Lead Finder in CSAR scoring challenge

Oleg Stroganov\textsuperscript{1}, Fedor Novikov\textsuperscript{1}, Viktor Stroylov\textsuperscript{1}, Val Kulkov\textsuperscript{2} and Ghermes Chilov\textsuperscript{1}

\textsuperscript{1} MolTech Ltd, Russian Federation, \textsuperscript{2} BioMolTech Corp, Canada

ACS Fall Meeting, Boston, USA

23 August 2010
Lead Finder in CSAR scoring challenge

\[ R^2 = 0.58 \]

\[ \text{RMSD} = 1.98 \text{ kcal/mol} \]
Scoring performance in CSAR challenge
Outline of the presentation

• Basic ingredients
  – Van der Waals and solvation
  – Electrostatics
  – Hydrogen bonds

• Magic ingredients

• Where do we go from here?
The Forcefield scoring functions (in Lead Finder)

AMBER, OPLS, CHARMM etc.

- van der Waals energy,
- Electrostatics,
- Hydrogen bonds,
- Dihedrals energy,
- Bonds, Angles
- Solvation

+ Magic ingredient

= best scoring function ever!
How to brew a scoring function: step 2

\[ \Delta G = k_{VdW} E_{VdW} + k_{Elec,i} E_{Elec,i} + k_{Hbondsi} E_{Hbondsi} + \ldots \]

1 for Van der Waals energy,
4 for electrostatics,
5 for hydrogen bonds,
1 for interaction with metals,
5 for Solvation,
4 for internal energy

20 coefficients

Training set: 230 structures (blue dots)
Test set: 100 structures (red dots)

RMSD of \( \Delta G = 1.75 \text{ kcal/mol} \)
Why do we need van der Waals energy?

- VdW-guided global search (docking)
- Optimization of given ligand poses
- Energy of the contact between ligand and protein

1e1v

*RMSD = 1.78 Å*

*Human CDK2 with inhibitor*
Solvation free energy

\[ \Delta G_{\text{solvation}} = \sum k_i \cdot S_{\text{contact},i} \]

<table>
<thead>
<tr>
<th></th>
<th>Polar</th>
<th>Non-polar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>-0.25</td>
<td>-0.40</td>
</tr>
<tr>
<td>Solvent</td>
<td>0.30</td>
<td>-0.01</td>
</tr>
</tbody>
</table>
Solvation and VdW energy are interchangeable

\[ k_{VdW} = 1, \quad k_{Sol} = 1 \]

\[ k_{VdW} = 0.5, \quad k_{Sol} = 1.7 \]

RMSD = 1.98 kcal/mol
\[ R^2 = 0.580 \]

RMSD = 2.06 kcal/mol
\[ R^2 = 0.57 \]
Solvation works!

Tyrosine protein phosphatase type 1

\[ \Delta G_{\text{calc}} = \Delta G_{\text{exp}} + E_{\text{sol}} + E_{\text{VdW}} \]

$kcal/mol$

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>0.68</td>
</tr>
<tr>
<td>Solvation + VdW</td>
<td>0.73</td>
</tr>
<tr>
<td>N(heavy)</td>
<td>0.80</td>
</tr>
<tr>
<td>N(all atoms)</td>
<td>0.87</td>
</tr>
</tbody>
</table>

\[ \begin{align*}
\Delta G_{\text{exp}} & = -5.3 \quad -8.7 \quad -10.3 \\
E_{\text{sol}} + E_{\text{VdW}} & = -6.7 \quad -10.1 \quad -11.0
\end{align*} \]
Solvation explains almost everything?

\[ E_{\text{VdW}} + E_{\text{Sol}} \]

\[ \Delta G_{\text{experiment}} \]

\[ R^2 = 0.47 \]

\[ R^2_{\text{Nheavy}} = 0.26 \]
Electrostatics pitfalls

• Long range interactions
  – Slowness of interaction energy decrease
  – Dependence of dielectric permittivity on (micro)environment

• Short range interactions
  – Calculations of atomic charges on ligand and protein
  – Polarization of interacting atoms
  – Competition between electrostatics and explicit interactions (h-bonds)

• Common pitfalls
  – Sampling of spatial distribution of charges
  – Sampling of ionization states of protein and ligand
Electrostatics in Lead Finder

Neuraminidase: surface contact

PPAR: buried contact
Electrostatics doesn’t always work...

\[ \Delta G_{\text{exp}} = -5.9 \text{ kcal/mol} \]
\[ \Delta G_{\text{calc}} = -3.7 \text{ kcal/mol} \]

\[ \Delta G_{\text{exp}} = -6.4 \text{ kcal/mol} \]
\[ \Delta G_{\text{calc}} = -8.7 \text{ kcal/mol} \]
H-bonds penalties and rewards

\[ \Delta G_{\text{HB}} = \Delta G_{\text{HB,complex}} - \Delta G_{\text{HB,solution}} \]

\[ E_{\text{H-bond}} = E_0 \left( r_{\text{HA}} \right) \cdot k_{\text{DHA}} \cdot k_{\text{LP}} \]

