Integration Of Machine Learning Approaches In Avian Flu Drug Discovery Workflows

Wilfred W. Li, Ph.D.

MURPA 2009
Calit2, April 22, 2009; Melbourne, April 23, 2009

National Biomedical Computation Resource
Center for Research in Biological Systems
San Diego Supercomputer Center
University of California, San Diego
Reported human cases of avian influenza (H5N1) infection

- 6 countries with continuous cases totals 356
- 15 countries with 412 cases to date
- Age ranged 3 months to 81 yr, 90% <39 yr
- Death rate varies by country, average of 63%, highest 83%
- People 10-19 years affected most; people > 50yr least affected.
- Gender neutral

Jan 13, 2009 statistics
Viral Replication Life Cycle

http://www.reactome.org/
http://www.wikipedia.org
http://library.thinkquest.org/05aug/01479/prevention1.html
## Influenza proteome and crystallome

<table>
<thead>
<tr>
<th>Protein</th>
<th>Selected functions</th>
<th>Crystal structural or NMR info</th>
<th>PDB ID and References</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA (Hemagglutinin)</td>
<td>Glycan receptor binding, membrane fusion</td>
<td>H5 trimeric complex</td>
<td>H5: 2FK0, 2IBX, 2ILF</td>
</tr>
<tr>
<td>NA (Neuraminidase)</td>
<td>Cleavage of SIA terminal residue, release of viral particles</td>
<td>N1 tetrameric complex</td>
<td>N1: 2HTY, 2HU0, 2L58</td>
</tr>
<tr>
<td>M2 (proton channel)</td>
<td>Uncoating of viral envelope in endosome</td>
<td>NMR or crystal structure of transmembrane domain in complex with amantadine</td>
<td>M2 transmembrane domain: 2RLF, 3B8D</td>
</tr>
<tr>
<td>NS1 (Nonstructural protein 1)</td>
<td>Host immune response modulation</td>
<td>RNA binding domain (RBD) and effector domain (ED) crystallized separately or in complex</td>
<td>NS1 RBD: 1AIL, 1AI2; NS1 ED: 2GS9, 2H5N1, NS1: 3FST</td>
</tr>
<tr>
<td>M1 (matrix protein)</td>
<td>Formation of ribonucleoprotein complexes with viral RNA</td>
<td>Two domains separated by linker region</td>
<td>M1 N-terminal domain: 1AA7</td>
</tr>
<tr>
<td>NS2, NEP (Nonstructural protein 2, nuclear export protein)</td>
<td>Nuclear export of viral ribonucleoproteins</td>
<td>IED available</td>
<td>NS2 IED: 1PD3, 2M9</td>
</tr>
<tr>
<td>NP (nucleoprotein)</td>
<td>Formation of viral capsid and packaging of RNA</td>
<td>Full length</td>
<td>NP: 2JQH, 2N7</td>
</tr>
<tr>
<td>PA (acidic protein)</td>
<td>Endonuclease and cap snatching</td>
<td>N-terminal domain; C-terminal domain bound to PB2</td>
<td>PA N-terminal domain: 2W69, 2EPA; PA C-terminal domain: 2ZN1, 2ZNL</td>
</tr>
<tr>
<td>PB1 (basic protein 1)</td>
<td>Polymerase catalytic subunit</td>
<td>In complex with PA</td>
<td>PB1-N-terminal domain: 2ZNL</td>
</tr>
<tr>
<td>PB1-F2 (basic protein 1 frame 2)</td>
<td>Pro-apoptosis</td>
<td>NMR structure</td>
<td>PB1-F2: 2HN8, 2FQ</td>
</tr>
<tr>
<td>PB2 (basic protein 2)</td>
<td>Nuclear import of RNA; capped RNA recognition</td>
<td>NMR and crystal structure in complex with importin</td>
<td>PB2 C-terminal domain: 2IDQ, 2CW4</td>
</tr>
</tbody>
</table>

http://www.pdb.org
Points of Intervention in Viral Infectious Cycle

- Adsorption
- Receptor containing sialic acid
- Endocytosis and fusion
- Uncoating
- RNA (+/-)
- mRNA
- GTP supply
- RNA polymerase inhibitors
- Neuraminidase inhibitors
- Glc5
- Gal4
- Nac3
- Gal2
- Sia1
- HA Binds to naturally occurring glycan receptors

