Big Data Science for Early-Stage Drug Discovery

Scope

Focus on early stages in the drug discovery pipeline
- How to identify most promising drug candidates from compound screenings using computational tools
- Data sources
  - Small molecules (e.g., structures, physicochemical properties, etc)
  - Protein targets (e.g., sequences, structures, families, etc)
  - Chemical screenings (e.g., primary, confirmatory, phenotypic, etc)
- Data properties
  - Big Data
  - Can be of relatively low quality
  - Compound potency, specificity, toxicity, etc

Why is this critical? It informs and directs the entire drug discovery process!

Xia Ning • Big Data Science in Drug Discovery and Development
Big Data Science for Early-Stage Drug Discovery

Problems and Computational Tools (I)

- Structure-Activity-Relationship (SAR) models
  - Computational tools to build SAR models
    - Classification, regression, ranking
  - Better single-target SAR
    - Pharmacophore identification
    - Multi-instance learning
  - SAR based on Chemogenomics
    - Multi-task learning
  - SAR beyond Chemogenomics
    - Multi-task learning, model ensembles
  - SAR for cancer drug selection
Big Data Science for Early-Stage Drug Discovery

Problems and Computational Tools (II)

- **Structure-Selectivity-Relationship (SSR) models**
  - Computational methods to build SSR models
    - Multi-task learning, cascade models

- **Structure-Property-Relationship (SPR) models**
  - Computational methods for compound toxicity
    - Deep learning for toxicity prediction
Outline

- Overview
- Background
- Structure-Activity-Relationship Modeling (SAR)
  - SAR Problem Formulation
  - Computational Tools to Build SAR
  - Better Single-Target SAR
  - SAR based on Chemogenomics
  - SAR beyond Chemogenomics
  - SAR for Cancer Drug Selection
- Structure-Selectivity-Relationship Modeling (SSR)
  - Computational Methods for SSR
- Structure-Property-Relationship Modeling (SPR)
  - Computational Methods for SPR
- References
- Q & A
Outline

- Overview
- **Background**
  - Structure-Activity-Relationship Modeling (SAR)
    - SAR Problem Formulation
    - Computational Tools to Build SAR
    - Better Single-Target SAR
    - SAR based on Chemogenomics
    - SAR beyond Chemogenomics
    - SAR for Cancer Drug Selection
  - Structure-Selectivity-Relationship Modeling (SSR)
    - Computational Methods for SSR
  - Structure-Property-Relationship Modeling (SPR)
    - Computational Methods for SPR
- References
- Q & A
Terminologies¹²

Assay

A biological test, measurement or analysis to determine whether compounds have the desired effect either in a living organism, outside an organism, or in an artificial environment

---

¹ http://www.combiochemistry.com/medical-chemistry-glossary.html
## Terminologies\(^1\)^2

<table>
<thead>
<tr>
<th>Assay</th>
<th>A biological test, measurement or analysis to determine whether compounds have the desired effect either in a living organism, outside an organism, or in an artificial environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligand</td>
<td>A small molecule that binds specifically to a larger one; for example, a hormone is the ligand for its specific protein receptor.</td>
</tr>
</tbody>
</table>

---

## Terminologies

<table>
<thead>
<tr>
<th>Assay</th>
<th>A biological test, measurement or analysis to determine whether compounds have the desired effect either in a living organism, outside an organism, or in an artificial environment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligand</td>
<td>A small molecule that binds specifically to a larger one; for example, a hormone is the ligand for its specific protein receptor.</td>
</tr>
<tr>
<td>Target</td>
<td>The naturally existing cellular or molecular structure involved in the pathology of interest that the drug-in-development is meant to act on.</td>
</tr>
</tbody>
</table>

---

## Terminologies

<table>
<thead>
<tr>
<th>Assay</th>
<th>A biological test, measurement or analysis to determine whether compounds have the desired effect either in a living organism, outside an organism, or in an artificial environment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligand</td>
<td>A small molecule that binds specifically to a larger one; for example, a hormone is the ligand for its specific protein receptor.</td>
</tr>
<tr>
<td>Target</td>
<td>The naturally existing cellular or molecular structure involved in the pathology of interest that the drug-in-development is meant to act on.</td>
</tr>
<tr>
<td>Drug</td>
<td>Any substance presented for treating, curing or preventing disease in human beings or in animals. A drug may also be used for making a medical diagnosis or for restoring, correcting, or modifying physiological functions.</td>
</tr>
</tbody>
</table>

Chemical Genetics (Genomics)

The research field that is designed to discover and synthesize protein-binding small organic molecules that can alter the function of all the proteins and use them to study biological systems.

