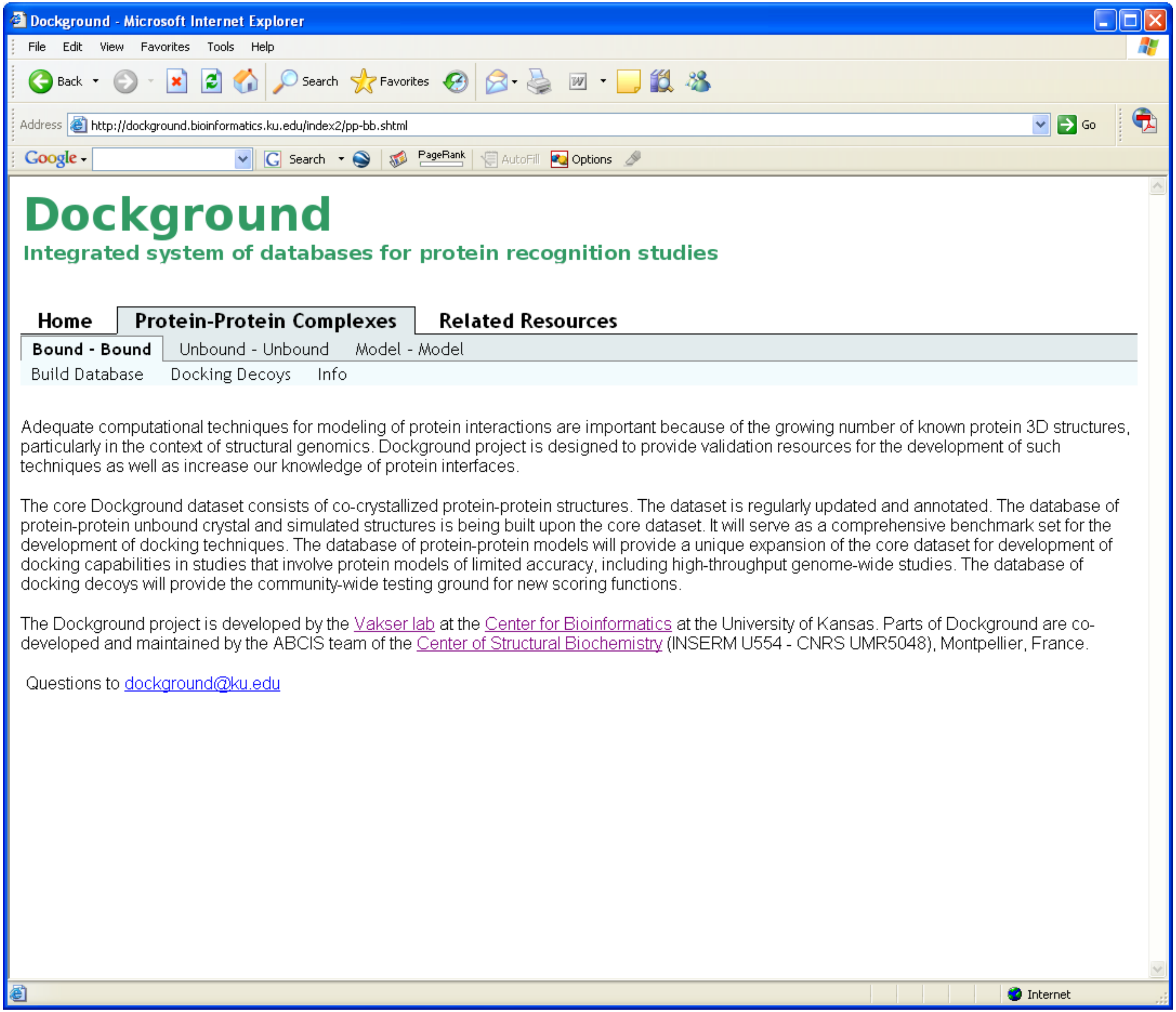
Protein Docking

Ilya Vakser
University of Kansas
vakser@ku.edu

Quaternary Structure



Dockground

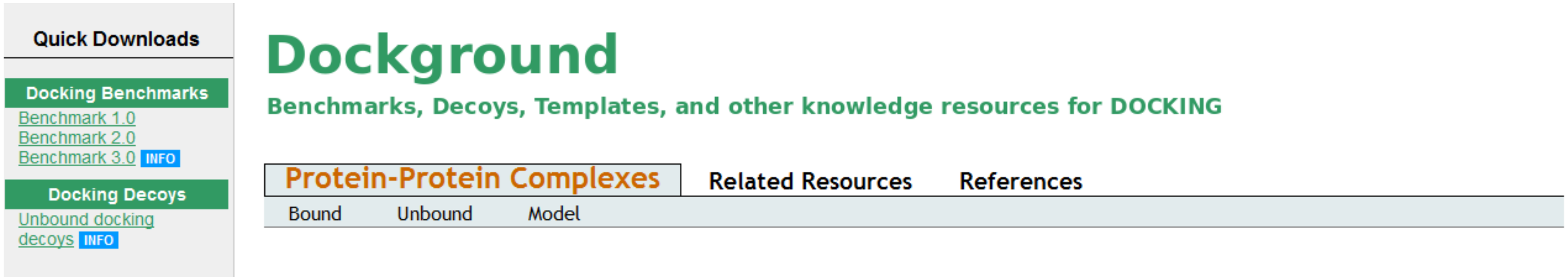
Integrated system of databases for protein recognition studies

Adequate computational techniques for modeling of protein interactions are important because of the growing number of known protein 3D structures, particularly in the context of structural genomics. Dockground project is designed to provide validation resources for the development of such techniques as well as increase our knowledge of protein interfaces.

The core Dockground dataset consists of co-crystallized protein-protein structures. The dataset is regularly updated and annotated. The database of protein-protein unbound crystal and simulated structures is being built upon the core dataset. It will serve as a comprehensive benchmark set for the development of docking techniques. The database of protein-protein models will provide a unique expansion of the core dataset for development of docking capabilities in studies that involve protein models of limited accuracy, including high-throughput genome-wide studies. The database of docking decoys will provide the community-wide testing ground for new scoring functions.

The Dockground project is developed by the Vakser lab at the Center for Bioinformatics at the University of Kansas. Parts of Dockground are co-developed and maintained by the ABCIS team of the Center of Structural Biochemistry (INSERM U554 - CNRS UMR5048), Montpellier, France.

Questions to dockground@ku.edu



Quick Downloads

Docking Benchmarks

Benchmark 1.0

Benchmark 2.0

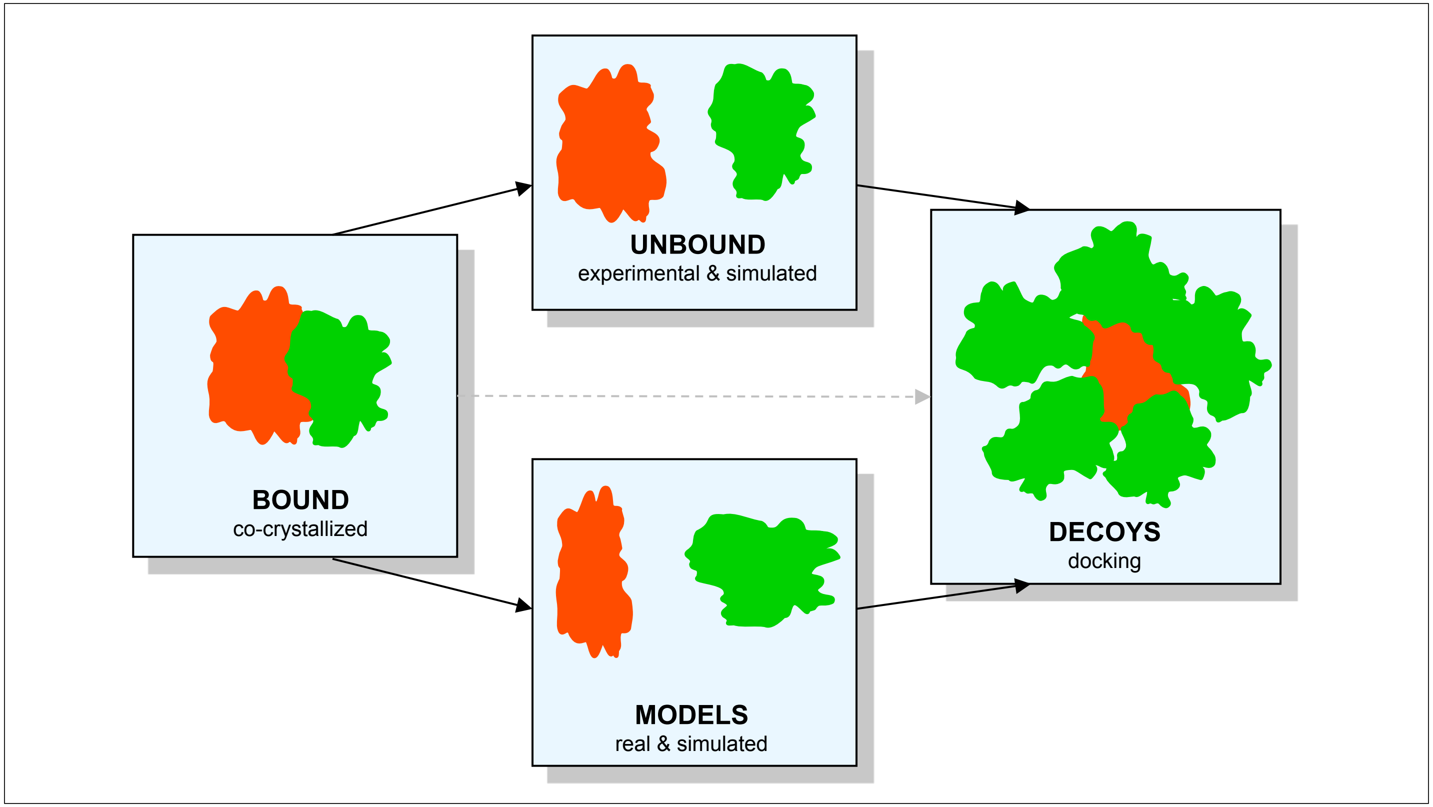
Benchmark 3.0 INFO

Docking Decoys

Unbound docking decoys INFO

Dockground

Benchmarks, Decoys, Templates, and other knowledge resources for DOCKING

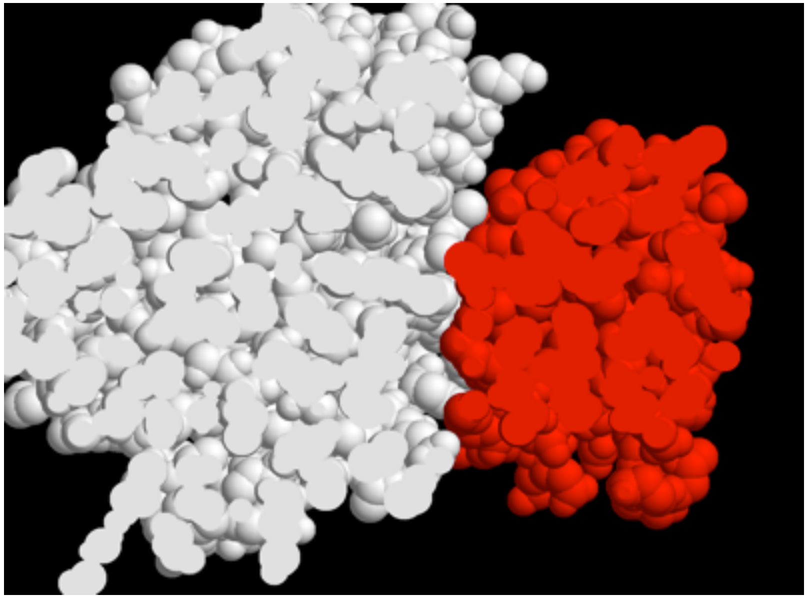


Complexes

- Homologous and Unique
- Obligate and Transient

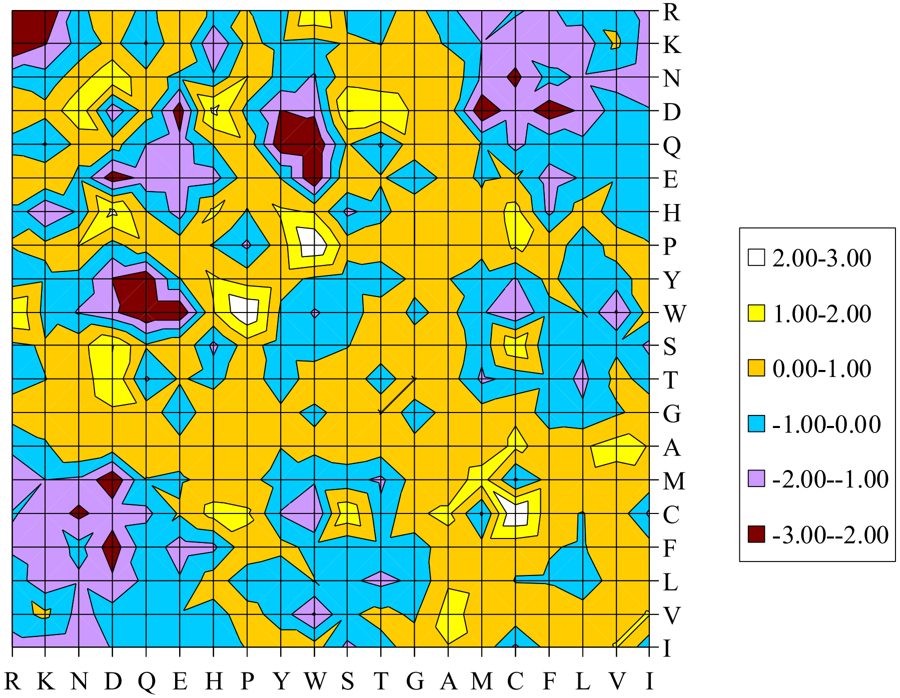
Complementarity

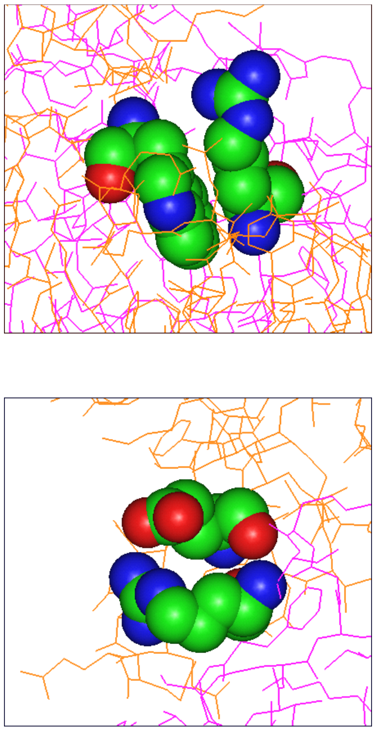
Steric



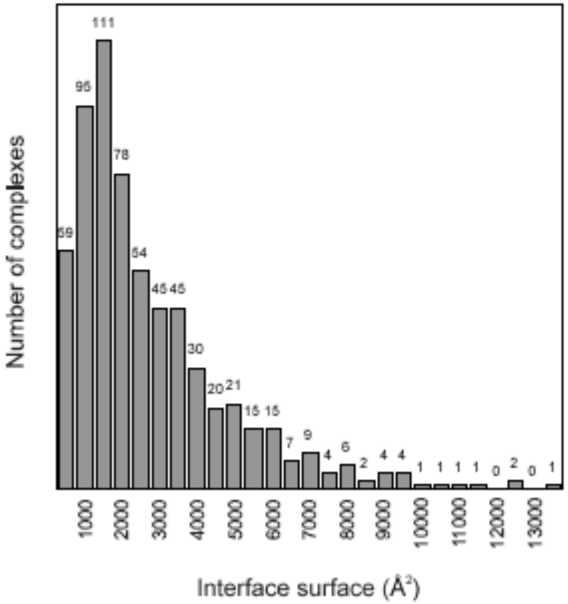
Complementarity

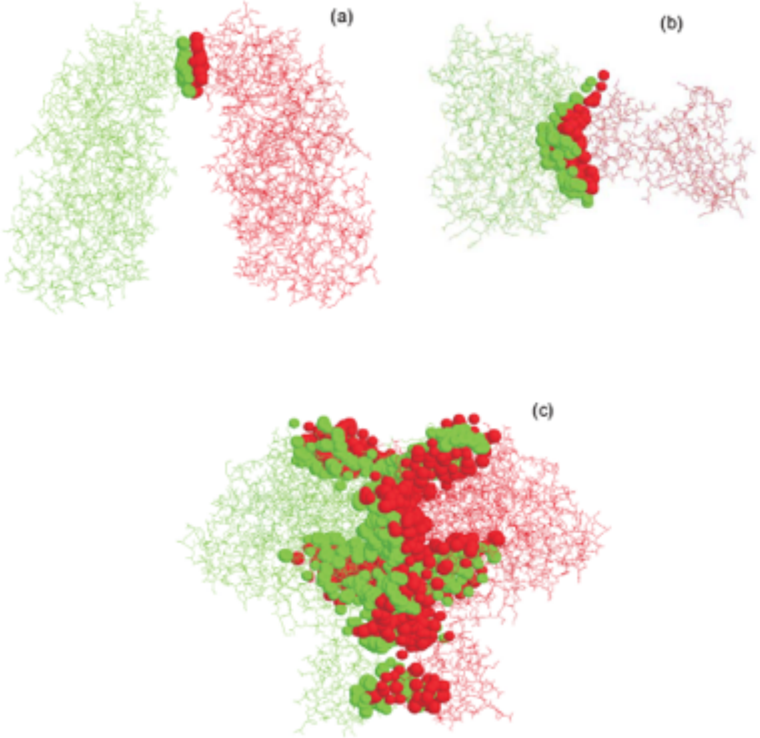
Physicochemical





Interface sizes





Complexes

	Oblig-Trans	Hydroph/ES	Packing	Size	Predictability
Crystal packing					
Multisubunit					
Enzyme-Inhibitor					
Electron transfer					
Antigen-Antibody					
Other…					

