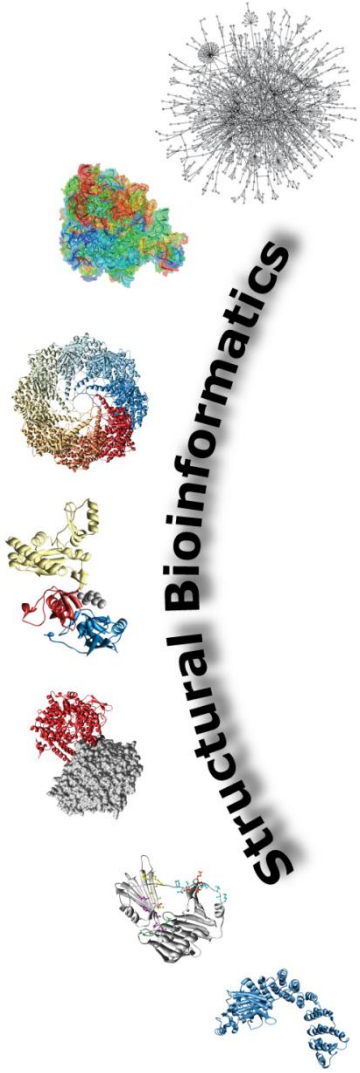
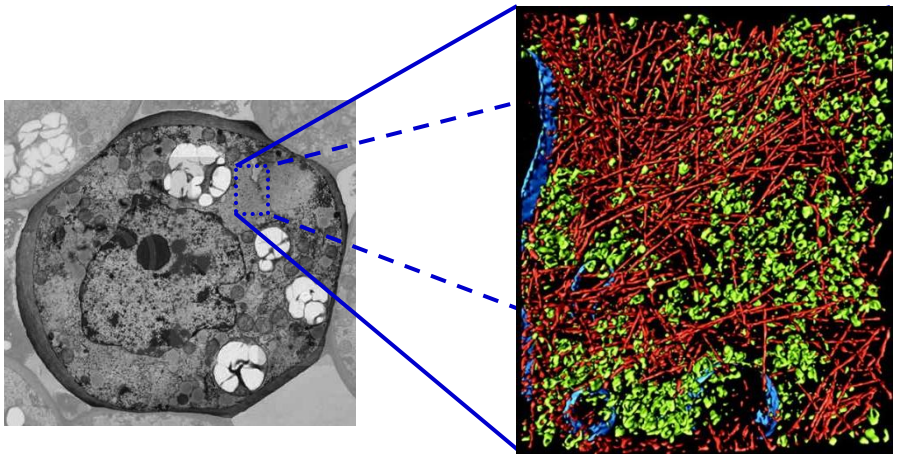


II8010. Lecture Materials

Dmitry Korkin, Ph.D.

Informatics Institute and
Department of Computer Science
University of Missouri, Columbia

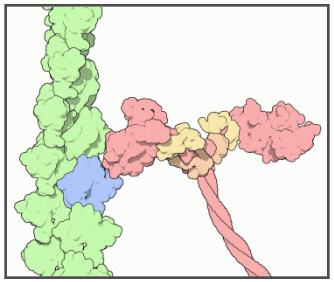
Protein assemblies are essential building blocks of a living cell



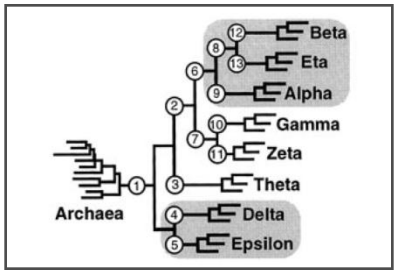
Cytoplasmic assemblies

- 1.7 interactions per protein
- 30,000 interactions in yeast
- proteins consist of ~2 domains

Function



Evolution



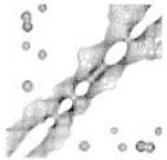
Medicine

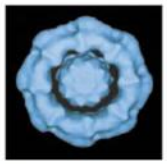
O. Medalia et al., Science (2002)

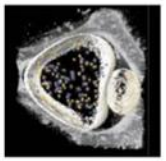
Methods for structural characterization of macromolecules

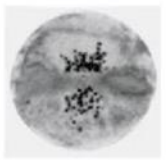
Experimental

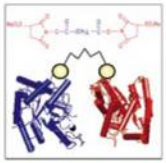

X-ray crystallography


NMR spectroscopy

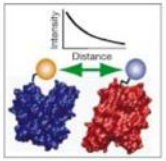

2D and single-particle EM


Electron tomography

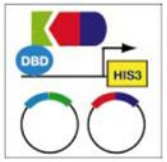

Immuno-electron microscopy


Chemical cross-linking

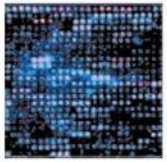

Affinity purification mass-spectroscopy



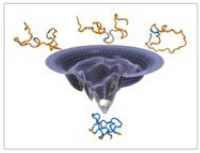
FRET



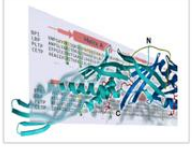
Yeast two-hybrid system

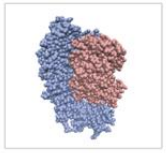

Gene/protein arrays

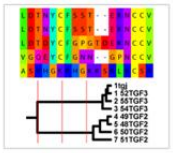
Computational


Ab-initio modeling

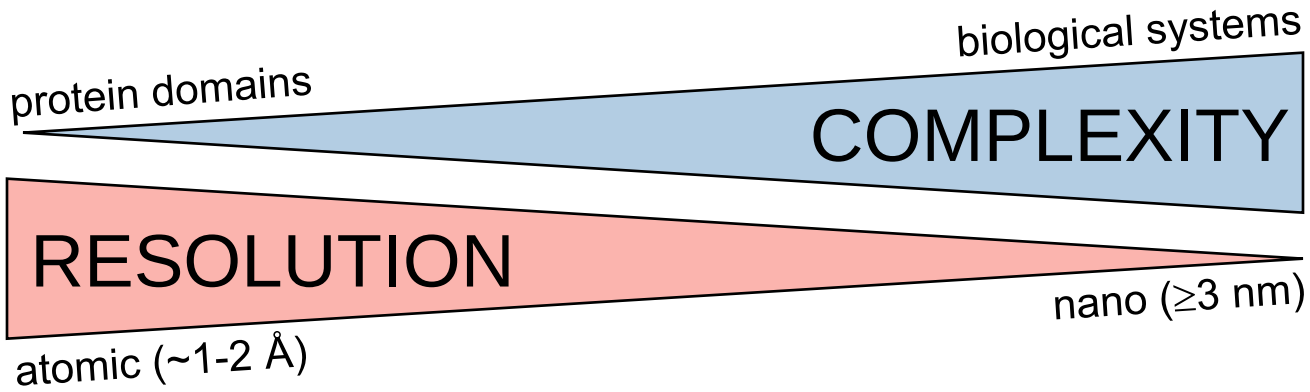

Comparative modeling


Protein threading


Molecular docking


Phylogenetic profiling

Knowledge-based information retrival

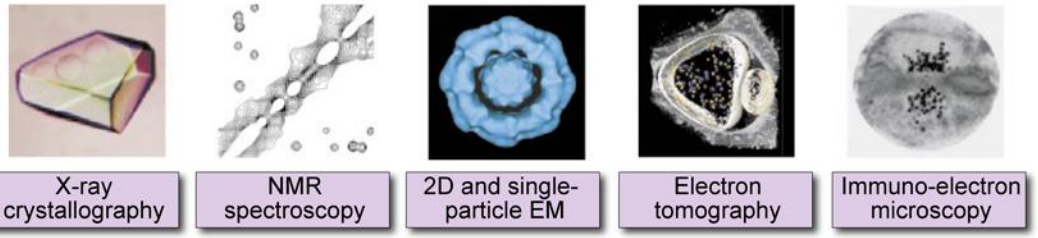


Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A, *Curr.Opin.Struct.Biol.*, 2004

