Docking & Scoring

Calculation of binding affinities using simulation techniques (MD/MC):

Calculate full thermodynamic cycle (FEP) or endpoints only (MM/PBSA, LIE, SIE)

<table>
<thead>
<tr>
<th>ΔG_1</th>
<th>ΔG_2</th>
<th>ΔG_3</th>
<th>ΔG_4</th>
</tr>
</thead>
</table>

BUT: Sampling very time consuming

Limitations:

• For FEP: only structurally similar compounds

• What is the overall binding mode? Standard MD simulations only locally sample ligand-protein conformations.

• Very time consuming (1+ days per compound)
**Docking & Scoring**

**Aim:** Predict binding mode and affinities in reasonable amount of time

Library of options
100’000 – 1’000’000 possible compounds

Synthesis
100 – 1000 compounds

Biological testing
100 – 1000 compounds
Docking & Scoring

For a given protein:

Which ligand is able to bind?

How does the ligand bind?
→ Orientation, position, conformation

How strong does the ligand bind?
→ Binding affinity

Docking & Scoring

FEP, MM/PBSA, LIE, SIE
General procedure

Generate a large set of possible ligand-protein configurations using an efficient search algorithm (posing)

Compute protein-ligand interaction using an efficient scoring function

Score = -6.2
Score = -8.4
Score = -9.3
Score = -7.2

Predicted pose
Docking & Scoring
Search algorithms

Characteristics:

• Extensive sampling of different poses

• Limiting duplicate visits of similar/identical poses
Search algorithm

Degrees of flexibility:

- Rigid ligand & rigid protein
- Rigid ligand, but multiple pre-generated conformations (library) & rigid protein
- Flexible ligand & rigid protein
- Flexible ligand & flexible protein
Search algorithm

Flexible ligand: Degrees of freedom

- Translation
- Rotation
- Torsion rotation
Search algorithm

General categories:

• **Systematic search:**
  Each degree of freedom is systematically explored in intervals in a combinatorial fashion

• **Stochastic search:**
  Conformations are changed by random steps under certain rules

• **Deterministic search** (via simulation methods):
  New conformation is generated from previous conformation following certain rules (e.g. forces) without randomness
Systematic search algorithm

Example: **Conformational search method**

Systematic scan for all degrees of freedom

- Fine scan $\rightarrow$ combinatorial explosion
- Coarse scan $\rightarrow$ correct binding mode is possibly not identified

15 rotatable bonds (torsions):

- Scan each $10^\circ$ $\rightarrow$ $36^{15} = 2.2 \cdot 10^{23}$ conformations
- Scan each $30^\circ$ $\rightarrow$ $12^{15} = 1.5 \cdot 10^{16}$ conformations
- Scan each $120^\circ$ $\rightarrow$ $3^{15} = 1.4 \cdot 10^7$ conformations

Usually used to pre-generate conformational library (e.g. for FLOG, SLIDE):

Fast, because only internal energy has to be calculated
Systematic search algorithm

Example: **Fragmentation methods** (e.g. DOCK 4.0, FlexX)

**Incremental approach:**
1. Dock core fragment
2. Add flexible regions incrementally

**Place-and-join approach:**
1. Dock several fragments
2. Link fragments covalently

**Assumption:** Core fragment position in the intact ligand is among \( n \) lowest docked poses → repeat process with \( n \) docked poses for each growing fragment.
Stochastic search algorithm

Example: **Monte Carlo algorithms** (e.g. ICM)

1. Random changes in translational, rotational and torsional degrees of freedom
2. Usually: Minimization
3. Accept new conformation based on Metropolis criteria

\[
\exp \left( - \frac{\Delta \text{Score}}{k_B T} \right)
\]

\Rightarrow \text{always accept}

\Rightarrow \text{accept with probability}
Stochastic search algorithm

Example: (Lamarckian) Genetic algorithm methods (e.g. AutoDock, GOLD)

Translation | Rotation | Torsions | Translation | Rotation | Torsions
---|---|---|---|---|---
0.7 | 0.2 | -0.4 | -0.2 | 0.5 | -1.1 | 0.3 | 1.3 | 0.0