Key advantages

- Small molecules can work rapidly
- Small molecules can disrupt protein-protein interactions
- If the target is pharmaceutical relevant, it can lead to the discovery of new drugs

- A laborious, multi-step, time-consuming and costly process
Data Sources

- Chemical compound libraries
  - Several millions of compounds (PubChem contains \( \sim 80M \) compounds)
  - Chemical space is estimated to be in the order of \( 10^{100} \).
    - Represent the potential space of compounds that need to be considered for drug/probe development

- Protein targets
  - Crystallographic & NMR information from in vitro experiments
  - Amino acid sequences, protein families, super-families, folds

- Bioactivity information
  - Screening results
    - High-throughput screening assays
    - Confirmatory assays
    - Dose-response assays
  - Target-ligand affinity information extracted from literature
Outline

- Overview
- Background
- **Structure-Activity-Relationship Modeling (SAR)**
  - SAR Problem Formulation
  - Computational Tools to Build SAR
  - Better Single-Target SAR
  - SAR based on Chemogenomics
  - SAR beyond Chemogenomics
  - SAR for Cancer Drug Selection
- **Structure-Selectivity-Relationship Modeling (SSR)**
  - Computational Methods for SSR
- **Structure-Property-Relationship Modeling (SPR)**
  - Computational Methods for SPR
- References
- Q & A
Structure-Activity-Relationship (SAR) Models

- Compound activity
  - Bind with high affinity to the protein target: $IC_{50}(c, t) < 1000nM$

**Half Maximal Inhibitory Concentration (IC₅₀)**

The concentration of a particular drug or other substance (inhibitor) that is needed to inhibit a given biological process (or component of a process, i.e. an enzyme, cell, cell receptor or microorganism) by half (50%) \(^3\)

- Computational foundation: The biological activity of a chemical compound can be mathematically expressed as a function of its physicochemical properties [1, 2]

- Problem setting
  - Given a set of active (and inactive) compounds to a particular target, build a model to predict the binding of other compounds against that target

---

\(^3\) [https://en.wikipedia.org/wiki/IC50](https://en.wikipedia.org/wiki/IC50)
Structure-Activity-Relationship (SAR) Models

Problem Formulation [3]

- A classification problem
  - Positive instances: active compounds
  - Negative instances: inactive compounds
    - Randomly sample negative instances from a diverse library of non-binding compounds

- A regression problem
  - Regression from compound features to IC\(_{50}\) values
  - Ordinal regression from compound features to activity ranges

- A ranking problem
  - Need to differentiate actives from inactives
  - Need to reproduce right orderings (not necessarily the exact IC\(_{50}\) values)
Compound Feature Representation

- Represent each compound as a frequency vector in a descriptor (feature) space
  - Descriptors: (simple) chemical substructures
  - Frequency: the number of times each descriptor exists in the molecular graph
Classification for SAR (I)

- Support Vector Machines
  - Tanimoto kernel (extended Jaccard-coefficient): a valid kernel

\[
\text{Tanimoto}(f_1, f_2) = \frac{|f_1 \cap f_2|}{|f_1 \cup f_2|}
\]

Figure 1: SVM Classification
Classification for SAR (II)

- **Neural Networks [4]**
  - Individual predictive model for each target
    \[
    P(\text{activity}|X, i) = \sigma(V \tanh(AX + B \tanh(We_i)))
    \]
  - Indirect predictive model for each target
    \[
    P(\text{activity}|X, i, T) = \sigma(V \tanh(AX + B \tanh(CTe_i)))
    \]

![Diagram of Neural Network for SAR][1]

**Figure 2**: Neural Network for SAR [4]

- **Deep architecture [5]**

---

[Xia Ning](#) • [Big Data Science in Drug Discovery and Development](#)
Regression for SAR (I)

- Partial Least Squares (PLS)
  - A limited number of compounds
  - Compound descriptors of high dimensionality

\[
X = TP^T + E, \quad Y = UQ^T + F
\]
\[
\Rightarrow Y = TBQ^T + F \quad (U = TB)
\]

- \(X\): compound descriptors, \(Y\): compound activities

**Figure 3**: PLS Regression

---

4 http://documentation.statsoft.com/STATISTICAHelp.aspx?path=mspc/PCAandPLSTechnicalDetails

Xia Ning • Big Data Science in Drug Discovery and Development
Ranking for SAR (I)

- Bipartite ranking [6]
- Active compounds should be ranked higher than inactive compounds
- Ranking error

\[
\frac{1}{|X^+||X^-|} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[ I(f(x_i^+) < f(x_j^-)) + \frac{1}{2} I(f(x_i^+) = f(x_j^-)) \right]
\]

- Convex upper bound on ranking error

\[
\frac{1}{|X^+||X^-|} \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - (f(x_i^+) - f(x_j^-)))_+, \quad (a)_+ = \max(a, 0)
\]

- In RKHS: RankSVM

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - (f(x_i^+) - f(x_j^-)))_+ + \frac{1}{2C} \|f\|_F^2
\]
Issues for SAR Modeling

- Unknown binding conformations
  - Protein-ligand binding is a dynamic process.
  - Binding mode is not necessarily the one with the lowest energy.

