Scoring functions for
of protein-ligand docking:

New routes towards old goals

Christoph Sotriffer

Institute of Pharmacy and Food Chemistry
University of Würzburg
Am Hubland
D – 97074 Würzburg

Key questions in structure-based drug design

- Where is the binding site?
- What is the structure of the complex?
- What is the energy of interaction?
- What is a suitable, tight-binding ligand?

Required: some sort of affinity prediction
Why is affinity prediction a challenge?

1.) Protein-ligand complexes are dynamic systems in aqueous solution
   • huge number of particles
   • simultaneous, unperiodic, continuously changing interactions
     Simulation methods required!
   Statistical thermodynamics: Calculation of $\Delta G^*$ needs integration over entire phase space!
     Computationally very expensive!

2.) The prediction methods need to be fast
   Database screens: $\sim 10^3 - 10^8$ molecules need to be compared
   Docking runs: $\sim 10^7 - 10^9$ configurations need to be evaluated
     „Scoring functions“ required:
     Fast, simplified, heuristic methods for prediction of binding strength

Scoring functions: Goals

The ultimate goals of an ideal function:
• accurate within less than 1 pK_a unit ($<1.4$ kcal/mol)
• generally valid (not system specific; large affinity range)
• robust (tolerant with respect to small structural uncertainties)
• widely applicable (docking, virtual screening)
• physically meaningful (interpretable)
• fast and easy to compute
**Scoring functions: Tasks and types**

Application tasks:

A) Identification of the correct binding mode for a given ligand  
   *Pose prediction in docking*

B) Identification of new active ligands  
   *Virtual screening*

C) Affinity ranking for compound series  
   *Ligand design, lead optimization*

Available approaches:

- Force field-based methods
- Knowledge-based scoring functions
- Empirical scoring functions

---

**Force field-based methods**

Molecular Mechanics (MM):

- atoms → charged spheres
- bonds → springs
- classical potentials
- no electrons → no bond formation / cleavage
- typically parameterized to reproduce molecular potential energy surface (→ conformational ΔH in the gas phase!)

---

Scoring protein-ligand complexes:

- + for pose prediction in docking
- – for ligand ranking by affinity

Terms accounting for (de)solvation & entropic factors required (cf. MM-PBSA)
**Knowledge-based scoring functions**

Derivation from crystal-structure data

\[ P_{ij}(r) = -\ln \frac{g_{ij}(r)}{g_{ref}} \]

- \( P_{ij} \): distance-dependent pair potential
- \( g_{ij} \): frequency distribution of atom-atom contacts
- \( g_{ref} \): reference distribution

**Empirical scoring functions**

Regression-based:

\[ pKi = \sum pKi_n f_n(\text{structure}) \]

- affinity
- weighting factors
- structure descriptors

Determined via regression analysis (MLR, PLS)

Data:

- Experimental binding affinities
- Experimental structures
Where do we stand with scoring?

A not too unusual result

*after over 20 years of scoring function development …*

Correlation with affinity for a test set of 800 known complexes:

*in general,*

\[ r < 0.55 \quad (r^2 < 0.3) \]


→ A more detailed look at scoring function performance …

**Performance of scoring functions**

**A) Pose prediction in docking**

Identification of near-native binding pose among a set of geometric decoys

- Test set of 195 complexes of 65 different targets
- 100 low-energy poses per complex (0-10 Å rmsd)
- 29 scoring functions tested

- native poses can be detected fairly well
- success rates of up to ~80%
- knowledge-based approaches work best

[DSCS 85%]

• native poses can be detected fairly well
• success rates of up to ~80%
• knowledge-based approaches work best

Detection of active compounds in screening databases

Problem: Testing scoring function performance in virtual screening is not trivial!

• significant enrichment can be obtained
• not always for the right reasons
• no function performs consistently well

B) Virtual screening

Performance of scoring functions

C) Affinity prediction

Correlation of scores with experimental binding affinities

Test set compiled by Cheng et al., 2009: 195 PDBbind complexes

Pearson correlation coefficient $R_p$

With most functions:

- poor correlation for generic data sets
- hardly possible to obtain correct ranking
- of limited use for ligand optimization

Functions tested by Cheng et al. 2009
C) Affinity prediction

Correlation of scores with experimental binding affinities

CSAR-NRC HiQ evaluation set: 343 (332) complexes


Table 1. Parametric and Nonparametric Measures of Correlation Between the Scores and Experimental Binding Affinities