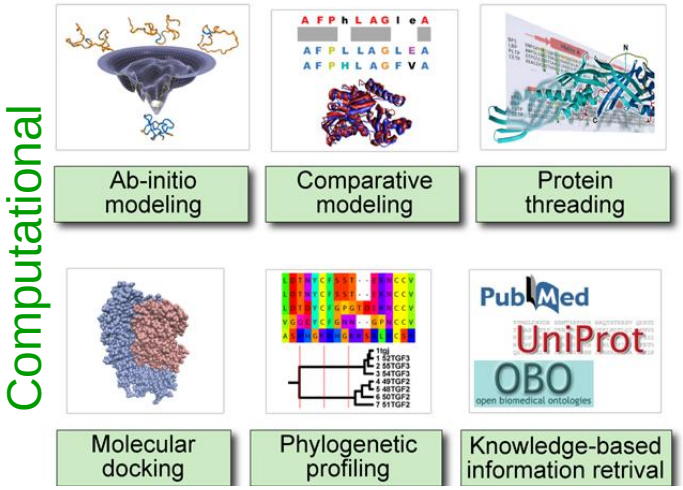
Alber F, Forster F, Korkin D, Topf M, Sali A, *Annu. Rev. Biochem.*, 2008, in press

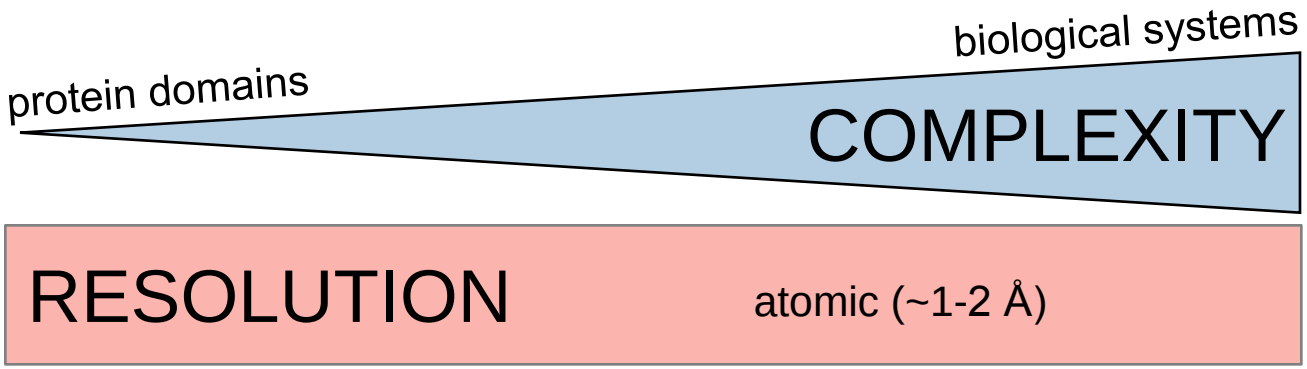
Methods for structural characterization of macromolecules

Experimental



Computational





Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A, *Curr.Opin.Struct.Biol.*, 2004

Alber F, Forster F, Korkin D, Topf M, Sali A, *Annu. Rev. Biochem.*, 2008, in press

Protein interactions in structural bioinformatics context

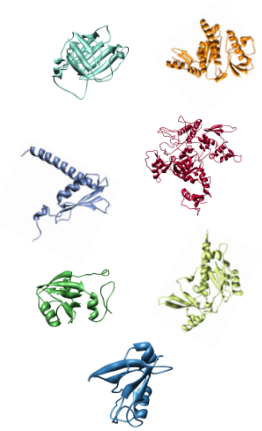
subunit sequences

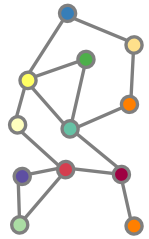
subunit structures & models

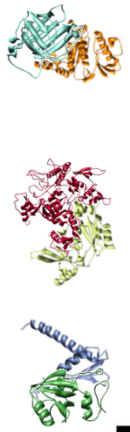
interaction network

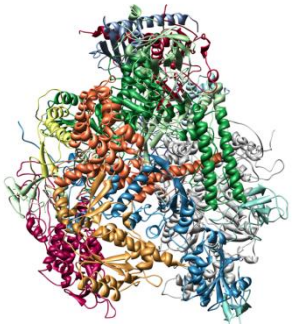
binary sub-complexes

assembly structure

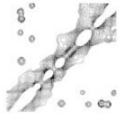






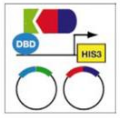


X-ray crystallography

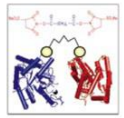


NMR spectroscopy

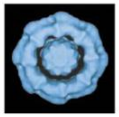
Affinity purification mass-spectroscopy



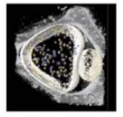
Yeast two-hybrid system



Chemical cross-linking



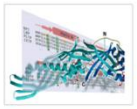
2D and single-particle EM



Electron tomography

Ab-initio modeling

Comparative modeling



Protein threading

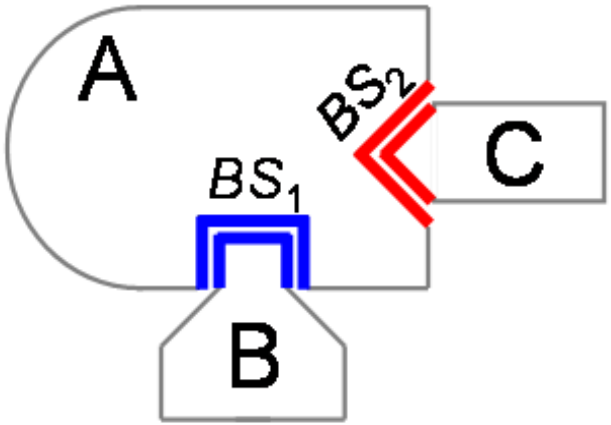
Protein–protein interaction vocabulary

- A protein–protein interaction (PPI), or binary protein interaction, is defined between any two structural subunits
- Structural subunits = small proteins, domains, peptides;
- Two residues (from two different subunits) are called the contact residues, if there is at least a pair of atoms, one from each residue that are in close proximity (usually ≤6Å)
- Two subunits interact if they have at least a pair of contact residues
- interaction = (S_1, S_2, Or)

Protein–protein interaction vocabulary

- Protein binding site for an interaction $I_{1=}$ (S_1, S_2, Or) is a set of all contact residues of either S_1 or S_2.

- Protein interface for an interaction $I_{1=}$ (S_1, S_2, Or) is a set of all pairs of contact residues, one from each protein binding site



How to characterize an interaction interface

- N of contact residues in a binding site
  - On average 20-30 residues per each binding site

- Buried surface = $Surf_1 + Surf_2 - Surf_{12}$
  - Usually >1100 $Å^2$
  - each of the interacting partners contributing at least 550 Å2 of complementary surface.
  - Average interface residue covers some 40 Å2.
  - dimers contribute 12% of their accessible surface area to the contact interface, trimers 17.4% and tetramers 20.9%.
  - variations are large

- Binding free energy
  - Energy required to dissociate two subunits

What causes two proteins interact?

