Outline

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

Goal: Demonstrate Current Methodologies & Capabilities in Protein-Protein Docking
Cellular Function Depends on Protein-Protein Interaction

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

Faulty interactions result in diseases

Protein docking tests our fundamental knowledge of biomolecular physics

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

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

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

FTDOCK: Fourier-Transform Docking (Rigid Body)

- Discretize the protein shape:
  \[ a_{i,\rho} = \begin{cases} 1, & \text{surface of molecule} \\ 0, & \text{outside of molecule} \end{cases} \]
  \[ b_{i,\varphi} = \begin{cases} 1, & \text{core of molecule} \\ 0, & \text{outside of molecule} \end{cases} \]

- And correlate the functions:
  \[ C_{\alpha,\beta,\gamma} = \sum_l \sum_m \sum_n a_{l,m,n} b_{l,m,n} \gamma \]

- \( l,m,n,\alpha,\beta,\gamma \to N^6 \)

FTDOCK

- Use a Discrete Fourier Transform
  \[ X_{\alpha,\beta,\gamma} = \sum_l \sum_m \sum_n \exp \left[ -\frac{2\pi}{N} (l\alpha + m\beta + n\gamma) \right] x_{l,m,n} \]

- Multiply in Fourier Space:
  \[ C_{\alpha,\beta,\gamma} = X_{\alpha,\beta,\gamma} B_{\alpha,\beta,\gamma} \]

- Invert:
  \[ x_{l,m,n} = \frac{1}{N^3} \sum_l \sum_m \sum_n \exp \left[ \frac{2\pi}{N} (l\alpha + m\beta + n\gamma) \right] C_{\alpha,\beta,\gamma} \]

- DFT \( \to N^3 \ln N^3 \)

- Then, search over rotation space: \( [\alpha,\beta,\gamma] \)
A wide variety of methods have been developed since Katchalski-Katzir

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

Protein Docking is Difficult!

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

Need to simplify!!
Our Approach to Modeling Proteins

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

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

**Low-Resolution Search**

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

Protein representation: backbone atoms + average centroids

**RosettaDock Algorithm Overview**

*Random Start Position*

1. Low-Resolution Monte Carlo Search
2. High-Resolution Refinement
3. Clustering

*Predictions*
Low-Resolution Search

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

Protein representation:
backbone atoms + average centroids

- Mimics physical diffusion process

Residue-scale scoring

<table>
<thead>
<tr>
<th>Score</th>
<th>Representation</th>
<th>Physical Force</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contacts</td>
<td>r_{center-center} &lt; 6 Å</td>
<td>Attractive van der Waals</td>
</tr>
<tr>
<td>Bumps</td>
<td>(r – R_{ij})^2</td>
<td>Repulsive van der Waals</td>
</tr>
<tr>
<td>Residue environment</td>
<td>-ln(P_{env})</td>
<td>Solvation</td>
</tr>
<tr>
<td>Residue pair</td>
<td>-ln(P_{ij})</td>
<td>Hydrogen bonding electrostatics, solvation</td>
</tr>
<tr>
<td>Alignment</td>
<td>-1 for interface residues in Antibody CDR</td>
<td>(bioinformatic)</td>
</tr>
<tr>
<td>Constraints</td>
<td>varies</td>
<td>(biochemical)</td>
</tr>
</tbody>
</table>

High-Resolution Refinement

- Simultaneous rigid-body and side-chain refinement
Side Chain Packing

- Build amino acid side chains
  - Choose side chains from Dunbrack’s backbone-dependent rotamer library
  - Vary $\chi_1$, $\chi_2$, $\chi_3$, $\chi_4$ angles
  - Minimize a full-atom energy function w.r.t. all rotamer combinations
  - With strict VdW parameters, extra angles are necessary (Chu Wang)

Phenylalanine rotamers
(Richardson, 2000)

Fig. 1. Observed frequency of the gauche$^+$ ($\chi_1 +60^\circ$), gauche$^-$ ($\chi_1 -60^\circ$), and trans ($\chi_1 180^\circ$), rotamers of valine (horizontally, respectively) in sheet, helix, and coil regions (vertically, respectively) of proteins as a function of the backbone dihedral angle $\psi$.
Data were taken from a list of 850 proteins, 1.7 Å resolution or better, and mutual sequence identity less than 50% (http://www.fccc.edu/research/labs/dunbrack/culledpdb.html). A B-factor cutoff of 40 was used as recommended by Lovell et al. [15].

