Fundamentals of Drug Design

Dima Kozakov
What is drug?
What is drug?

Pharmaceutical drug, also called a medication or medicine, is a chemical substance (typically organic molecule) used to treat, cure, prevent, or diagnose a disease or to promote well-being.
What is drug?

Pharmaceutical drug, also called a medication or medicine, is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being.

Zanamivir
History of Drug Design

1806

A. Cherkasov
# History of Drug Design

<table>
<thead>
<tr>
<th>Time</th>
<th>New Sources</th>
<th>Testing Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>ancient &amp; middle ages</td>
<td>plants, poisons (Paracelsus)</td>
<td>humans</td>
</tr>
<tr>
<td>1806</td>
<td>morphine (first extracted)</td>
<td>humans</td>
</tr>
<tr>
<td>1850</td>
<td>chemicals (chinin)</td>
<td>humans (prisoners)</td>
</tr>
<tr>
<td>1890</td>
<td>synthetics, pigments</td>
<td>animals</td>
</tr>
<tr>
<td>1920</td>
<td></td>
<td>animals, isolated organs</td>
</tr>
<tr>
<td>1970-1980</td>
<td></td>
<td>enzymes, cell lines (HeLa)</td>
</tr>
<tr>
<td>1990</td>
<td>High throughput libraries</td>
<td>recombinant proteins</td>
</tr>
<tr>
<td>2000</td>
<td>chemical libraries</td>
<td>chips, virtual screening, ADMET testing</td>
</tr>
</tbody>
</table>
Sources of Drugs

All drugs by source, registered 01/1981 - 06/2006, FDA, n = 1184

- S*; 4%
- V; 4%
- N; 5%
- S*/NM; 10%
- S/NM; 10%
- B; 14%
- ND; 23%
- S; 30%

B – biologicals,
N – nature compounds,
ND – nature compounds derivatized,
S – synthetic compounds,
S/NM – synthetics mimicking natural compounds,
S* - synthetic, with pharmacophore from natural compounds
V - vaccines

D. J. Newman and G. M. Cragg,
What is the drug target?
What is the drugs target?
Drug Targets

How do drugs work?
How do drugs work?

Most Typical Mechanism of Drug Action

• Lock and Key Analogon, 1894

"Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glykosid zueinander passen müssen, wie Schloss und Schlüssel, um eine chemische Wirkung aufeinander ausüben zu können."

Emil Fischer, Nobel Laureate 1902
Drug Design Project

Illness

Isolation of cause (many years)

Identification of lead compound (2-5 years)

Preclinical testing on animals (1-3 years)

Formulation (1-3 years)

Human clinical trials (2-10 years)

File IND

File NDA

FDA approval (2-3 years)

Heal and Sell
Vocabulary

• **Target**
  – Biomolecule interacting with the drug

• **Lead**
  – Base molecular structural motif of developed drug

• **Hit**
  – Compound with positive hit in initial screening

• **Candidate compounds**
  – Selected compounds used for next testing

• **Efficacy**
  – Qualitative property – (drug heals or not)

• **Activity**
  – Quantitative property – dosage needed for effect to happen
    (pM – great, nM – excellent, μM – sufficient, mM – well...)

• **Bioavailability**
  – Availability of compound in site of target in necessary concentration
Drug Design

Identification of new drug:

• Expensive problem
  – Expenditures per 1 drug development - 1 300 000 000 USD¹
  + expenses for production, patents, distribution...
  \[\Rightarrow\] New drugs are expensive >1 000 USD/dose of drug²

• Hard problem
  – Identification of target-drug pair is not simple
  – ADMET
  – Side-effects

¹ - Tufts Center for the Study of Drug Development, 2012
² – SÚKL, 3Q 2011, average price tag for most expensive drug category in CZ (over 10kCZK)
What tools do we need to make Drug for a target rationally?
What tools do we need to make Drug for a target rationally?