De Clercq, Nat Rev, 2006
Xu et al, JMB, 2009

α-2,3 linked α-2,6 linked

De Clercq, Nat Rev, 2006
<table>
<thead>
<tr>
<th>Rank</th>
<th>NSC</th>
<th>Mean Energy</th>
<th>Predicted Kd (μM)</th>
<th>Chemical Structure</th>
<th>Binding Site</th>
<th>Apo Crystal Rank</th>
<th>Holo Crystal Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109836</td>
<td>-10.63</td>
<td>0.016</td>
<td><img src="image1.png" alt="Chemical Structure" /></td>
<td>SA-cavity</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>211332</td>
<td>-10.34</td>
<td>0.026</td>
<td><img src="image2.png" alt="Chemical Structure" /></td>
<td>SA-cavity</td>
<td>212</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>45583</td>
<td>-10.09</td>
<td>0.040</td>
<td><img src="image3.png" alt="Chemical Structure" /></td>
<td>SA-cavity 150-cavity 430-cavity</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>-</td>
<td>Oseltamivir</td>
<td>-9.82</td>
<td>0.063 (0.3 – 1.0)</td>
<td><img src="image4.png" alt="Chemical Structure" /></td>
<td>SA-cavity</td>
<td>238</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>Zanamivir</td>
<td>-9.38</td>
<td>0.133 (0.5 – 2.5)</td>
<td><img src="image5.png" alt="Chemical Structure" /></td>
<td>SA-cavity</td>
<td>230</td>
<td>12</td>
</tr>
</tbody>
</table>

- Patent filed
- Cheng et al, JMC, 2008
Mechanism of influenza virus receptor binding specificity

- What is the determinant for HA binding specificity switch and its gain of human transmissibility?
- Traditional view: $\alpha_2$-3 and $\alpha_2$-6 linkage preference
- New Topological view: Cone-like, Umbrella-like

- How do glycan length, composition, topology and inducible conformation affect the HA-glycan binding?
- What are the effect of HA amino acid mutation and glycosylation?

- The nature of these questions is challenging because of the variety and complexity of HA (16 sub-types and evolving mutants) and glycans.

- Addressing these questions is essential for pandemics monitoring and prevention as well as HA-based drug discovery.

Comparative Molecular Dynamics Simulations of HA’s

HA Trimers: avian H3, H5 and swine H9, apo and LSTa/LSTc bound/unbound
Explicit solvent, 0.15M NaCl, pH 7.4
~350,000 atoms
NAMD2, Amber99SB/Glycam06 force field
1 fs timestep, PBC, PME

Xu, et al. JMB, 2008 (pending).
Visualization of HA-Glycan Interactions
Glycan composition and species specific binding of H1 A
Glycan topology and induced conformational changes

Table 4. Glycan conformational entropic changes between bound and free states.

<table>
<thead>
<tr>
<th></th>
<th>H3</th>
<th>H5</th>
<th>H9</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTa</td>
<td>-TAS_{CA}</td>
<td>1.59 (0.34)</td>
<td>0.28 (0.01)</td>
</tr>
<tr>
<td></td>
<td>-TAS_{2,3}</td>
<td>0.79 (0.18)</td>
<td>0.13 (0.03)</td>
</tr>
<tr>
<td>% Contribution</td>
<td>50%</td>
<td>46%</td>
<td>77%</td>
</tr>
<tr>
<td>LSTb</td>
<td>-TAS_{CA}</td>
<td>2.47 (0.80)</td>
<td>5.52 (0.11)</td>
</tr>
<tr>
<td></td>
<td>-TAS_{2,5}</td>
<td>1.39 (0.45)</td>
<td>3.55 (0.06)</td>
</tr>
<tr>
<td>% Contribution</td>
<td>56%</td>
<td>64%</td>
<td>38%</td>
</tr>
<tr>
<td>Δ(-TAS_{CA})</td>
<td>0.88 (0.43)</td>
<td>5.24 (0.06)</td>
<td>2.39 (0.82)</td>
</tr>
</tbody>
</table>