Selection: Sexual partners according to their fitness

Mutation | Crossover

3-D structure

Local optimization

Interaction energy ligand - receptor

-15 kcal/mol

-10 kcal/mol
Deterministic search methods

Example: **Molecular dynamics simulations**

**Problem**: MD simulations can not overcome large energy barriers → only local sampling

**Possible Solution**: MD at high temperature or reduce barriers

Scoring functions

Characteristics:

• **Accurate** inclusion of dominant protein-ligand interaction terms to assign the most favorable scores to the experimentally determined complex
  – Steric complimentarity
  – Physico-chemical complimentarity

• **Fast** to allow scoring of many different poses

⇒ Compromises necessary
Scoring functions

Number of non-bonded interactions e.g. for N=10000 atoms: $N(N-1)/2 \sim 5 \cdot 10^7$

$$V(r_1, \ldots, r_N) = \sum_{i<j} \frac{q_i q_j}{4 \pi \varepsilon_0 r_{ij}} + \sum_{i<j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right)$$

$$+ \sum_{bonds} \frac{1}{2} k^b_{ij} (r_{ij} - r^0_{ij})^2$$

$$+ \sum_{angles} \frac{1}{2} k^a_{ijk} (\theta_{ijk} - \theta^0_{ijk})^2$$

$$+ \sum_{dihedrals} \frac{1}{2} k^d_{ijkl} \left( 1 + \cos(n(\phi_{ijkl} - \phi^0_{ijkl})) \right)$$
A scoring function ranks the different ligand-protein configurations.

**Aims:**
1. Calculate binding affinity.
2. Enable distinction between true binding mode and all the other alternative modes explored during configurational search.
3. Efficient calculation of ligand-protein interactions to allow for extensive configurational search.

**General categories:**

- **Force-field based scoring functions** (e.g. AutoDock 3.0, GoldScore)

\[
\Delta G = \Delta G_{vdW} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + \Delta G_{HBond} \sum_{i,j} E(\theta) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + \Delta G_{el.st.} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}
\]

\[+ \Delta G_{Tors} N_{Tors} + \Delta G_{Solv} \sum_{i,j} \left( S_i V_j + S_j V_i \right) \exp \left( - \frac{r_{ij}^2}{2 \sigma^2} \right)\]

Stored on grid → Efficiency

**Issue:** Steep potential function ↔ protein flexibility
Scoring function

- **Empirical scoring functions** (e.g. LUDI, ChemScore)
  also called: **Regression-based scoring functions**:
  Optimize pre-factors ($\Delta G_0$, $\Delta G_{\text{HBond}}$, etc.) e.g. with multi-linear regression on dataset of experimental binding data (i.e. binding modes, affinity)

$$\Delta G = \Delta G_0 + \Delta G_{\text{HBond}} \sum_{\text{HBonds}} f(\Delta r, \Delta \theta) + \Delta G_{\text{ionic}} \sum_{\text{ionic contacts}} f(\Delta r, \Delta \theta)$$

$$+ \Delta G_{\text{hydrophob}} A_{\text{hydrophob}} + \Delta G_{\text{Tors}} N_{\text{Tors}}$$

Typical difference between empirical scoring functions:
pre-factors, explicit aromatic term, function of hydrophobic term

**Issues:**
- Scoring function (i.e. pre-factors) only as good as the experimental data set on which it was optimized (experimental uncertainty in $\Delta G$ !!!)
- Is data set representative?
- Additivity of $\Delta G$?
Scoring function

- **Knowledge-based scoring functions** (e.g. PMF, DrugScore)

Idea: - Represent ligand and protein by individual atom-types
  - Frequency of occurrence of individual contacts (atom-type j ligand – atom-type i protein) is a measure of their energetic contribution to binding

\[
\Delta G = \sum_{\text{pairs}} E_{i,j}(r)
\]

with

\[
E_{i,j}(r) = -k_B T \ln \left( \frac{\rho_{i,j}(r)}{\rho(r)} \right)
\]

\( \rho_{i,j}(r) \) : normalized probability of atom pair i,j to be in contact at distance r
\( \rho(r) \) : normalized reference probability

Issues:
- Is data set representative?
- Directionality of hydrogen bonds?
Docking & Scoring

References


Applications/Validation

Types of applications:

• **Prediction of binding modes**
  Generate large set of binding poses (ligand orientation and conformation) and use scoring function to select most likely pose (which should be similar to that seen crystallographically) \(\rightarrow\) base pose for lead optimization.
  