Experimental answer (expensive)

Target Structure (using X-RAY, NMR)

Molecular complex structure
The Structural Genomics Pipeline (X-ray Crystallography)

Basic Steps

Target Selection

Crystallomics
- Isolation,
- Expression,
- Purification,
- Crystallization

Data Collection

Structure Solution

Structure Refinement

Functional Annotation

Publish
How to determine structure? NMR

Summary of solution NMR spectroscopy

Experiment → Spectra processing → Spectra assignment

Model generation → Distance restraints → NOE assignment
Status - Numbers and Complexity of PDB

(a) myoglobin (b) hemoglobin (c) lysozyme (d) transfer RNA
(e) antibodies (f) viruses (g) actin (h) the nucleosome
(i) myosin (j) ribosome

Courtesy of David Goodsell, TSRI
What tools do we need to make Drug for a target rationally?

Computational Modelling answer (much cheaper)

Sequence-> Structure (Folding)
How do we do it? Both structure and complex–global minimum of free energy

\[
G = U + PV - TS
\]

\[
G \rightarrow \text{min}
\]
Modelled molecular representation – physics based model + knowledge of structures

\[ G = U + PV - TS \]

\[ G \rightarrow \text{min} \]
The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel "for the development of multiscale models for complex chemical systems".
Metropolis Monte Carlo

• Start with configuration A (energy $E_A$)
• Make random move to configuration B (energy $E_B$)
• Accept move when:
  $$E_B < E_A \text{ or if }$$
  $$E_B > E_A \text{ except with probability } P:$$

$$P = \exp\left(-\frac{[E_A - E_B]}{kT}\right)$$
Molecular Dynamics

- force-field is used to calculate forces on each atom of the simulated system
- following *Newton* mechanics, calculate accelerations, velocities and new coordinates from the forces.
  
  (Force = mass *times* acceleration)
- The atoms are moved slightly with respect to a given time step
\[ E_{\text{binding}}(\alpha, \beta, \gamma) = \sum_{p=1}^{P} \sum_{lmn} R_p(l, m, n)L_p(l + \alpha, m + \beta, n + \gamma) \]

\[ E_{\text{binding}}(\alpha, \beta, \gamma) = \text{IFT}(\sum_{p=1}^{P} \text{FT}^*(R_p(l, m, n))\text{FT}(L_p(l, m, n)))(\alpha, \beta, \gamma) \]

\[ O(N^6) \rightarrow O(N^3 \ln N^3) \]
Steps of rational Lead discovery

If we already have natural ligands
Screening Tools using ligand similarity

- ROCKS (commercial)
- Ligsift (open source)
Steps of rational Lead discovery

If we don’t

Asses druggability

• Does the target have a site where a drug can bind, and with appropriate affinity?
Experimental observation (D. Ringe, 1996)

Amount of diverse small fragment sized molecules
Strongly corellates with druggability

C. Mattos and D. Ringe (1996)
Locating and characterizing binding sites on proteins, Nature Biotech.
14: 595-599
Part I: Finding hot spots by computational mapping

Find consensus sites: regions with several probes binding
FTMAP Step 1: “Soft” grid based docking using FFT
FTMAP Step 2: Refinement and filtering
FTMAP Step 3: Clustering and scoring of clusters
FTMAP Step 4: Consensus Clustering
FTMAP Step 4: Consensus Clustering
Small set of 16 probes used for mapping
Computational and experimental mapping of elastase

Experimental mapping

Computational mapping
Largest consensus sites
FTMap server: over 2000 registered users

Computational solvent mapping is a powerful tool to understand interactions between proteins and solvent molecules. It docks small organic molecules on a protein surface, finds favorable binding positions, clusters the conformations of all prediction, and ranks the clusters on the basis of their average free energy. The low energy clusters are grouped into consensus sites and the largest consensus sites are able to identify active or ligand binding sites. The docked fragments can also be served as the building blocks for fragment-based drug design.

Identification of “hot spots” in traditional drug targets: Renin

• Major target for the treatment of hypertension.

• Over 25 years of research into small molecule inhibitors.

• Most inhibitors are peptidomimetics.

