Modeling HTS Screening for Drug discovery

Ying-Ta Wu
Genomics Research Center, Academia Sinica, Taiwan.
e-mail: ywu@gate.sinica.edu.tw
Outlines

• Drug discovery
• HTS Screening
• Computational Screening
  • Structure-based approach
  • Compound-based approach
• Large-Scale Virtual Screening on GRID
• Bioactive Compound Profile
Drug Discovery

- Disease Targets
- Chemical Entity ------ > drug
Drug Targets

Human

Non-Human

parasites, bacteria, viruses, …
Drug Development by chemist intuition ...
**Drug Discovery** by chemist intuition …

Target

![Molecular structure](image)

Assay

SAR (structure Activity relationship)

optimization cycle

Candidate
Time and Cost

How many compounds can a chemist synthesize per years? -- few hundreds

\[
\text{R1} - \text{R2} - \text{R3} \quad \rightarrow \quad \text{R2} - \text{NH}_2
\]

how many? \(10^9\)

Library – chemical landscape \(10^{100}\)

Screening – speed vs cost

Solution: automation and miniaturize. HTS

Alternative? VS
New Paradigm in Drug Discovery

- Target selected
- Assay developed
- HTS
- HTS hits confirmed
- Chemistry begins
- Target structure obtained
- Development candidate is taken forward

Screen Strategy

Refer to Walters et al. DDT, 3, 160-178 (1998)
New Paradigm in Drug Discovery

Target selected
Assay developed
HTS
HTS hits confirmed
Chemistry begins
Target structure obtained
Development candidate is taken forward

Database clustering
Similarity analysis/
Virtual screening
QSAR
Pharmacophores
Structure -based design/
lead optimizing

Homology modeling
library selecting

library selecting

Target selected
Assay developed
HTS
HTS hits confirmed
Chemistry begins
Target structure obtained
Development candidate is taken forward

Refer to Walters et al. DDT, 3, 160-178 (1998)
Drug Discovery by Screening Strategy

Example: Broad (random) Screen

1. HTS → library
2. HTS → reconfirmation assay
3. confirmed hits → cluster/MCS/mode
4. cluster/MCS/mode → hit series
5. hit series → SAR/ADME/IP
6. SAR/ADME/IP → prioritized hits
7. prioritized hits → selected hits
8. selected hits → extended hits
9. extended hits → substructure similarity
10. substructure similarity → repository
Drug Discovery by Screening

Strategy

Random Screen (> 1,000,000)
- whole library screening at initial stage
- no biased, novel
- need uHTS if library is large
- low hit rate, cost

Focused Screen (~10,000)
- specific collection screening
- manageable, efficient
- need prior bioactive information
- novelty

Sequential Screen (5000~10,000)
- representative subset or explore hit series, which may be recruited after other two screen procedures
- need clustering, data-mining, etc.
- initial selection
Sequential/Focused Screen

virtual screening
medicinal chemistry

Initial library → HTS → data analysis → lead opt

new library ← active model

- SD docking
- LB filtering, similar searching
- subset diversity (features)

• clustering
Computational modeling

When to apply

Information drives drug discovery
-- the more of it, the sooner, the better.

What methods and tools

virtual library, cluster, screen, or score…
Examples: computational methods and tools

• Target structure-based
  - Target structure-based approaches
  - Protein-ligand docking
  - Active site-directed pharmacophores
  - Molecule-based queries
    - 2D substructures
    - 3D pharmacophores
    - Complex molecular descriptors (e.g., electrotopological)
    - Volume- and surface-matching algorithms
  - Molecular fingerprints
    - Keyed 2D fingerprints (each bit position is associated
      with a specific chemical feature)
    - Hashed 2D fingerprints (properties are mapped to
      overlapping bit segments)
    - Multiple-point 3D pharmacophore fingerprints
  - Compound classification techniques
    - Cluster analysis
    - Cell-based partitioning (of compounds into sub-
      sections of n-dimensional descriptor space)
    - 3D/4D-QSAR models

• Compound-based

• Score and cluster
  - Statistical methods
    - Binary QSAR/QSPR
    - Recursive partitioning
Study Case

Develop inhibitor of SARS coronavirus main protease-3CL\textsuperscript{pro}
A: Pre-screening