  Speed: -  Accuracy: +

• **Lead identification (Virtual screening)**
  Dock large database of compounds to target protein and rank compounds \(\rightarrow\) The early part of list of top ranked compounds should be enriched with active compounds.
  
  Speed: +  Accuracy: - (often different settings as in binding mode prediction)

• **Lead optimization (Prediction of binding affinity)**
  Score accurate binding mode to predict relative potencies among compounds.
  
  Speed: --  Accuracy: ++
Applications/Validation

**Issues:**

- How well do current docking algorithms perform with respect to these three tasks?

- What needs to be improved?
Prediction of binding modes

Percentage of successfully docked ligands (compared to X-ray):

RMSD < 2.0Å
RMSD < 4.0Å
Prediction of binding modes

Success rates (7 protein systems): RMSD < 2.0Å

<table>
<thead>
<tr>
<th></th>
<th>Any pose</th>
<th>Top pose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best docking/scoring method:</td>
<td>38 - 94%</td>
<td>8 - 74%</td>
</tr>
<tr>
<td>e.g. GLIDE</td>
<td>8 - 74%</td>
<td>0 - 55%</td>
</tr>
</tbody>
</table>

In general:

- Docking methods can generate poses close to X-ray data for about 40-50% of all protein-ligand systems. However, RMSD for best scored pose shows significant drop in success rates. Scoring functions are not always successful in separating X-ray conformation from decoy conformations.
- Best docking/scoring method depends strongly on protein system
- Success rate drops with
  - increasing ligand size (in particular >8 rotatable bonds)
  - increasing size of binding pocket
  - decreasing number of specific non-hydrophobic contacts
Virtual screening


A. Chk1 Kinase

G. Gyrase B

B. PPARγ

H. HCV Polymerase

= Percent predicted as actives (top ranked)
Virtual screening


**Enrichment factor:**

\[
EF = \frac{\text{% real actives found (identified in experiment)}}{\text{% predicted as actives (highest ranked)}}
\]

---

**Table 4. Enrichment Factor for Actives (≤ 1 μM) Found at 10% of the Docking-Score-Ordered List**

<table>
<thead>
<tr>
<th>program</th>
<th>Chk1</th>
<th>FXa</th>
<th>gyrase B</th>
<th>HCVP</th>
<th>MRS</th>
<th>E. coli PDF</th>
<th>Strep PDF</th>
<th>PPARδ</th>
</tr>
</thead>
<tbody>
<tr>
<td>ideal</td>
<td>10.0</td>
<td>9.8</td>
<td>10.0</td>
<td>9.5</td>
<td>10.0</td>
<td>7.6</td>
<td>8.3</td>
<td>8.6</td>
</tr>
<tr>
<td>Dock4</td>
<td>1.4</td>
<td>4.1</td>
<td>1.7</td>
<td>1.8</td>
<td>4.2</td>
<td>0.9</td>
<td>0.8</td>
<td>1.7</td>
</tr>
<tr>
<td>DockIt</td>
<td>4.2</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.2</td>
<td>0.0</td>
<td>3.2</td>
</tr>
<tr>
<td>FlexX</td>
<td>7.0</td>
<td>2.2</td>
<td>5.8</td>
<td>0.9</td>
<td>3.9</td>
<td>0.8</td>
<td>0.8</td>
<td>5.2</td>
</tr>
<tr>
<td>Flo+</td>
<td>5.6</td>
<td>2.7</td>
<td>2.3</td>
<td>3.4</td>
<td>1.7</td>
<td>1.5</td>
<td>0.8</td>
<td>3.6</td>
</tr>
<tr>
<td>Fred</td>
<td>2.9</td>
<td>4.1</td>
<td>1.9</td>
<td>2.0</td>
<td>0.6</td>
<td>3.2</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>Glide</strong></td>
<td><strong>6.3</strong></td>
<td><strong>3.4</strong></td>
<td><strong>1.0</strong></td>
<td><strong>1.0</strong></td>
<td><strong>5.3</strong></td>
<td><strong>0.6</strong></td>
<td><strong>0.4</strong></td>
<td><strong>4.8</strong></td>
</tr>
<tr>
<td>Gold</td>
<td>0.1</td>
<td>4.1</td>
<td>4.0</td>
<td>0.0</td>
<td>0.8</td>
<td>1.0</td>
<td>0.1</td>
<td>5.5</td>
</tr>
<tr>
<td>LigandFit</td>
<td>3.3</td>
<td>1.9</td>
<td>2.8</td>
<td>1.8</td>
<td>2.9</td>
<td>2.9</td>
<td>1.7</td>
<td>1.2</td>
</tr>
<tr>
<td>MOEDock</td>
<td>3.9</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>2.1</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>MVP</strong></td>
<td><strong>7.2</strong></td>
<td><strong>5.8</strong></td>
<td><strong>5.3</strong></td>
<td><strong>3.6</strong></td>
<td><strong>6.4</strong></td>
<td><strong>6.7</strong></td>
<td><strong>6.9</strong></td>
<td><strong>3.9</strong></td>
</tr>
</tbody>
</table>