• Novartis got FDA approved drug Aliskiren, a novel non-peptidomimetic renin inhibitor.

http://www.merck.com/mmhe/sec03/ch022/ch022a.html
Aliskiren and a peptidomimetic inhibitor of renin
Aliskiren (violet) and peptidomimetics (green) over mapping results (cyan)

Aliskiren (violet) and peptidomimetics (green) over mapping results (cyan)
Protein-protein interactions: mapping of IL-2

A. Mapping of inhibitor-bound IL-2. B. Mapping of ligand-free IL-2.

Kozakov D et al. PNAS 2011
Non druggable example - ZipA protein

Small clusters

A. Mapping of inhibitor-bound ZipA. B. Mapping of ligand-free ZipA.

Kozakov D et al. PNAS 2011
New Frontiers in Druggability

Dima Kozakov,*† David R. Hall,‡ Raeanne L. Napoleon,§ Christine Yueh,∥ Adrian Whitty,*§ and Sandor Vajda*§∥

†Department of Applied Mathematics & Statistics, Stony Brook University, Stony Brook, New York 11794, United States
‡Acpharis Inc., Holliston, Massachusetts 01746, United States
§Department of Chemistry and ∥Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, United States

Supporting Information
<table>
<thead>
<tr>
<th>Druggability class</th>
<th>Strength (S) (Number of probe clusters)</th>
<th>Center-to-center distance (CD), Å</th>
<th>Maximum dimension (MD), Å</th>
<th>Druggability subclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Druggable</td>
<td>S ≥ 16</td>
<td>CD &lt; 8</td>
<td>MD ≥ 10</td>
<td>Druggable using druglike compounds</td>
</tr>
<tr>
<td>Not druggable</td>
<td>S &lt; 13</td>
<td>Any CD</td>
<td>Any MD</td>
<td>Not druggable due to weak hot spots</td>
</tr>
<tr>
<td></td>
<td>Any S</td>
<td>Any CD</td>
<td>MD &lt; 7</td>
<td>Not druggable due to small hot spot ensemble</td>
</tr>
<tr>
<td>Non-Canonically Druggable - Large</td>
<td>S ≥ 16</td>
<td>CD ≥ 8</td>
<td>MD ≥ 10</td>
<td>Druggable only by large chemotype such as macrocycle or foldamer</td>
</tr>
<tr>
<td>Non-Canonically Druggable - Small</td>
<td>S ≥ 16</td>
<td>CD &lt; 8</td>
<td>7 ≤ MD &lt; 10</td>
<td>Druggable only by peptide, macrocycle, or charged compound</td>
</tr>
</tbody>
</table>
Docking programs

- DOCK (Developed at Stony Brook!)
- AutoDOCK (open)
- GOLD (commercial)
- Glide (commercial)
Example Ligand Docking Protocol overview

Padhorny et. al. JCAMD 2017. In press
FFT Based sampling

$$E_{binding}(\alpha, \beta, \gamma) = \sum_{p=1}^{P} \sum_{l} \sum_{m} \sum_{n} R_p(l, m, n) L_p(l + \alpha, m + \beta, n + \gamma)$$

$$E_{binding}(\alpha, \beta, \gamma) = \text{IFT}(\sum_{p=1}^{P} \text{FT}^*(R_p(l, m, n)) \text{FT}(L_p(l, m, n)))(\alpha, \beta, \gamma)$$

$$E_{full} = E_{vdw} + E_{elec} + E_{pair}$$

$$E_{vdw} = E_{rep} + E_{attr}$$

$$E_{elec} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{q_i q_j}{r_{ij}^2 + D^2 \exp \left( \frac{-r_{ij}^2}{4D^2} \right)^{1/2}}$$

$$E_{pair} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \epsilon(i, j) = \begin{cases} 0 & r_{ij} > D \\ \epsilon_{IJ} & d < r_{ij} < D \quad a_i \in I, a_j \in J \end{cases}$$

Kozakov D. et. Al. eLIFE 2014
Padhony et. al., PNAS 2016
Energy as correlation

\[ E_{elec} = \sum_i q_i \phi_r \]

Kozakov et. al. 2006; Brenke et. al. 2009; Bohnuud et. al. 2012; Kozakov et. al, 2014
Finding most probable region

- Low energy conformations below a given threshold will cluster.
- Clusters are representative of the energy minima.
- Clusters size represents the “region of attraction” and thus some entropic contributions to the free energy.


\[ P_k = \frac{Z_k}{Z}, \quad Z = \sum_j \exp(-E_j / RT), \quad Z_k = \sum_j \exp(-E_j / RT) \]
\[ Z = N \exp(-E / RT), \quad Z_k = K \exp(-E / RT), \quad P_k = \frac{Z_k}{Z} = \frac{K}{N} \]
Monte Carlo Minimization

\[ \begin{pmatrix} \omega_1, \omega_2, \omega_3, x_1, x_2, x_3 \end{pmatrix} \]

\[ T(x) = (x_1, x_2, x_3) \]

\[ R(\omega) = \exp \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \]