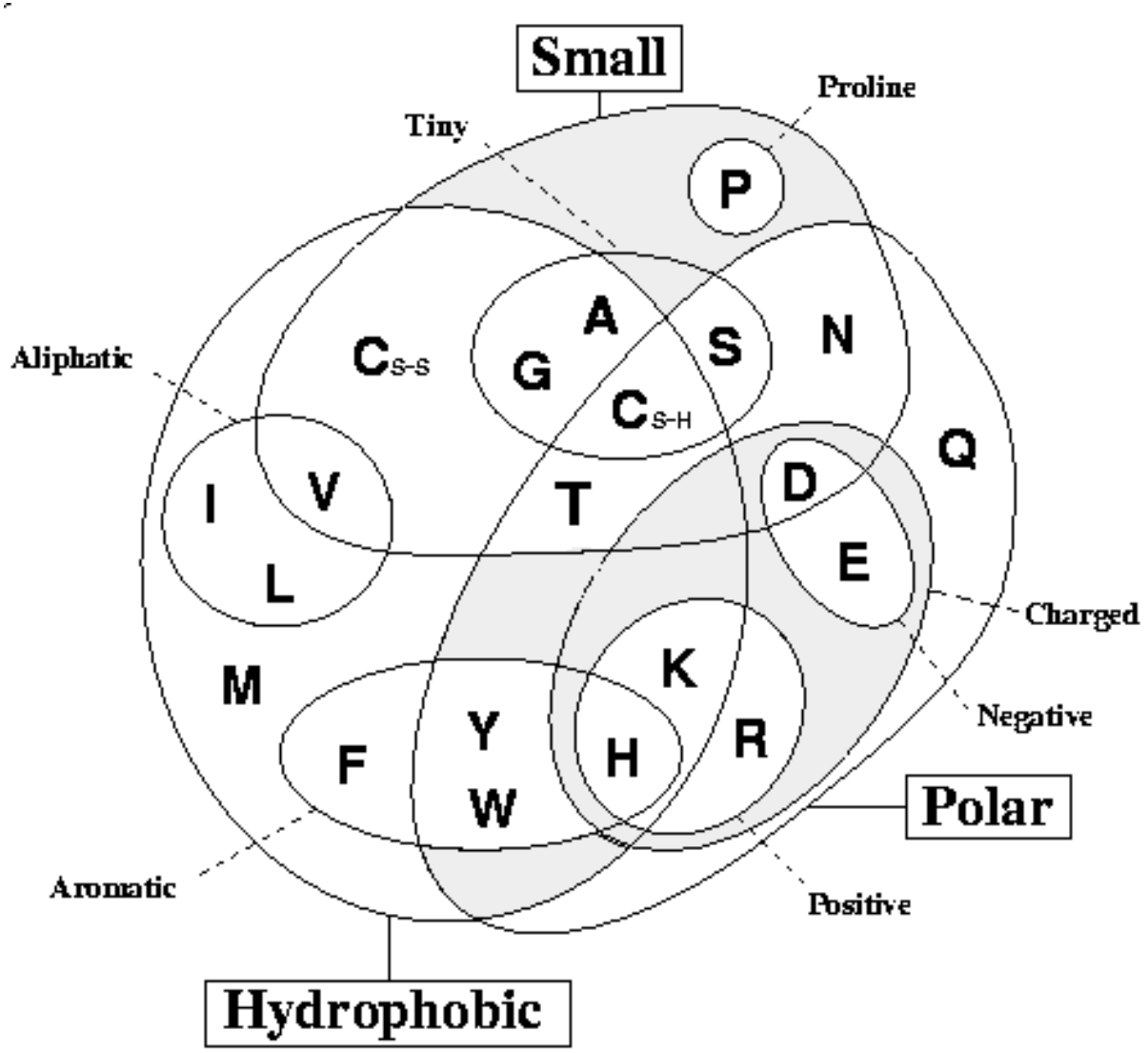
- Geometrical complementarity
  - Do they have to be completely complementary? Not necessarily!

- Physico-chemical complementarity
  - Electrostatic interactions
  - Hydrogen bonds
  - van der Waals attraction

- Interaction with water
  - Hydrophobic effect: Hydrophobic residues tend to be buried in the interface

A standard size interface (~ 1600 Å^2) buries about 900 Å^2 of the non-polar surface, 700 Å^2 of polar surface, and contains 10 (± 5) hydrogen bonds.

Amino acid residues. Basic classes



Electrostatic interactions and hydrophobic efeect

- The average protein-protein interface is not less polar or more hydrophobic than the surface remaining in contact with the solvent
- Water is usually excluded from the contact region
- Non-obligate complexes tend to be more hydrophilic in comparison, as each component has to exist independently in the cell.

Van der Waals interactions

- Van der Waals interactions occur between all neighboring atoms
- These interactions at the interface are no more energetically favorable than those made with the solvent
- However, they are more numerous, as the tightly packed interfaces are more dense than the solvent and hence they contribute to the binding energy of association.

Hydrogen bonds

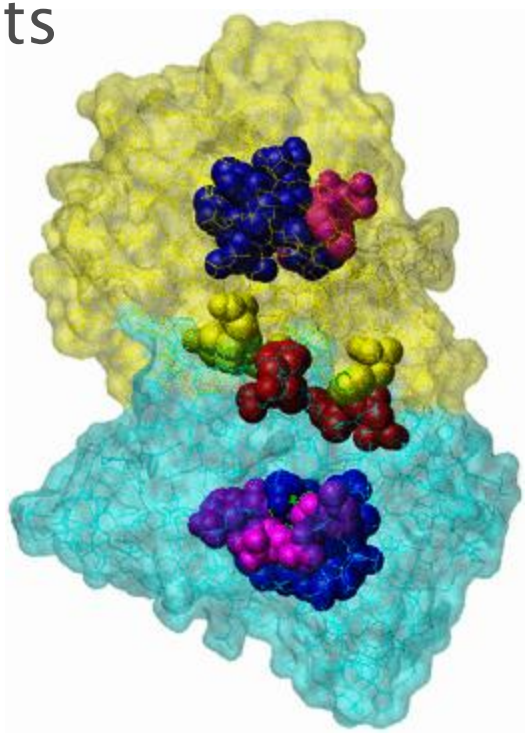
- Hydrogen bonds between protein molecules are more favorable than those made with water
- Interfaces in permanent associations tend to have fewer hydrogen bonds than interfaces in transient associations
- The number of hydrogen bonds is about 1 per 170 $Å^2$ buried surface
- In a set of reasonably stable dimers there are, on average, 0.9 to 1.4 hydrogen bonds per 100 $Å^2$ of contact area buried (interfaces covering > 1000 $Å^2$)
- The number of hydrogen bonds varies from 0 to 46
- Side-chain hydrogen bonds represent approximately 76-78% of the interactions.