For MVP: Manually defined pharmacophore points → MedChem knowledge can improve results significantly.
Virtual screening

In general:
- Docking methods yield enrichment above random for most protein systems
- Best docking/scoring method depends strongly on protein system
- Success rate drops with
  - increasing similarity of actives and decoys
Prediction of binding affinities

**Prediction of binding affinities (rank order):** G.L. Warren J. Med. Chem. 2006, 49, 5912

Weak correlation between predicted and experimental binding affinities ($r$ value ranges from 0 to -0.57 → $r^2$ from 0 to 0.33)

**Table 7.** Best Correlation Coefficient $r$ between the $-\log$ Affinity ($p$Affinity) and Docking Score for All Programs across All Targets

<table>
<thead>
<tr>
<th>program</th>
<th>Chk1</th>
<th>FXa</th>
<th>gyrase B</th>
<th>HCVP</th>
<th>MRS</th>
<th>E. coli PDF</th>
<th>Strep PDF</th>
<th>PPARδ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dock4</td>
<td>−0.33</td>
<td>−0.31</td>
<td>−0.39</td>
<td>0.00</td>
<td>−0.13</td>
<td>−0.38</td>
<td>−0.34</td>
<td>0.07</td>
</tr>
<tr>
<td>DockIt</td>
<td>−0.49</td>
<td>−0.19</td>
<td>−0.37</td>
<td>0.04</td>
<td>−0.28</td>
<td>−0.13</td>
<td>−0.30</td>
<td>−0.34</td>
</tr>
<tr>
<td>FlexX</td>
<td>−0.57</td>
<td>−0.31</td>
<td>−0.39</td>
<td>−0.12</td>
<td>−0.01</td>
<td>−0.42</td>
<td>−0.25</td>
<td>−0.36</td>
</tr>
<tr>
<td>Flo+</td>
<td>−0.44</td>
<td>−0.38</td>
<td>−0.36</td>
<td>−0.09</td>
<td>0.05</td>
<td>−0.27</td>
<td>−0.39</td>
<td>−0.42</td>
</tr>
<tr>
<td>Fred</td>
<td>−0.14</td>
<td>0.01</td>
<td>−0.13</td>
<td>−0.07</td>
<td>0.13</td>
<td>0.07</td>
<td>0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>Glide</td>
<td>−0.47</td>
<td>−0.08</td>
<td>−0.21</td>
<td>−0.04</td>
<td>0.08</td>
<td>−0.13</td>
<td>−0.12</td>
<td>−0.43</td>
</tr>
<tr>
<td>Gold</td>
<td>−0.42</td>
<td>−0.05</td>
<td>−0.14</td>
<td>−0.09</td>
<td>0.04</td>
<td>−0.13</td>
<td>−0.12</td>
<td>−0.11</td>
</tr>
<tr>
<td>LigandFit</td>
<td>−0.45</td>
<td>−0.13</td>
<td>−0.39</td>
<td>−0.06</td>
<td>−0.15</td>
<td>−0.21</td>
<td>−0.49</td>
<td>−0.10</td>
</tr>
<tr>
<td>MOEDock</td>
<td>−0.29</td>
<td>0.00</td>
<td>0.07</td>
<td>−0.01</td>
<td>−0.13</td>
<td>0.08</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>MVP</td>
<td>−0.26</td>
<td>0.10</td>
<td>−0.33</td>
<td>−0.01</td>
<td>−0.18</td>
<td>−0.17</td>
<td>−0.16</td>
<td>−0.18</td>
</tr>
</tbody>
</table>
Prediction of binding affinities

In general: Regression coefficient $r^2$ usually well below 0.5.