Mirzaei H et al. JCTC 2012; Mirzaei et al, JCTC 2015; Mamonov et. al. JCC 2015; Padhony et. al JCAMD 2017. in press
Results: O1-17-OHP

Best RMSD Top 1 Pose
Definition of Biological Product

• US:
  – The term “biological product” or biologics means a "any virus, therapeutic serum, toxin, antitoxin or analogous product applicable to the prevention, treatment or cure of diseases or injuries of man”

• EU:
  – 'biological medicinal products' as "a protein or nucleic acid–based pharmaceutical substance used for therapeutic or in vivo diagnostic purposes, which is produced by means other than direct extraction from a native (nonengineered) biological source"

Ronald A Rader (Re)defining biopharmaceutical Nature Biotechnology 26, 743 - 751 (2008) doi:10.1038/nbt0708-743
Small Molecules vz Biologicals

**Chemical medicines** are chemicals made by chemists out of other chemicals.

**Biologics** are *grown* from living things. Biologics are highly sensitive to manufacturing conditions.
FDA Approved New Chemical Entities and Biological Derivatives

What are Biologicals?

• made of sugars, proteins, or nucleic acids or complex combinations of these substances, or may be living entities such as cells and tissues

• Like drugs, biological products are used to either:
  – treat or cure diseases and medical conditions,
  – prevent diseases, or
  – diagnose diseases

• Biological products are made from a variety of natural sources.
Types of Biological Products

• Blood Derivatives
• Whole Blood
• Blood Components
• Proteins
• Human Tissues
• Xenotransplantation Products
• Cellular & Gene Therapies
• Vaccines
• Allergenic Extracts
Size and Complexity of Biologicals in Comparison with Small Molecules

Aspirin 180 Da

Monoclonal Antibody ~150,000 Da
Computational – biologics design – docking
State of the art: multiscale docking approach

FFT-based global sampling on a grid using “smooth” potentials: PIPER

A clustering approach to finding near-native conformations

Refinement: Monte Carlo Semidefinite Underestimator (SDU)

Submit models to CAPRI
I. Global search

Cartesian FFT- $\mathbb{R}^3 \times \text{SO}(3)$

$$E_{\text{binding}}(\alpha, \beta, \gamma) = \sum_{p=1}^{P} \sum_{lmn} R_p(l, m, n) L_p(l + \alpha, m + \beta, n + \gamma)$$

$$E_{\text{binding}}(\alpha, \beta, \gamma) = \text{IFT}(\sum_{p=1}^{P} \text{FT}^*(R_p(l, m, n)) \text{FT}(L_p(l, m, n)))(\alpha, \beta, \gamma)$$

$$O(N^6) \rightarrow O(N^3 \ln N^3)$$
Energy Model

\[ E_{full} = E_{vdw} + E_{elec} + E_{pair} \]

\[ E_{vdw} = E_{rep_{trunc}} + E_{attr} \]

\[ E_{elec} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{q_i q_j}{\left( r_{ij}^2 + D^2 \exp \left( \frac{-r_{ij}^2}{4D^2} \right) \right)^{1/2}} \]

\[ E_{pair} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \epsilon(i, j) = \begin{cases} 
0 & r_{ij} > D \\
\epsilon_{IJ} & d < r_{ij} < D \quad a_i \in I, a_j \in J 
\end{cases} \]

\[ \epsilon_{IJ} = \sum_{t=1}^{K} \epsilon_{tI} \lambda_t \epsilon_{tJ} \]

\[ E_{VdW}, E_{coul}, E_{pb}, E_{solv}, E_{hb} \]

II. Finding most probable region

- Low energy conformations below a given threshold will cluster

- Clusters are representative of the energy minima

- Clusters size represents the “region of attraction” and thus some entropic contributions to the free energy

Kozakov, D., et. al. 2013

\[ P_k = \frac{Z_k}{Z}, \quad Z = \sum_j \exp\left(-\frac{E_j}{RT}\right), \quad Z_k = \sum_j \exp\left(-\frac{E_j}{RT}\right) \]

\[ Z = N \exp\left(-\frac{E}{RT}\right), \quad Z_k = K \exp\left(-\frac{E}{RT}\right), \quad P_k = \frac{Z_k}{Z} = \frac{K}{N} \]
ClusPro server

https://cluspro.org/

Welcome to Cluspro 2.0

Recent news: ClusPro server featured on the cover of February 2017 issue of Nature Protocols