- Ideas for better SAR
  - Predict binding mode: pharmacophore identification
  - Use multiple possible conformations: Multi-Instance Learning (MIL)

Figure 4: Protein-Ligand Binding
Pharmacophore Identification

- Pharmacophore
  - An abstract description of molecular features that are necessary for molecular recognition of a ligand by a biological macromolecule

- Identify pharmacophores
  - Superimposition: align conformers to a rigid conformation of reference ligand [7]
  - Graph mining: identify common substructures from active compounds [8]

Figure 5: Pharmacophores

https://en.wikipedia.org/wiki/Pharmacophore
Multi-Instance Learning for SAR

- Classify bags of instances
  - A bag is positive if one of its instances is positive.
  - A bag is negative when all of its instances are negative.
  - A compound ⇒ a bag; conformers ⇒ instances

- Multi-Instance Learning (MIL) [9]
  - Instance-based embedding

\[
D(M_i, C^r) = \min_j D(C_{ij}, C^r)
\]

- \( \ell_1 \)-norm SVM

**Figure 6**: MIL with \( \ell_1 \)-norm SVM [9]

Xia Ning • Big Data Science in Drug Discovery and Development
Chemical Genomics for SAR

- Biological foundation: proteins from a same family (e.g., GPCRs, nuclear receptors, kinases, proteases, etc.) tend to bind to similar compounds [10].
- Key idea: leverage protein and protein family information
  - Multi-Task Learning (MTL)
    - Multiple related tasks learned simultaneously
    - Domain-specific information contained in the related tasks transferred across
    - SAR for each target ⇒ a task; proteins of a same family ⇒ task relatedness
- Features of protein targets
  - Protein structures [11], amino acid sequences [12], binding site descriptors [13, 14], topological descriptors [15], etc.
Chemical Genomics Methods for SAR (I)

- **Multi-Task Collaborative Filtering [4]**
  - Collaborative Filtering (CF) in recommender systems
    - “users” ⇔ “targets”, “items” ⇔ “ligands”, “ratings” ⇔ “activities”
    - “rating prediction” ⇔ “activity prediction”
  - Kernel perceptron
    \[
    F(t, c; w) = \Psi(t, c)^T w
    \]
    - \(\Psi(t, c)\): map a target-compound pair \((t, c)\) to a size-\(D\) feature vector
      \[
      K((t, c), (t', c')) = \Psi(t, c)^T \Psi(t', c')
      \]
      \[
      = K((t, c), (t', c')) = K_t(t, t') \cdot K_c(c, c')
      \]
  - Representer Theorem to CF:
    \[
    F(t, c; \alpha) = \sum_{(t', c')} \alpha_{t', c'} K((t, c), (t', c'))
    \]
Chemical Genomics Methods for SAR (II)

- Multi-task SVM classification [12]

\[ K((t, c), (t', c')) = K_t(t, t') \cdot K_c(c, c') \]

- Various target kernels

\[ K(t, t') = \langle \Phi_h(t), \Phi_h(t') \rangle \]

- \( \Phi_h(t) \): features in protein family hierarchy
SAR beyond Chemical Genomics

Key ideas
- Data-driven
- Utilize proteins outside protein families
- Utilize the target-ligand activity matrix

Hypothesis
- The ligands of “related” targets can be used to build a better model for the target under consideration.

Multi-assay based SAR [16]
- Step 1: identify a set of related targets to the target under consideration
- Step 2: leverage the activity information of these targets to build a better SAR model
SAR Methods beyond Chemical Genomics (I)
Identify Related Proteins from Protein-Protein Similarities

- Similarities from proteins
  - Binding site similarity
    - Predict binding site structures/ligand-binding residues
  - Sequence similarity
    - A profile-based alignment score of entire sequences
    - Restrict the profile-based similarities to only (predicted) ligand-binding residues

- Similarities from ligands
  - The idea: if two proteins bind to ligands that are similar, then their binding sites should exhibit similar compatible characteristics.
  - Use ligand similarities to determine protein-protein similarities
SAR Methods beyond Chemical Genomics (II)
Leverage Additional Information

- **Semi-Supervised Learning**
  - Select unlabeled ligands from related proteins
  - Build a model from both labeled and unlabeled ligands