P. Ferrara *J. Med. Chem.* **2004**, *47*, 3032:

Exception: Serine and metallo-proteases

BUT: Correlation between log(MW) and affinity:

SerPr: 0.81
MetPr: 0.58
Prediction of binding affinities


Linear correlation between experimental binding affinity and number of heavy atoms in ligand (up to 15 heavy atoms):
Applications/Validation

Issues:
• How well do current docking algorithms perform with respect to these three tasks?

• What needs to be improved?
Future directions

• **Scoring function**
• Entropic contributions
• Solvation effects (direct interactions, desolvation)

• Protein flexibility (see next lecture)

• Quality/consistent binding data
Post-processing

Store best N poses (for M best compounds in enrichment experiments) from docking simulation

Re-rank poses (and compounds) using higher-accuracy scoring/free energy methods (e.g. MM/PBSA, MM/GBSA, LIE, SIE)

Refine poses (e.g. minimization) → protein flexibility

Generate ensemble of configurations (e.g. MD)
Take home message

Don’t believe every result from the docking simulation.

Docking might give you good ideas about binding poses and active lead compounds,

BUT always use your medicinal chemistry knowledge:

• Check if similar compounds are all predicted as actives or in the same binding mode [Still be careful: Algorithm might predict pose or activity wrong for all similar compounds]
• Test known compounds with known binding modes first → Estimate accuracy of selected system: Docking/Scoring method & Target protein
• Do experimental validation: X-ray, NMR, site-directed mutagenesis etc.
Further references:


Protein flexibility

In about half of all protein systems protein flexibility plays an important role in ligand binding → rigid-protein assumption results in deleterious effect on docking results.

Types of protein flexibility:
• side-chain motion
• small backbone motion ("breathing motion")
• loop flexibility
• large conformational changes
Protein flexibility

• Side-chain flexibility

Palmitoleic acid
N-palmitoylmethionine

P450 BM-3
Protein flexibility

- Side-chain flexibility
- Breathing motion
Protein flexibility

- Side-chain flexibility
- Breathing motion
- Loop flexibility
Protein flexibility

- Side-chain flexibility
- Breathing motion
- Loop flexibility
- Large conformational changes

Estrogen receptor
Computational approaches

Side-chain flexibility:

• Rotamer libraries
  
  Large set of PBD structures
  
  → bin statistically preferred rotamer states

During docking search for optimal rotamer states for each docking pose.

Example for list of rotamers [Lys: 4 rotatable bonds = torsions]:

<table>
<thead>
<tr>
<th>LYS</th>
<th>-177.1</th>
<th>177.0</th>
<th>-176.6</th>
<th>-67.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYS</td>
<td>179.6</td>
<td>-178.0</td>
<td>-75.8</td>
<td>70.6</td>
</tr>
<tr>
<td>LYS</td>
<td>-175.0</td>
<td>-177.3</td>
<td>-70.2</td>
<td>-176.8</td>
</tr>
<tr>
<td>LYS</td>
<td>-175.4</td>
<td>-178.7</td>
<td>-71.0</td>
<td>-65.1</td>
</tr>
<tr>
<td>LYS</td>
<td>-162.4</td>
<td>-81.9</td>
<td>82.2</td>
<td>74.2</td>
</tr>
<tr>
<td>LYS</td>
<td>-163.0</td>
<td>-83.1</td>
<td>95.7</td>
<td>-178.9</td>
</tr>
<tr>
<td>LYS</td>
<td>-164.5</td>
<td>-82.8</td>
<td>88.3</td>
<td>-80.2</td>
</tr>
<tr>
<td>LYS</td>
<td>-161.2</td>
<td>-77.8</td>
<td>-177.5</td>
<td>66.4</td>
</tr>
<tr>
<td>LYS</td>
<td>-165.2</td>
<td>-87.2</td>
<td>-177.2</td>
<td>-178.9</td>
</tr>
<tr>
<td>LYS</td>
<td>-166.1</td>
<td>-83.0</td>
<td>-177.8</td>
<td>-67.3</td>
</tr>
<tr>
<td>LYS</td>
<td>-160.1</td>
<td>-90.5</td>
<td>-80.9</td>
<td>68.3</td>
</tr>
<tr>
<td>LYS</td>
<td>-166.8</td>
<td>-96.6</td>
<td>-69.4</td>
<td>178.4</td>
</tr>
<tr>
<td>LYS</td>
<td>-168.0</td>
<td>-93.8</td>
<td>-76.0</td>
<td>-65.7</td>
</tr>
<tr>
<td>LYS</td>
<td>-88.0</td>
<td>81.2</td>
<td>68.0</td>
<td>69.3</td>
</tr>
<tr>
<td>LYS</td>
<td>-80.9</td>
<td>81.3</td>
<td>66.9</td>
<td>-179.6</td>
</tr>
<tr>
<td>LYS</td>
<td>-86.8</td>
<td>86.0</td>
<td>84.3</td>
<td>-83.2</td>
</tr>
<tr>
<td>LYS</td>
<td>-80.7</td>
<td>75.7</td>
<td>176.2</td>
<td>62.3</td>
</tr>
<tr>
<td>LYS</td>
<td>-81.0</td>
<td>74.0</td>
<td>-178.3</td>
<td>177.9</td>
</tr>
<tr>
<td>LYS</td>
<td>-87.6</td>
<td>74.5</td>
<td>178.0</td>
<td>-61.8</td>
</tr>
<tr>
<td>LYS</td>
<td>-86.6</td>
<td>83.2</td>
<td>-85.6</td>
<td>68.2</td>
</tr>
<tr>
<td>LYS</td>
<td>-82.6</td>
<td>71.2</td>
<td>-96.3</td>
<td>-172.0</td>
</tr>
</tbody>
</table>
Comparison of rigid protein docking versus docking with flexible sidechains

Table 3. Ligand-free structures and their corresponding ligand-bound complexes used in testing the minimal rotation hypothesis

<table>
<thead>
<tr>
<th>No.</th>
<th>PDB code</th>
<th>Protein/ligand complex</th>
<th>Resolution (Å)</th>
<th>Best ligand RMSD (Å)</th>
<th>Template size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Free Bound</td>
<td>Flexible docking</td>
<td>Rigid docking</td>
</tr>
<tr>
<td>1M</td>
<td>1ahe</td>
<td>Alpha-momorcharin/Formycin 5'-monophosphate</td>
<td>2.0 2.2</td>
<td>0.94</td>
<td>—</td>
</tr>
<tr>
<td>1M</td>
<td>1ajb</td>
<td>Dihydropteroate synthase/Dihydropterine-diphosphate</td>
<td>2.0 2.0</td>
<td>0.75</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>3cox</td>
<td>Cholesterol oxidase/3-beta-hydroxy-5-androsten-17-one</td>
<td>1.8 1.8</td>
<td>1.61</td>
<td>1.63</td>
</tr>
<tr>
<td>4</td>
<td>1gmr</td>
<td>RNAse SA/Guanosine-2'-monophosphate</td>
<td>1.8 1.8</td>
<td>1.28</td>
<td>1.87</td>
</tr>
<tr>
<td>5</td>
<td>1gra</td>
<td>Glutathione reductase/Glutathione disulfide</td>
<td>1.5 2.0</td>
<td>0.69</td>
<td>1.88</td>
</tr>
<tr>
<td>6</td>
<td>1kel</td>
<td>Catalytic antibody 28B4 FAB Fragment/AAH</td>
<td>2.2 1.9</td>
<td>0.46</td>
<td>—</td>
</tr>
<tr>
<td>7M</td>
<td>2hvm</td>
<td>Hevamine/endochitinase/N-Acetyl-d-allosamine</td>
<td>1.8 1.9</td>
<td>0.67</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>1nsb</td>
<td>Neuraminidase/N-acetyl neuraminic acid(sialic acid)</td>
<td>2.2 1.7</td>
<td>0.40</td>
<td>0.67</td>
</tr>
<tr>
<td>9</td>
<td>1swa</td>
<td>Streptavidin/Biotin</td>
<td>2.0 1.9</td>
<td>0.62</td>
<td>0.67</td>
</tr>
<tr>
<td>10M</td>
<td>1tp5</td>
<td>Trypsin/Inhibitor A90720A</td>
<td>1.5 1.9</td>
<td>0.93</td>
<td>—</td>
</tr>
<tr>
<td>11</td>
<td>1xid</td>
<td>D-Xylose isomerase/L-Ascorbic acid</td>
<td>1.6 1.7</td>
<td>2.28</td>
<td>—</td>
</tr>
<tr>
<td>12</td>
<td>1ycd</td>
<td>Carbonic anhydrase II/Azetazolamide</td>
<td>2.0 1.9</td>
<td>1.42</td>
<td>—</td>
</tr>
<tr>
<td>13</td>
<td>2cht</td>
<td>Chorismate mutase/Endo-oxabicyclic inhibitor</td>
<td>1.9 2.2</td>
<td>1.02</td>
<td>—</td>
</tr>
<tr>
<td>14</td>
<td>3apr</td>
<td>Acid proteinase/Reduced peptide inhibitor</td>
<td>1.8 1.8</td>
<td>0.54</td>
<td>—</td>
</tr>
<tr>
<td>15</td>
<td>3tmn</td>
<td>Thermolysin/VAL-TRP</td>
<td>2.0 1.7</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>16</td>
<td>5e4a</td>
<td>Concanavalin A/alpha-methyl-D-mannopyranoside</td>
<td>2.0 2.0</td>
<td>1.99</td>
<td>—</td>
</tr>
<tr>
<td>17</td>
<td>5ga</td>
<td>Proteinase A/Tetrapeptide ACE-PRO-ALA-PRO-TYR</td>
<td>1.5 1.8</td>
<td>0.59</td>
<td>1.90</td>
</tr>
<tr>
<td>18M</td>
<td>7taa</td>
<td>Fam. 13 alpha amyrase/Modified acarbose hexasaccharide</td>
<td>2.1 2.0</td>
<td>0.82</td>
<td>—</td>
</tr>
</tbody>
</table>