Use Without an Account

Use the server without the benefits of your own account

--or--

Login

Username: midas
Password: 

Login

--or--

Sign up for an account

Forgot Password?
Reset Password

Forgot Username?
Retrieve Username

ClusPro should only be used for noncommercial purposes.
ABC Group and Structural Bioinformatics Lab
Boston University and Stony Brook University

Kozakov et. al. Nature Protocols 2017
ClusPro usage: 8,000 users:

~100,000 jobs run, hundreds of research papers
CAPRI – Blind Protein-protein docking

Kozakov, D., et. al (2010, 2013),
## ClusPro and CAPRI

### Servers (2013)

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CLUSPRO</td>
<td>6/4**/2*</td>
</tr>
<tr>
<td>2</td>
<td>HADDOCK</td>
<td>4/1***/2**</td>
</tr>
<tr>
<td>3</td>
<td>SWARMDOCK</td>
<td>4/1***/3*</td>
</tr>
<tr>
<td>4</td>
<td>PIE-DOCK</td>
<td>3/1***/2*</td>
</tr>
</tbody>
</table>

### Servers (2016)

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CLUSPRO</td>
<td>9/3**</td>
</tr>
<tr>
<td>2</td>
<td>PyDockWeb</td>
<td>6/2**</td>
</tr>
<tr>
<td>3</td>
<td>LzerD</td>
<td>4/1***/3**</td>
</tr>
<tr>
<td>4</td>
<td>HADDOCK</td>
<td>4/2**</td>
</tr>
</tbody>
</table>

### Human/Server (2013)

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bonvin</td>
<td>9/1***/3**</td>
</tr>
<tr>
<td>2</td>
<td>Bates</td>
<td>8/2**/6*</td>
</tr>
<tr>
<td>3</td>
<td>Vakser</td>
<td>7/1***/6*</td>
</tr>
<tr>
<td>4</td>
<td>Kozakov/ Vajda</td>
<td>6/2***/4**</td>
</tr>
<tr>
<td>5</td>
<td>Shen</td>
<td>6/1***/3**</td>
</tr>
<tr>
<td>6</td>
<td>Fernandez-Recio</td>
<td>6/1***/3**</td>
</tr>
<tr>
<td>7</td>
<td>CLUSPRO</td>
<td>6/4***/2*</td>
</tr>
<tr>
<td>8</td>
<td>Zou</td>
<td>6/1***/2**</td>
</tr>
<tr>
<td>8</td>
<td>Zacharias</td>
<td>6/1***/5*</td>
</tr>
</tbody>
</table>

### Human/Server (2016)

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Geroix</td>
<td>10/1***/8**</td>
</tr>
<tr>
<td>2</td>
<td>Zacharias</td>
<td>10/3***/2**</td>
</tr>
<tr>
<td>3</td>
<td>CLUSPRO</td>
<td>9/3**</td>
</tr>
<tr>
<td>4</td>
<td>Kozakov/ Vajda, Seok</td>
<td>8/3***/2**</td>
</tr>
<tr>
<td>5</td>
<td>Fernandez-Recio</td>
<td>7/1***/3**</td>
</tr>
<tr>
<td>6</td>
<td>Zou</td>
<td>7/1***/2**</td>
</tr>
<tr>
<td>7</td>
<td>Weng</td>
<td>6/1***/4**</td>
</tr>
<tr>
<td>8</td>
<td>Vakser</td>
<td>6/2***/2**</td>
</tr>
<tr>
<td>9</td>
<td>Bates</td>
<td>6/3**</td>
</tr>
</tbody>
</table>
Take Home Message

• Drugs comes from various sources
• Drug design is hard and expensive problem
  – Mainly due to the biology!
• Most typical drug targets are:
  – GPCRs, ion channels, nuclear receptors, kinases
  – But - long tail of other drug targets
• Biologicals are more complex than small molecules
• There is no gold path for drug design – the methods have to be tied up to the current project
Search problems

Analytical

Systematic enumeration of

Random sampling