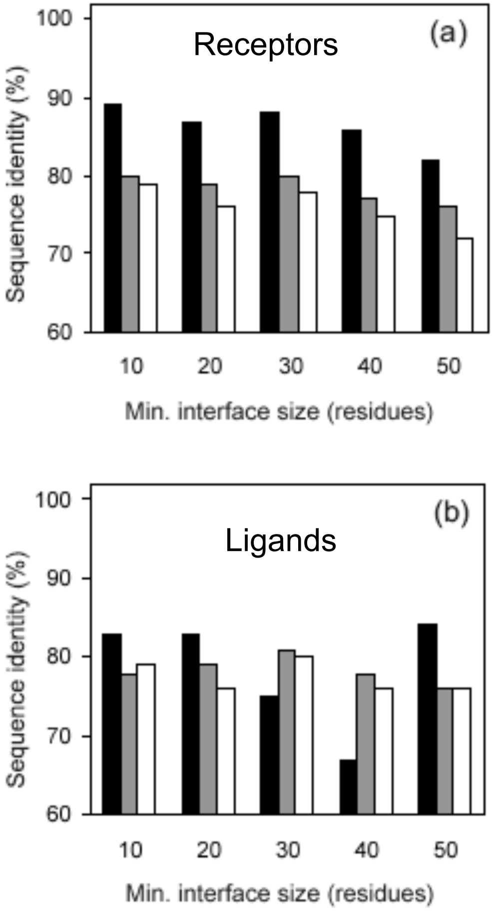
Hot spots

Experimentally determined by alanin mutations within interface and measuring binding affinity

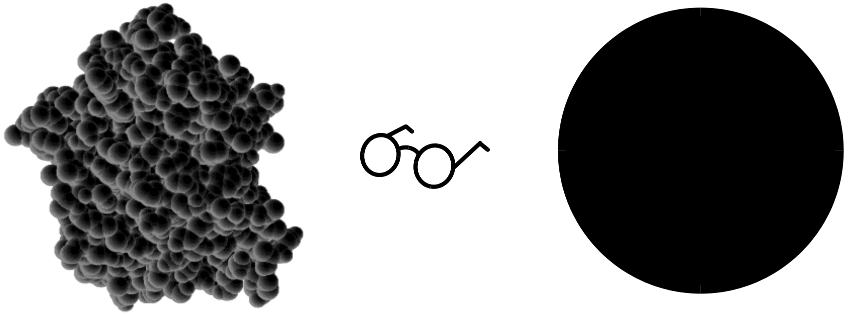
- Near the center of the interface
- Solvent inaccessible
- Complementary to each other across the interface
- More conserved
- Not many
- Bound and unbound conformations same
  (anchor residues on smaller protein)

Binding site conservation

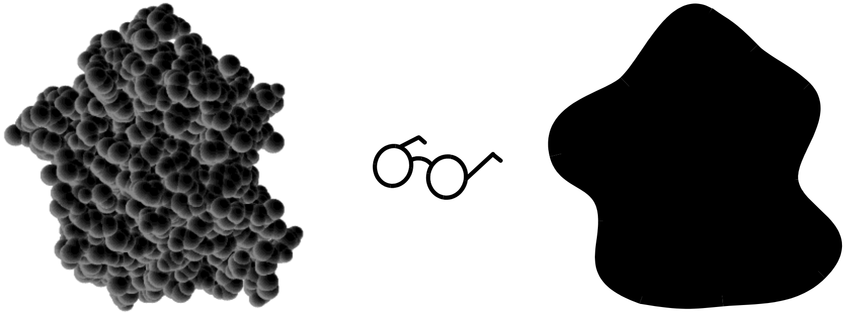
Sequence identity of the core, surface interface, and surface non-interface residues



Large recognition factors

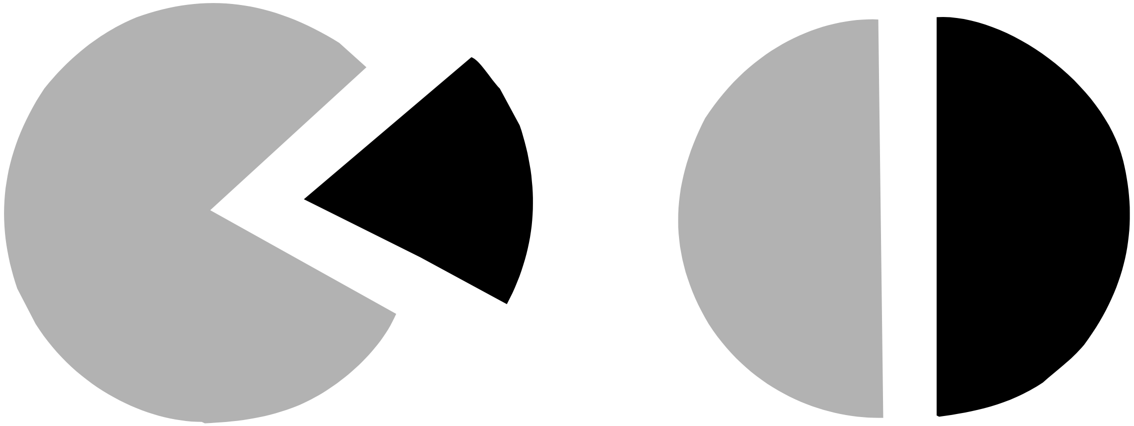


No...



Yes!

Large recognition factors



?

Larger protein b/site is concave

Evidence:

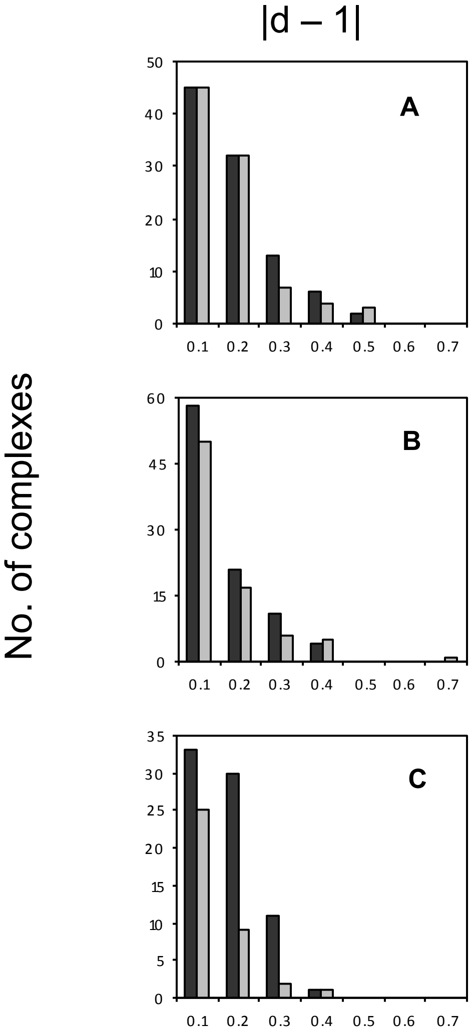
Observation of co-crystallized structures

Docking

Surface residues distances from CM

$d = \langle d_{in} \rangle / \langle d_{out} \rangle$

■ less than average ☐ more than average



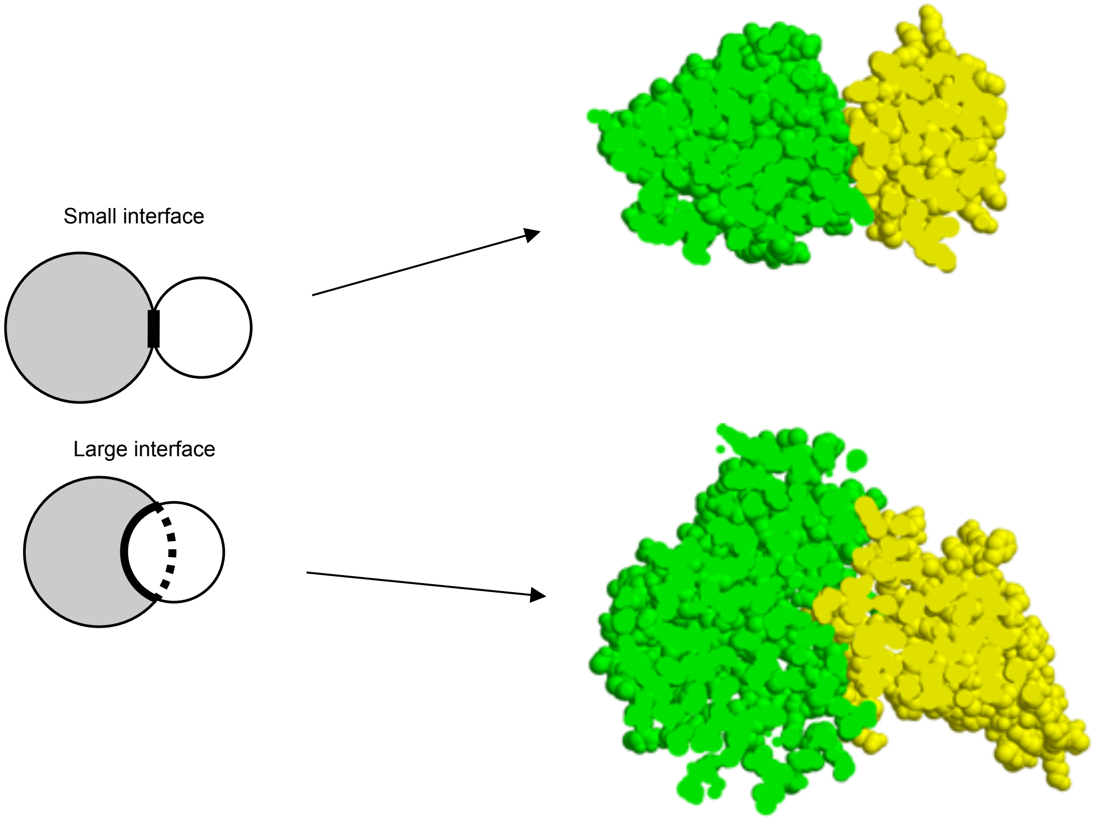
Interface:

Small (1000-2000Å^2)

Medium (2000-4000Å^2)

Large (>4000Å^2)

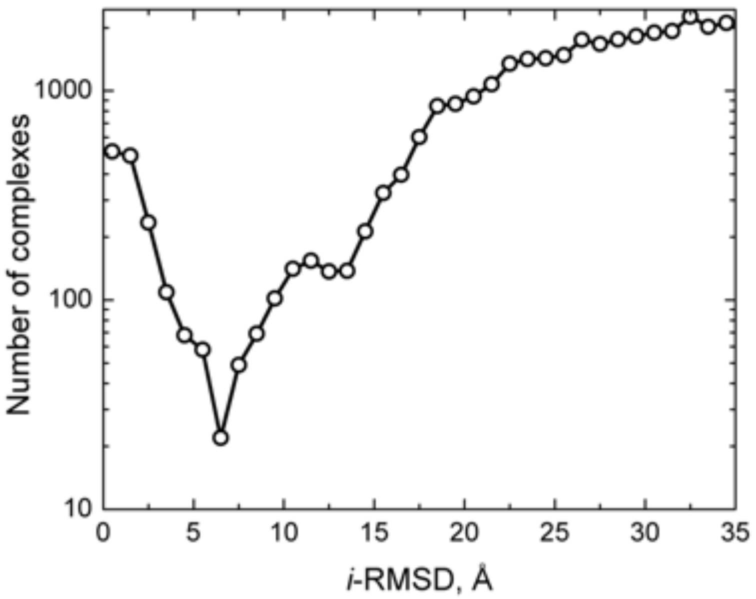
Nicola G, Vakser IA. A simple shape characteristic of protein-protein recognition. *Bioinformatics*. 2007, 23:789-92



Small interface

Large interface

Distribution of complexes generated by structural alignment with homologous templates

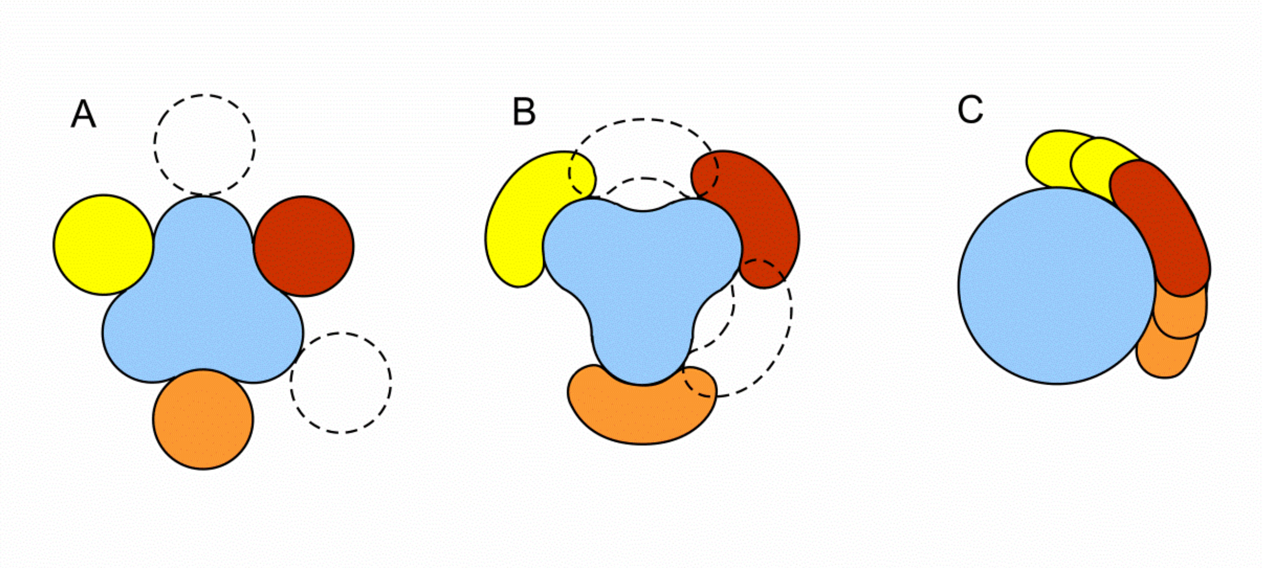


Targets: 372 non-redundant (< 30% SeqID) complexes

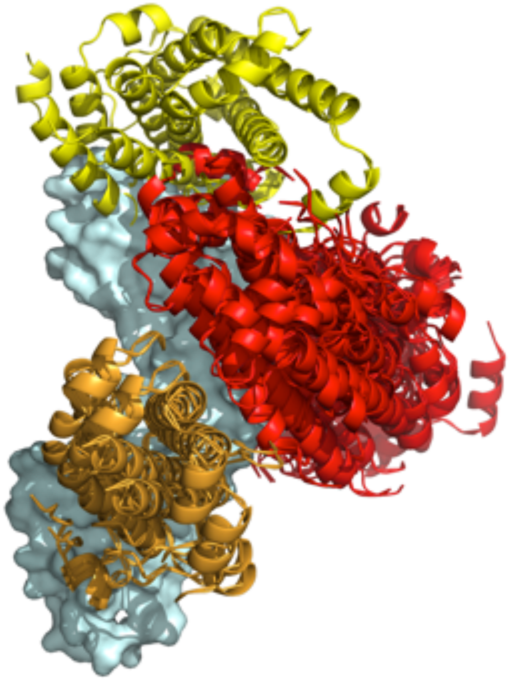
Templates: 11,932 redundant (< 90% SeqID) complexes from biounit