Rotamer Statistics

Minimization

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

• Simultaneous rigid-body displacement and side chain minimization

Full-Atom scoring

<table>
<thead>
<tr>
<th>Score</th>
<th>Form / Source</th>
<th>Discriminatory z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repulsive van der Waals</td>
<td>Modified Lennard-Jones 6-12</td>
<td>73.0</td>
</tr>
<tr>
<td>Attractive van der Waals</td>
<td>Lennard-Jones 6-12</td>
<td>49.0</td>
</tr>
<tr>
<td>Surface area solvation</td>
<td>Surface area (see Tsai 2003)</td>
<td>28.5</td>
</tr>
<tr>
<td>Gaussian solvent-exclusion</td>
<td>Lazaridis &amp; Karplus, 1999</td>
<td>27.2</td>
</tr>
<tr>
<td>Rotamer probability</td>
<td>Dunbrack &amp; Cohen, 1997</td>
<td>19.6</td>
</tr>
<tr>
<td>Residue pair probability</td>
<td>Empirical, Kuhlman &amp; Baker 2000</td>
<td>6.9</td>
</tr>
<tr>
<td>Electrostatics</td>
<td>Coulomb model with simple charges</td>
<td>0.4-15.1 (LR rep)</td>
</tr>
</tbody>
</table>

Scoring Weights

<table>
<thead>
<tr>
<th>Score</th>
<th>Weight (P)</th>
<th>Weight (M)</th>
<th>Weight (O)</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repulsive van der Waals</td>
<td>0.80</td>
<td>0.338</td>
<td>0.08</td>
<td>73.0</td>
</tr>
<tr>
<td>Attractive van der Waals</td>
<td>0.80</td>
<td>0.338</td>
<td>0.338</td>
<td>45.0</td>
</tr>
<tr>
<td>Surface area solvation</td>
<td>-</td>
<td>0.344</td>
<td>-</td>
<td>28.5</td>
</tr>
<tr>
<td>Gaussian solvent-exclusion</td>
<td>0.80</td>
<td>0.279</td>
<td>0.279</td>
<td>27.2</td>
</tr>
<tr>
<td>Rotamer probability</td>
<td>0.79</td>
<td>0.069</td>
<td>0.069</td>
<td>19.6</td>
</tr>
<tr>
<td>Hydrogen bonding</td>
<td>2.1</td>
<td>0.441</td>
<td>0.441</td>
<td>-</td>
</tr>
<tr>
<td>SC/SC + SC/BB</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14.9</td>
</tr>
<tr>
<td>BB/BB</td>
<td>-</td>
<td>-</td>
<td>6.8</td>
<td>6.8</td>
</tr>
<tr>
<td>Residue pair probability</td>
<td>0.66</td>
<td>0.164</td>
<td>0.164</td>
<td>6.9</td>
</tr>
<tr>
<td>Simple electrostatics</td>
<td>-</td>
<td>0.025</td>
<td>0.025</td>
<td>3.2</td>
</tr>
<tr>
<td>Short-range repulsive</td>
<td>-</td>
<td>0.025</td>
<td>0.025</td>
<td>8.3</td>
</tr>
<tr>
<td>Short-range attractive</td>
<td>-</td>
<td>0.098</td>
<td>0.098</td>
<td>15.1</td>
</tr>
<tr>
<td>Long-range repulsive</td>
<td>-</td>
<td>0.0020</td>
<td>0.0020</td>
<td>6.4</td>
</tr>
<tr>
<td>Long-range attractive</td>
<td>-</td>
<td>0.0020</td>
<td>0.0020</td>
<td>-</td>
</tr>
</tbody>
</table>
Hydrogen Bonding Energy
(Kortemme, Morozov & Baker 2003 JMB)