Secondary structure of protein-protein interfaces

- Can vary drastically
- In one study the loop interactions contributed, on average, 40% of the interface contacts
- In another study (involving 28 homodimers), 53% of the interface residues were α-helical, 22% β -sheets, and 12% $\alpha\beta$, with the rest being coils

Hot spots

- Residues that make significant contribution to the binding free energy are generally clustered together
- The clusters are called the hot-spots
- Introduced by Jim Wells

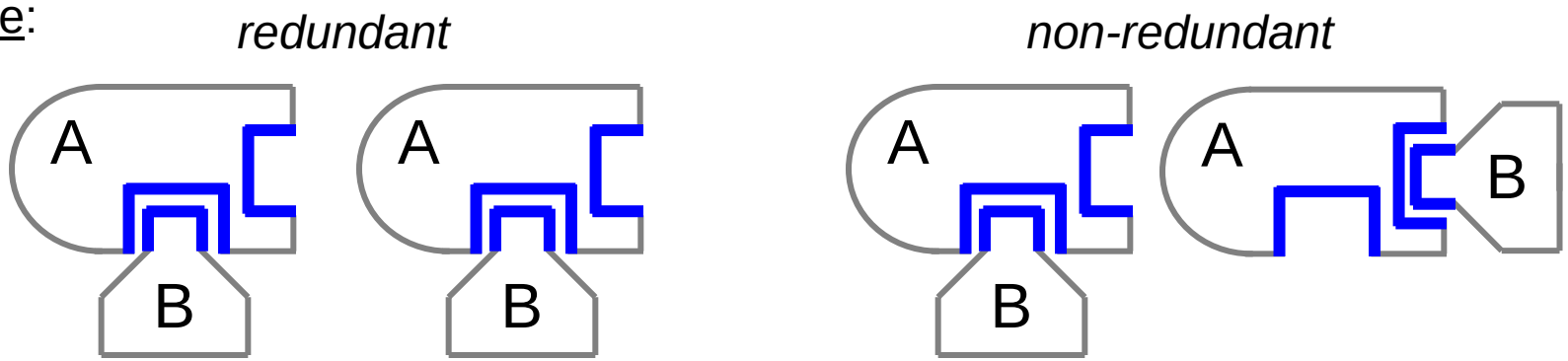


Redundant interactions

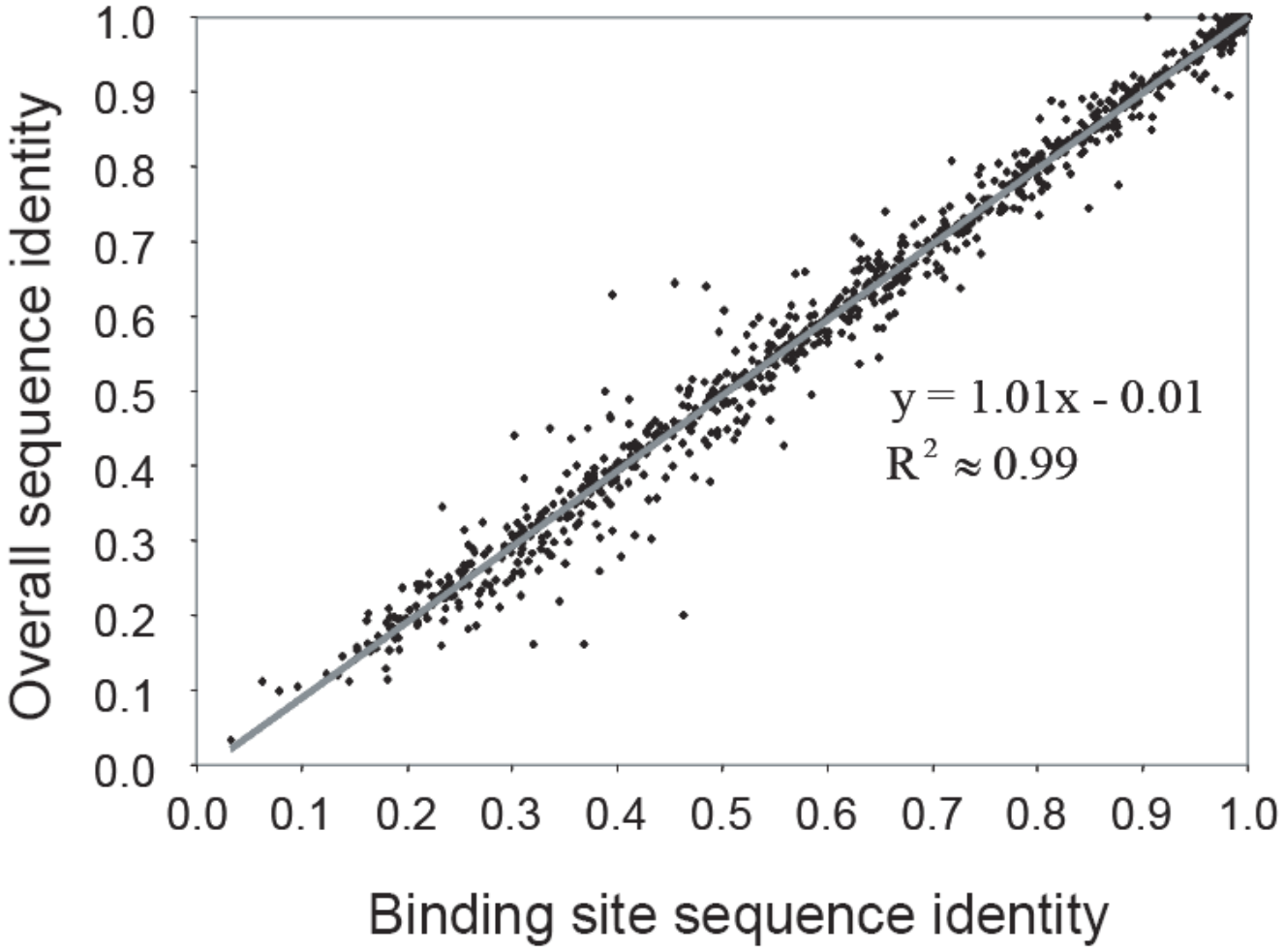
Definition. Two interactions, (A_1, B_1) and (A_2, B_2) are *redundant* if both pairs of partners are similar AND interfaces are similar :

1. Sequence identity between A_1 and A_2 is more then 90% ;
2. Sequence identity between B_1 and B_2 is more then 90%;
3. The interfaces of (A_1, B_1) and (A_2, B_2) are in the same PIBASE cluster.

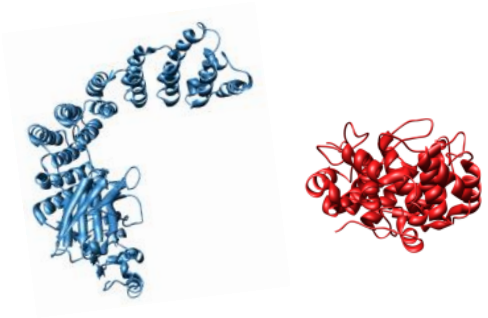
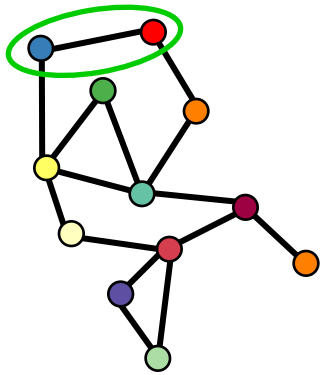
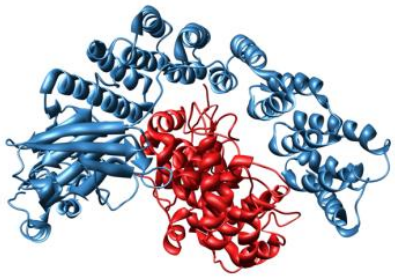
Example:



Are PPI interfaces more conserved in sequence than the rest of the protein?



