Machine Learning to Predict Protein–Protein Interactions

Sergei Grudinin, 24 Apr 2012
Bacterium

Model of a signalling array.
Protein Structure

Amino acid residues  α Helix  Polypeptide chain  Assembled subunits

images are courtesy of Uppsala Universitet, Sweden, http://www.uu.se
**Protein Structure**

- **Protein folding** (prediction of protein *tertiary* structure)
  - works very well if there are homologous structures available
  - many web-servers available
  - CASP competitions

- **Protein docking** (prediction of protein *quaternary* structure)
  - currently much less mature
  - CAPRI competitions

Images are courtesy of Uppsala Universitet, Sweden, http://www.uu.se
Types of Interactions

- Bond vibration
- Angle vibration
- Torsion potentials

van der Waals interactions
Electrostatics
Typical Shape of a Force Field

\[
U = \sum_{\text{bonds}} \frac{1}{2} k_{ij}^b \left( \mathbf{r}_{ij} - \mathbf{r}_{ij}^0 \right)^2 \\
+ \sum_{\text{angles}} \frac{1}{2} k_{ijk}^\theta \left( \mathbf{\theta}_{ijk} - \mathbf{\theta}_{ijk}^0 \right)^2 \\
+ \sum_{\text{torsions}} \left( \sum_n k_\theta \left[ 1 + \cos(n \mathbf{\phi} - \mathbf{\phi}^0) \right] \right) \\
+ \sum_{\text{improvers}} k_\xi \left( \mathbf{\xi}_{ijkl} - \mathbf{\xi}_{ijkl}^0 \right)^2 \\
+ \sum_{i,j} \frac{q_i q_j}{4\pi \varepsilon_0 \mathbf{r}_{ij}} \\
+ \sum_{i,j} 4\varepsilon \left[ \left( \frac{\sigma}{\mathbf{r}} \right)^{12} - \left( \frac{\sigma}{\mathbf{r}} \right)^6 \right]
\]
Sampling: Bottom-Up approach

Remember: all quantities we care about are averages in phase space:

\[ \langle F \rangle = \int dr^N dp^N F(r^N, p^N) e^{-\beta H(r^N, p^N)} / Z \]

- Integral over momenta may be evaluated analytically
- The difficult problem is the computation of the average of \( F(r^N) \)
- Potential looks like
- Typically we use MD or MC
Sampling: Bottom–Up approach

- Binding free energy $W(r^6)$ can be then evaluated as:
  $$\exp^{-\beta W(r^6)} \sim \int dx^{N-6} \exp^{-\beta U(r^N)} / Z$$

- Then, the minimum value of $W(r^6)$ will correspond to the native complex

Example of 6 degrees of freedom for rigid body docking as it used by HEX algorithm, from Dave Ritchie, INRIA Nancy
Problems

- Number of degrees of freedom (DOF) in a protein $N \sim 10,000$.
- We have to include solvation with DOF $\sim 100,000$.
- Long-range interactions, each atom feels each other atom.
- Extremely computationally expensive. Might take years on a supercomputer.

- Standard forcefields are very limited. They do not work for systems with polarization (ion channels) and in reactive centers.
- Forcefields errors accumulate in big systems.
- Forcefields exist only for a limited number of molecules.
- Small molecules must be parametrized separately.
Possible Solution: Rigid-Body Docking
Rigid–Body Docking

Find the minimum of potential function as fast as possible

\[ E = \int \phi(r) \rho(r) dV \]

For 2 proteins

\[ \phi(r) = \phi_A(r) + \phi_B(r) \]
\[ \rho(r) = \rho_A(r) + \rho_B(r) \]

Therefore,

\[ E = \int (\phi_A(r) \rho_B(r) + \phi_B(r) \rho_A(r)) dV \]

- In the rigid body approximation we have 6 DOFs
- For middle–size proteins we need about 30 points in each direction
- Complexity will be \( \sim 10^9 \) of such integrals
- Modern algorithms simultaneously treat several such terms

from Dave Ritchie’s presentations, INRIA Nancy, http://loria.fr/~richtied
FFT in Cartesian System

\[ f_{A_{i,m,n}} = \begin{cases} 1 : & \text{surface of molecule} \\ \rho : & \text{core of molecule} \\ 0 : & \text{open space} \end{cases} \]

\[ f_{B_{i,m,n}} = \begin{cases} 1 : & \text{inside molecule} \\ 0 : & \text{open space} \end{cases} \]

\[ f_{C_{i,i'\alpha,\beta,\gamma}} = \sum_{i=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} f_{A_{i,m,n}} \times f_{B_{i+i'\alpha,\beta,\gamma+n}} \]

\[ \alpha, \beta, \gamma \text{ - shift vectors of A relative to B} \]

\[ N \text{ - number of points in each direction} \]

\[ F_A = \text{DFT}(f_A) \]

\[ F_B = \text{DFT}(f_B) \]

\[ F_C = (F_A^*) (F_B) \]

\[ f_C = \text{IFT}(F_C) \]