For the most cases \( \Delta N_{\text{hbonds}} \approx 0 \)

\[ E_{\text{penalty}} = k \cdot N_{\text{lost,ligand}} \]

\[ E_{\text{penalty}} = k \cdot N_{\text{lost,protein}} \]

H-bonds penalties serve to sieve out bad poses and poor binders.
H-bonds extra energy

\[ \Delta G_{\text{calc}} \]

\[ \Delta G_{\text{experiment}} \]

**R\(^2\) without extra H-bonds** = 0.47
**R\(^2\) with extra H-bonds** = 0.62

on CSAR subset of 48 structures, where systems of correlated H-bonds were found
Quest for new molecular interactions

• Thoroughly inspect complexes with discrepancies between experimental and calculated free energies

• Point out “interaction X”

• Estimate energy of the interaction

• Add interaction to the program, avoiding overfitting and false positives
Weak & rare interactions

• Weak hydrogen bonds
  – Aromatic rings as hydrogen bonds acceptors
  – Polarized C-H bond (Cα)
  – F as acceptor: CF ⋯ HX (O,N)

• Specific halogen interactions
  – Orthogonal multipolar interactions (C-X ⋯ C=O)
  – Interactions of halogens with nuclephils and electrophils

• Specific aromatic contacts
  – π-cationic interactions
  – Specific orientations

Weak hydrogen bonds

CSAR has 13 cases of weak H-bonds (Hα),
Average O-Hα distance is 2.15 Å

<ΔΔG> = -1.3 kcal/mol
## Halogen interactions

<table>
<thead>
<tr>
<th>Halogen</th>
<th>N structures</th>
<th>Error, kcal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>29</td>
<td>+0.6</td>
</tr>
<tr>
<td>Cl</td>
<td>21*</td>
<td>+0.1</td>
</tr>
<tr>
<td>Br</td>
<td>7**</td>
<td>+1.3</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>+2.6</td>
</tr>
</tbody>
</table>

* 10 of 21 structures with Cl are coagulation factor X with inhibitors. R² within this subset is 0.8

** 6 of 7 structures with Br are tyrosine protein phosphatase type 1 with inhibitors
Stacking and $\pi$-cationic interactions

$\Delta G_{\text{exp}} = -12.4 \text{ kcal/mol}$

$\Delta G_{\text{calc}} = -7.0 \text{ kcal/mol}$

1q0y
Anti-morphine antibody complexed with morphine
Are we missing something?

<table>
<thead>
<tr>
<th>PDB id</th>
<th>Error, kcal/mol</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1duv</td>
<td>5.2</td>
<td>parameterization</td>
</tr>
<tr>
<td>2c1q</td>
<td>2.3</td>
<td>biotin</td>
</tr>
<tr>
<td>2i0d</td>
<td>5.4</td>
<td>HIV-protease</td>
</tr>
<tr>
<td>2qi5</td>
<td>2.7</td>
<td>HIV-protease</td>
</tr>
<tr>
<td>2qi6</td>
<td>2.4</td>
<td>HIV-protease</td>
</tr>
<tr>
<td>2fv5</td>
<td>2.8</td>
<td>??</td>
</tr>
<tr>
<td>1swk</td>
<td>3.6</td>
<td>biotin</td>
</tr>
<tr>
<td>1y1m</td>
<td>-4.9</td>
<td>protein conformation?</td>
</tr>
<tr>
<td>1y1z</td>
<td>-3.0</td>
<td>protein conformation?</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Explicit water

26 of 28 HIV protease inhibitors from CSAR set interact with conservative water molecule
Loops and sidechains flexibility

1y1m
Glutamate receptor
TSAR – a new algorithm for multistate calculations

Thermodynamic Sampling of Amino acid Residues

• Represent interactions between residues as graph

• Invoke belief networks theory to reduce complexity of graph

• Find global minima using Dead-End Elimination technique

Or

• Calculate energy difference between states

\[
\Delta G = RT \ln \frac{\sum_{\text{ligand enabled}} e^{-E_i/RT}}{\sum_{\text{ligand disabled}} e^{-E_i/RT}}
\]
Future directions of mastering scoring

• Improvements of sampling
  – Thermodynamic integration over ligand and protein conformations
  – Sampling of flexible loops

• Explicit treatment of water
  – Conservative water molecules
  – Replaced by ligand
  – Water networks rearrangements energy evaluation