The template sizes are given as the number of template points to indicate the differences in the sizes of the binding sites of the proteins.

a AAH = 1-[N-4'-nitrobenzyl-N-4'-carboxybutylaminomethylphosphononic acid.

M Proteins existing as biological monomers both in their ligand-free and ligand-bound states.
Small backbone and side-chain motions

- **Ensemble docking:**
  Dock ligand into each member of ensemble of protein structures individually
  Source: X-ray, NMR, MD/MC simulations

- **Soft docking:**
  Reduce repulsive van der Waals interactions

\[
V(r) = \frac{a}{r^9} - \frac{b}{r^6}
\]

\[
V(r) = \frac{A}{r^{12}} - \frac{B}{r^6}
\]
Small backbone and side-chain motions


**Ensemble docking versus soft docking versus rigid protein docking:**

Enrichment for two target systems:
(red: rigid protein  
blue: soft docking  
green: ensemble of X-ray structures)

**Details for ensemble docking:**
Overlay several structures of same protein with different co-crystallized ligands.  
Identify flexible regions. Overall template:  
Rigid core of protein plus flexible parts in alternative conformations.
Small backbone and side-chain motions

**Ensemble docking:**
If not multiple X-ray protein structures are available → run MD/MC simulations to generate ensemble of target structures.

**Simulation on apo structure:**
Often the protein has to accommodate its shape (and properties) to the bound ligand. Apo form might be in different conformation than necessary to form energetically favorable complex with ligands → Docking to trajectory of halo form is often (but not always) more successful.

**Simulation on a single ligand-bound structure:**
Docking into the ligand bound trajectory might bias docking towards the ligand used in the MD simulation → Docking of structurally diverse compounds might give worse results than apo form.