- for each orientation of B we need \( O(N^6) \) computations of correlation using the direct method
- or \( O(N^3 \log N^3) \) using FFT

Our Approach

- Hermite Space
- Hermite - Fourier Space
Problems

- Potential function is too simple and in many cases unrealistic.
- 6 DOFs are obviously not sufficient.
- We often start predictions with protein structures in their bound conformations. However, upon binding they adopt different, “unbound” states.
Knowledge-Based Protein Docking: Top-Down Approach
Protein Docking

How to find Binding Free Energy of a protein complex?
- have to make several assumptions
Assumptions I: Interface

- **Binding energy** depends only on the interface between the proteins within a certain *cutoff distance*
Assumptions II: Atom Types

• Protein molecule is represented by a set of $M$ discrete interaction sites that are located at the sites of the atomic nuclei.

• Protein Folding - individual types for all atoms.

• Protein Docking - a set of types, about 20.
Assumptions III:

\[
F(n(r)) \equiv F(n_{11}(r), \ldots, n_{kl}(r), \ldots, n_{MM}(r)) = \sum_{k=1}^{M} \sum_{l=k}^{M} \int_{0}^{r_{\text{max}}} n_{kl}(r) U_{kl}(r) \, dr
\]

• **Binding energy** \( F \) depends only on the distributions \( n_{kl}(r) \) of distances between the interaction sites (the number of site pairs at a certain distance)

• **Binding energy** \( F \) is a linear functional

Given a set of \( n_{kl}(r) \) and constants \( U_{kl}(r) \) we can find the binding free energy \( F(n(r)) \)!
Knowledge-base

- Native: 850 non-homologues complexes from PDB
- Non-native: generated by rolling one over another
- Non-native: generated using NMA
**How do we compute $U_{kl}(r)$?**

- Expand $U_{kl}(r)$ and $n_{kl}(r)$ in an orthogonal basis.
- Compute distance distributions $n_{kl}(r)$ for native and nonnative structures.
- Find energy expansion coefficients $\mathbf{w}$ by solving convex quadratic problem with about $10^5 - 10^6$ linear constraints.

**minimize:**

$$\frac{\mathbf{w} \cdot \mathbf{w}}{2} + \sum_{i=0}^{m} C_{ij} \xi_{ij}$$

**subject to:**

$$y_{ij} [\mathbf{w} \cdot \mathbf{x}_{ij} + b] - 1 + \xi_{ij} \geq 0$$

$$\xi_{ij} \geq 0$$
Algorithm

1) Projection
   - Repeat for all decoy sets
   - Projection:
     - Native
     - Decoy #1
     - Decoy #2
     - ...

2) Formulation
   - Linear combination:
     - II
     - III
     - ...

3) Quadratic Programming
   - Score vs. Distance, Å
RESULTS

One of the simplest functionals \( F_{mn} \) fulfilling these assumptions has the following structure:

\[
F_{mn} \equiv F_{m11}^{nn}, \ldots, n_{kl}^{m1r}\n
\text{MM}_{mn} = M_k \sum_{l=1}^{r} r_{\text{max}} \sum_{n_{kl}} U_{kl}^{m1r} dr_{mwn}
\]

It contains unknown functions \( U_{kl} \) that are to be determined from the training set of native complexes. From now on, we will call these functions scoring functions \( s \).

Once the scoring functions are known, in order to compute the value of \( F \), we need to specify site number densities \( n_{kl}^{m1r} \). In practice, we calculate them as a sum of all distances in a given protein complex by equation 6:

\[
 n_{kl}^{m1r} = \sqrt{\pi/\sigma^2} e^{-\left(r_{ij} - r_{kl}^{m1r}\right)^2/2\sigma^2}
\]

where each distance distribution is represented by a Gaussian distribution centered at \( r_{ij} \) with the variance of \( \sigma^2 \). The sum is taken over all site pairs \( i \) and \( j \) separated by the distance \( r_{ij} \) smaller than \( r_{\text{max}} \) with site \( k \) on the first protein of the complex and site \( l \) on the second protein. In the limiting case of variance tending to zero, equation 6 turns into a sum over Dirac delta functions. In our study, we fix the value of \( \sigma \) to \( \text{˚A} \) for all site distances. However, if one has additional information about individual distance distributions, e.g., Debye-Waller factors, molecular dynamics trajectories, etc., it can be used for more precise parametrization of variance or even instead of the Gaussian approximation in equation 6.