Kundrotas PJ, Vakser IA. Protein-protein alternative binding modes do not overlap. *Protein Sci*, 2013, 22:1141-1145

The concept of alternative binding modes

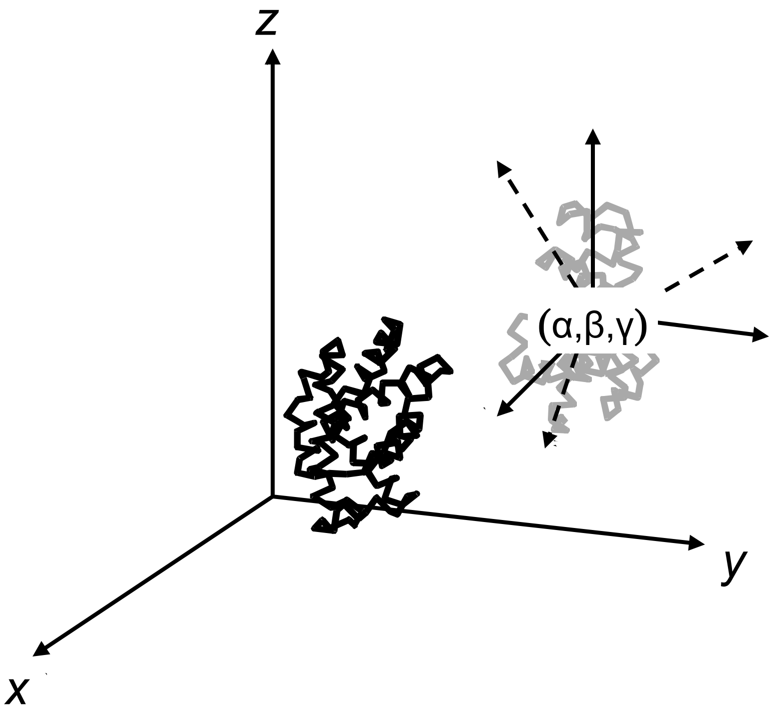


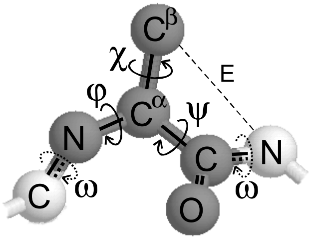
Example of alternative binding modes



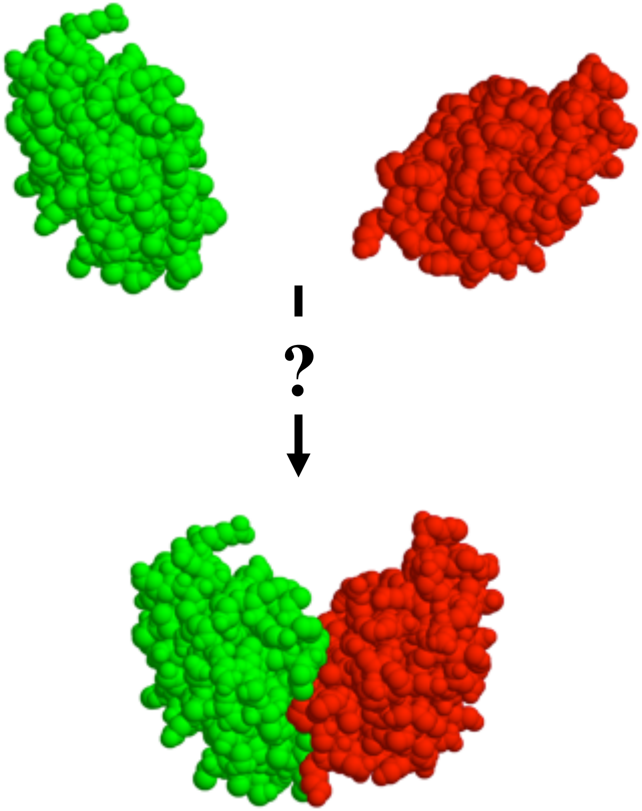
Docking Foundations

Degrees of freedom





Approach to Docking

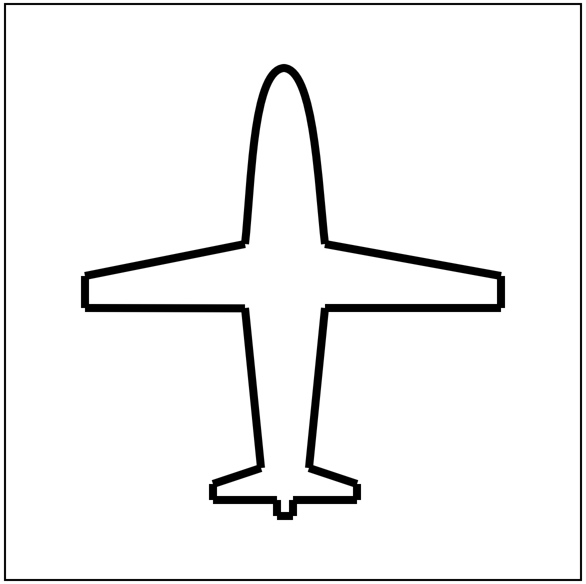


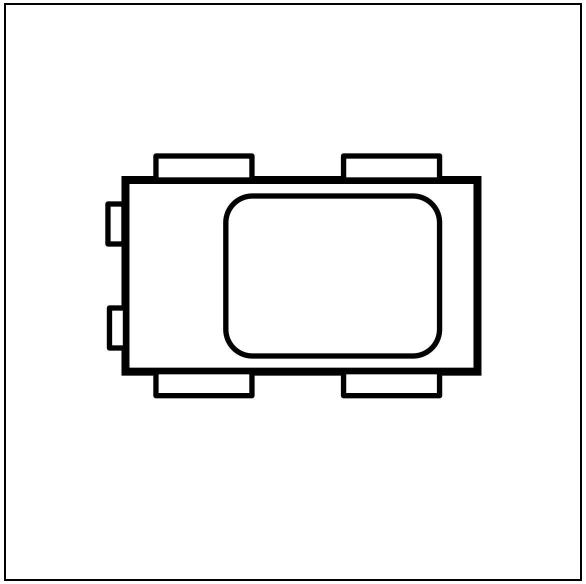
Manually	**No…**
MM or MD	**No…**
Engineering	**Yes!**

image recognition

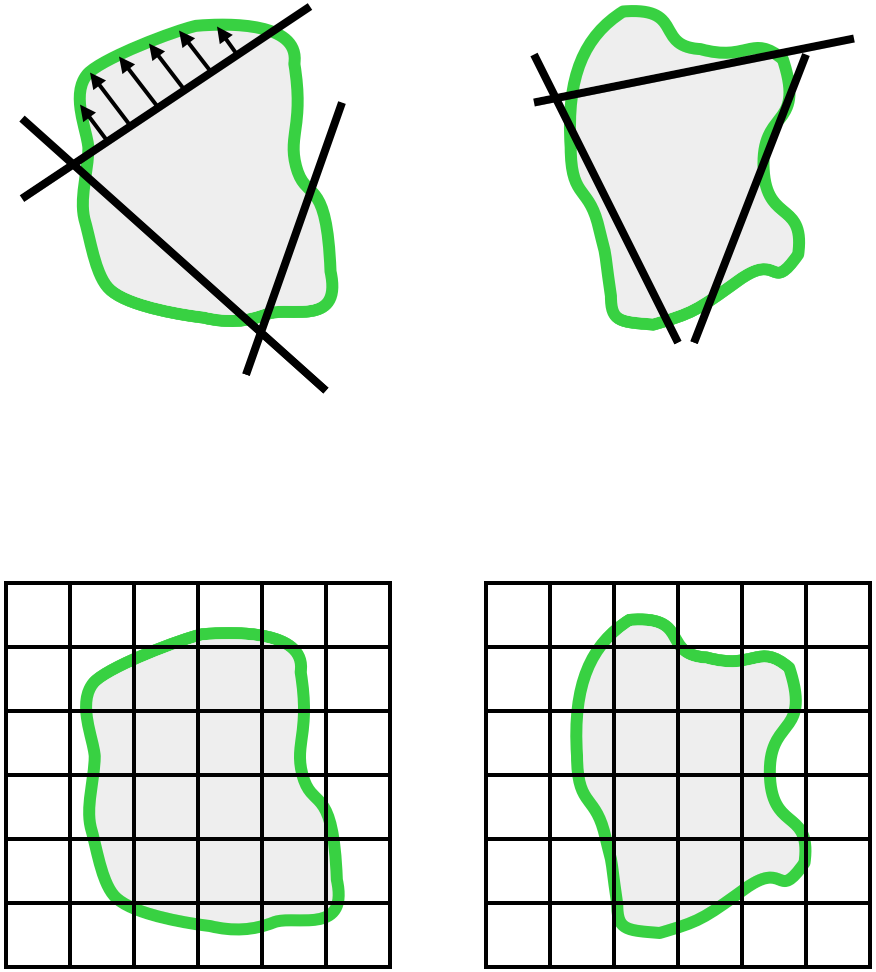
spacecraft docking

Pattern Recognition



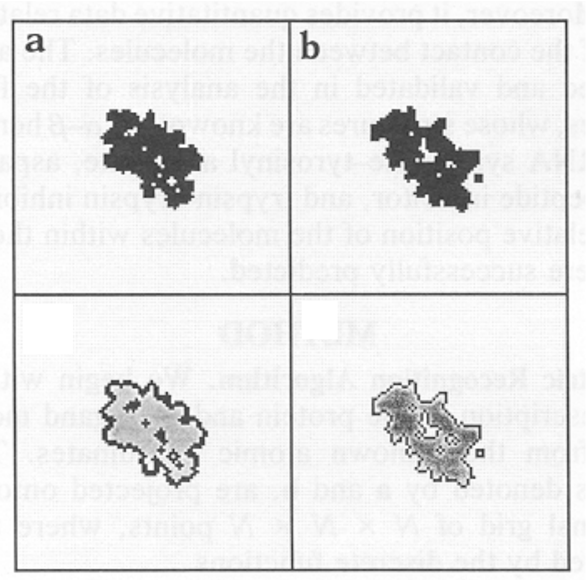


How to represent proteins



Structure digitization

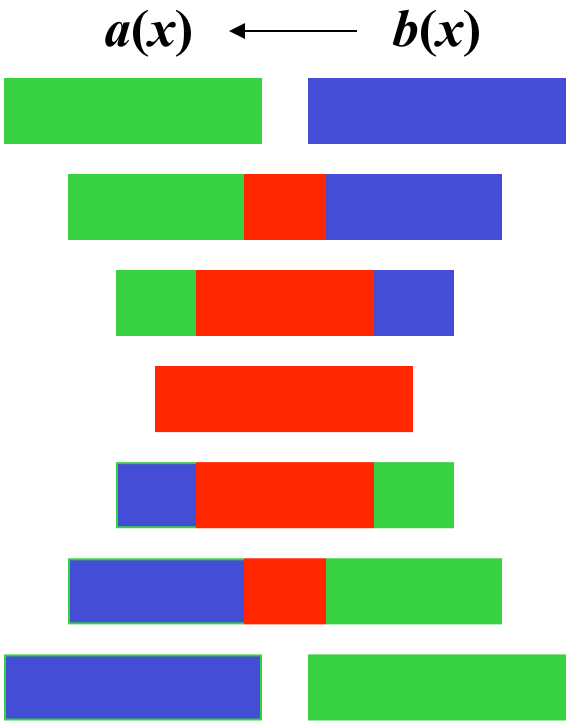
$N \times N \times N$ grid

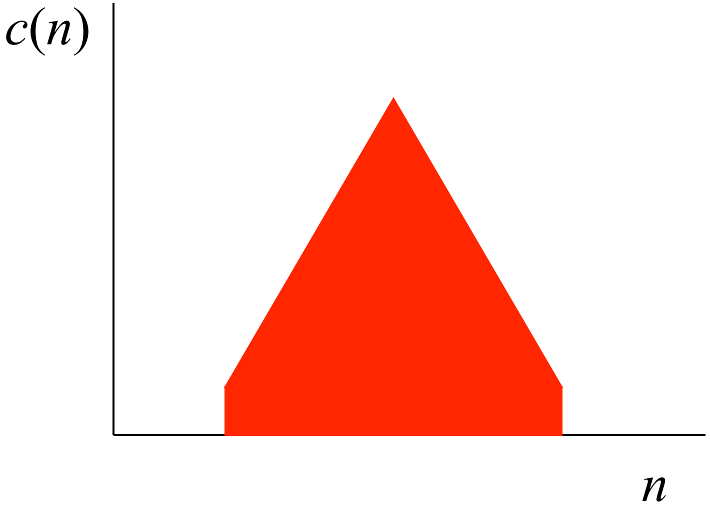


$$a_{l,m,n} = \begin{cases} 1 & \text{inside} \\ 0 & \text{outside} \end{cases} \qquad b_{l,m,n} = \begin{cases} 1 & \text{inside} \\ 0 & \text{outside} \end{cases}$$