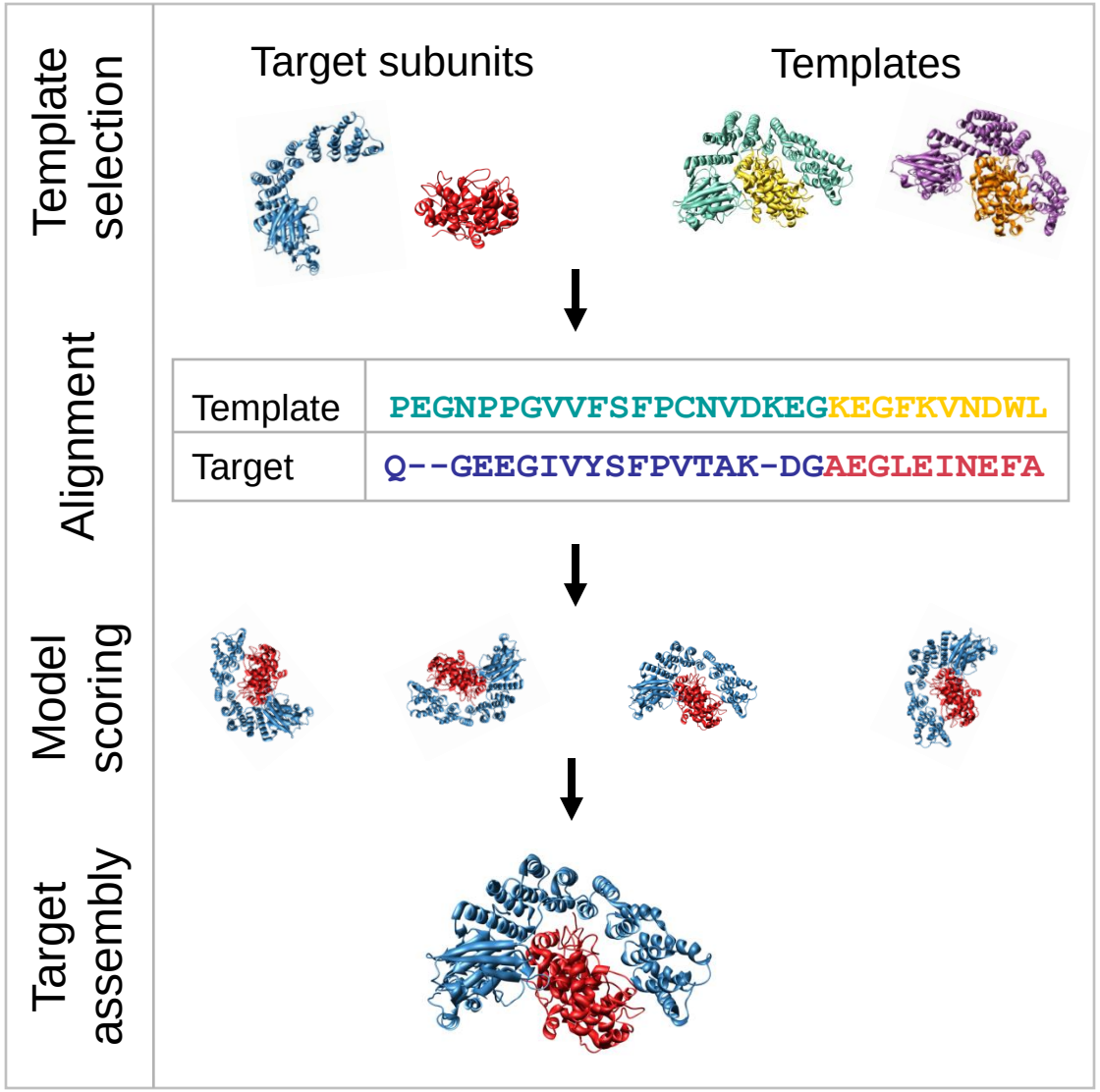
Modeling structures of subunit-subunit interactions

INPUT	Subunit structures	Binary interaction	OUTPUT	Assembly structure
				

Methods usually address three common aspects:

1. Efficient representation of structures
2. Comprehensive enumeration of all candidate models
3. Accurate selection of the native model

Homology modeling of protein assemblies



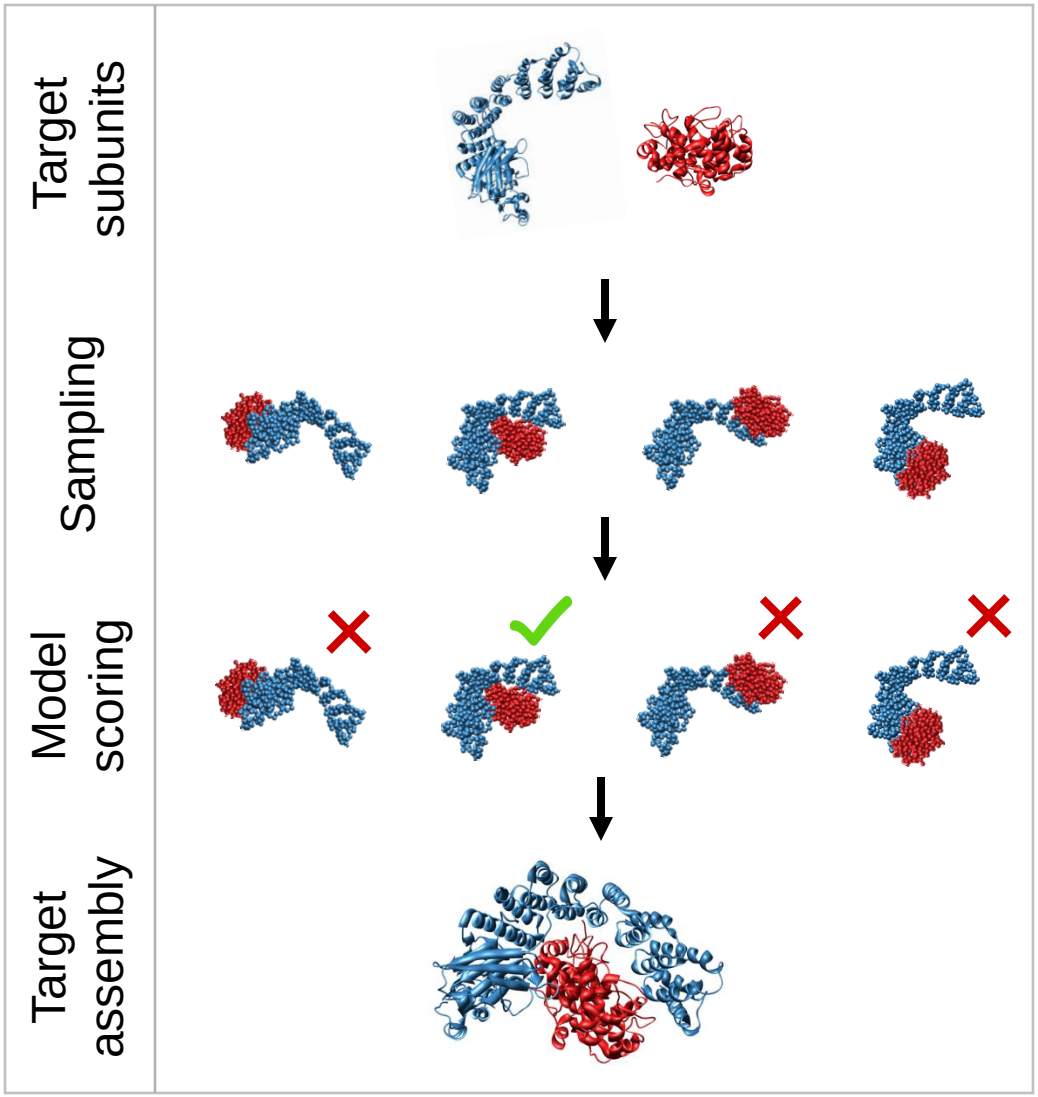
Pros:

- high accuracy
- fast

Cons:

- low coverage
- predicts only existing binding modes

Protein docking



Pros:

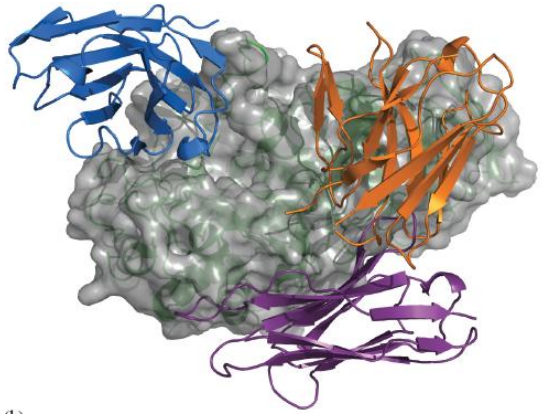
- high coverage
- can predict novel binding modes

Cons:

- low accuracy
- slow

Homology modeling: Can we improve it?