NCI 25K compounds

commercial available

modified RO5

structural diversity

2K compounds
B: Docking Methods

Docking Engine: AutoDock 3.0.5

create energy maps/element on active-site

evaluate $1.5 \times 10^6$ energies/molecule

carry out $50 \times 10$ runs/molecule

cluster on RMSD=1Å and $max \Delta G_{\text{binding}}$

test compounds with
- $\Delta G_{\text{binding}} > 12$ kcal/mol

1. Prepare the Target Protein
   -- add polar hydrogen atoms
   -- assign charges to atoms
   -- decide range of binding site
2. Run AutoGrid
3. Prepare the Ligand
   -- assign charges to atoms
   -- decide flexible bonds (run AutoTors)
4. Run AutoDock
5. Evaluate Results and Rank Score

Structure-Based Design: Modeling Thiazin derivatives

Example:

$$-13 \text{ kcal/mol}$$

$$3 \times 10^{-10}$$
<table>
<thead>
<tr>
<th>Compound</th>
<th>Case</th>
<th>IC50 (uM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Compound Image]</td>
<td>JMF310</td>
<td>10</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF311</td>
<td>4</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF312</td>
<td>30</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF313</td>
<td>16</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF314</td>
<td>37</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF315</td>
<td>21</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF316</td>
<td>17</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF317</td>
<td>18</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF318</td>
<td>23</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF319</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF320</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF321</td>
<td>20</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF322</td>
<td>15</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF323</td>
<td>33</td>
</tr>
<tr>
<td>![Compound Image]</td>
<td>JMF309</td>
<td>&gt; 50</td>
</tr>
</tbody>
</table>
Screening by Molecular Docking
Application Characteristics

- Computational screening based on molecular docking is the most time consuming part in structure-based drug design workflow.

- The requirement of CPU power and storage space increases proportional to the number of compounds and target proteins involved in the screening.
  \[
  \text{Number of docking tasks} = N \times M
  \]
  - \(N\): number of compounds
  - \(M\): number of target structures

The Challenge

- CPU-bound application, huge amount of output, no communication between tasks.
Grid-Enabling Virtual Screening
DCI: against Influenza A
EuAsiaGrid: Dengue virus


http://en.wikipedia.org/wiki/Aedes

Replication cycle of Flaviviridae

1. Entry into host cells
2. Receptor-mediated endocytosis
3. Fusion/uncoating
4. Translation & processing
5. Membrane associated RNA replication
6. Virion morphogenesis
7. Virion maturation
8. Virion release
Molecular docking method is commonly used to predict potential interacting complexes of a small molecule and a target protein. Using molecular docking method for compound screening purpose, however, is restricted by the availability of computing resources. In this work, Grid Application Platform (GAP) and GVSS Virtual Screening Service (GVSS) were developed to enable users to get access to the Grid technology and worldwide-scale computing resources seamlessly. Working with production e-infrastructures (such as EGEE and EUAsiaGrid), GVSS presents intensive computing power and effective data management, which provides opportunities for in-silico drug discovery on the neglected and emerging diseases, for instance, Avian Influenza and Dengue Fever.

Pitfalls
- In the molecular docking based screening, the requirement of CPU power and storage space increases proportional to the number of compounds (N) and target proteins (M).
  \[ \text{Number of docking tasks} = N \times M \]
- CPU-bound application, huge amount of output, no communication between tasks
- Task complexity is unpredictable

on Grid side
- difficult to apply trivial domain decomposition method in splitting the tasks
  - Need extra works to manage the efficient job handling and result gathering
  - Need efforts to handle transient network or site problems
  - Need application oriented GUI to hide Grid complexities from end users.
- Due to the significant system overhead:
  - Grid only benefit to those jobs with long computing time.
  - not suitable for pilot jobs (required for decision making).

Grid Application Platform (GAP)
a high level grid application framework developed by ASGC

DIANE/-autoDock framework
- A lightweight framework for parallel scientific applications in master-worker model.
- The framework takes care of all synchronization, communication, and workflow management details on behalf of applications.

GVSS Summary
- Grid-enabled Virtual Screening Service (GVSS), incorporating the docking engine of the Autodock 3.0.5, was developed on Grid Application Platform (GAP).
- GVSS uses the DIANE framework to take control of Grid failures and isolate Grid system latency, leading to efficiency, stability and usability.
- Thanks to DIANE framework and GAP, the GVSS has the ability to efficiently utilize the available Grid resource, which allows submitted jobs to be split into multiple independent subtasks and run to complete.
- GVSS includes JAVA-based graphical user interface, which allows end-users to specify target and compound library, set up docking parameters, monitor docking jobs and computing resources, visualize and refine docking results, and finally download the final results. GVSS hides the complexity of deploying large-scale molecular docking on Grid while provides users more flexible control over their works on Grid.
- Two applications demonstrated that modeling compound-protein complexes can be speeded up by distributing molecular docking processes on production Grids. Large-scale compound library can therefore be effectively enriched by executing docking tasks on Grid.