- **Multi-Task Learning**
  - Learn models for all targets in parallel

- **Multi-Ranking**
  - Build models for the target of interest and its related targets, respectively
  - Combine the predictions from the models with weights
SAR Methods beyond Chemical Genomics (III)

Figure 7: Connectivity Pattern between Related Proteins [16]

- Conclusion: data-driven approaches can outperform Chemogenomics approaches
Cancer drug response prediction [17]

- Cell line features: Copy Number Variations (CNV), methylation, exome sequencing, gene expression, RNA-seq
- A set of drugs
- Cell line sensitivity to drugs
- Cancer drug selection towards precision medicine

Figure 8: Cancer Drug Response Prediction Problem [17]
SAR Methods in Cancer (I)

- Neural Networks [18]:
  - Concatenate cell line features and drug features

Figure 9: Neural Networks for Sensitivity Prediction [18]
**SAR Methods in Cancer (II)**

- Bayesian Multitask Multiple Kernel Learning (BMTMKL) [17]
  - Multiple tasks: each drug as a task
  - Multiple kernels: a kernel for each genomic measurement
  - Drug-specific weights on kernels
  - Shared kernel weights

**Figure 10:** Bayesian Multitask Multiple Kernel Learning Framework [17]
SAR Methods in Cancer (III)

- Kernelized Bayesian Matrix Factorization (KBMF) [19]
  - Multiple kernels for each cell line/drug
  - Project kernels into composite components (low-dimension latent factors)
  - Sensitivities from composite components (matrix factorization)

**Figure 11:** Drug and Cell Line Features [19]

**Figure 12:** Kernelized Bayesian Matrix Factorization Framework [19]
Outline

- Overview
- Background
- Structure-Activity-Relationship Modeling (SAR)
  - SAR Problem Formulation
  - Computational Tools to Build SAR
  - Better Single-Target SAR
  - SAR based on Chemogenomics
  - SAR beyond Chemogenomics
  - SAR for Cancer Drug Selection
- **Structure-Selectivity-Relationship Modeling (SSR)**
  - Computational Methods for SSR
- Structure-Property-Relationship Modeling (SPR)
  - Computational Methods for SPR
- References
- Q & A
Structure-Selectivity-Relationship (SSR) Models

- Compound selectivity
  - Bind with high affinity to only the protein target: act efficaciously and minimize side effects
  
  \[(i)\quad c \text{ is active for } t_i, \text{ and}\\
  \quad (ii)\quad \min_{\forall t_j \in T_i} \frac{IC_{50}(c, t_j)}{IC_{50}(c, t_i)} \geq 50.\]

- Testing of drug candidates: activity $\rightarrow$ selectivity
- Failure of selectivity testing leads to significant wastes

- Key idea
  - Leverage the activity information of a compound against all the proteins
Ranking based SSR

- Multiclass SVM Ranking [20]
  - One-vs-rest approaches, two-step approaches

Figure 13: Ranking based Selectivity [20]
Multi-Task SSR

- MTL for compound selectivity [21]
  - Compound $c$’s selectivity for $t_i$ defined with respect to challenge set $T_i$
  - Useful information from $T_i$’s compounds help determine $c$’s selectivity for $t_i$
  - Selectivities for $T_i$ learned simultaneously so as to facilitate information transfer
    - All the SAR and SSR models learned together

**Figure 14**: A Multi-Task Neural Network for Compound Selectivity [21]
Outline

- Overview
- Background
- Structure-Activity-Relationship Modeling (SAR)
  - SAR Problem Formulation
  - Computational Tools to Build SAR
  - Better Single-Target SAR
  - SAR based on Chemogenomics
  - SAR beyond Chemogenomics
  - SAR for Cancer Drug Selection
- Structure-Selectivity-Relationship Modeling (SSR)
  - Computational Methods for SSR
- Structure-Property-Relationship Modeling (SPR)
  - Computational Methods for SPR
- References
- Q & A
Compound Toxicity Prediction (I)

- Deep Learning (DL) [22]
  - Rectify linear units: enforce sparsity and counteract vanishing gradient
  - Dropout for regularization
  - Cross-entropy objective with softmax or sigmoid activation

\[ - \sum_{k=1}^{n} m_k (t_k \log(y_k) + (1 - t_k) \log(1 - y_k)) \]

**Figure 15**: Deep Neural Networks [22]
Compound Toxicity Prediction (II)

- Toxicity prediction using DL [22]
  - A lot of different compound features
  - Reconstruct toxicophores

Figure 16: Toxicophores [22]
References I


References II


References III


Xia Ning • Big Data Science in Drug Discovery and Development
Thank You!

Questions?