- What if we have only substructures as templates, not the entire structure?
- What if we have two structural templates that observe two different binding modes?



- Can we use additional data to improve the accuracy?

A recent approach addressing this questions

Nucleic Acids Research, 2006, Vol. 34, No. 10 ***2943–2952***
doi:10.1093/nar/gkl353

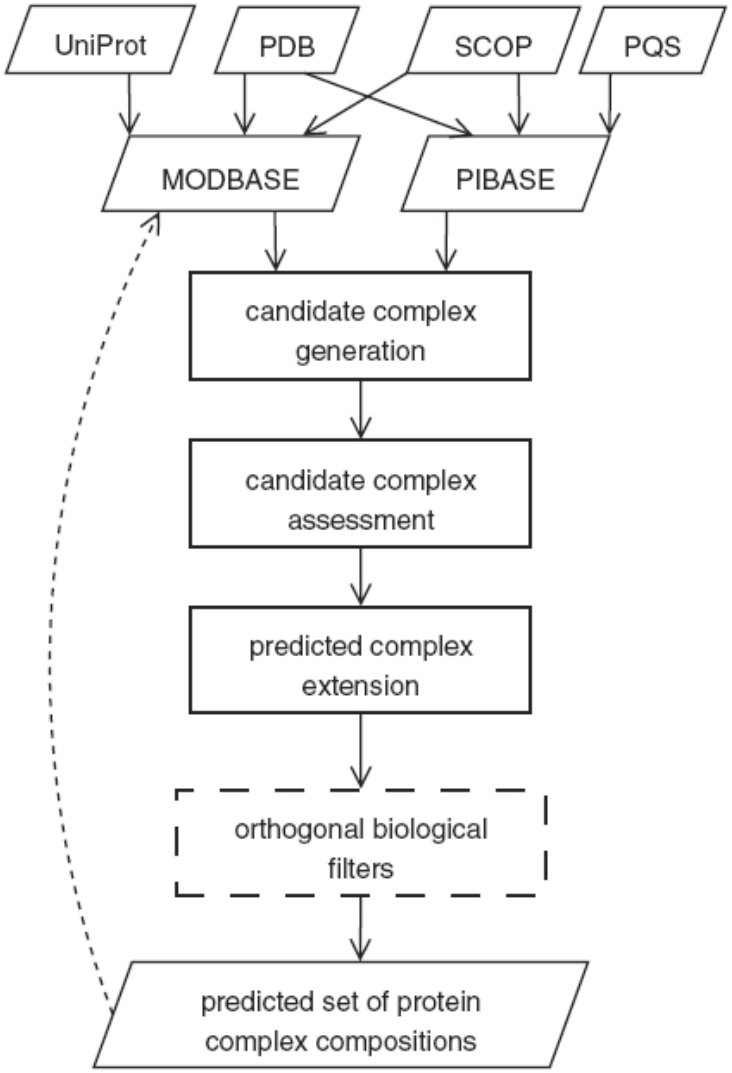
Protein complex compositions predicted by structural similarity

Fred P. Davis[1,2], Hannes Braberg[1,2], Min-Yi Shen[1,2], Ursula Pieper[1,2], Andrej Sali[1,2,*] and M.S. Madhusudhan[1,2,*]

[1]Department of Biopharmaceutical Sciences and [2]Department of Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, 1700 4th Street, Byers Hall, San Francisco, CA 94143-2552, USA

Received March 15, 2006; Revised April 1, 2006; Accepted April 20, 2006

Methods flowchart



Statistical potential to evaluate protein interfaces

- A series of statistical potentials was built using the binary domain interfaces in PIBASE
- Extracted from structures at or above 2.5 s resolution, randomly excluding 100 benchmark interfaces
- 24 statistical potentials were built using different values of 3 parameters:
  - The contacting atom types (main chain–main chain, main chain–side chain, side chain–side chain or all)
  - The relative location of the contacting residues (inter- or intra-domain)
  - Distance threshold for contact participation

Statistical potential to evaluate protein interfaces

$$g_{ij} = \frac{\sum_{p=1}^{N} \sum_{c=1}^{\Delta n_{ij}^{(p)}(R_o)} \mathrm{cifa}_{ci,\,cj} n_p}{\sum_{p=1}^{N} n_{ij}^{(p)} \max\left(\mathrm{cifa}_{i,j}\right)}$$

$$\mathrm{cifa}_{x,\,y} = \min\left(\frac{\text{interacting atoms}_x}{\text{atoms}_x}, \frac{\text{interacting atoms}_y}{\text{atoms}_y}\right),$$

$$n_{ij}^{(p)} = \begin{cases} n_i^{(p)} n_j^{(p)} & \text{intra-domain potential.} \\ n_i^{(d1)} n_j^{(d2)} + n_i^{(d2)} n_j^{(d1)} & \text{inter-domain potential.} \end{cases}$$

$$w_{ij} = -\ln\left[\frac{g_{ij}}{\frac{1}{400}\sum_{k=1}^{20}\sum_{l=1}^{20} g_{kl}}\right].$$

i, j: residue types in protein *p*

Adding experimental data

		Experimental Overlap		
	Predicted	All	BIND	Cellzome
Binary Interactions				
experimental		19,424	13,191	6,942
Z-score ≤ -1.7	12,867	409	324	151
Z + Co-Function	6,808	390	311	145
Z + Co-Localization	4,606	278	220	102
Z + Co-Loc + Co-Func	3,387	270	217	97
Higher-Order complexes				
experimental		783	296	491
Z-score	12,702	66	54	35
Z + Co-Function	3,544	51	45	28
Z + Co-Localization	2,189	14	7	10
Z + Co-Loc + Co-Func	1,234	8	4	7