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

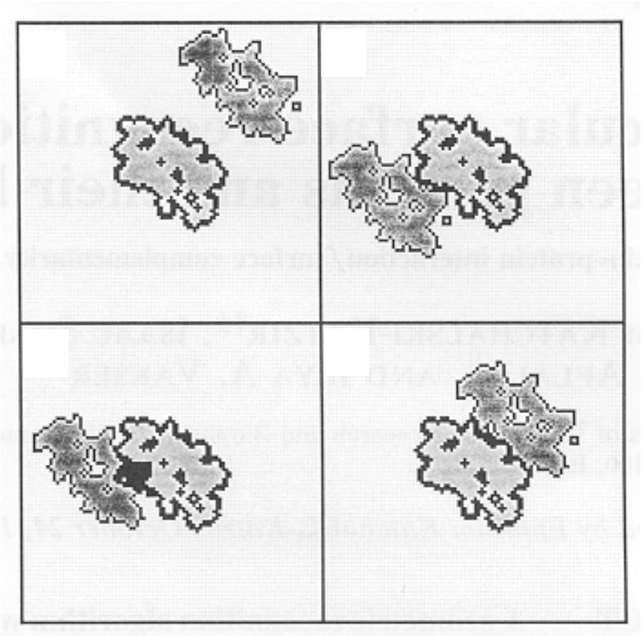
Correlation





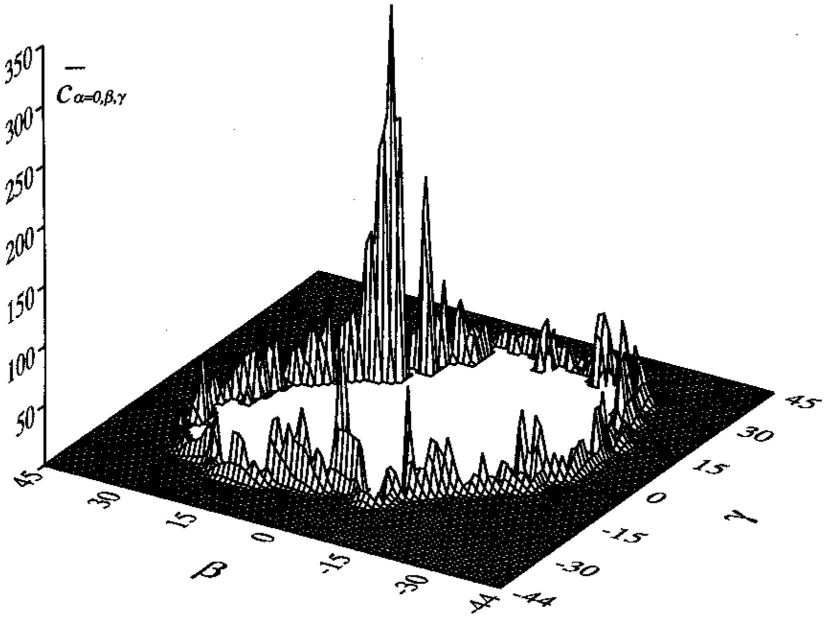
$$c(n) = \sum_{n=-\infty}^{+\infty} a(x) \cdot b(x + n\Delta x)$$

Matching by correlation



$$\bar{c}_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \bar{a}_{l,m,n} \cdot \bar{b}_{l+\alpha,m+\beta,n+\gamma}$$

Correlation diagram



Fast Fourier Transformation

$$\bar{c}_{\alpha,\beta,\gamma} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} \bar{a}_{l,m,n} \cdot \bar{b}_{l+\alpha,m+\beta,n+\gamma}$$

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

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

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

Algorithm

(*i*) derive *a* from atomic coordinates of molecule **a**, by projecting them on a grid

(*ii*) $A^* = [\mathrm{DFT}(a)]$ (A^* is the complex conjugate of A)

(*iii*) derive *b* from atomic coordinates of molecule **b**, by projecting them on a grid

(*iv*) $B = \mathrm{DFT}(b)$

(*v*) $C = A^* \cdot B$

(*vi*) $c = \mathrm{IFT}(C)$

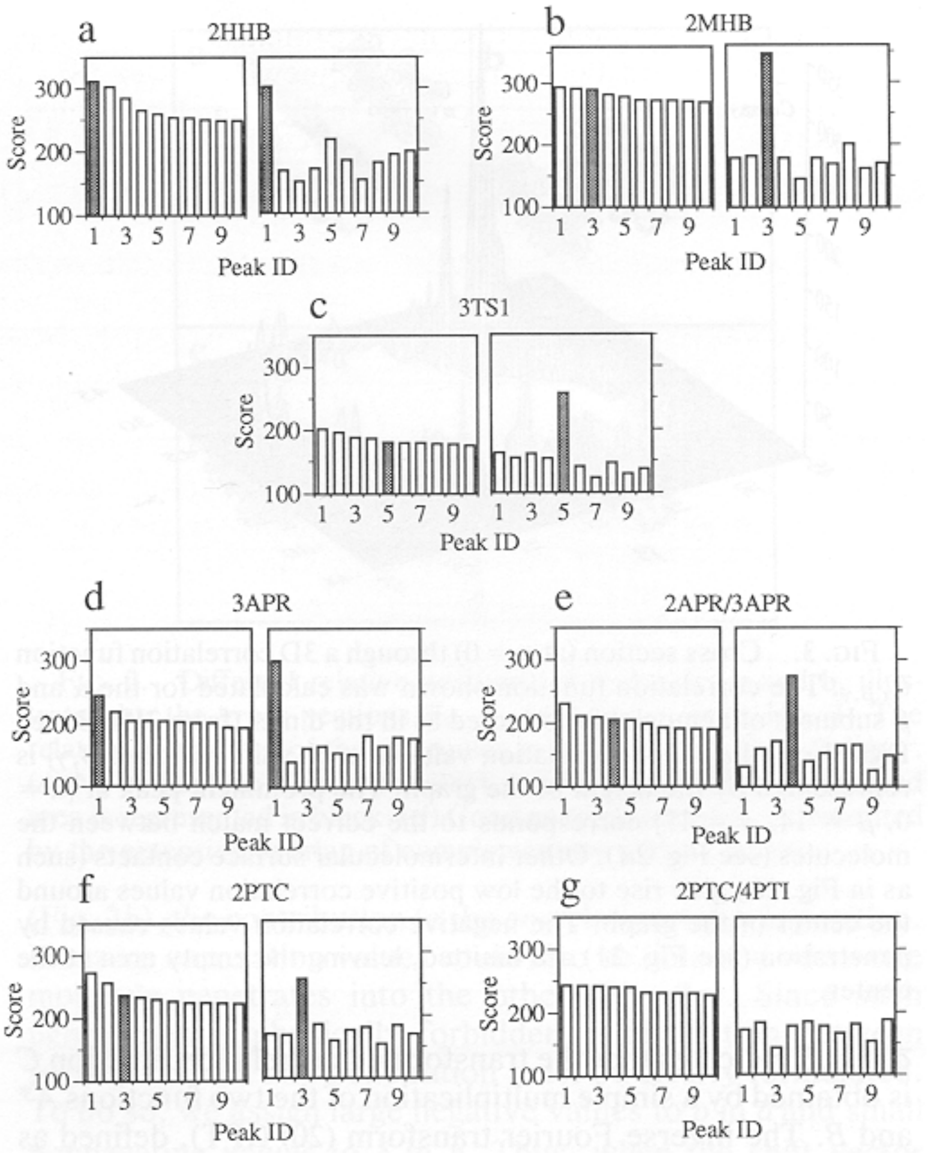
(*vii*) look for a high peak of *c*

(*viii*) rotate molecule **b** to a new orientation

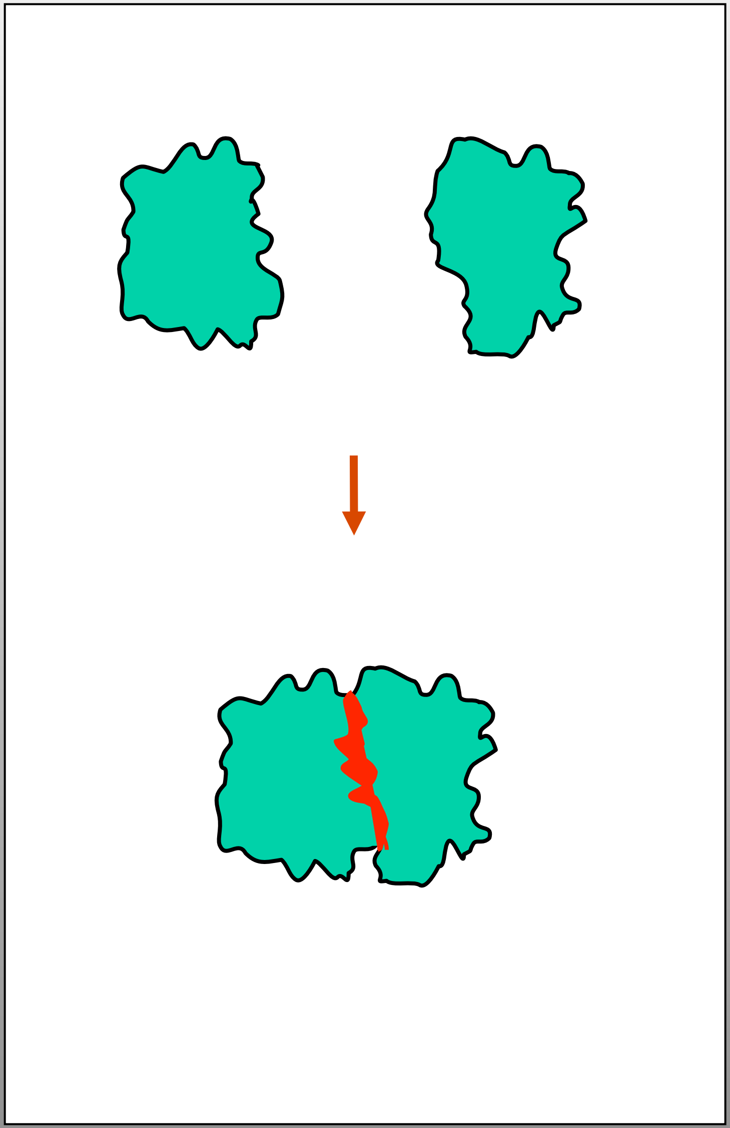
(*ix*) repeat steps *iii*-*viii* and end when the orientation scan is completed

(*x*) sort all of the peaks by their height

Matching results



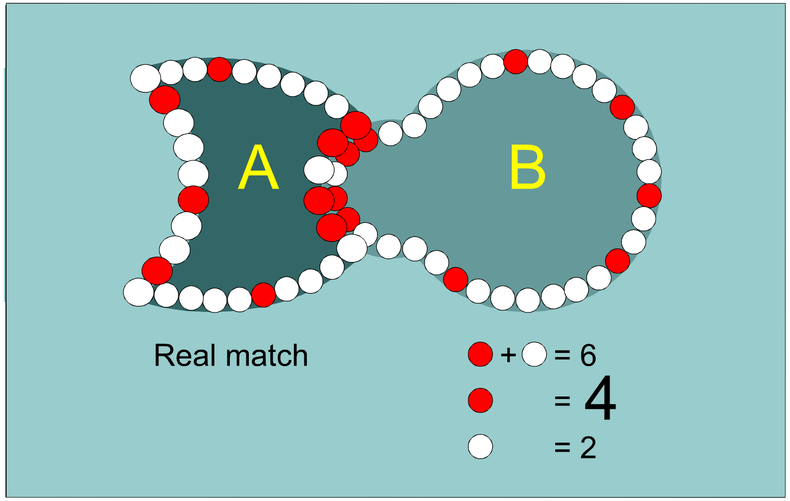
Protein-protein interaction



Hydrophobicity factor

Protein-protein interfaces are more hydrophobic than non-binding surface

Hydrophobic Docking Rationale

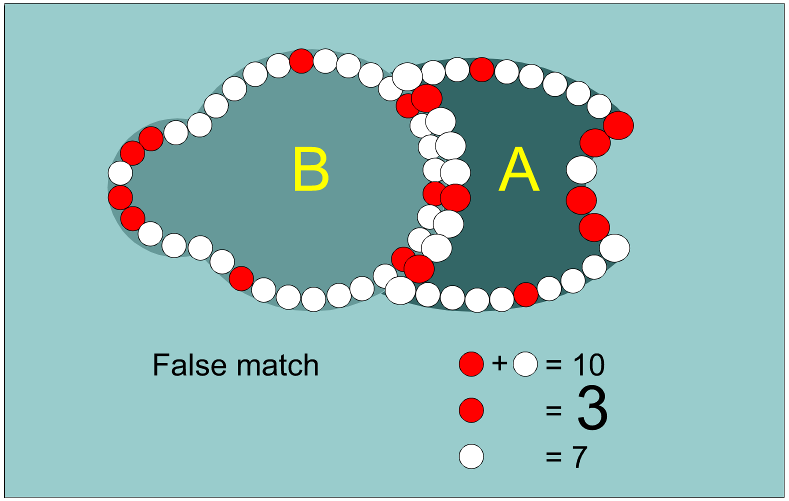


Real match

● + ○ = 6

● = 4

○ = 2



False match

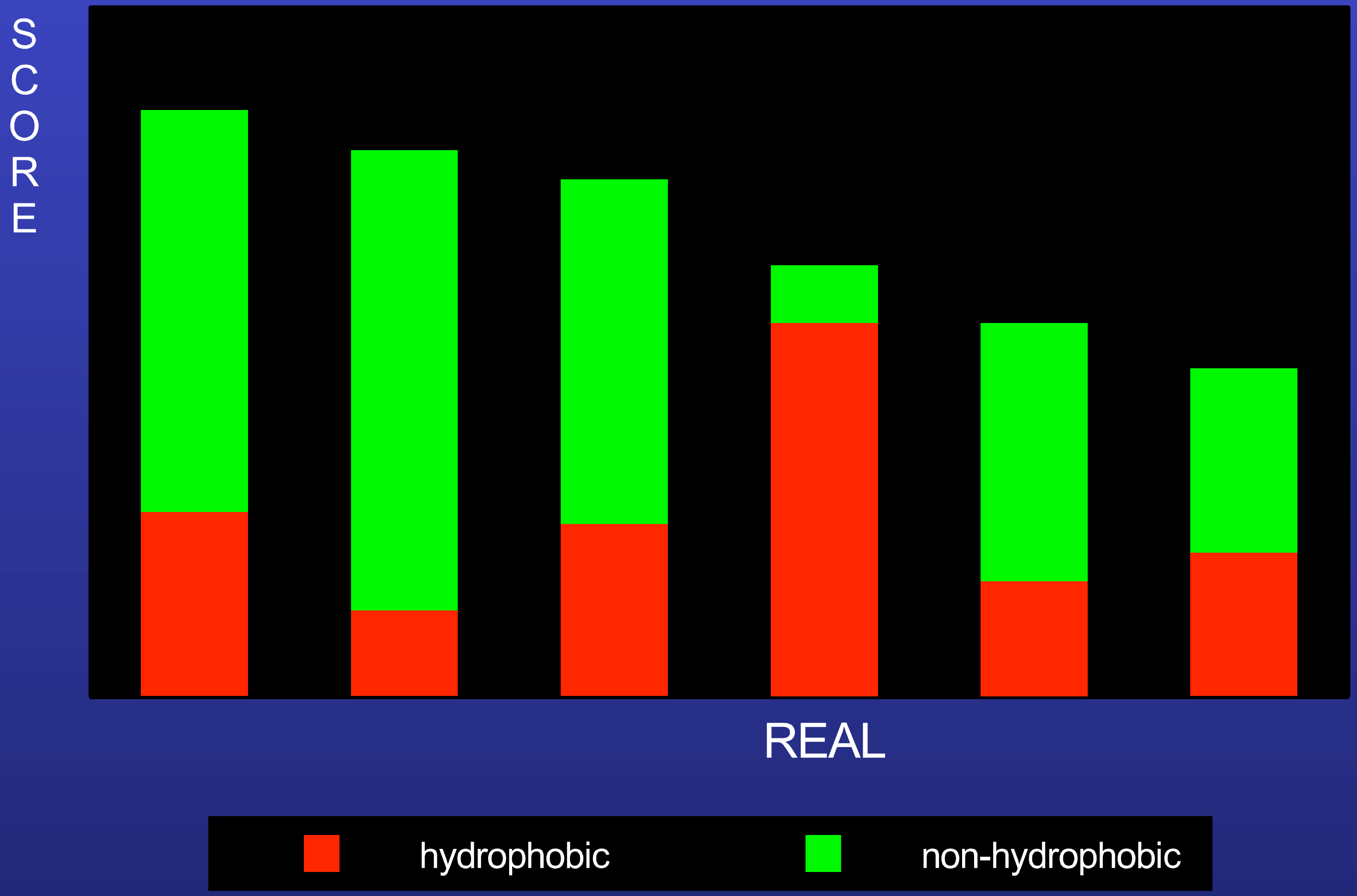
● + ○ = 10

● = 3

○ = 7

● hydrophobic element ○ non-hydrophobic element

Hydrophobic Docking Signal/Noise



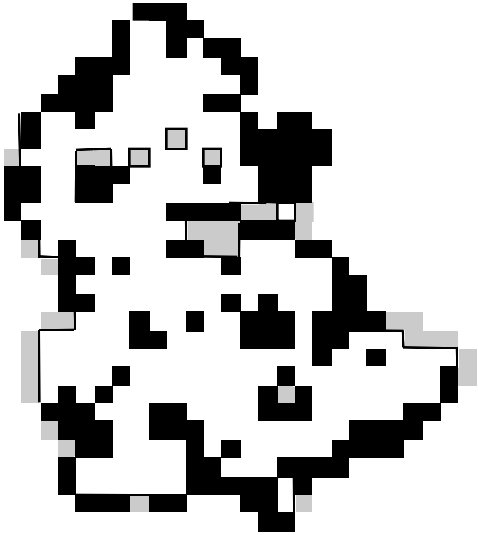
Hydrophobic groups

	Group	Solvation Energy kJ/mole
1	N	-3.8
2	C(CONH)	-2.5
3	O(C=O)	-8.2
4	H(CONH)	-1.4
5	H(OH)	-0.4
6	CH2	6.6
7	CH	4.1
8	CH(arom)	2.0
9	O(COO-)	-58.1
10	CH3	12.0
11	O(OH)	-11.0
12	H(NH3+)	-13.7
13	S	4.3