Finally, we compute the score of each protein complex by equation 7:

\[
\text{Score} = \sum_{ij} \Upsilon_{kl}^{m1r} n_{ij}^{m1r}
\]

where the sum is taken over all pairs of atoms \( i \) and \( j \) separated by the distance \( r_{ij} \) smaller than \( r_{\text{max}} \) with atom \( i \) of type \( k \) on the first protein of \( 1 \).

Though the scoring functional 5 is similar by the structure to e.g. the excess internal energy \( ?? \), generally, our scoring functions \( U_{kl}^{m1r} \) are not equal to the potential energy functions between sites \( k \) and \( l \).

Generally, if distance distributions have a non-Gaussian shape, \( n_{kl}^{m1r} = f(r_{ij} - r_{kl}^{m1r}) \), functions \( \Upsilon_{kl}^{m1r} \) will be computed as a convolution \( \Upsilon_{kl}^{m1r} = f \ast U_{kl}^{m1r} \).

aliphatic carbons – \( C_a \) carbons amide nitrogens – oxygens \quad \text{N+ - O-}
CAPRI Blind Predictions

Influenza virus with hemagglutinin protein trimers (HA) on the surface of the viral capsid

Prediction of the complex of HA with the designed protein HB36

X-Ray  Us  Baker’s group
Validation

Rosetta Unbound Benchmark

- Training set of 850 complexes is predicted with 100% accuracy
- Top 1 predictions on Standard Benchmarks (1000 complexes of different qualities, contact side chains rebuilt)
  - Rosetta Unbound 83%
  - Rosetta Bound 89%

<table>
<thead>
<tr>
<th>PDB</th>
<th>Quality</th>
<th>Rank</th>
<th>PDB</th>
<th>Quality</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A0O</td>
<td>3</td>
<td>1</td>
<td>1MAH</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1ACB</td>
<td>2</td>
<td>1</td>
<td>1MDA</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1AHW</td>
<td>3</td>
<td>1</td>
<td>1MEL</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1ATN</td>
<td>1</td>
<td>1</td>
<td>1MLC</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1AVW</td>
<td>2</td>
<td>1</td>
<td>1NCA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1AVZ</td>
<td>-</td>
<td>&gt;10</td>
<td>1NMB</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1BQL</td>
<td>3</td>
<td>1</td>
<td>1PPE</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1BRC</td>
<td>2</td>
<td>1</td>
<td>1QFU</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1BRS</td>
<td>3</td>
<td>1</td>
<td>1SPB</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1BTH</td>
<td>3</td>
<td>1</td>
<td>1STF</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1BVK</td>
<td>3</td>
<td>1</td>
<td>1TAB</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1CGI</td>
<td>3</td>
<td>3</td>
<td>1TGS</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1CHO</td>
<td>1</td>
<td>1</td>
<td>1UDI</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1CSE</td>
<td>2</td>
<td>1</td>
<td>1UGH</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1DFJ</td>
<td>2</td>
<td>1</td>
<td>1WEJ</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>1DQJ</td>
<td>3</td>
<td>1</td>
<td>1WQ1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1EFU</td>
<td>-</td>
<td>&gt;10</td>
<td>2BTF</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1EO8</td>
<td>3</td>
<td>1</td>
<td>2JEL</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1FBI</td>
<td>3</td>
<td>1</td>
<td>2KAI</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>1FIN</td>
<td>-</td>
<td>&gt;10</td>
<td>2PCC</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1FQ1</td>
<td>3</td>
<td>4</td>
<td>2PTC</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1FSS</td>
<td>2</td>
<td>1</td>
<td>2SIC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1GLA</td>
<td>2</td>
<td>1</td>
<td>2SNI</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1GOT</td>
<td>3</td>
<td>1</td>
<td>2TEC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1IAI</td>
<td>2</td>
<td>1</td>
<td>2VIR</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1IGC</td>
<td>3</td>
<td>1</td>
<td>3HHR</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1JHL</td>
<td>3</td>
<td>4</td>
<td>4HTC</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Top1 ITScore 59.3% RosettaDock 66.7% Us 83.3%
# Results

**CapsPepRi Assessment, 2010–2012**

```
<table>
<thead>
<tr>
<th>Rank</th>
<th>Group</th>
<th>T46</th>
<th>T47 (Water-mediated interactions)</th>
<th>T48</th>
<th>T49</th>
<th>T50</th>
<th>T51.1</th>
<th>T51.2</th>
<th>T51.3</th>
<th>T52 (Not assessed)</th>
<th>T53</th>
<th>T54</th>
<th>Summary: #Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bonvin</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>**</td>
<td>7/3</td>
<td>**</td>
<td>+ 4 *</td>
</tr>
<tr>
<td>2</td>
<td>Shen</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td>**</td>
<td>6/3</td>
<td>**</td>
<td>+ 3 *</td>
</tr>
<tr>
<td>3</td>
<td>Bates</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>6/1</td>
<td>**</td>
<td>+ 5 *</td>
</tr>
<tr>
<td>4</td>
<td>Vaida</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>***</td>
<td>5/1</td>
<td>***</td>
<td>+ 3 * + 1 *</td>
</tr>
<tr>
<td>5</td>
<td>Eisenstein</td>
<td>**</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>*</td>
<td>5/2</td>
<td>**</td>
<td>+ 3 *</td>
</tr>
<tr>
<td>6</td>
<td>Fernandez-Recio</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>**</td>
<td>5/1</td>
<td>***</td>
<td>+ 4 *</td>
</tr>
<tr>
<td>7</td>
<td>Zacharias</td>
<td>***</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>5/1</td>
<td>***</td>
<td>+ 4 *</td>
</tr>
<tr>
<td>8</td>
<td>Vakser</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>5/1</td>
<td>**</td>
<td>+ 4 *</td>
</tr>
<tr>
<td>9</td>
<td>ClusPro</td>
<td>***</td>
<td>**</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>*</td>
<td>4/3</td>
<td>**</td>
<td>+ 1 *</td>
</tr>
<tr>
<td>10</td>
<td>Nakamura</td>
<td>***</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>4/1</td>
<td>***</td>
<td>+ 1 * + 3 *</td>
</tr>
<tr>
<td>11</td>
<td>Zou</td>
<td>***</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>4/1</td>
<td>***</td>
<td>+ 3 *</td>
</tr>
<tr>
<td>12</td>
<td>Wang</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>**</td>
<td>4/1</td>
<td>**</td>
<td>+ 3 *</td>
</tr>
<tr>
<td>13</td>
<td>Grudinin</td>
<td>–</td>
<td>**</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td>–</td>
<td>3/2</td>
<td>**</td>
<td>+ 1 *</td>
</tr>
<tr>
<td>14</td>
<td>HADDOCK</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>*</td>
<td>3/1</td>
<td>**</td>
<td>+ 2 *</td>
</tr>
<tr>
<td>15</td>
<td>PIE/DOCK</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>*</td>
<td>3/1</td>
<td>**</td>
<td>+ 2 *</td>
</tr>
<tr>
<td>16</td>
<td>Wolfson</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>3/1</td>
<td>**</td>
<td>+ 2 *</td>
</tr>
<tr>
<td>17</td>
<td>Zhou</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>3/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Seok</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>2/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Elber</td>
<td></td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td>*</td>
<td>2/1</td>
<td>**</td>
<td>+ 1 *</td>
</tr>
<tr>
<td>20</td>
<td>Fernandez-Fuentes</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>2/1</td>
<td>**</td>
<td>+ 1 *</td>
</tr>
<tr>
<td>21</td>
<td>Gray</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>2/1</td>
<td>**</td>
<td>+ 1 *</td>
</tr>
<tr>
<td>22</td>
<td>SwarmDock</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>2/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Camacho</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Cui</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>LZeRd</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Ritchie</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Ten Eyck</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Wang</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Luethy</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Pal</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Poupon</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>SurfFit</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Zhang</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>1/1</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>About 24 Others</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0/0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

*CAPRI Assessment, 2010–2012*

http://web.mit.edu/sheny/capri.html

Wednesday, May 2, 2012
Problems

• Protein flexibility must be taken into account
  Collective motions with Normal Modes

• Sidechain flexibility must be taken into account
