Computational Drug Discovery

David Ryan Koes

7/9/2018

http://bits.csb.pitt.edu
What is a drug?

According to the Food, Drug, and Cosmetic Act (1) : a substance recognized in an official pharmacopoeia or formulary (2) : a substance intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease (3) : a substance other than food intended to affect the structure or function of the body (4) : a substance intended for use as a component of a medicine but not a device or a component, part, or accessory of a device

http://www.merriam-webster.com/dictionary/drug

A small molecule intended to affect the structure/function of macromolecules
COX-1
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)
1. Does the compound do what you want it to?

2. Does the compound **not** do what you **don’t** want it to?

3. Is what you want it to do the right thing?
Drug Discovery

Omics

Target Identification

Screening

Lead Identification

Lead Optimization

Compounds

Hits

Leads

Clinical Candidates

Cost
Computational Drug Discovery

Omics
Target Identification
Compounds
Virtual
Screening
Hits
Lead Identification
Leads
Cost
Clinical Candidates
Modeling
Lead Optimization
Leading
7
Kinds of Virtual Screening

ADMET

Ligand Based
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based
- dock and score
- simulation
ADMET

Absorption
Distribution
Metabolism
Excretion
Toxicity

Will this be a usable drug?

Screening for ADMET:
Cytochrome P450 interaction
Lipinski’s Rule of Five
QSPR: Quantitative Structure Property Relationship
Kinds of Virtual Screening

ADMET

Ligand Based
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based
- dock and score
Ligand Based: Similarity

Fingerprint Methods

- map molecules to a descriptor space:
  
  1D: molecule weight, #h-bonds, etc.
  2D: paths, bond distances between atom-pairs

- similarity is “distance” between descriptors
- for bit vectors, Tanimoto distance used

\[
T(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]
Topological Fingerprints

Daylight/FP2 Fingerprints

- all paths up to 7 bonds long
- each path corresponds to bit position (**hashing**)
- fast similarity checking (Tanimoto)
Topological Fingerprints

ECFP4
- all substructures with diameter 4 around every atom
Ligand Based: Similarity

Superposition Methods

- compute “overlap” between molecules
- consider shape, electrostatics, pharmacophores

http://www.cresset-group.com/
Ligand Based: QSAR

Quantitative Structure/Activity Relationships

<table>
<thead>
<tr>
<th>Compd Number</th>
<th>Cmpd Name</th>
<th>X</th>
<th>Log EC\textsubscript{50}</th>
<th>n</th>
<th>Calculated Log EC\textsubscript{50}</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6a</td>
<td>H</td>
<td>1.07</td>
<td>0</td>
<td>0.79</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>6b</td>
<td>Cl</td>
<td>0.09</td>
<td>0.71</td>
<td>0.21</td>
<td>-0.12</td>
</tr>
<tr>
<td>3</td>
<td>6d</td>
<td>NO\textsubscript{2}</td>
<td>0.66</td>
<td>-0.28</td>
<td>1.02</td>
<td>-0.36</td>
</tr>
<tr>
<td>4</td>
<td>6e</td>
<td>CN</td>
<td>1.42</td>
<td>-0.57</td>
<td>1.26</td>
<td>0.16</td>
</tr>
<tr>
<td>5</td>
<td>6f</td>
<td>C\textsubscript{6}H\textsubscript{5}</td>
<td>-0.62</td>
<td>1.96</td>
<td>-0.81</td>
<td>0.19</td>
</tr>
<tr>
<td>6</td>
<td>6g</td>
<td>N(CH\textsubscript{3})\textsubscript{2}</td>
<td>0.64</td>
<td>0.18</td>
<td>0.65</td>
<td>-0.01</td>
</tr>
<tr>
<td>7</td>
<td>6h</td>
<td>I</td>
<td>-0.46</td>
<td>1.12</td>
<td>-0.12</td>
<td>-0.34</td>
</tr>
</tbody>
</table>

Properties

Biological Activity = Learned linear function of properties

3D-QSAR: includes geometric/structural properties
**Ligand Based: QSAR**

Quantitative Structure/Activity Relationships

\[
f(\vec{x}) = w_1 \vec{x}_1 + w_2 \vec{x}_2 + w_3 \vec{x}_3 + \ldots + b
\]

What's a QSAR Model?