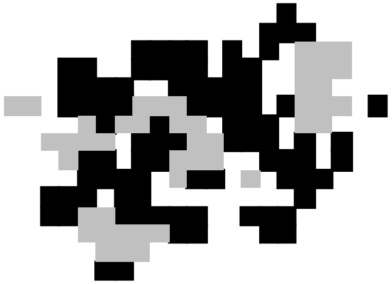
Hydrophobic representation

□ Core
■ Hydrophobic surface element
▨ Non-hydrophobic surface element

■ Hydrophobic element
▨ Non-hydrophobic element

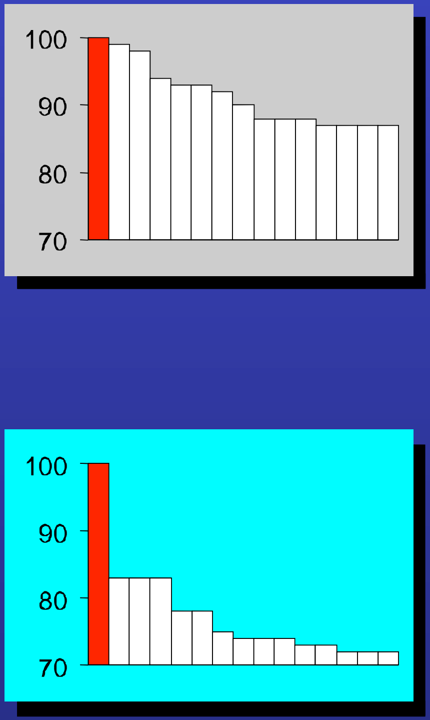


Molecule A

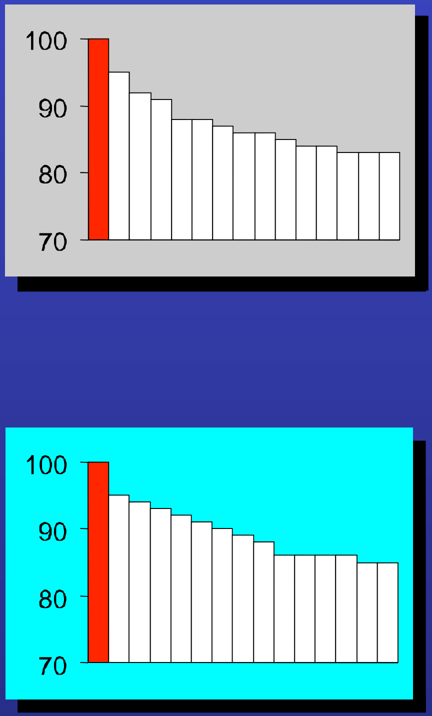


Molecule B

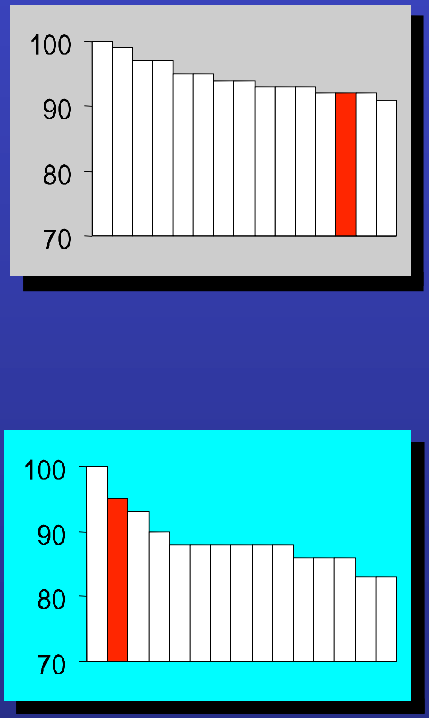
Hydrophobic Docking Scores



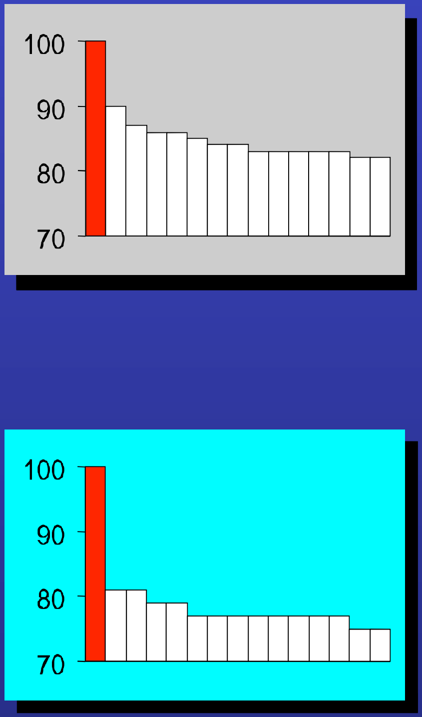
Human hemogolobin
a and b subunits



Horse hemogolobin
a and b subunits



Trypsin - trypsin inhibitor

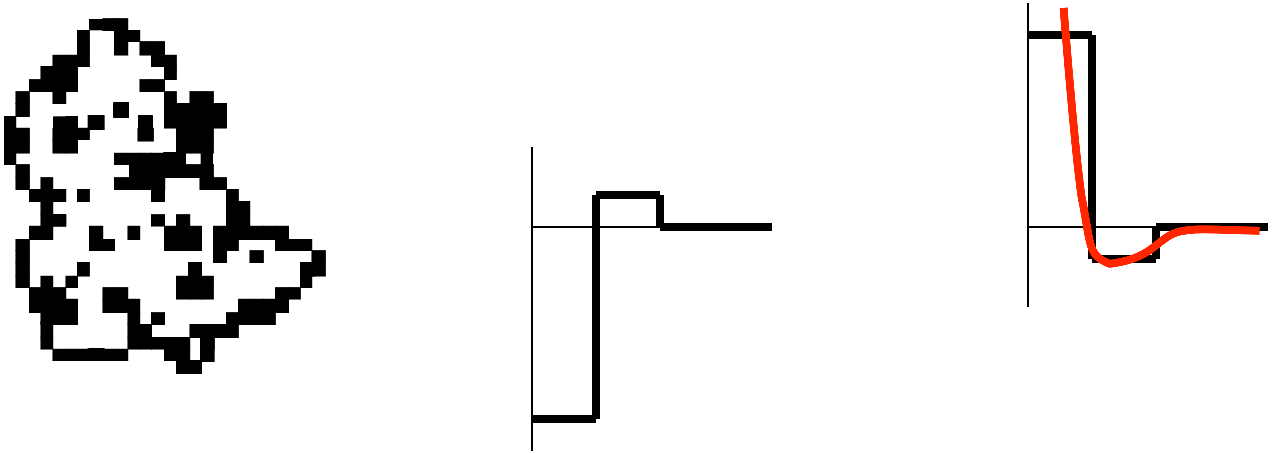


Aspartic proteinase -
peptide inhibitor

🟥 *correct match* ⬜ *false-positive*

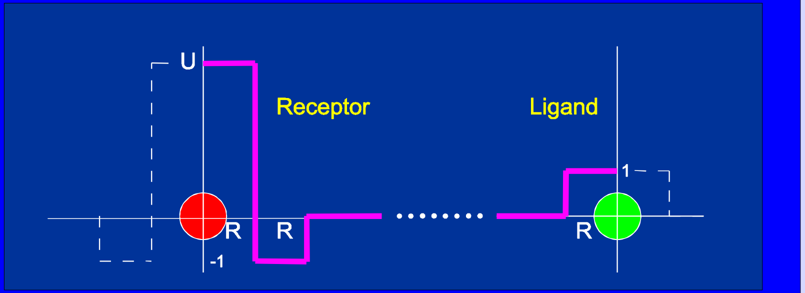
Scoring game

Physics is still there

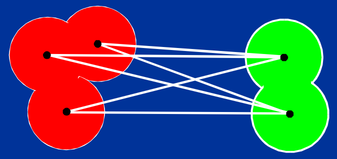


Correlation as energy calculation

Step-function potential

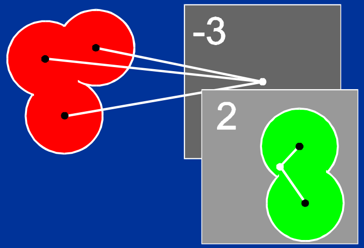


Energy calculations with explicit atom-atom interactions



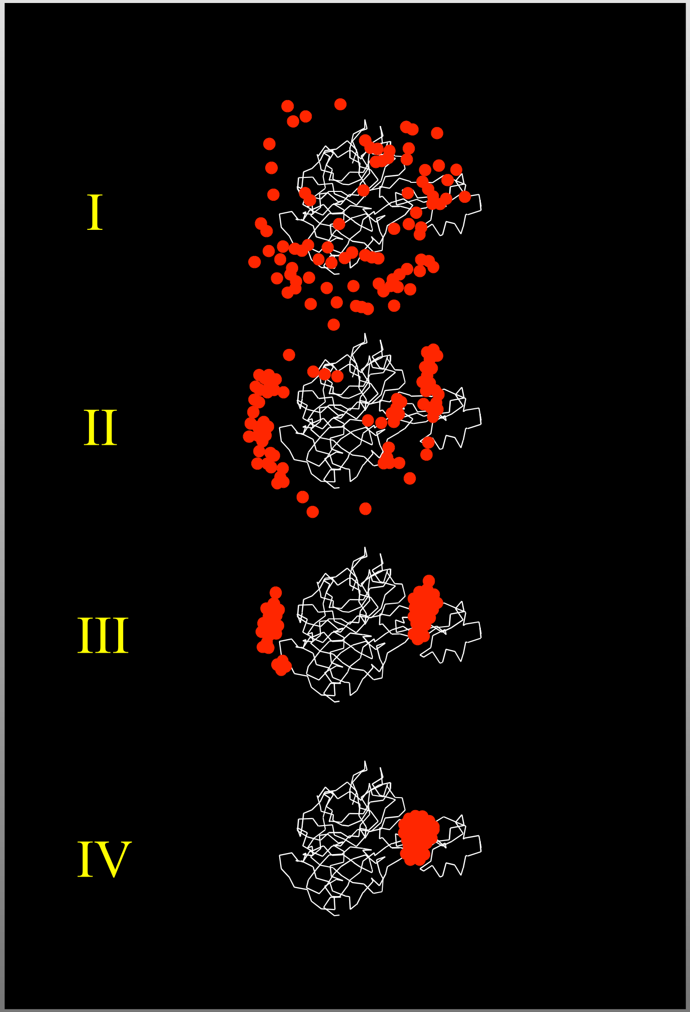
E = -1 … -1 = -6

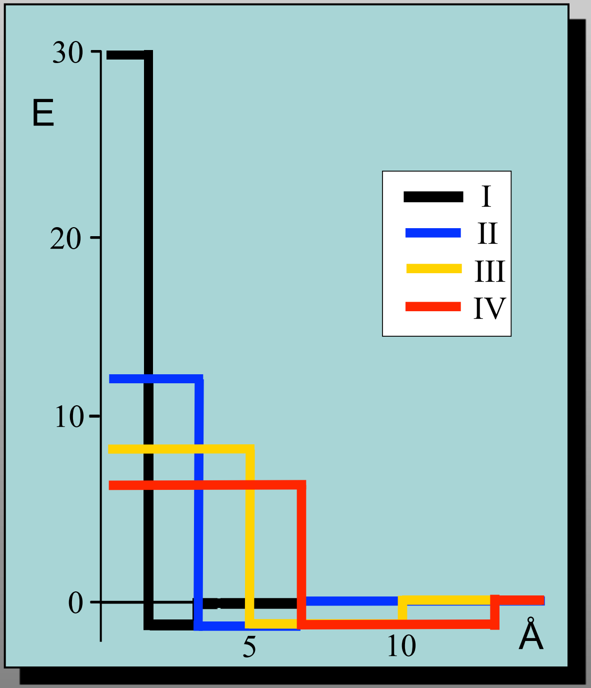
Energy calculations by a correlation technique