<table>
<thead>
<tr>
<th>method</th>
<th>Pearson ρ</th>
<th>Spearman ρ</th>
<th>Kendall τ</th>
<th>R²</th>
<th>p</th>
<th>RMSE</th>
<th>Med</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>code 1</td>
<td>0.76 (0.68–0.77)</td>
<td>0.74 (0.69–0.78)</td>
<td>0.51 (0.40–0.64)</td>
<td>0.18 (0.04–0.30)</td>
<td>0.13</td>
<td>1.49</td>
<td>1.30</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>code 2</td>
<td>code 3</td>
<td>code 4</td>
<td>code 5</td>
<td>code 6</td>
<td>code 7</td>
<td>code 8</td>
<td>code 9</td>
<td>code 10</td>
<td>code 11</td>
</tr>
<tr>
<td>code 12</td>
<td>code 13</td>
<td>code 14</td>
<td>code 15</td>
<td>code 16</td>
<td>code 17</td>
<td>focused (0.84–0.93)</td>
<td>0.25 (0.15–0.36)</td>
<td>0.11 (0.05–0.18)</td>
<td>1.09</td>
</tr>
<tr>
<td>YARFSCD (Mean and &quot;null&quot; Correlation)</td>
<td>0.67 (0.57–0.74)</td>
<td>0.14 (0.05–0.26)</td>
<td>0.26 (0.11–0.40)</td>
<td>0.02</td>
<td>0.95</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heavy atoms</td>
<td>0.11 (0.08–0.042)</td>
<td>0.49 (0.37–0.60)</td>
<td>0.41 (0.31–0.60)</td>
<td>0.26 (0.11–0.18)</td>
<td>1.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log P</td>
<td>0.46 (0.34–0.58)</td>
<td>0.30 (0.23–0.41)</td>
<td>0.34 (0.40–0.55)</td>
<td>0.22 (0.30–0.14)</td>
<td>1.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance across 17 core methods:
- R_p in the range 0.35 – 0.76 (only 3 >0.65)
- RMSE in the range 2.99 – 1.51 (pK_d units)
- correlation with heavy atom count: R_p 0.51

How to improve current scoring functions?

Empirical scoring functions

Regression-based: \( pKi = \Sigma pKi_i f_i(\text{structure}) \)

Data:

- training sets
- descriptors
- regression methods
The SFCscore approach

• Training sets: SFC: Scoring Function Consortium
  Data collection from public & industry sources
  up to 855 complexes with affinity data

• Descriptors:

• Regression method: MLR + PLS

Example: SFCscore function „sfc_290m“

\[ p\text{Ki} = - p\text{Ki}_1 \times \text{n_rot_bonds} \]
\[ + p\text{Ki}_2 \times \text{neutral_H_bonds} \]
\[ + p\text{Ki}_3 \times \text{metal_interaction} \]
\[ + p\text{Ki}_4 \times \text{AHPDI} \]
\[ + p\text{Ki}_5 \times \text{ring-ring_interaction} \]
\[ + p\text{Ki}_6 \times \text{ring-metal_interaction} \]
\[ + p\text{Ki}_7 \times \text{total_buried_surface} \]
\[ + p\text{Ki}_8 \]

Statistical parameters for training set (n = 290):

<table>
<thead>
<tr>
<th>R</th>
<th>R²</th>
<th>s</th>
<th>F</th>
<th>Q²</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.843</td>
<td>0.711</td>
<td>1.09</td>
<td>99.2</td>
<td>0.692</td>
<td>1.12</td>
</tr>
</tbody>
</table>

Sotriffer et al., Proteins 73 (2008), 395
Performance of SFCscore functions:
Cheng test set (195 complexes)

Pearson correlation coefficient $R_p$

Remaining limitations:
- data set issues ($IC_{50}$ etc.)
- implicit model assumptions (i.e., functional form of descriptors, linear regression techniques)

Overcoming the limitations

- Training sets:
  growth of PDBbind $\rightarrow$ 1105 complexes with $K_i$ data
  (not overlapping with Cheng test set)

- Regression methods:
  Non-parametric machine-learning methods:
  (not imposing any particular functional form)

  in particular: Random Forest
Random Forest

Decision Tree (or Recursive Partitioning)

Advantages:
- handles high-dimensional data well
- has ability to ignore irrelevant descriptors
- handles multiple mechanisms of action
- is amenable to model interpretation

Disadvantage:
- Relatively low prediction accuracy
  can be overcome by using ensembles of trees
  one ensemble method: Random Forest (RF)

Svetnik et al., JCICS 43 (2003), 1947

Random Forest

RF: outputs of all trees are aggregated to produce one final prediction

for classification:
class predicted by majority of trees

for regression:
average of the individual tree predictions

Training of a Random Forest:

1) Draw a random sample of the training data

2) For each sample, grow a tree to maximum size (no pruning) as follows:
   at each node choose the best split among a randomly selected subset of \( m_{\text{try}} \) descriptors

3) Repeat the above steps until a sufficiently large number of trees are grown

Svetnik et al., JCICS 43 (2003), 1947
Random Forest for scoring functions

First scoring function trained with Random Forest:

**RF-Score**  (Ballester & Mitchell, *Bioinformatics* 2010)

- Training set: 1105 PDBbind complexes
- Descriptors: count of protein-ligand atom type pair contacts within 12 Å
  - 9 atom types (C, N, O, S, P, F, Cl, Br, I) → 36 pairs
  - each complex characterised by vector of 36 contact counts

⇒ RF-Score yields much higher $R_p$ for Cheng test set!

**BUT:**  *Do the pure contact counts sufficiently well capture the physicochemical interaction features?*

Random Forest for scoring functions: SFCscore$^{RF}$

⇒ use SFCscore descriptors to train Random Forest model!

**SFCscore$^{RF}$**

- Training set: 1105 PDBbind complexes
- Descriptors: 63 SFCscore descriptors

**Test set (Cheng)**

$R_p = 0.787$  \ RMSE = 1.53

**Relative descriptor importance**

**Increase of the mean squared error** when randomly permuting the descriptor values
Performance comparison: Cheng test set (195 complexes)

Pearson correlation coefficient $R_p$

Performance on CSAR-NRC set

Complete CSAR-NRC (343 complexes)
- overlap: 100 complexes
  - $R_p = 0.80$
  - RMSE = 1.35

Reduced CSAR-NRC (243 complexes)
- no overlap
  - $R_p = 0.74$
  - RMSE = 1.53
Performance on CSAR-NRC set

- Complete CSAR-NRC (343 complexes)
  - overlap: 100 complexes
  - \( R_p = 0.80 \) \( \text{RMSE} = 1.35 \)

- Reduced CSAR-NRC (243 complexes)
  - no overlap
  - \( R_p = 0.74 \) \( \text{RMSE} = 1.53 \)

**Where are the limits?**

**Inherent experimental error**

limits the possible correlation between scores and measured affinity.

\[ R_p \text{ is limited to:} \]

\[ \sim 0.91 \]

when fitting to the data set

\[ \sim 0.83 \]

when scoring the data set with a method trained on outside data

(estimate based on error with \( \sigma = 1.0 \text{ log} K \))


---

**Fundamental limitations of scoring functions (I)**

- Accuracy of experimental data!
  - Structural data (mainly X-ray) of protein-ligand complexes
    - multiple conformations (highly dynamic systems)
    - hydrogen atom positions (protonation states) not observable
    - side-chain orientation may be ambiguous (Asn, Gln, His)
    - water molecules are only partially observable
    - binding modes may depend on crystallization conditions and crystal packing
  - Affinity data of protein-ligand complexes
    - depend highly on pH, buffer, salt concentration, temperature
    - enzyme kinetics: inhibition mechanism must be known
    - \( IC_{50} \leftrightarrow K_i \leftrightarrow K_d \)

Knowledge-based and empirical scoring methods cannot be better than the exp. data they are based on!
Leave-Cluster-Out (LCO) Validation: Target-dependent performance

**Limitations**

The TGT example - or: Limitations of scoring functions

- **pKᵢ = 5.08**
  - **Backbone flip**
  - Hardly accounted for by current scoring function!

- **pKᵢ = 4.08**
  - **Inclusion of water molecule**
Fundamental limitations of scoring functions (II)

- $\Delta G^0 = RT \ln K_D = \Delta H^0 - T \Delta S^0$

depending on the entire accessible phase space

difference between two states (bound/unbound)


yet scoring functions in general ...
... consider only the complexed state
... consider only a single (or very few) configurations
... attempt to provide $\Delta G^0$ also for arbitrary non-equilibrium states (poses)

„Dynamics – Water – Entropy“

Overall, the simplistic scoring functions work surprisingly well!!

Acknowledgement

David Zilian
Daniel Cappel
Michael Hein
Manuel Krug
Monika Nocker
Ulrich Peinz
Benjamin Schaefer
Johannes Schiebel
Martin Sippel
Constanze Waltenberger
Armin Welker

Hans Matter (Sanofi-Aventis)
Gerhard Klebe (Univ. of Marburg)
Paul Sanschagrin
Gerd Neudert

DFG (SFB 630, KFO 216)