- Regression model used to estimate relationships between variables.
- Model identifies relationship between 2D chemical structures and bioactivity.

Karla Robles, Jeffrey Madura, David Koes

University of Pittsburgh, Duquesne University
Ligand/Receptor Based: Pharmacophore

Pharmacophore:
IUPAC: The ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response.

Common Features:
aromatic ring
hydrophobic area
positive ionizable
negative ionizable
hydrogen bond donor
hydrogen bond acceptor
Pharmacophore Features

- Hydrogen Donor
- Hydrogen Acceptor
- Positive
- Hydrophobic
- Negative
- Aromatic
Charge-Charge

\[ \vec{F}_{Q-q} \]

\[ \vec{F}_{q-Q} \]

\[ +q \]

\[ +Q \]

\[ +q \]

\[ -Q \]

\[ r \]

\[ |\vec{F}_{Q-q}| = |\vec{F}_{q-Q}| = k \frac{|q| \times |Q|}{r^2} \]
Charge-Charge

2QWK.pdb

Salt Bridge

Inhibitor of the influenza virus neuraminidase (antiviral agent)
Hydrogen Bond
Hydrogen Bond

Distance:
D-A: 2.5Å – 3.5Å (4.0Å?)
H-A: 1.5Å – 2.5Å

Angle:
Depends on context

Turkey Ovomucoid Inhibitor

1HJA.pdb
Hydrophobic
Hydrophobic

MDM2 (over expressed in >50% of cancers) down-regulates p53 (guardian of the genome)
Aromatic
Aromatic

Rings offset
Interplanar distance: 3.3-3.8Å

Human liver glycogen phosphorylase a complexed with caffeine
Kinds of Virtual Screening

ADMET

Ligand Based
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based
- dock and score
Pharmacophores Aren’t Enough

.2μM

50μM

n.i.
Docking

Determine the **conformation** and **pose** of a ligand at a docking site

Challenge is to find conformation and pose with the best **score**
Two Phase Docking

1. Global Pose Estimation

- Stochastic
- Minimization
Scoring Goals

Affinity Prediction
- how well does it bind?

Inactive/Active Discrimination
- does it bind?

Pose Prediction
- how does it bind?

Speed

Approximations:
Rigid or semi-rigid receptor
Implicit water model
Scoring Types

Force-field based
- inter- and intra- molecular forces
  - van der Waals, electrostatic, torsional

Empirical
- parameterized function is fit to binding energy data

Knowledge based
- scoring function based on data, not physical principles

Consensus
- combine multiple scoring functions
Force Field: Dock 4.0

\[ E = \sum_{i=1}^{\text{lig}} \sum_{j=1}^{\text{rec}} \left( \frac{A_{ij}}{r_{ij}^a} - \frac{B_{ij}}{r_{ij}^b} + 332 \frac{q_i q_j}{D r_{ij}} \right) \]

van der Waals
a = 12, b = 6
Lennard-Jones potential

Coulomb’s Law
q: partial charges
D: dielectric constant
Empirical: AutoDock Vina

\[ \text{gauss}_1(d) = w_{\text{guass}_1} e^{-(d/0.5)^2} \]
\[ \text{gauss}_2(d) = w_{\text{guass}_2} e^{-(d-3)^2} \]
\[ \text{repulsion}(d) = \begin{cases} w_{\text{repulsion}} d^2 & d < 0 \\ 0 & d \geq 0 \end{cases} \]

\[ \text{hydrophobic}(d) = \begin{cases} w_{\text{hydrophobic}} & d < 0.5 \\ 0 & d > 1.5 \\ w_{\text{hydrophobic}} (1.5-d) & \text{otherwise} \end{cases} \]

\[ \text{hbond}(d) = \begin{cases} w_{\text{hbond}} & d < -0.7 \\ 0 & d > 0 \\ w_{\text{hbond}} (-\frac{10}{7} d) & \text{otherwise} \end{cases} \]
Knowledge Based: RF-Score