The percent contribution is -TAS_{2,3} or -TAS_{2,5} relative to -TAS_{CA}. Standard errors are given in parenthesis. All units are in kcal/mol. Temperature is 310K. Δ(-TAS_{CA}) is the difference between -TAS_{CA} of LSTc and LSTa bound to the same HA. Standard errors are given in parenthesis.
Transparent access of applications on Avian Flu Grid through middleware

- H5N1 related glycan conformation analysis using M*Grid and Glyco-M*Grid
- Relaxed Complex Method Molecular Dynamics Simulation Data Sets & Database
- Virtual Screening Data Sets & Database
- HPC Clusters, NBCR, TeraGrid, MHPCC
- PRAGMA Portal My WorkSphere CSF4 Server
- Virtual Directory Tree /gfs/$USER
- CNIC VSDB
- Konkuk Glyco-M*Grid
- NBCR CADD
- applications
- databases
- Zinc
- NAMD
- AutoDock
- Mr. Bayes
Ensemble-based Virtual Screening with Relaxed Complex

Receptor Ensemble
Receptor crystal structure (or homology model), apo or holo complex
MD GB-MD SMD High T MD TMD Accl. MD
Snapshot 10 ps
Reduction of structures
clustering RMSD QR manual selection

Ligand Ensemble
available N/A
ZINC NCI ACD org. synth.
Including explicit water molecules
ligand PDBs
AutoDock (full ligand flexibility)

Post-Processing
Set of docked complexes
AD4 MM-PBSA single step perturb. LIE FEP, TI
Experimental verification
Known test set
Cheminformatics
CPU cost, accuracy confidence level

NCI Diversity Set: 3.3 MB, 2000 compounds; Required at each site ZINC subset: 200,000. A few hundred MB

Multiple targets: HA, NA subtypes
Each target: 30~50 MD snapshots, 1~2 MB each

Simulation Data: hundreds of GB

Total data: ~5 TB per year in long term storage.
Each experiment is about 1 Petaflops accumulative in computation cost.
NBCR Opal Web Service Toolkit v1

Web Portals

Continuity

MGL Tools

Kepler

State Mgmt

Application Services

Security Services (GAMA)

Globus

DRMAA

Globus

Condor pool

SGE Cluster

PBS Cluster
Opal 2 for SaaS
Vision Based Grid Workflow Environment

Clementi et al, IEEE eScience 2008
Service as Software: Workflows based upon distributed services
CADD Pipeline
Typical Drug Discovery Process

IND: investigational new drug
**Phase I:** pharmacokinetics (ADME) and safety
**Phase II:** controlled trials for safety and efficacy
**Phase III:** Expanded trials involving hundreds to thousands
ADME: adsorption, metabolism, distribution and excretion
**NDA:** new drug application
**FDA:** federal drug administration

Reform: fast track and priority review – less than 1 year for FDA approval

http://www.netsci.org/Courseware/Drugs/Intro/
Informatics for Biology and Chemistry: bioinformatics and cheminformatics

<table>
<thead>
<tr>
<th>Drug-like</th>
<th>Lead-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>MW &lt; 500</td>
<td>MW &lt; 350</td>
</tr>
<tr>
<td>ClogP &lt; 5</td>
<td>ClogP &lt; 3.0</td>
</tr>
<tr>
<td>Hydrogen bond donors &lt; 5</td>
<td>Chemically stable</td>
</tr>
<tr>
<td>Hydrogen bond acceptors &lt; 10</td>
<td></td>
</tr>
<tr>
<td>Number of rotatable bonds ≤ 10</td>
<td></td>
</tr>
<tr>
<td>PSA ≤ 140Å²</td>
<td></td>
</tr>
<tr>
<td>Peptides not suitable</td>
<td>Non-substrate peptides suitable</td>
</tr>
<tr>
<td>Eliminate reactive functional groups, promiscuous inhibitors, and metabolically unstable compounds</td>
<td></td>
</tr>
</tbody>
</table>

Bunin et al, Chemoinformatics, 2007
**Molecular Descriptors, Finger Prints, Pharmacophores**