E = -3 x 2 = -6

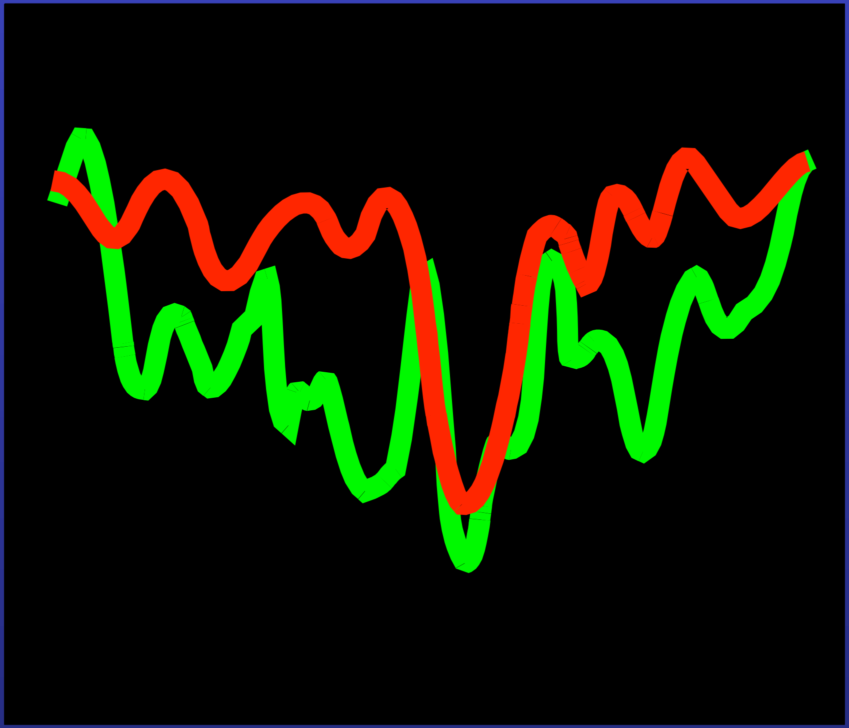
Transition to longer potential ranges



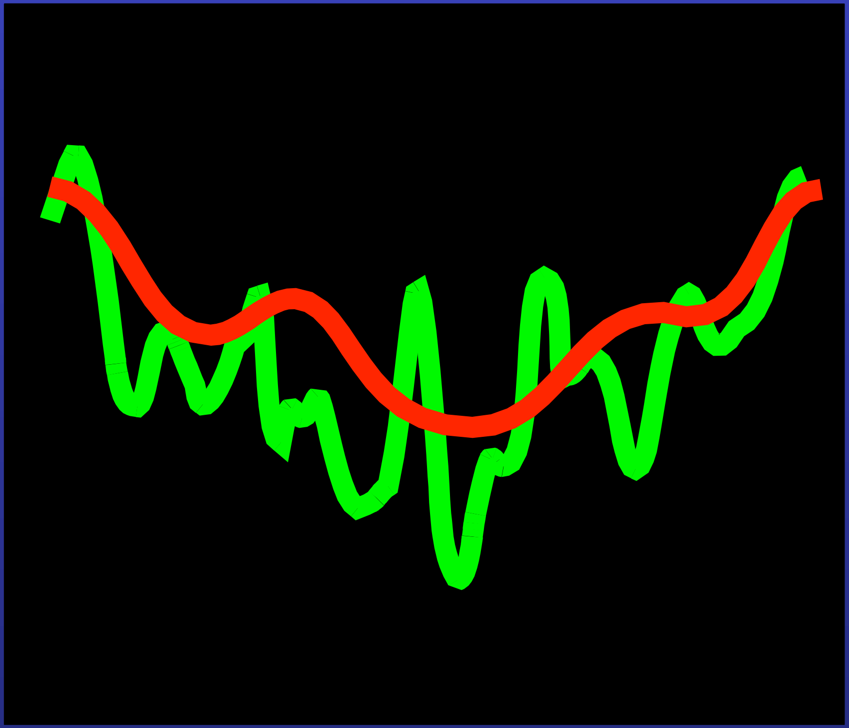


Energy landscape

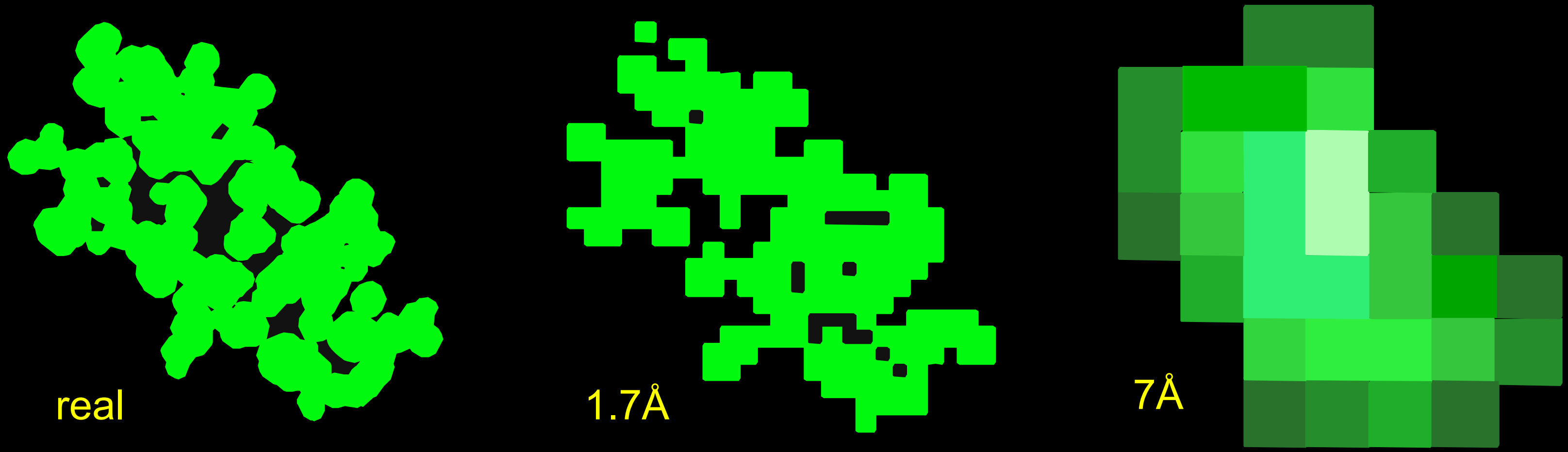
Hydrophobic Docking



Low-resolution Docking



Low-resolution representation

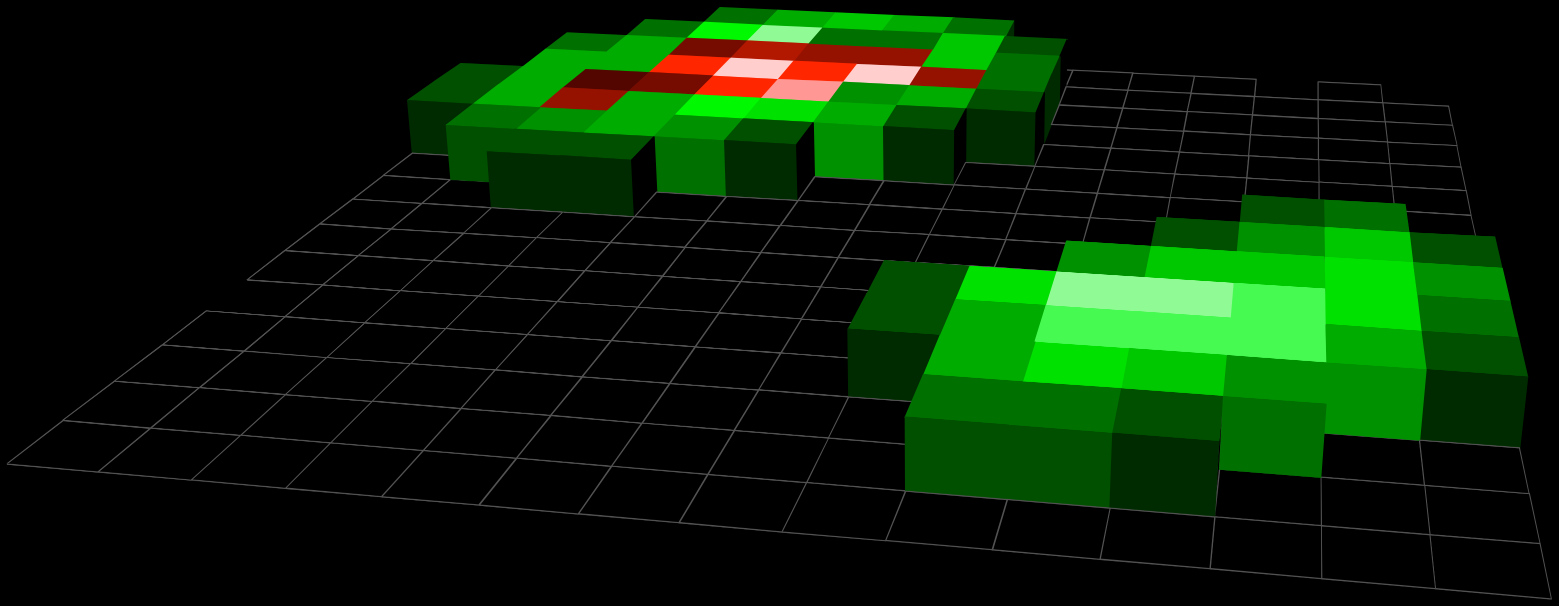


real

1.7Å

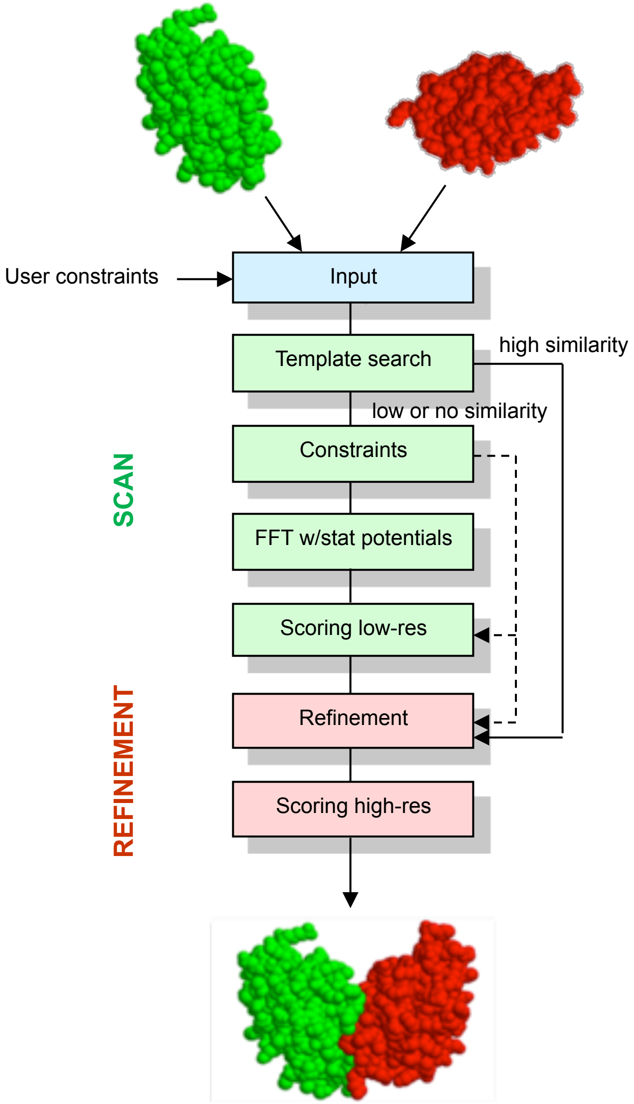
7Å

b subunit of human hemoglobin



Human hemoglobin a and b subunits

Docking Current and Future



User constraints

Input

Template search

high similarity

low or no similarity

Constraints

FFT w/stat potentials

Scoring low-res

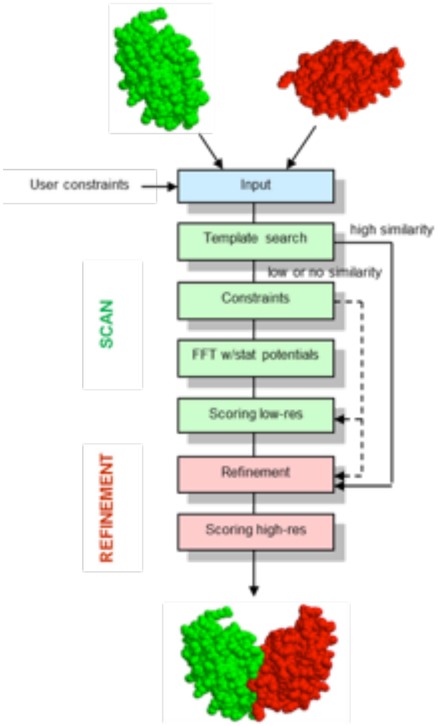
Refinement

Scoring high-res

SCAN

REFINEMENT

Scoring



- Low-resolution (Unrefined)
- High-resolution (Refined)

In the future may disappear, divided between **Scan** and **Refinement** search procedures

GRAMM-X Scoring

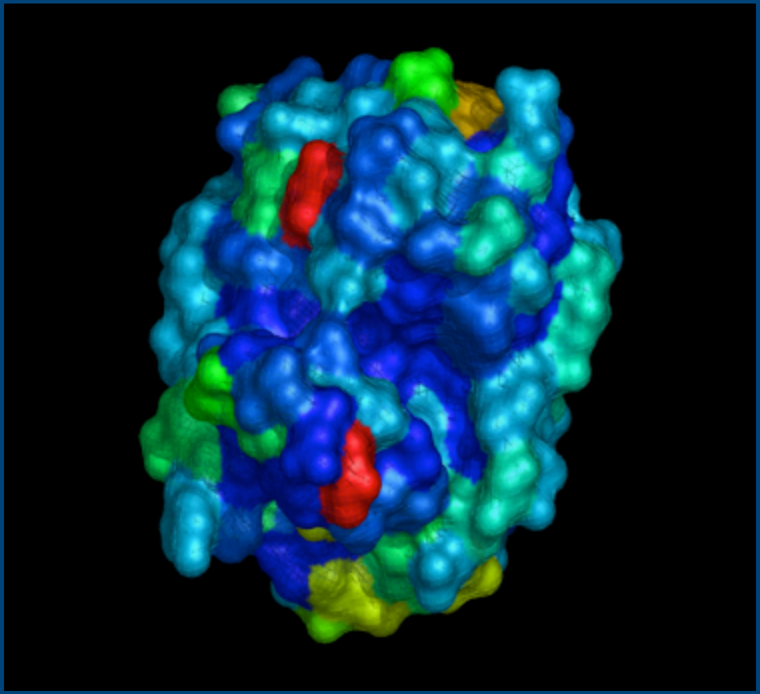
[VW soft] +

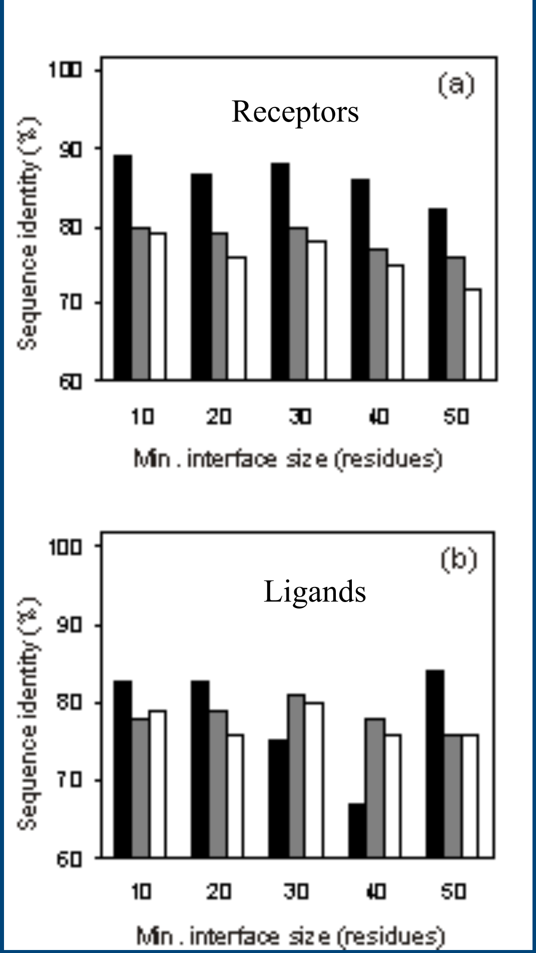
[Evolutionary Residue Conservation] +

[Local Minimum Volume] +

[Statistical Residue Contact Preference]

Evolutionary Residue Conservation



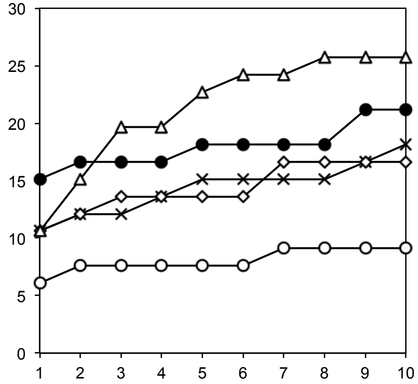


Interfaces are more conserved

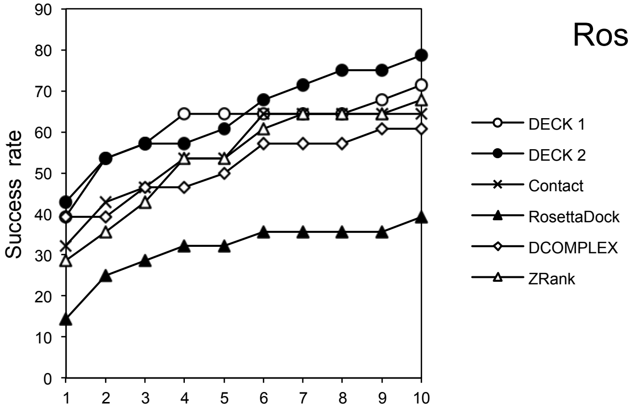
Conservation scores from ConSurf server
http://consurf.tau.ac.il

Total score for the prediction is the sum of residue scores for the putative interface normalized by protein sizes

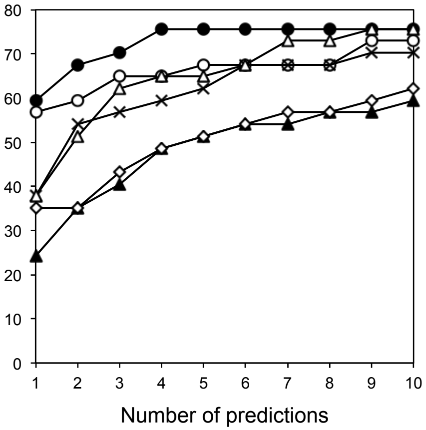
Statistical Potentials



ZRANK decoys



RosettaDock decoys RMSD < 5 Å



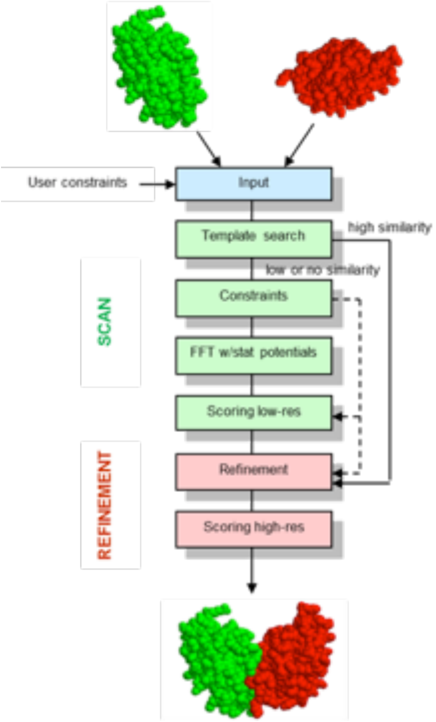
RosettaDock decoys RMSD < 10 Å

Future

- Higher order across-interface statistical propensities – environment dependent residue-residue potentials, structural motifs coupling, etc.