References

Acknowledgements
- ASGC/Taiwan
- His Kai Wang
- Mason Hsiung
- Li-Yung Ho
- Hung-Chun Lee
- Wei-Long Ueng
- Hsiung Yen Chen
- Eric Yen
- Simon Lin
- LCG-ARDA, CERN
- Jakub Moscicki
- Massimo Lamanna
- Supports from Genomics Research Center, Academia Sinica, Taiwan
- Academia Sinica, Taiwan
- Genomics Research Center, Academia Sinica, Taiwan
- Taiwan
Informatics Implications

• Need to be able to store chemical structure and biological data for millions of datapoints
  – *Computational representation of 2D structure*

• Need to be able to organize thousands of active compounds into meaningful groups
  – *Use cluster analysis or machine learning methods to group similar structures together and relate to activity*

• Need to learn as much information as possible from the data (data mining)
  – *Apply statistical methods to the structures and related information*
post-screen data analysis

Extract structural and interacting information
To give suggestion:
• Lead structures
• Optimization strategy
Similarity Clustering
Prioritized Filter
Rescoring

Using a Prioritized filter
Control compounds

Applied to validate screening quality and decide the hit rate

2qwf (G20)  2qwe (GNA)  2qwh (G39)

1f8e (49A)  1f8c (4AM)  1f8b (DAN)
prioritized hits with controls
Optimization

1,5-dihydro-2H-pyrrol-2-one

Figure 1. Two strategies for structure-based molecule assembly from fragments. The solid line represents a ligand-binding pocket on the surface of a protein. (a) Sequential growth technique and (b) fragment-placing and linking. Adapted from Ref. [58].

Figure 6. The optimization of hits based on the structural information can be done in two ways: decoration, in which the analogs are identified or synthesized, and fusion, in which fragments binding to adjacent sites in the target are linked together.
Data Analysis

Evaluation of docked poses

<table>
<thead>
<tr>
<th>compd</th>
<th>( K_i ) (nM)(^\text{a} )</th>
<th>( EC_{50} ) (nM)(^\text{b} )</th>
<th>( CC_{50} ) (µM)(^\text{c} )</th>
<th>S.I.(^\text{d} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.90 (± 0.30)</td>
<td>31.3 (± 3.5)</td>
<td>&gt;100</td>
<td>&gt;3200</td>
</tr>
<tr>
<td>3(^e)</td>
<td>0.15 (± 0.02)</td>
<td>4.67 (± 0.68)</td>
<td>74 (± 5.7)</td>
<td>15800</td>
</tr>
<tr>
<td>13a</td>
<td>2.02 (± 0.25)</td>
<td>5.60 (± 1.2)</td>
<td>&gt;100</td>
<td>&gt;17800</td>
</tr>
<tr>
<td>13b(^e)</td>
<td>0.06 (± 0.01)</td>
<td>0.09 (± 0.02)</td>
<td>~5</td>
<td>~56000</td>
</tr>
</tbody>
</table>

\(^{a}\)Neuraminidase inhibition activity.
\(^{b}\)EC\(_{50}\) values for anti-influenza activity.
\(^{c}\)CC\(_{50}\) values for cytotoxicity.
\(^{d}\)S.I. = \( EC_{50} \) / \( CC_{50} \).

Shie et al. J. AM. CHEM. SOC. 2007, 129, 11892-11893
**Summary** Computational Screening for Drug Discovery

- As a complementary to bio-assay, computational modeling enable the evaluation of a large compound library.

Accuracy and validity of the computational method is much important than integration at present stage.

- Grid computation is proved to increase the screening speed enable complete screening of massive compounds.

- GUI make the use of Grid computing resources easier.
Bioactive Compound Profiling

Enable the prediction of potential synergic effects or side-effects.

Given a search compound…, return with potential interacting targets.
Acknowledgements

ASGC, Taiwan
Hsin-Yen Chen
Eric Yen
Dr. Simon Lin
His-Kai Wang
Mason Hsung*
Li-Yung Ho*
Hurng-Chun Lee*

Genomics Research Center, Academia Sinica
National Science Council, Taiwan