Protein docking

doi:10.1016/S0022-2836(03)00670-3 *J. Mol. Biol.* (2003) **331**, 281–299

JMB

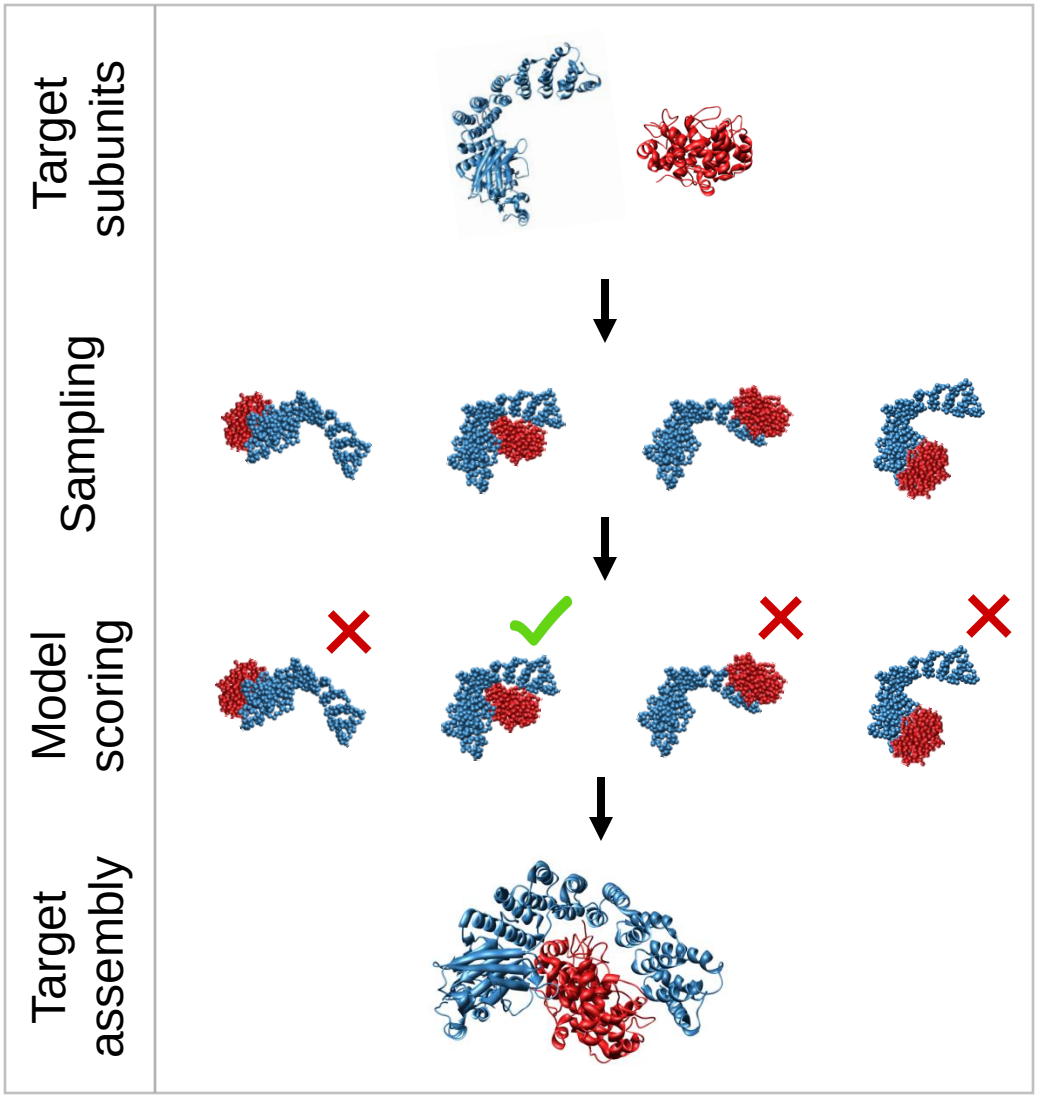
Available online at www.sciencedirect.com

SCIENCE DIRECT®

Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations

Jeffrey J. Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman
Brian Kuhlman, Carol A. Rohl and David Baker*

Protein docking



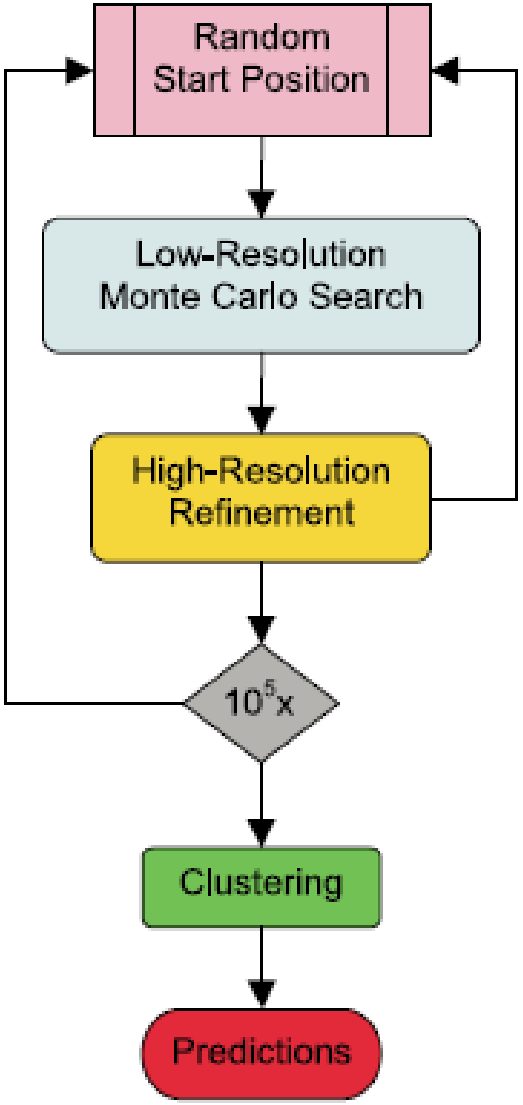
Pros:

- high coverage
- can predict novel binding modes

Cons:

- low accuracy
- slow

Docking protocol



Protocol details

1. Create a decoy set: start with random orientation of each partner + translation of one partner along the line of protein centers to create glancing contact

2. Perform Monte-Carlo simulation: 500 attempts of translating and rotating one partner around surface of another one. 50% acceptance rate. Each step is chosen randomly with a mean value of 0.7 Å (translation) and 5 (rotation)

3. Low-resolution residue scale potentials are calculated based on Bayesian expansion that estimates the probability of correctness for each decoy

4. High resolution refinement: explicit side-chains are added using a backbone-dependent rotamer packing algorithm; use fixed number of multiple rotamers; select an optimal configuration using simulated annealing Monte-Carlo search

Protocol details (contd.)

5. Rigid body is minimized using a full-atom scoring function

6. Select the best-scoring candidates and cluster them using pair-wise RMSD using hierarchical clustering algorithm with 2.5 Å clustering threshold

7. Clusters with the most members are selected as final

All-atom scoring function

Terms included:

• van der Waals interactions

• solvation using a pair-wise Gaussian solvent-exclusion model

• hydrogen bonding energies using an orientation-dependent function derived from high-resolution protein structures

• a rotamer probability term

• residue–residue pair interactions derived statistically from a database of protein structures

• a simple electrostatic term

•a surface area and atomic solvation term (for decoy discrimination only, due to the expense of calculation)

All-atom scoring function

General form of all-atom scoring function:

$$
\begin{aligned}
S = {} & w_{\text{atr}} S_{\text{atr}} + w_{\text{rep}} S_{\text{rep}} + w_{\text{sol}} S_{\text{sol}} + w_{\text{sasa}} S_{\text{sasa}} + w_{\text{hb}} S_{\text{hb}} \\
& + w_{\text{dun}} S_{\text{dun}} + w_{\text{pair}} S_{\text{pair}} + w_{\text{elec}}^{\text{sr-rep}} S_{\text{elec}}^{\text{sr-rep}} + w_{\text{elec}}^{\text{sr-atr}} S_{\text{elec}}^{\text{sr-atr}} \\
& + w_{\text{elec}}^{\text{lr-rep}} S_{\text{elec}}^{\text{lr-rep}} + w_{\text{elec}}^{\text{lr-atr}} S_{\text{elec}}^{\text{lr-atr}} \qquad (7)
\end{aligned}
$$

Weights are learned using a statistical approach: a logistic regression was used to define the weights that maximally separates good decoys from others

All-atom scoring function

General form of all-atom scoring function:

$$S = w_{\text{atr}}S_{\text{atr}} + w_{\text{rep}}S_{\text{rep}} + w_{\text{sol}}S_{\text{sol}} + w_{\text{sasa}}S_{\text{sasa}} + w_{\text{hb}}S_{\text{hb}} + w_{\text{dun}}S_{\text{dun}} + w_{\text{pair}}S_{\text{pair}} + w_{\text{elec}}^{\text{sr-rep}}S_{\text{elec}}^{\text{sr-rep}} + w_{\text{elec}}^{\text{sr-atr}}S_{\text{elec}}^{\text{sr-atr}} + w_{\text{elec}}^{\text{lr-rep}}S_{\text{elec}}^{\text{lr-rep}} + w_{\text{elec}}^{\text{lr-atr}}S_{\text{elec}}^{\text{lr-atr}} \quad (7)$$

Weights are learned using a statistical approach: a logistic regression was used to define the weights that maximally separates good decoys from others

A potential problem: what if the contribution of some members is not linear?