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

\[ \Delta G = -kT \ln P \]

\[ E_{HB} = W_{HB}(E(H) + E(\Theta) + E(\Psi) + E(X)) \]

Score correlates with Binding Energy

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

\[ \Delta \text{score for bound backbone docking} \]

Clustering

- Compare all top-scoring decoys pairwise

\[ \text{RMSD} = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \]

- Cluster decoys hierarchically

- Decoys within 2.5Å form a cluster
Benchmark Studies

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

- Bound-Bound
  - Start with bound complex structure, but remove the side chain configurations so they must be predicted

- Unbound-Unbound
  - Start with the individually-crystallized component proteins in their unbound conformation

- Bound-Unbound (Semibound)

Binding Funnels

Decoys: graylab.jhu.edu

Anticholinesterase / Fasciculin II

Antibody Fab 5G9 / Tissue Factor
Benchmark Results

<table>
<thead>
<tr>
<th></th>
<th>Bound Perturbations</th>
<th>Unbound Perturbations</th>
<th>Global Searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme/Inhibitor</td>
<td>21/22</td>
<td>19/22</td>
<td>17/22</td>
</tr>
<tr>
<td>Antibody/Antigen</td>
<td>10/16</td>
<td>9/16</td>
<td>9/16</td>
</tr>
<tr>
<td>Other</td>
<td>5/10</td>
<td>5/10</td>
<td>3/10</td>
</tr>
<tr>
<td>Difficult</td>
<td>6/6</td>
<td>0/6</td>
<td>0/6</td>
</tr>
<tr>
<td>TOTAL</td>
<td>42/54</td>
<td>32/54</td>
<td>28/54</td>
</tr>
</tbody>
</table>

Number of successful dockings, starting from either bound or unbound protein backbones and searching either near the native structure or globally. Benchmark set assembled by R. Chen et al., see Proteins 2003

ZDOCK / RosettaDock
(Vajda & Camacho 2004)
CAPRI: Critical Assessment of PRotein Interactions

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

Hemagglutinin + Antibody
α-amylase + Camelid Antibody
T-Cell Receptor + Strep. pyrogenic exotoxin A (superantigen)

Target 6 (Round 2, Mar 2002)
- α-amylase + VHH, model #1:
  - 48/65 contacts, distance 1.33Å, rotation 3º, rmsd 1.5Å

Target 8 (Round 3, M. Daily, Jan 2003)
- Laminin + Nidogen, model #2:
  - 53% contacts, rmsd 4.6 Å, interface rmsd 0.66 Å

Xtal by C. Cambillau, CNRS
Xtal by T. Springer, Harvard
Docking a Homology Model (Round 4, Sep 2003)

CAPRI T11/12: Cohesin + Dockerin
Model #6 (T11): 42% contacts, 6.1 Å rmsd, 1.9 Å interface rmsd
- Dockerin coordinates modeled by homology via the Robetta server
- RosettaDock produced the best model by correct contacts

Target 19: prion + Fab, model #2
64% contacts, rmsd 3.64 Å, interface rmsd 1.27 Å

Prion constructed manually from a 95% identical homologue

Targets 4 and 5 (Round 2)
- α-amylase + VHH
  - Incorrectly assumed binding occurs at CDRs
Target 7 – “Homology Target”

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

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

<table>
<thead>
<tr>
<th>Target</th>
<th>Complex</th>
<th>Type</th>
<th>Nres Model</th>
<th>Final L rmsd</th>
<th>I rmsd Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Collin D – immD</td>
<td>BB-BB</td>
<td>194 7</td>
<td>0.88 0.547 0.243 ***</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Coheisin-dockern</td>
<td>U-B</td>
<td>196 1</td>
<td>0.87 0.99 0.51 ***</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Coheisin-dockern</td>
<td>U-H</td>
<td>196 5</td>
<td>0.42 6.11 1.93 ***</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Oliva prim – fab</td>
<td>H-B</td>
<td>312 2</td>
<td>0.64 3.84 1.27 ***</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Laminin-ridigen</td>
<td>U-B</td>
<td>427 2</td>
<td>0.53 4.63 0.66 ***</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>GH11 xylanase – XIP</td>
<td>H-O</td>
<td>464 5</td>
<td>0.07 12.91 8.78 -</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>sag1-fab</td>
<td>U-B</td>
<td>474 NP</td>
<td>NP NP NP -</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>GH11 xylanase – TAXI</td>
<td>U-B</td>
<td>552 NP</td>
<td>NP NP NP -</td>
<td></td>
</tr>
<tr>
<td>16*</td>
<td>GH10 xylanase – XIP</td>
<td>H-U</td>
<td>575 7</td>
<td>0.14 8.13 11.64 *</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>myp1-PP1</td>
<td>U-B</td>
<td>600 NP</td>
<td>NP NP NP -</td>
<td></td>
</tr>
</tbody>
</table>

9 LiT homodimer | U-U | 412 | |
10 TBEV envelope trimer | U-U | 1146 | |

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

Many Docking Players (Vajda/Camacho 2004)

Table 1: Algorithms of some common protein-protein docking methods

<table>
<thead>
<tr>
<th>Method (Year)</th>
<th>Regularity search</th>
<th>Accepting/rejecting</th>
<th>Accuracy of DMMs (and submission)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRM (2016)</td>
<td>Four breeding Monte Carlo with anti-</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
<tr>
<td>Coelho and</td>
<td>Regularity search</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
<tr>
<td>Vajda (2016)</td>
<td>Regularity search</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
<tr>
<td>Zhong (2016)</td>
<td>Regularity search</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
<tr>
<td>ZYMM (2016)</td>
<td>Regularity search</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
<tr>
<td>GRIDOX (2016)</td>
<td>Regularity search</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
<tr>
<td>SWMM (2016)</td>
<td>Regularity search</td>
<td>Accepting/rejecting</td>
<td>Accuracy of DMMs (and submission)</td>
</tr>
</tbody>
</table>
RosettaDock Assumptions

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

What RosettaDock study tells us about Proteins

- Packing dominates free energy
- Solvation, hydrogen bonding also important
- Electrostatics not important?
- Energy function is closer to correct than past models
- A short list of probable best docking structures
What it doesn’t tell you about Proteins

• THE energy function
• Unambiguously the “best” conformation
• How specificity is achieved
• Binding affinities

Side chain movement
(Camacho 2004 PNAS)

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

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

Three-step mechanism
Grüenberg, Leckner & Nilges 2004 Structure

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

Docking Ensembles
Grüenberg, Leckner & Nilges 2004 Structure

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

Loop Flexibility

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

Rohl, CA et al 2004 to appear
Target 1: HPr + HPr Kinase:
(Round 1, Sep 2001)
- Model #8 among the closest:
  2/52 contacts
distance 2.6Å
rotation 55°
RMSD 8.8Å
Distance constraint between
Ser157C and Asp46A
Xtal by Fieulaine et al., CNRS

Backbone Conformational Change
CAPRI T01: HPr + HPr Kinase (Round 1, Sep 2001)
Terminal helix swings upon docking,
nuzzling HPr in a pocket
No energy funnel for binding the unbound components

Torsion Angle Perturbation
Torsion angle movement in residues 290-292 would allow
the correct conformation to be observed.
Flexible Docking Results
With torsion angle perturbations and explicit minimizations

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

CAPRI T24: Homology model flexible docking (March 2006)

105 residue domain modeled from 30% sequence identity homolog

Prediction by Monica Berrondo and Aroop Sircar
Summary

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

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

Recommended References


Docking into EM maps