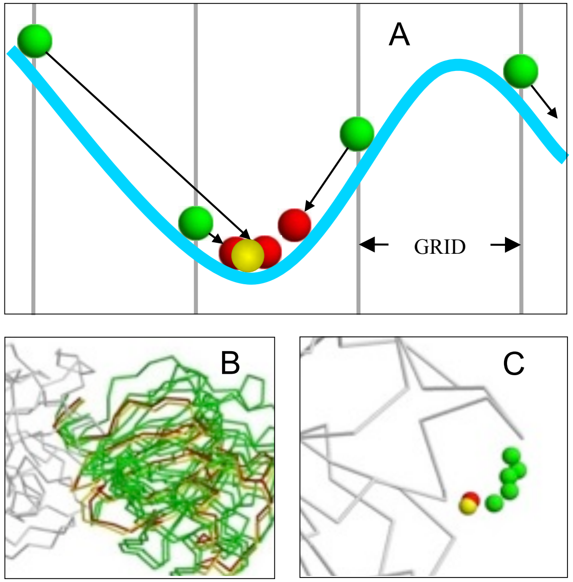
- Energy basin characteristics that would help identify the funnel – size, ruggedness, etc.

Refinement

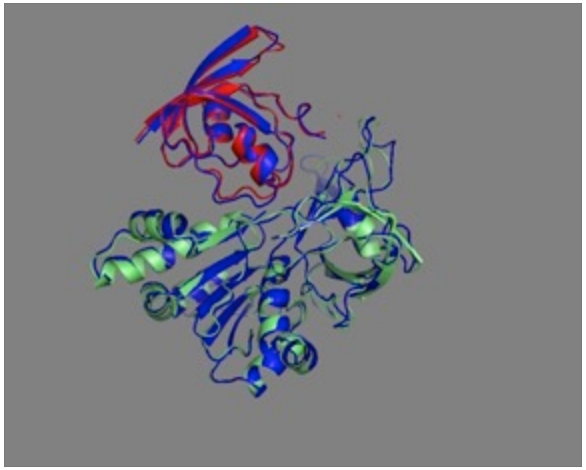

User constraints
Input
Template search
high similarity
low or no similarity
Constraints
FFT w/stat potentials
Scoring low-res
Refinement
Scoring high-res
SCAN
REFINEMENT

GRAMM-X Refinement

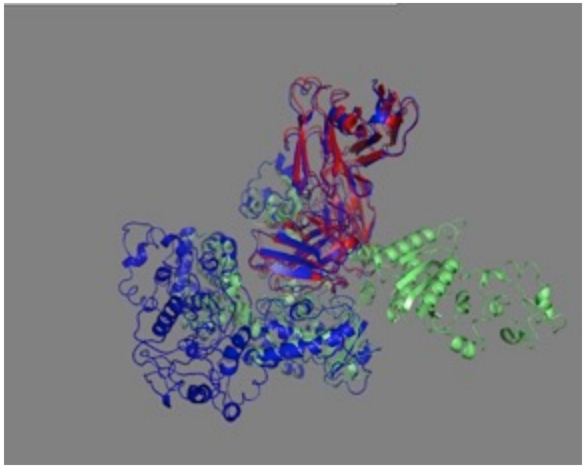
Conjugate Gradient Minimization with [VW soft]



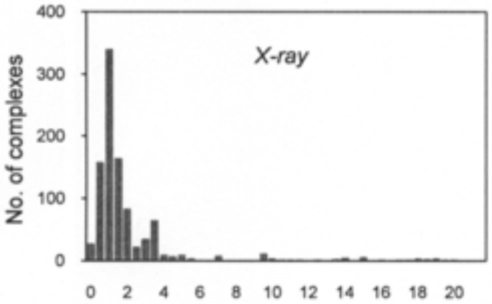
Comparison of bound and unbound structures



Typical RMSD ~1.00-1.25Å



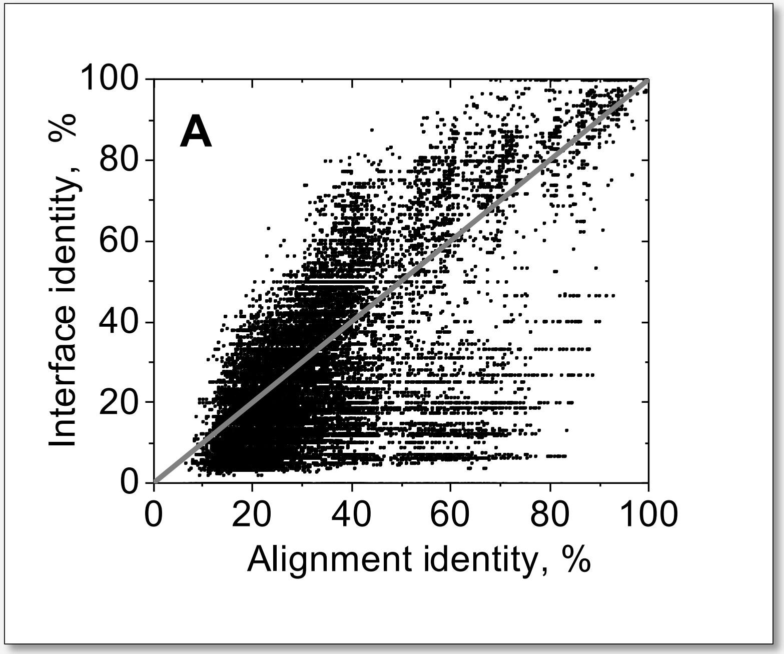
Large change due to domain movement



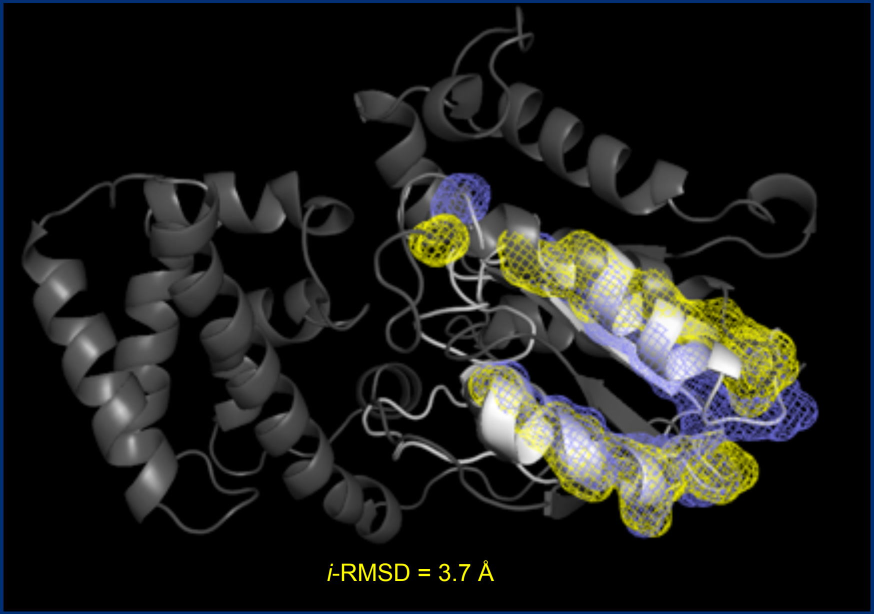
Distribution of biological, non-obligate complexes according to bound/unbound all atom RMSD

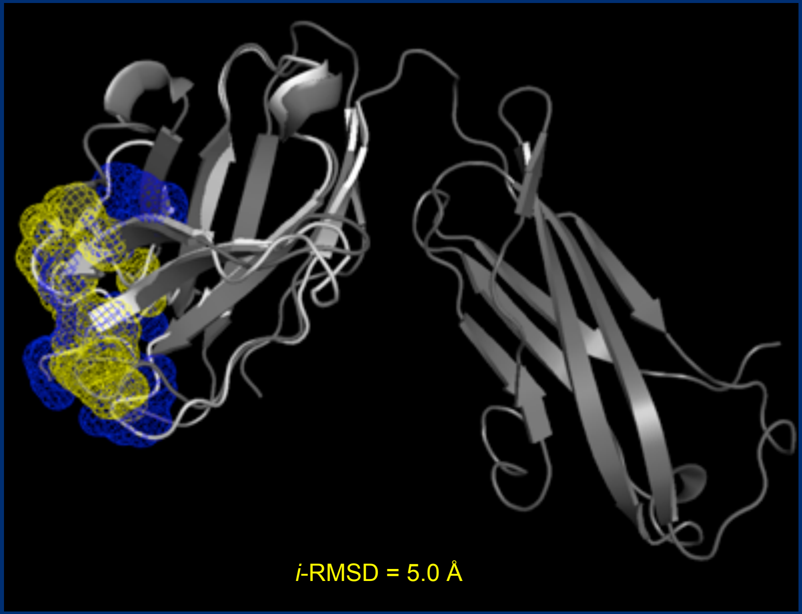
Binding Site Structural Accuracy

Interface is more conserved



Binding Site Structural Accuracy


i-RMSD = 3.7 Å


i-RMSD = 5.0 Å

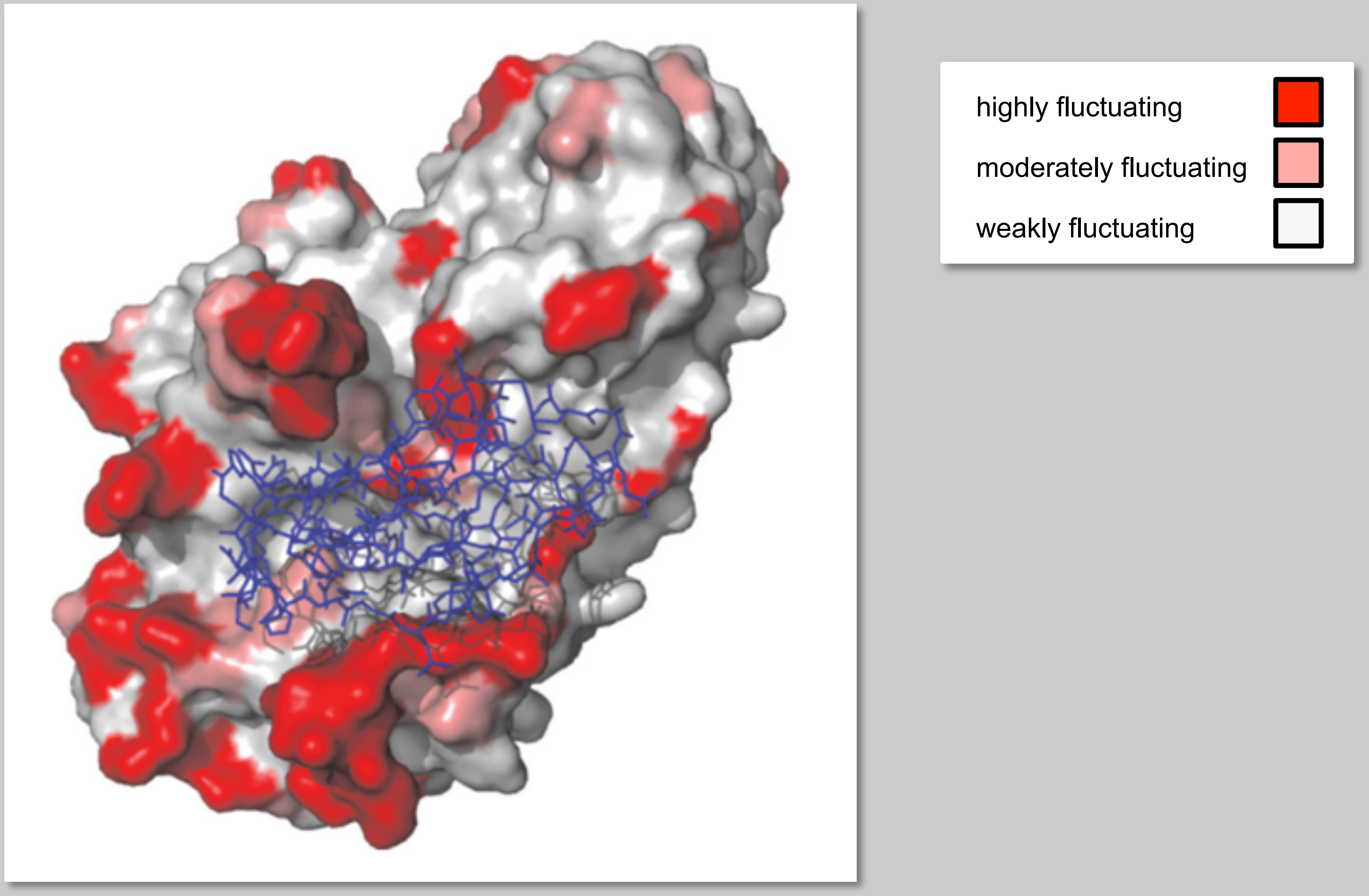
□ Partial homology model

■ Target native structure

~ 50% of complexes with interfaces modeled by high-throughput techniques have accuracy suitable for docking