<table>
<thead>
<tr>
<th>Descriptor category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical properties</td>
<td>Molecular weight, logP(o/w)</td>
</tr>
<tr>
<td>Atom and bond counts</td>
<td>Number of nitrogen atoms, Number of aromatic atoms, Number of rotatable bonds</td>
</tr>
<tr>
<td>Pharmacophore features</td>
<td>Number of hydrogen bond acceptors, Sum of van der Waal surface areas of basic atoms</td>
</tr>
<tr>
<td>Charge descriptors</td>
<td>Total positive partial charge, Dipole moment from partial charges</td>
</tr>
<tr>
<td>Connectivity and shape descriptors</td>
<td>Kier and Hall molecular shape indices</td>
</tr>
<tr>
<td>Surface area and volume</td>
<td>Solvent-accessible surface area</td>
</tr>
</tbody>
</table>

**1D**

- \( \text{C}_{22}\text{H}_{32}\text{ClF}_{2}\text{N}_{12}\text{O}_{3} \) → Number of carbon atoms

**2D**

- Number of rotatable bonds
- \( \log(P(\text{o/w})) \)
- Molecular connectivity index

**3D**

- Solvent-accessible surface area
- Van der Waals volume
Information Representation and Similarity Search

Graph Reduction

3D pharmacophore representation

Substructure Search
Classification and Similarity Search

Decision Tree

\[
Tc(Mol1, Mol2) = \frac{16}{(19+18)-16} = 0.76
\]

\[
Tc(Mol1, Mol3) = \frac{11}{(19+13)-11} = 0.52
\]

Tanimoto coefficient
\[
Tc = \frac{n_{ij}}{n_i + n_j - n_{ij}}
\]

Dice coefficient
\[
Dc = \frac{2n_{ij}}{n_i + n_j}
\]

Cosine coefficient
\[
Cc = \frac{n_{ij}}{\sqrt{n_i n_j}}
\]

- \(n_{ij}\) is shared features between the two molecules
Cheminformatics Workflow For Virtual Screening

- Diversity-based selection
  - Screening
    - Hit?
      - yes
        - Statistical analysis
        - Predictive model
        - Compound analysis
        - Activity-based selection
          - Hit expansion
          - Hit to lead
    - no
      - Proprietary target 2 (belonging to the GPCR superfamily)
        - 300,000 synthesized compounds
          - HTS
            - 300,000 tested compounds
              - No validated hits
        - 4.5 M virtual compounds
          - Ligand-based VS
            - 16 of 30 tested compounds
              - One hit: IC_{50} ~ 5 μM
Tools to generate Molecular Descriptors

- **OpenEye Filter + Weka**
  - Generate 1D molecular descriptors on the known dataset which got from the Virtual Screening, and build classification model using Weka on the 1D molecular descriptors.

- **Molprint2D**
  - Generate 2D molecular descriptors and Molprint2D has its own classifier generating system which can classify the data.

- **OpenEye Rocs**
  - Generate 3D molecular descriptors, and ranking by the similarity depend on these 3D molecular descriptors.
Performance of three classification models built by three tools

- Lots of the experiments detailed results in the report, just highlight the final result here:
  - Sensitivity: active molecules are correctly classified as active molecules.
  - Selectivity: inactive molecules are incorrectly classified as active molecules.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Sensitivity (true positive)</th>
<th>Selectivity (false positive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenEye Filter + Weka (ADTree –B 15 –E -3)</td>
<td>94.2</td>
<td>0.111</td>
</tr>
<tr>
<td>Molprint-2D</td>
<td>97.4</td>
<td>0.06</td>
</tr>
</tbody>
</table>

- **OpenEye Rocs**
  - Top four hits in the ranking have the probability of 97.9% being same type with the active known inhibitors.
MURPA 2008 sets the standard
Avian Flu Grid: The A Team

Science Team:
• Irene Newhouse
• Rommie Amaro
• Dong Xu
• Youngjin Choi
• Hsing Pao
• Jung-Hsin Lin
• Habibah Wahab
• Macsudul Alam
• Doman Kim

Technology Team:
• Zhaohui Ding
• Guanyuan Liu
• Osamu Tatebe
• Yusuke Tanimura
• Jonghyun Lee
• Kai Nan
• Xiaohui Wei
• Suntae Hwang
• Aimee Li, MURPA

Principal Investigator:
Peter Arzberger
Wilfred Li

Advisory Committee:
• J. Andrew McCammon
• Arthur Olson
• Fang-Pang Lin
• Jysoo Lee
• Satoshi Sekiguchi

http://avianflugrid.pragma-grid.net