  Rotamers optimization
Predicting Positions of Water Around a Protein
Assumptions:

\[ F(n(r)) \equiv F(n^1(r), \ldots, n^M(r)) = \sum_{k=1}^{M} \int_{0}^{r_{\text{max}}} n^k(r) U^k(r) \, dr, \]

- **Solvation free energy** \( F \) depends only on the distributions of distances between the interaction sites (the number of site pairs at a certain distance).
- **Solvation free energy** \( F \) is a linear functional.

Given a set of \( n^k(r) \) and constants \( U^k(r) \) we can find the solvation free energy \( F(n(r)) \)!
RESULTS

Potentials $U^k(r)$

- aromatic carbon
- carboxyl oxygen
- charged nitrogen

Wednesday, May 2, 2012
On The Way
Minimization with a KB-Potential

<table>
<thead>
<tr>
<th>Set without native structures</th>
<th>Top1 (q = 1,2,3)</th>
<th>Top10 (q = 1,2)</th>
<th>Top1Q1*</th>
<th>Top10Q1*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Before minimization</strong></td>
<td>422 (52.88%)</td>
<td>502 (62.90%)</td>
<td>351 (77.31%)</td>
<td>417 (91.85%)</td>
</tr>
<tr>
<td><strong>After minimization</strong></td>
<td>652 (81.70%)</td>
<td>679 (85.09%)</td>
<td>611 (95.17%)</td>
<td>639 (99.53%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set without native and near-native structures</th>
<th>Top1 (q = 1,2,3)</th>
<th>Top10 (q = 1,2)</th>
<th>Top1Q1*</th>
<th>Top10Q1*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Before minimization</strong></td>
<td>248 (31.07%)</td>
<td>311 (38.97%)</td>
<td>171 (76.00%)</td>
<td>204 (90.67%)</td>
</tr>
<tr>
<td><strong>After minimization</strong></td>
<td>563 (70.55%)</td>
<td>593 (74.31%)</td>
<td>504 (95.64%)</td>
<td>525 (99.62%)</td>
</tr>
</tbody>
</table>
Flexibility

- Combination with internal-angle mechanics
- CG orientation-dependent potential
- QP optimization
The hybrid B3LYP density functional procedure was used, with the 6–311+G(d) basis set for CF₃X, CF₃Cl and CF₃Br, and DGDZVP for CF₃I. In the molecules CF₃X, each atom X is involved in a C–X bonding orbital, σₓCX, and also has three unshared pairs of electrons, two of which are in p-orbitals perpendicular to the C–X axis. The third unshared pair is in what is largely an s-orbital, but with some degree of p-hybridization along the C–X axis; Table 1 shows that the extent of this is much greater when X = F (24.9%) than when X is Cl, Br or I (12.0 to 8.4%). The lower level of sp-hybridization for the latter three means that their unshared electron pairs approximate the s²p⁰x⁰p⁰y distribution of a Cl⁺ ion constrained to have an empty p-orbital. The contribution of each X to the bonding σₓCX is primarily a p-orbital, although the situation is again somewhat different when X = F, for which there is 25% s-character (more than double what is found for Cl, Br and I). Also notable is that only for X = F (by far the most electronegative atom) is there a considerable shifting of the bonding electrons toward X; its share is 71.4%, compared to approximately 50% for Cl, Br and I. The σ-holes seen in Figs. 2, 3 and 4, and in other molecules, simply reflect the fact that the positive Vₛ(r) that totally encompasses the free, ground-state spherically symmetric atom X has not been countered in that region by an influx of electronic charge. The atom X in these molecules is similar to a single atom X with an s²p²ₓp²ᵧp¹z electronic configuration (where the C–X bond lies along the z-axis), in which two p-orbitals are filled and one is half-filled. An example is shown in Fig. 5; a σ-hole is clearly visible, surrounded by a belt of negative potential. Why is a σ-hole not found when X = F, as well as in some other instances, e.g. CH₃Cl? Firstly, the higher electronegativity of fluorine gives it a disproportionately large share of the σₓCX bonding electrons, which helps to neutralize the σ-hole. This also applies as chlorine in CH₃Cl, which does not have a σ-hole and does not halogen...
Protein-Ligand Interactions

- Pairwise-additive KB function
- ~ 50 atom types
- QP optimization
- Bachelor project of a MIPT student
- Currently tested
Open Problems

A, 6 conformations

- We have A with $N_1$ conformations and B with $N_2$ conformations
- All states of A and all states of B are accessible
- Then, partition function is given by
  \[ Z = \sum_k \exp^{-x_k \circ w} \]
- And Helmholtz free energy is
  \[ F = -\log Z = -\log \sum_k \exp^{-x_k \circ w} \]
- So, optimization problem is:
  \[ \log \sum_k \exp^{-x_k \circ w} > \log \sum_k \exp^{-x'_k \circ w} \]