Results

Four classes of complexes:

- Enzyme/Inhibitor
- Antibody/Antigen
- Difficult
- Other

Two types of structural conformations:

- Semibound
- Unbound-unbound

Overview of correct predictions

Table 2. Correct predictions by interface type

	All			Unbound–unbound			Semibound		
	BB-p	XU-p	XU-g	BB-p	UU-p	UU-g	BB-p	BU-p	BU-g
Enzyme/inhibitor	21/22	18/22	17/22	15/16	12/16	11/16	6/6	6/6	6/6
Antibody/antigen	10/16	9/16	8/16	3/5	1/5	3/5	7/11	8/11	5/11
Other	5/10	5/10	3/10	1/4	1/4	0/4	4/6	4/6	3/6
Difficult	6/6	0/6	0/6	4/4	0/4	0/4	2/2	0/2	0/2
Total	42/54	32/54	28/54	23/29	14/29	14/29	19/25	18/25	14/25

How can we combine both approaches?

- Search locally, not globally
- Use known data about related proteins/assemblies
- Knowledge of subunit binding sites is crucial when associating them into assembly

Do similar proteins use similar binding sites?

Do similar proteins use similar binding sites?

Family *F* of protein homologs

↓

$\forall\ p_i \in F$: determine binding sites

↓

select a representative structure $p_R \in F$

↓

$\forall\ p_i \in F$: map its binding sites onto p_R

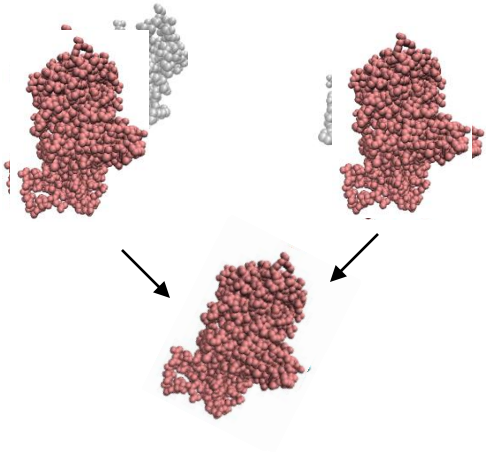
using structural alignment of p_i and p_R

↓

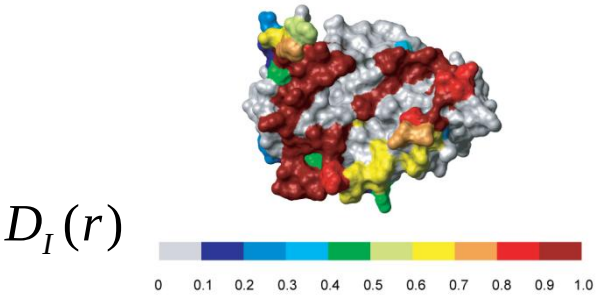
generate a random sample of binding sites for p_R

↓

calculate binding site localization



$$Loc = \frac{\sum_{r \in CR} D_I(r)}{N_{CR}}, \quad D_I(r) = \frac{N_{BS}(r)}{\max_r(N_{BS}(r))}$$



Yes: 72% of 1,847 families have binding sites with co-localization greater than expected by chance

Korkin D, Davis FP, Sali A, *Protein Sci.*, 2005

Comparative patch analysis

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Structural Modeling of Protein Interactions by Analogy: Application to PSD-95

Dmitry Korkin[1,2,3], Fred P. Davis[1,2,3], Frank Alber[1,2,3], Tinh Luong[4], Min-Yi Shen[1,2,3], Vladan Lucic[5], Mary B. Kennedy[4], Andrej Sali[1,2,3*]

1 Department of Biopharmaceutical Sciences, University of California San Francisco, San Francisco, California, United States of America, 2 Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, 3 California Institute for Quantitative Biomedical Research, University of California San Francisco, San Francisco, California, United States of America, 4 Division of Biology, California Institute of Technology, Pasadena, California, United States of America, 5 Department of Structural Biology, Max Planck Institute of Biochemistry, Martinsried, Germany

Basic steps of comparative patch analysis

Target domains

Korkin D, Davis FP, Alber F, Lucic V, Kennedy MB, Sali A., *PLoS Comput. Biol.*, 2006

Comparative patch analysis: Methods

- Protein binding sites were extracted from PiBASE (Davis FP and Sali A, 2005)
  - database of non-redundant protein interactions
  - proteins are clustered into families based on SCOP classification

- PatchDOCK was used for the local docking (Shneidman-Duhovny et al, 2005)
  - rigid body docking
  - restrained to maximize geometrical complementarity of the given binding sites

- Scoring function is composite:

$$f_{SCORE} = f_{DOPE} + f_{PATCHDOCK}$$

  - DOPE score: an atomic distance-dependent pairwise statistical potential (Shen MY and Sali A, 2006)

Comparative patch analysis: Computational challenges

Time complexity of the method:

$$T(N,M) = \underbrace{O(N) + O(M)}_{\text{alignment}} + \underbrace{O(NM)}_{\text{loc. docking}} + \underbrace{O(NM)}_{\text{scoring}}$$

N, *M* are the numbers of family members for the input subunits

- Running time for a benchmark set of 20 protein assemblies, with 50 non-redundant members on average: ~800 CPU-hours.

- *N* and *M* could be large (up to 3,000). Can we reduce them?

The number of binding sites can be reduced

Benchmark: Comparative patch analysis converges to a native configuration, if the residue overlap of the input and native binding sites is $\geq$ 75%

Idea: No need to try all sites with the high co-localization

Solution (work in progress):

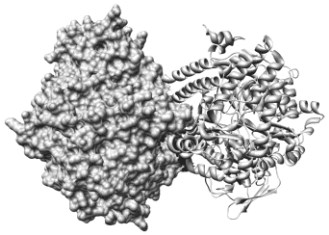
- cluster all binding sites based on their mutual overlap;
- use one representative binding site per cluster as the input to comparative patch analysis;

Performance on a benchmark set

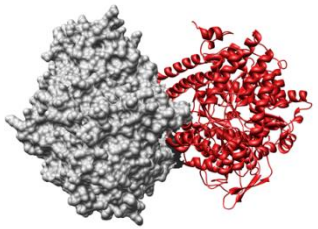
- benchmark set: 20 binary protein complexes (9 multidomain proteins, 11 protein assemblies)
- the method was evaluated using three different measures:

$$(1)\ O_B = \frac{1}{2}\left[\frac{N\ \ B_1^{pred} \cap B_1^{exp}}{N\ \ B_1^{pred} \cup B_1^{exp}} + \frac{N\ \ B_2^{pred} \cap B_2^{exp}}{N\ \ B_2^{pred} \cup B_2^{exp}}\right]; \quad (2)\ O_I = \frac{N(I_{pred} \cap I_{native})}{N(I_{pred} \cup I_{native})}; \quad (3)\ \textit{RMS error}$$

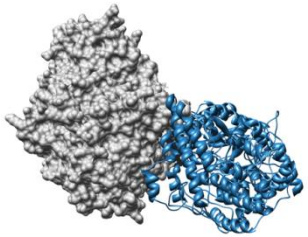
native



2 binding sites



1 binding site



conventional docking