Pairwise Distance Counts (<12Å)

Random Forest

Protein

<table>
<thead>
<tr>
<th>C</th>
<th>N</th>
<th>O</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cl</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Br</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ligand
Can we do better?

Accurate pose prediction, binding discrimination, and affinity prediction without sacrificing performance?

Key Idea: Leverage “big data”

- 231,655,275 bioactivities in PubChem
- 125,526 structures in the PDB
- 16,179 annotated complexes in PDBbind
Machine Learning

Features $X \rightarrow$ Model $\rightarrow y$ Prediction
Neural Networks

\[ \text{output} = \sigma \left( \sum_i w_i x_i + b \right) \]
The universal approximation theorem states that, under reasonable assumptions, a feedforward neural network with a finite number of nodes can approximate any continuous function to within a given error over a bounded input domain.
Deep Learning

\[
\delta^l = \left((\omega^{l+1})^T \delta^{l+1}\right) \odot \sigma'(z^l)
\]

\[
\frac{\partial L}{\partial \omega^l_{jk}} = \alpha^{l-1}_k \delta^l_j \quad \text{and} \quad \frac{\partial L}{\partial b^l_j} = \delta^l_j
\]
Image Recognition

Image Recognition

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

ILSVRC top-5 error on ImageNet

Convolutional Neural Networks

https://devblogs.nvidia.com
Convolutional Neural Networks

- Convolution
- Feature Maps
- Fully Connected
- Traditional NN

Dog: 0.99
Cat: 0.02
CNNs for Protein-Ligand Scoring

- Pose Prediction
- Binding Discrimination
- Affinity Prediction
Protein-Ligand Representation

(R,G,B) pixel \rightarrow (\text{Carbon, Nitrogen, Oxygen, \ldots}) \text{ voxel}

The only parameters for this representation are the choice of grid resolution, atom density, and atom types.
Model

2x2 Max Pooling
3x3x3 Convolution
2x2 Max Pooling
3x3x3 Convolution
2x2 Max Pooling
3x3x3 Convolution
Fully Connected

Rectified Linear Unit
Rectified Linear Unit
Rectified Linear Unit
Rectified Linear Unit

48x48x48x35
24x24x24x35
12x12x12x32
12x12x12x64
6x6x6x64
6x6x6x128
6x6x6x128

Fully Connected
Pseudo-Huber Loss

Affinity
Pose
Score
Results

Affinity Prediction

- CNN (R=0.74, RMSE=1.44)
- Vina (R=0.55, RMSE=1.86)

Pose Prediction

- CNN (AUC=0.89)
- Vina (AUC=0.61)

Trained on PDBbind refined; tested on CSAR 😐
Beyond Scoring

Deep Dreams

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Beyond Scoring

2Q89

More Oxygen Here

Less Oxygen Here

\[
\frac{\partial L}{\partial A} = \sum_{i \in C^A_{X_{\text{data}}}} \frac{\partial L}{\partial D_{i}} \frac{\partial G_{i}}{\partial D_{i}} \frac{\partial D_{i}}{\partial A_{\text{label}}}
\]
Key Concepts

Ligand-Based Virtual Screening
Identifying new active compounds based on similarity to known active compounds; fingerprint is a bit vector representation of a molecule

Pharmacophore
A spatial arrangement of molecular features essential for biological activity - hydrogen bonding, hydrophobic, charged, etc.

Docking
Predict the position, pose and affinity of a molecule using the receptor structure

Scoring
force field … empirical … knowledge based
http://pharmit.csb.pitt.edu