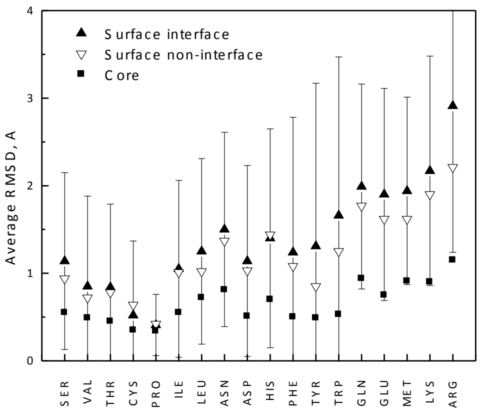
Elastic Network Models

Binding site is more rigid

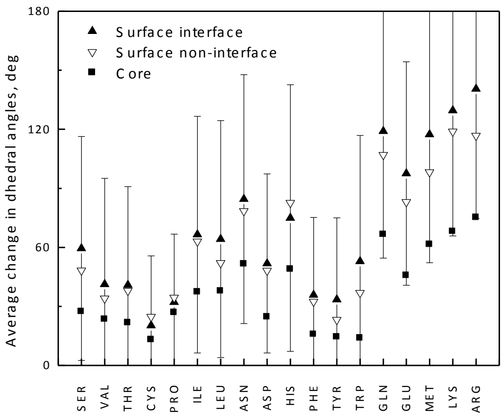


Unbound/bound change

Change in Cartesian coordinates



Change in dihedral angles

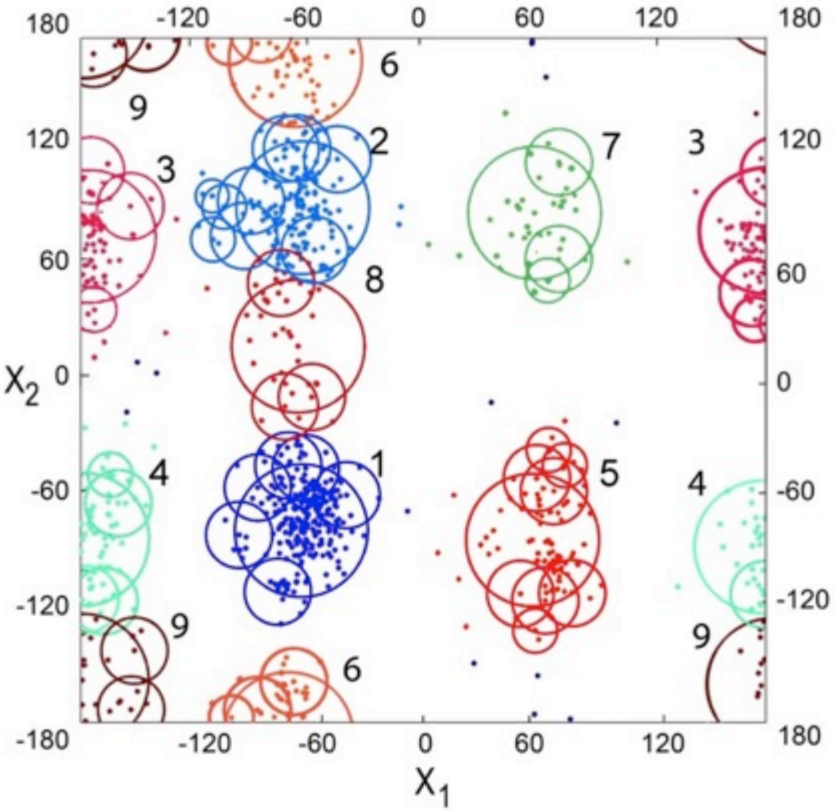


Residues are sorted by the number of χ angles (and increasing mass, if the number of χ is the same)

- Conformational changes increase with the number of χ angles
- Conformational changes are greater at interface than at non-interface

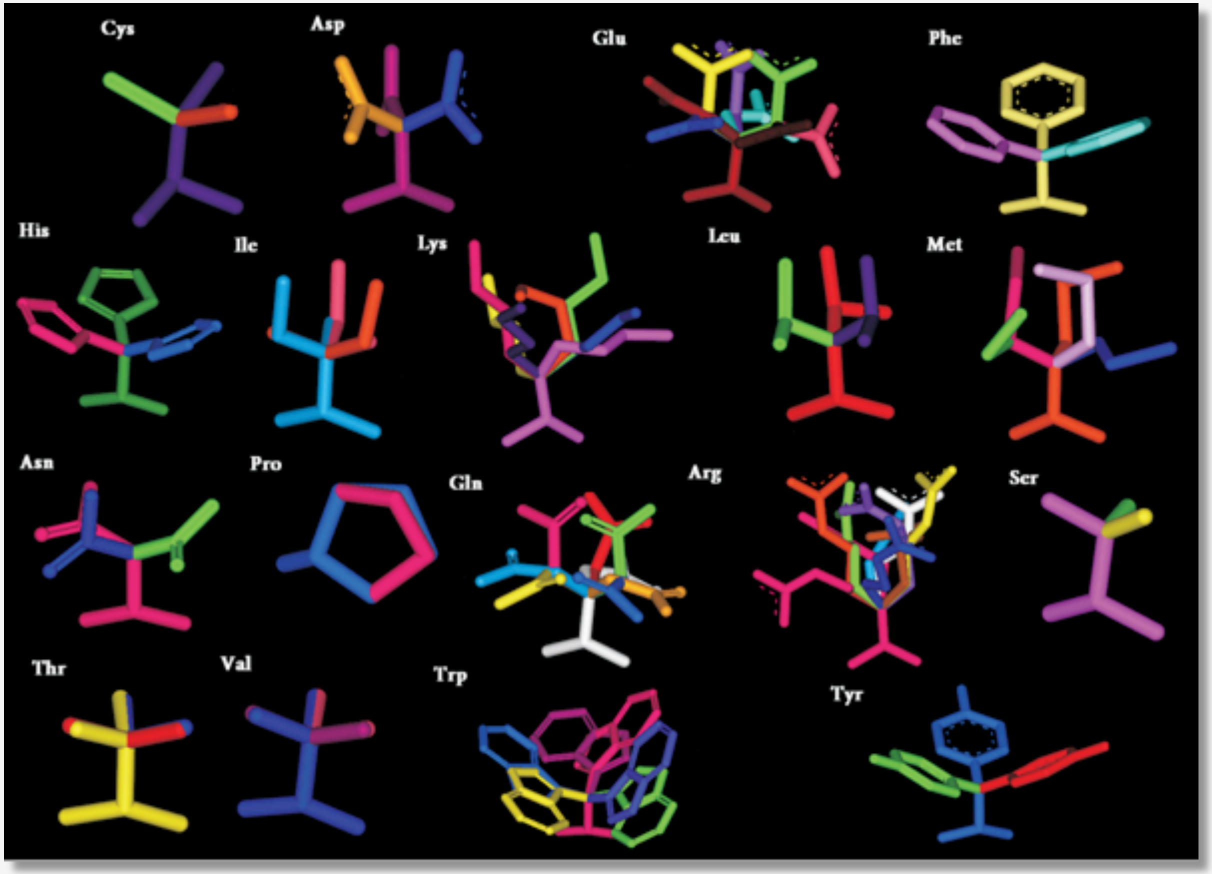
Side Chain Rotamers

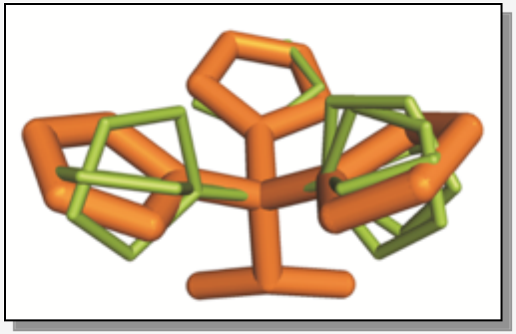
Hierarchical clustering with a variable radius in the torsional space to reduce the share of non-clustered conformations



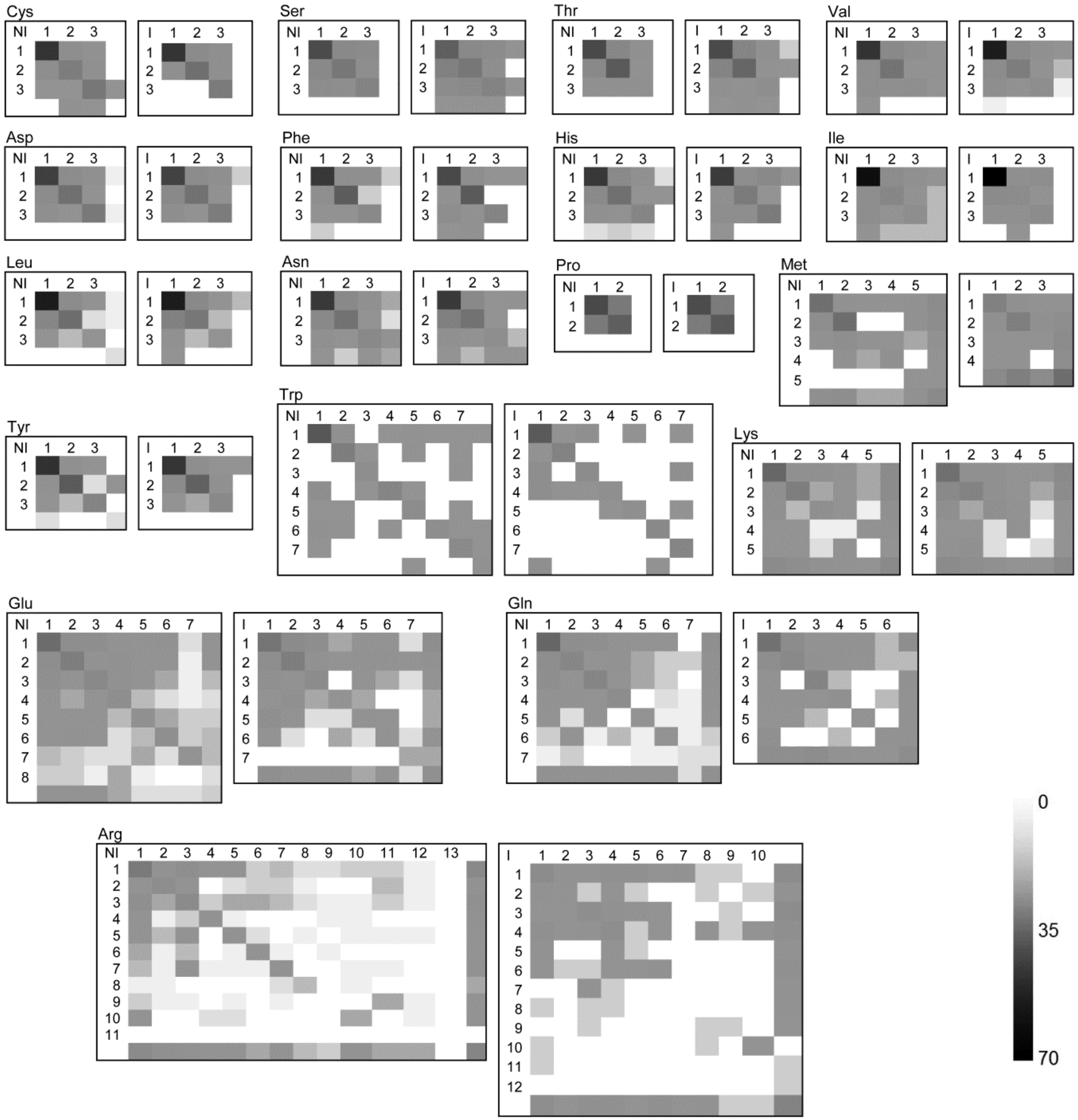
Histidine rotamers at interface

Surface Interface Rotamer Library

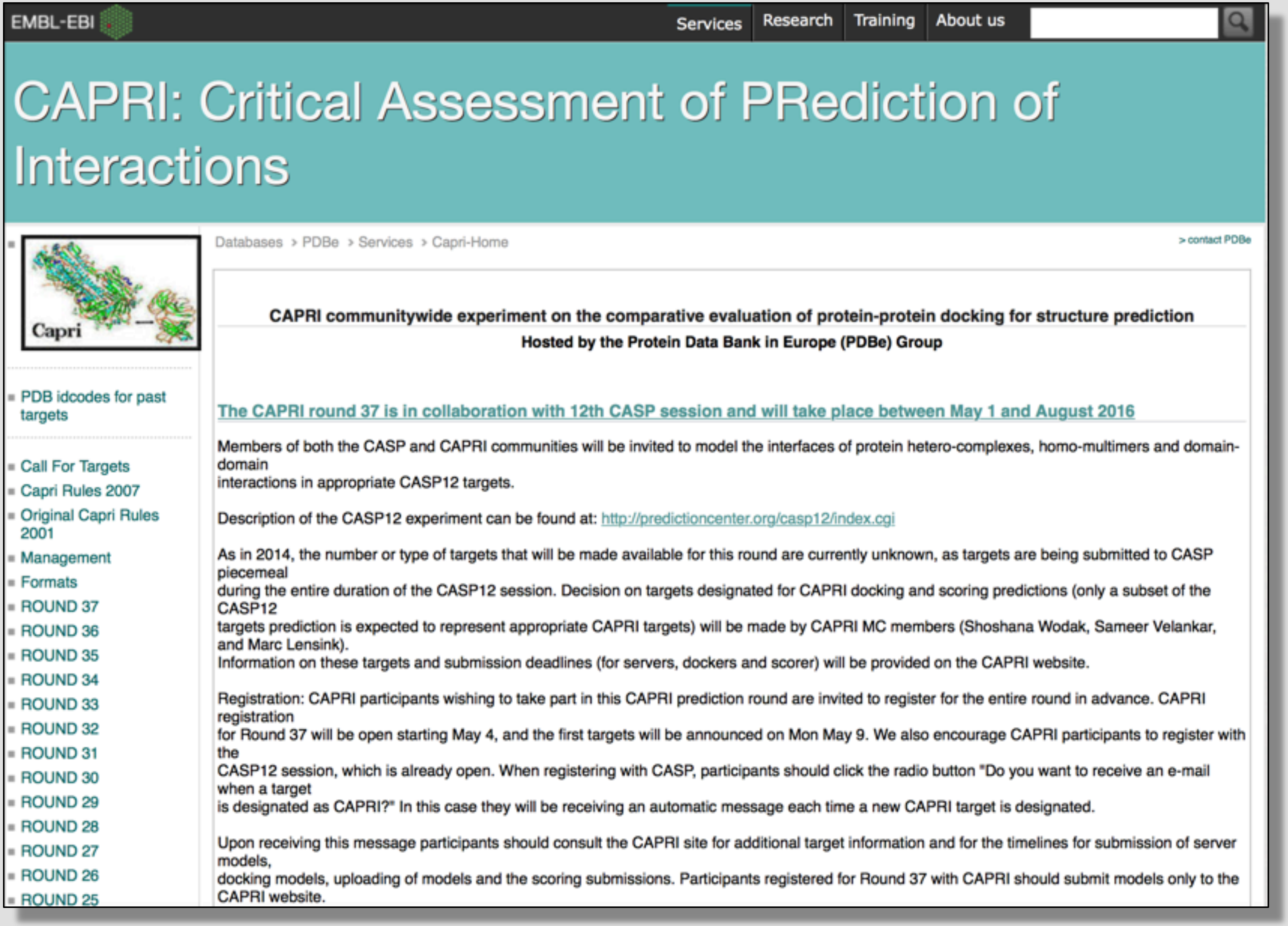




Unbound To Bound Transitions



- Largest values are on diagonal - side chains prefer to stay within same rotamer upon binding
- In case of transition to a different rotamer, side chains prefer closest, often most occupied rotamer
- Side chains with 3 and 4 χ are least stable. Most stable are with 2 χ (1 χ are less stable)
- At interface, non-polar amino acids are more stable than polar



CAPRI: Critical Assessment of PRediction of Interactions

CAPRI communitywide experiment on the comparative evaluation of protein-protein docking for structure prediction

Hosted by the Protein Data Bank in Europe (PDBe) Group

The CAPRI round 37 is in collaboration with 12th CASP session and will take place between May 1 and August 2016

Members of both the CASP and CAPRI communities will be invited to model the interfaces of protein hetero-complexes, homo-multimers and domain-domain
interactions in appropriate CASP12 targets.

Description of the CASP12 experiment can be found at: http://predictioncenter.org/casp12/index.cgi

As in 2014, the number or type of targets that will be made available for this round are currently unknown, as targets are being submitted to CASP piecemeal
during the entire duration of the CASP12 session. Decision on targets designated for CAPRI docking and scoring predictions (only a subset of the CASP12
targets prediction is expected to represent appropriate CAPRI targets) will be made by CAPRI MC members (Shoshana Wodak, Sameer Velankar, and Marc Lensink).
Information on these targets and submission deadlines (for servers, dockers and scorer) will be provided on the CAPRI website.

Registration: CAPRI participants wishing to take part in this CAPRI prediction round are invited to register for the entire round in advance. CAPRI registration
for Round 37 will be open starting May 4, and the first targets will be announced on Mon May 9. We also encourage CAPRI participants to register with the
CASP12 session, which is already open. When registering with CASP, participants should click the radio button "Do you want to receive an e-mail when a target
is designated as CAPRI?" In this case they will be receiving an automatic message each time a new CAPRI target is designated.

Upon receiving this message participants should consult the CAPRI site for additional target information and for the timelines for submission of server models,
docking models, uploading of models and the scoring submissions. Participants registered for Round 37 with CAPRI should submit models only to the CAPRI website.