“Holo form to specialized for specific class of ligands”
Small backbone and side-chain motions

Ensemble docking:
MD simulation in frequently changing ligand model (Xu, Lill
Loop flexibility

Loop regions are usually significantly less stable than $\alpha$-helices or $\beta$-sheets → regularly observed conformational changes for different ligands bound. e.g. kinases:

Possible approach:
Predict alternative loop conformations (e.g. with PLOP, LOOPY) → use alternative templates for docking (similar to ensemble docking)

Large scale conformational changes

Possible approaches:

• Identify hinge regions in protein
  Use reaction coordinates for hinge-bending motion as additional degrees of freedom

• Run MD simulation
  Do principal component analysis (PCA) \(\rightarrow\) essential modes of protein dynamics (= reaction coordinates for main global conformational changes)
  Use essential modes as additional degrees of freedom in docking
References


Virtual screening

- Library of compounds
- Target

Virtual Screening
- Similarity based screening
- Pharmacophore search
- Docking

Hit

Lead

Drug
Combinatorial chemistry

1990s: “Invention” of combinatorial chemistry
→ Synthesis of large combinatorial libraries possible

Problem:
Number of possible compounds >>
Number of compounds that could be synthesized in large enough quantities or that could be screened

Which compounds to screen? → Library design
1. Properties of ligands

Initial HTS experiments did not lead to the expected increase in new lead compounds.

Reason: “General organic molecules” are not “drug-like”
Properties of libraries

Solution:
Select compounds with drug-like properties

a, Remove compounds with reactive groups (→ unspecific binding); important for experimental HTS

b, Remove compounds with poor absorption (“Rule-of-five”)

c, More sophisticated descriptor-based characterization of drug-likeness of compounds


OR better screening lead-like compounds, because during optimization phase of lead compound molecule usually becomes more complex (H-bond donors/acceptors, logP, MW)
Examples of libraries

**ZINC database**

http://zinc.docking.org

Total number of compounds: 21,603,031

Drug-like (logP < 5, MW < 500, MW >150,
    #donors <= 5, #acceptors <= 10): 9,497,542

Lead-like (logP < 3.5, MW < 350,
    #rotatable bonds <= 7): 1,949,828

Still a large number!!!

Theoretical estimates:
Possible number of drug-like compounds $\sim 10^{20} - 10^{60}$
→ Selection of compounds necessary
Properties of libraries

2. “Diverse” vs. “focused” libraries

Diverse library \(\rightarrow\) Validation difficult (true positive?)

Focused library \(\rightarrow\) Validation using the similarity principle possible, i.e. similar compounds should also be identified as active; however, smaller chemical space

\(\rightarrow\) Find balance between diversity and focus

In general: The more information about target, the more focused the library.

- Binding site known \(\rightarrow\) size of ligands \(\leq\) size of binding pocket
- Actives known \(\rightarrow\) ligands similar properties (2D descriptors, pharmacophores etc.)
- More frequently: Specific libraries for target classes (e.g. kinases) or disease states (e.g. cancer)
Properties of libraries

How to select diverse set of compounds?
1. Cluster analysis
2. Dissimilarity-based selection method
3. Cell-based methods
Cluster analysis

Aim: Divide compounds in clusters such that ligand in a cluster are similar but ligands in different clusters are dissimilar
Cluster analysis

General process:
1. Generate descriptors for each ligand
2. Calculate dissimilarity between all ligands based on descriptors
3. Use clustering algorithm to group ligands within data set
4. Select one (or more) representative ligand(s) for each cluster
Cluster analysis

Example: Hierarchical clustering
Iterative merging of most similar clusters until criterion is fulfilled (e.g. total number of clusters, minimum dissimilarity value, etc.); initially each cluster contains exactly one compound.
Dissimilarity-based selection method

Aim: Identify set of dissimilar compounds directly

Algorithm:
1. Select initial compound (e.g. random, largest/smallest sum of similarities to all other ligands) → subset D
2. Calculate dissimilarity between all remaining ligands and all ligands in D based on descriptors
3. Choose compound $i$ that is most dissimilar to D
   a. MaxSum: $\sum_{j \in D} d_{i,j}$
   b. MaxMin: $\min(d_{i,j} | j \in D)$
4. Repeat 2-3 until stop criterion is reached (see clustering algorithm)
Dissimilarity-based selection method

Example:

[Diagram showing a plot with principal components PC1 and PC2. The points represent compounds in D, with numbers 1, 2, 3, 4, and 5 highlighted.]
Cell-based method

Aim: Identify set of dissimilar compounds without calculating pair-wise similarity values

Algorithm:
1. Divide compounds into bins for each property (or principal component)
2. Choose one (or more) representative(s) for each cell in n-dimensional space (n: number of properties/principal components)
Cell-based method

Example: