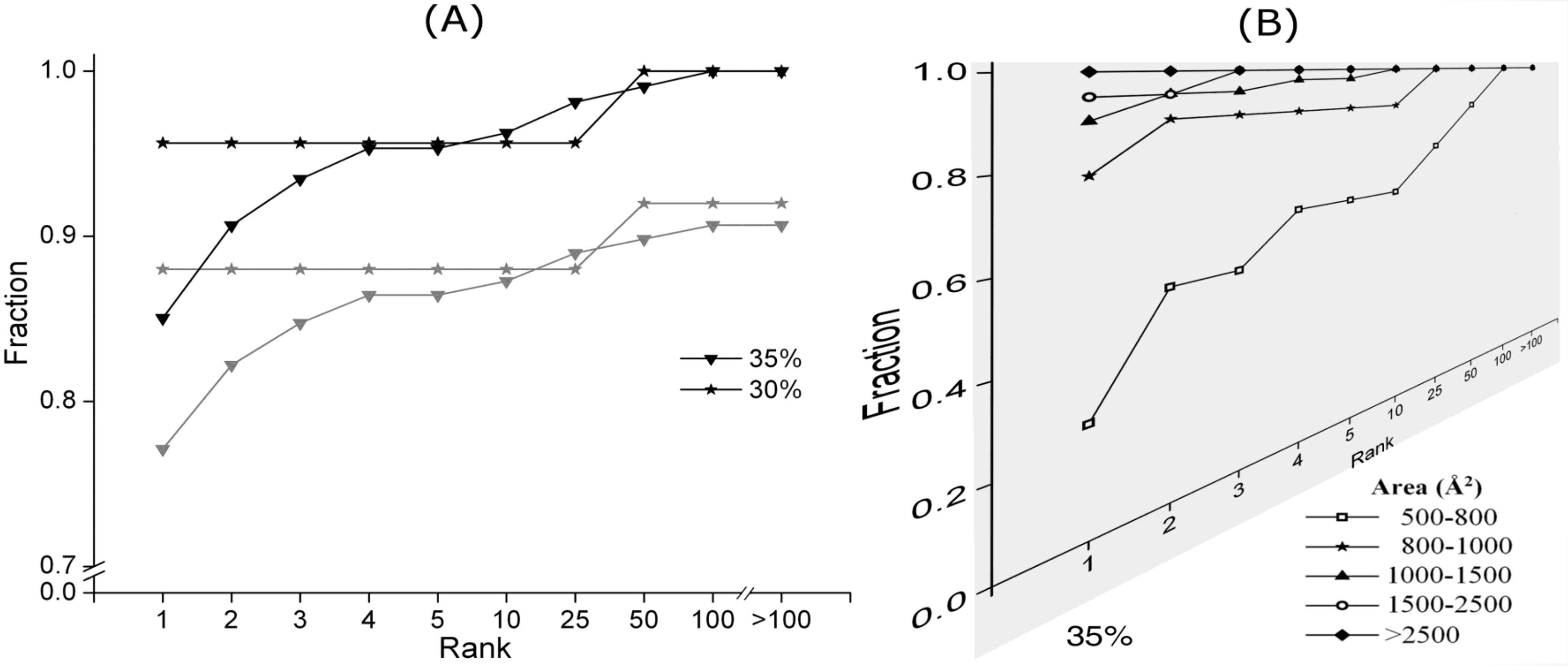
Reject PDB if both the subunits are not of size >25

Make it non-redundant at 90%

828 homodimers + 119 heterodimers = 947 protein dimers

26 unbound-unbound + 6 bound-unbound = 32 protein hetero dimers

Evaluating bound dataset



(A) Variation of accuracy with rank . The darker curve shows the accuracy where the dimers could be successfully screened by IA filter. The lighter curve shows the accuracy over the whole dataset.

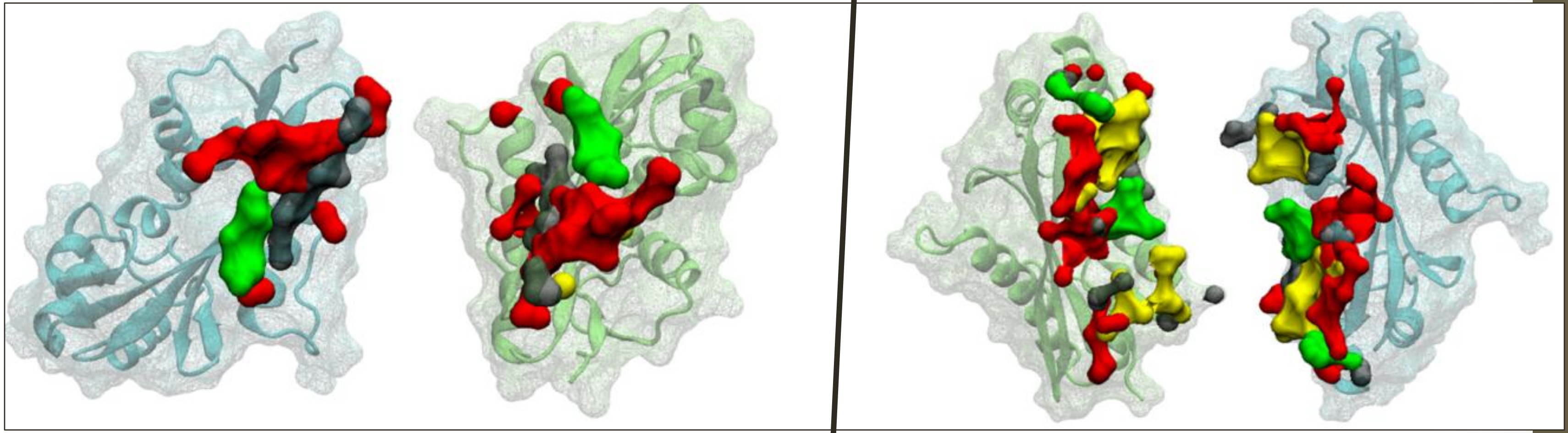
(B) Variation of accuracy with rank when the cases screened by IA filter was divided into various interface area categories.

*Mitra and Pal (2011) *J. Comput. Chem.*

Example prediction (PDB: 1EX2)

🟥 Charged 🟩 Aromatic 🟨 Hydrophobic 🟫 Polar

Our prediction



PDB and PQS structure

Residue property at the interface of the protein
- a conserved *Bacillus subtilis* protein Maf

*Mitra and Pal (2011) *J. Comput. Chem.*

ZRANK

$$Score = w_{\mathrm{vdW\_a}}E_{\mathrm{vdW\_a}} + w_{\mathrm{vdW\_r}}E_{\mathrm{vdW\_r}} + w_{\mathrm{elec\_sra}}E_{\mathrm{elec\_sra}} + w_{\mathrm{elec\_srr}}E_{\mathrm{elec\_srr}} + w_{\mathrm{elec\_lra}}E_{\mathrm{elec\_lra}} + w_{\mathrm{elec\_lrr}}E_{\mathrm{elec\_lrr}} + w_{\mathrm{ds}}E_{\mathrm{ds}}$$

$$E_{\mathrm{vdW}}(i,j) = \varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right)$$

Van der Wall interaction

$$E_{\mathrm{elec}}(i,j) = 332\frac{q_i q_j}{r_{ij}^2}$$

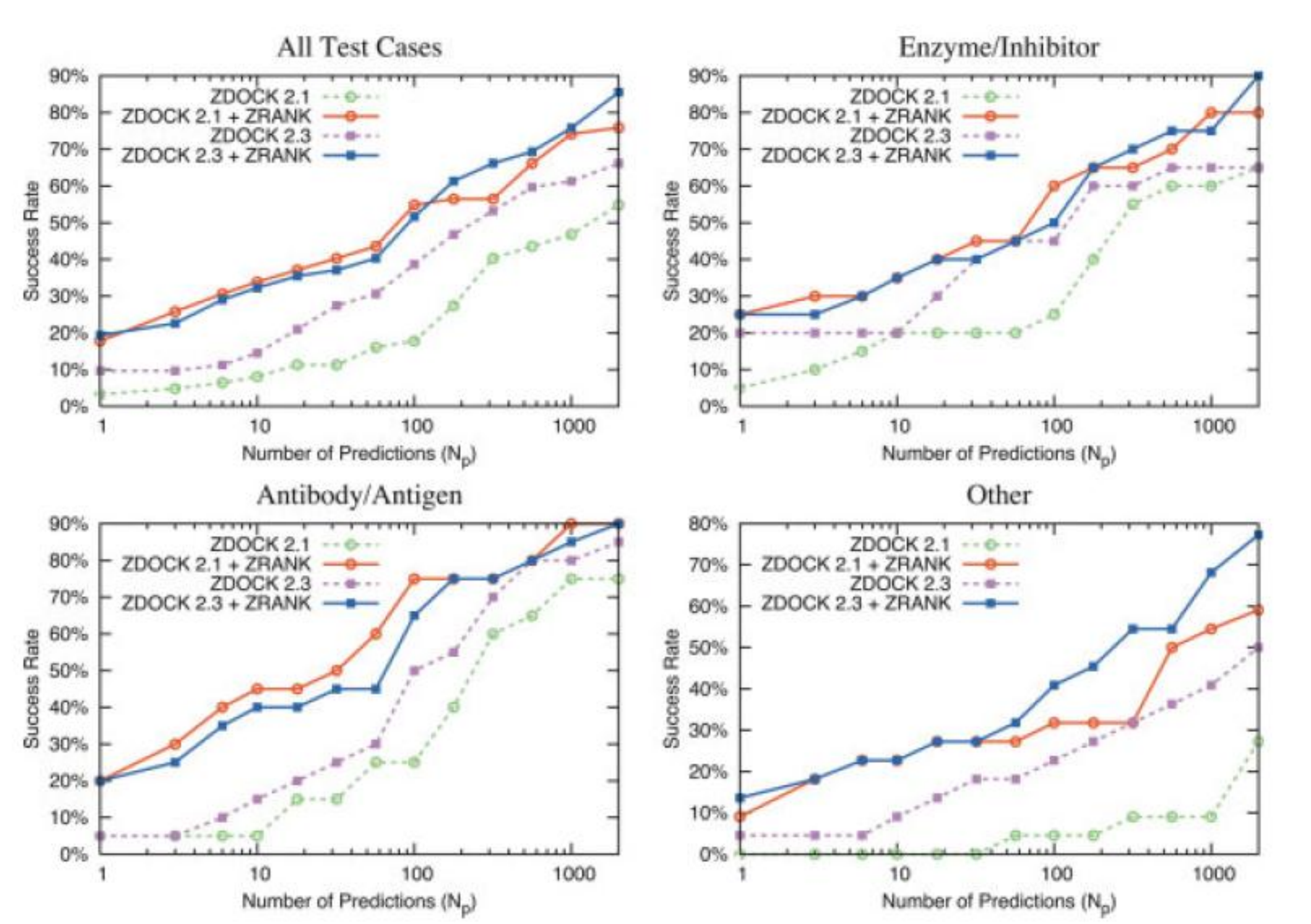
Electrostatic Interaction

$$E_{\mathrm{ds}}(i,j) = a_{ij}$$

Desolvation energy

Pierce and Weng (2007)
Proteins, 67:1078–1086

ZRANK



Pierce and Weng (2007)
Proteins, 67:1078–1086

PatchDock and FireDock

- **PatchDock:** Molecular Docking Algorithm Based On Shape Complementarity Principles
- **FireDock:** Includes three main steps:

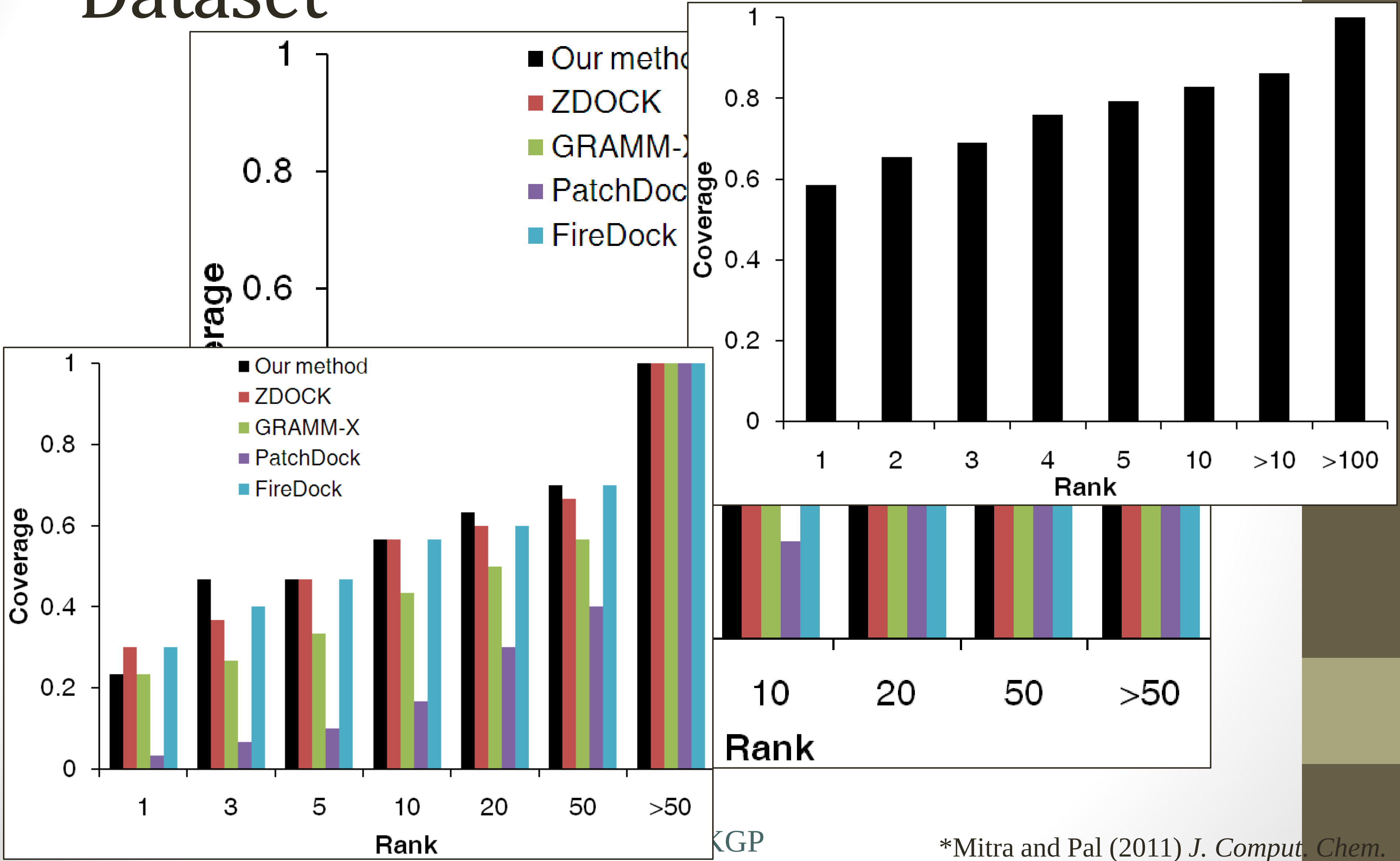
(1) Side-chain optimization: The side-chain flexibility of the receptor and the ligand is modeled by a rotamer library. The optimal combination of rotamers for the interface residues is found by solving an integer LP problem.

(2) Rigid-body minimization: This minimization stage is performed by a MC technique that attempts to optimize an approximate binding energy by refining the orientation of the ligand structure.

(3) Scoring and ranking: This final ranking stage attempts to identify the near-native refined solutions. The ranking is performed according to a binding energy function that includes a variety of energy terms: desolvation energy, van der Waals interactions, partial electrostatics, hydrogen and disulfide bonds, *p*-stacking and aliphatic interactions, rotamer's probabilities and more.

Schneidman-Duhovny et al. (2005)
Nucleic Acids Research, 33, W363–W367

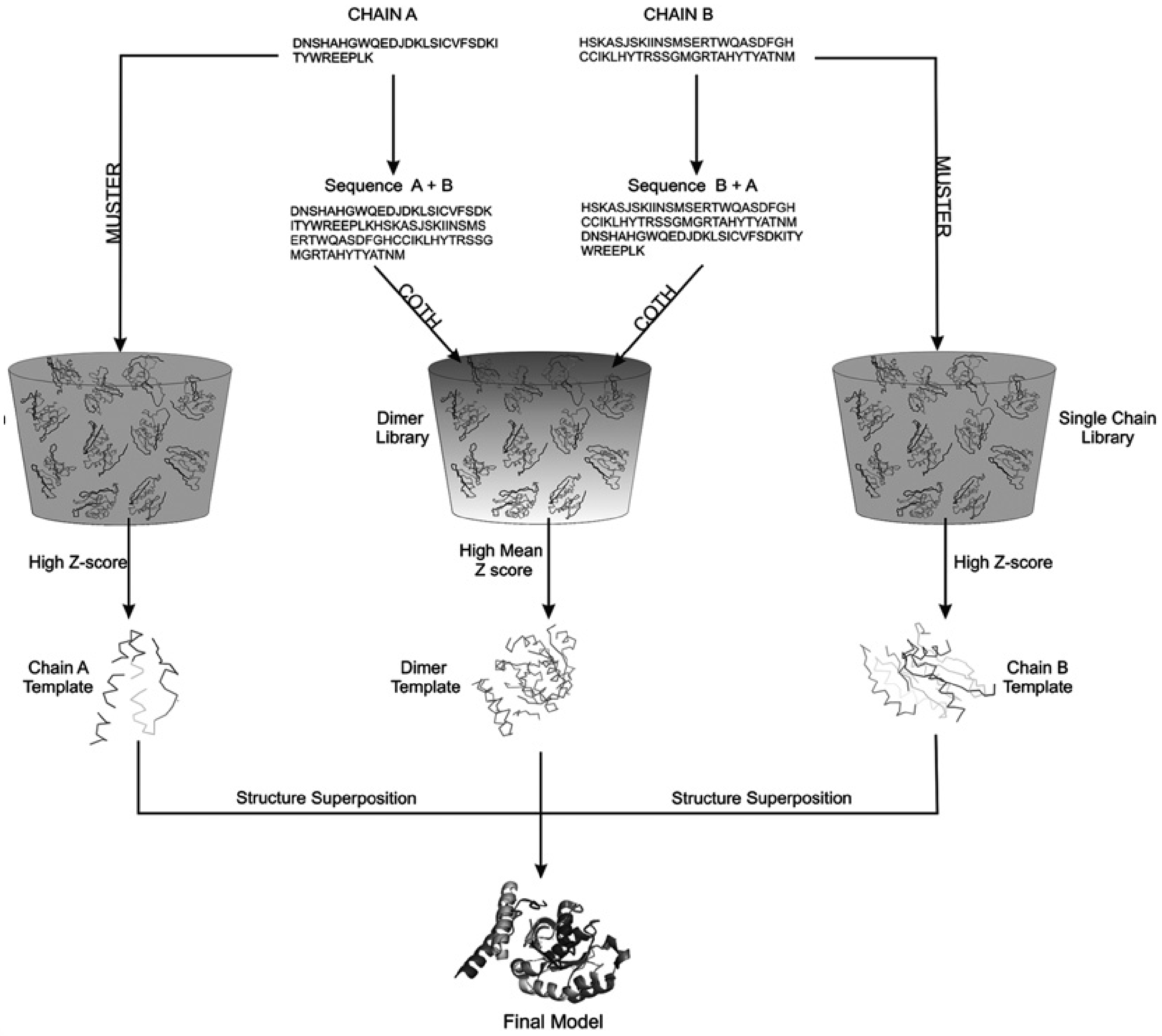
Predictive Docking - The Unbound Dataset*



*Mitra and Pal (2011) _J. Comput. Chem._

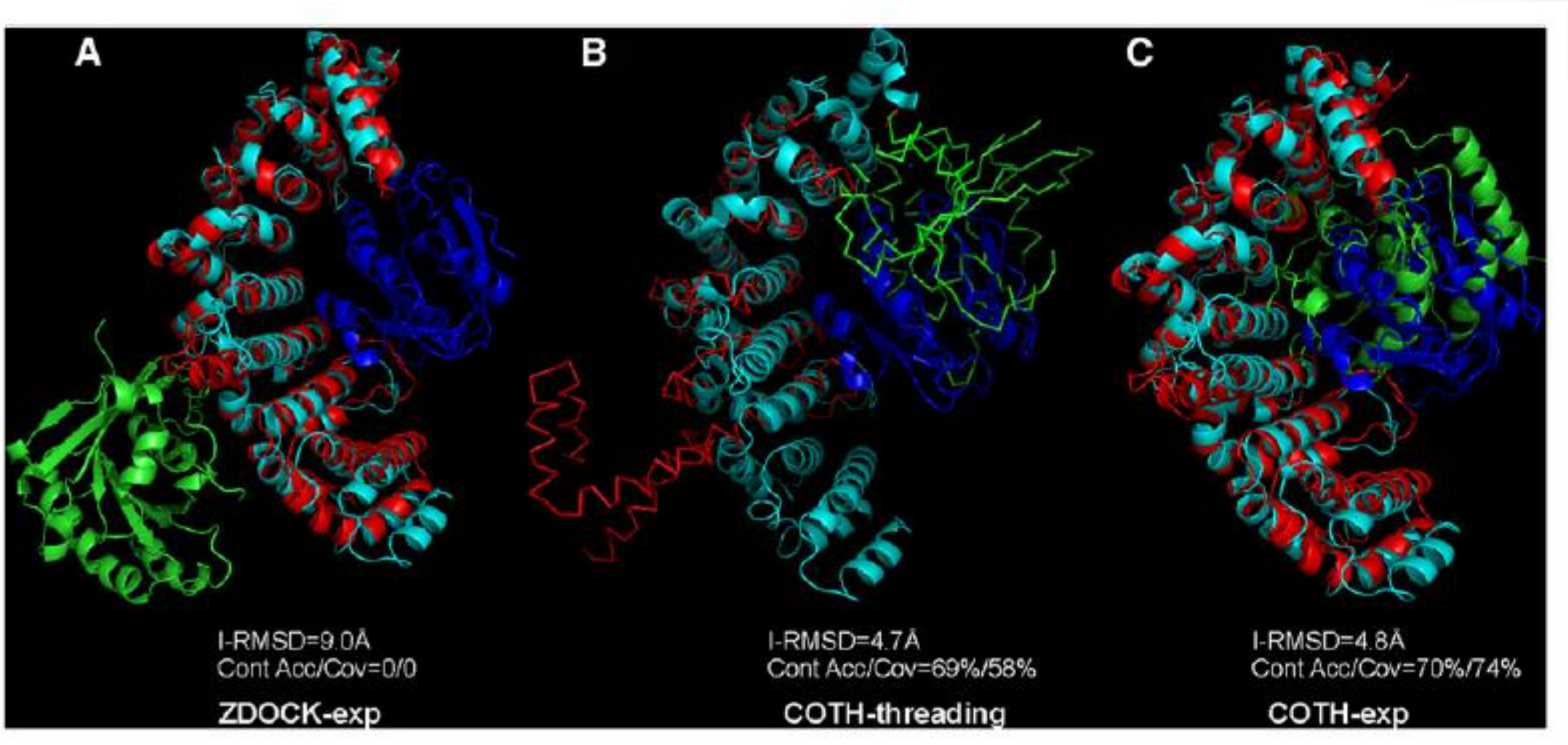
Docking from sequence
Application to Genome-wide scale

COTH – docking from sequence



Mukherjee and Zhang (2011)
Structure 19, 955–966

COTH – docking from sequence



The native complex (Ran-Importin β complex) is represented in cyan.

Mukherjee and Zhang (2011)
Structure 19, 955–966

Critical Assessment of PRediction of Interactions (CAPRI)

Critical Assessment of PRediction of Interactions (CAPRI)

Predictor	***Affiliation***	***Software***	***Algorithm***
Abagyan	Scripps	ICM	Force Field
Camacho/Vajda	Boston	CHARMM	Force Field Refinement
Gardiner	Sheffield	GAPDOCK	Shape+Area GA
Sternberg/Smith	Imperial	FTDOCK	FFT
Bates/Fitzjohn	ICRF	Guided Docking	Force Field
Ten Eyck/Mitchell	SDSC	DOT	FFT
Vakser/Tovchigrechko	SUNY/MUSC	GRAMM	FFT
Olson	Scripps	Harmony	Spherical Harmonics
Weng/Chen	Boston	ZDOCK	FFT
Eisenstein	Weizmann	MolFit	FFT
Wolfson/Nussinov	Tel Aviv	BUDDA/PPD/FireDock	Geometric Hashing
Iwadate	Kitasato	TSCF	Force Field+Solvent
Ritchie/Mustard	Aberdeen	Hex	Spherical Polar Fourier
Palma	Lisbon	BIGGER	Geometric+Electrostatic
Gray/Baker	Washington/JHU	RosettaDock	Monte Carlo+Flexibility
Mitra and Pal	**IISc**	**PROBE/PRUNE**	**FFT**

T50, T53

Parallel Implementation

- At the generation phase:
  - The protein can be divided into different parts that are mutually exclusive.
- At the scoring phase:
  - All the decoys are mutually independent; thus they can be processed separately on different processors.

Summary

✓ The bound test set is easy to predict, but the real benchmark set is unbound data set.

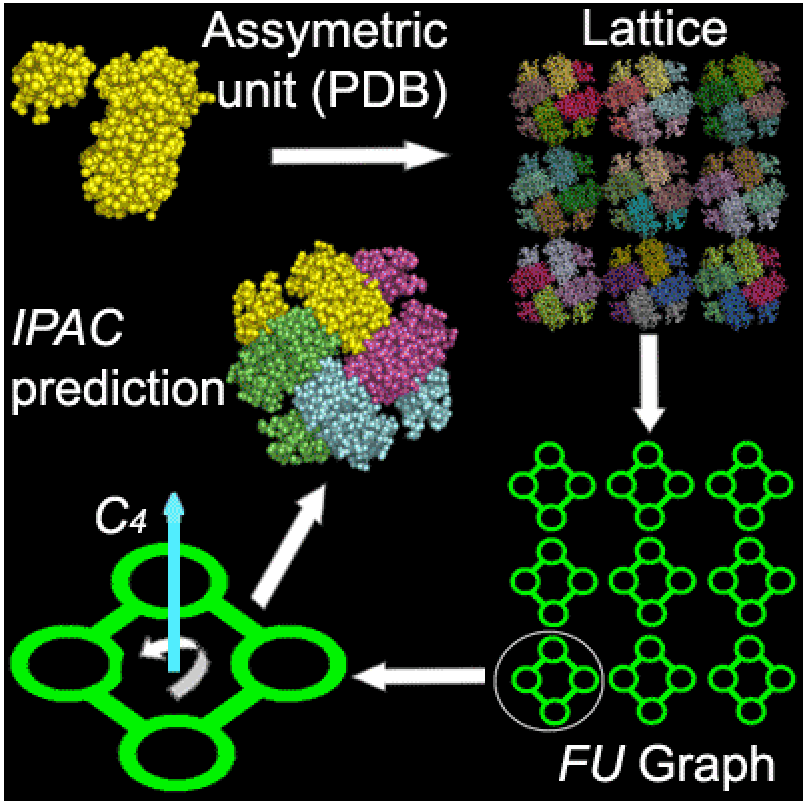
✓ Refining the side chain of the unbound docked complexes are still an active area of research.

✓ Computationally flexible docking is more challenging than rigid body docking.

✓ Evolutionary information can be integrated to improve the performance of the method.

Thank you for your attention

http://cse.iitkgp.ac.in/~pralay/


Assymetric unit (PDB)
Lattice
IPAC prediction
C_4
FU Graph