Elucidating Molecular Overlays from Pairwise Alignments

Gareth Jones, Yinghong Gao and Carleton Sage
DIFGAPE

- OE Tools
  - Omega, generates conformations
  - ROCS, rigid pairwise alignments

- In projects, frequently we have a set of actives and no protein structure:
  - Can use OMEGA to generate conformer libraries
  - Can use ROCS to align conformer pairs

- Unable to predict binding mode

- If we select conformers, can we use ROCS pairwise overlays to construct binding mode?

- Conformer selection highly combinatorial:
  - 7 ligands 50 conformers/ligand $51^7 = 9.0E11$
  - 10 ligands $51^{10} = 1.2E17$

- Therefore use GA to pick conformers
DIFGAPE is a Tool For Alignment

Input: 2D ligands without binding site information

2D Structures

Omega: 3D Conformation Library

ROCS: Pairwise Alignments

Conformer Selection Using GA

Build Alignment From the Selected Conformers

Output: 3D alignment

Crystal alignment

FXa example
Operator-based Genetic Algorithm

1. A set of reproduction operators (crossover, mutation etc) is chosen. Each operator is assigned a weight
2. An initial population is randomly created and the fitness’s of its members determined
3. An operator is chosen using roulette wheel selection based on operator weights
4. The parents required by the operator are chosen using roulette wheel selection based on scaled fitness
5. The operator is applied and child chromosomes produced. Their fitness is evaluated
6. The children replace the least fit members of the population
7. Repeat steps 3-6 for a suitable number of iterations
**Chromosome Encoding and Fitness Function**

*Chromosome Encoding:* An integer string is used to encode molecular conformations:

**Example: Chromosome \{2,13,0,9\}**

- conformer 2 for molecule 1
- conformer 13 for molecule 2
- no conformer being selected for molecule 3
- conformer 9 for molecule 4

*Possible Fitness function:*
For \( m \) molecules in the overlay
\( c_x \) is the selected conformer for molecule \( x \).

\[
\text{similarity\_score} = \frac{1}{np} \cdot \sum_{i=1}^{m} \sum_{j=i+1}^{m} \text{similarity}(c_i, c_j)
\]
Unable to overlay ligands from 1YGT, 4HVP and 1YGT in a manner consistent with the three pairwise alignments
Penalty based on Triangle Inequalities

- Triangle Distance Constraints: $|AB| \leq |BC| + |CA|$
- Incorporate penalty for violations of triangle inequality into GA fitness function.
- $\text{tri\_penalty}(A,B,C) = (|AB| - |BC| - |CA|)^2$ if the triangle constraint is broken or 0 otherwise.
- $a_{i,x}$ is hetero-atom number $i$ in molecular conformer $x$.

$$\text{penalty} = \frac{1}{nt} \cdot \sum_{x=1}^{m} \sum_{y=x+1}^{m} \sum_{z=y+1}^{m} \sum_{i=1}^{k_x} \sum_{j=1}^{k_y} \sum_{k=1}^{k_z} \text{tri\_penalty}(a_{i,x}, a_{j,y}, a_{k,z})$$

$$\text{fitness} = \text{similarity\_score} - \text{penalty}$$
It Works.

Without triangle penalty

With triangle penalty
Building the Overlay

- GA selects a set of molecular conformations
- Needs to create overlay from conformations
- Can create multiple overlays, one for each conformer using ROCS results
- Choose a base overlay using the best ROCS score and fit all other overlays onto it.
- Create averaged coordinates for each molecule
- If the averaged coordinates are distorted, build consensus coordinates by leaving out the most distant orientations
- Note that severe distortions of bond angles and lengths are indicative of a poor solution
Building the Overlay 1: CDK2 example.
Building the Overlay 2: CDK example
Consensus: HIV Example

1CPI

Average

Consensus

1YGT

Average

Consensus
Performance

• **Java 1.5**
  - Rocs Parser to serialized Java object
  - GA run

• **FXa_Focused**
  - 11 compounds
  - 1620 conformers
  - 2.5 million pairwise alignments
  - Linux 64 bit workstation
    - AMD Opteron 246 1GHz
    - 8GB RAM
  - Omega (50 conformers/structure) 43s
  - ROCS 22m
  - DIFGAPE 44m, 7.1GB
Method Validation

- Select different protein families with multiple known ligand bound crystal structures
- For each target select a set of active compounds
- Align proteins and extract crystallographically observed overlay
- Using ROCS (or another tool) create pair-wise alignments between the ligand binding conformations
- Run DIFGAPE in exhaustive search mode on the ligand binding conformations and pair-wise alignments
- Compare the DIFGAPE overlay with the crystallographically observed overlay. If the survival rate $\geq 0.5$ and the RMSD $< 2\AA$ then continue, otherwise stop here
- Using Omega (or another tool) create conformer libraries for the active compounds
  - Used MAXCONFS of 50
  - Evaluated 20, 50, 150
    - Saw little improvement at 150
- Using ROCS (or another tool) exhaustively create pair-wise alignments and scores for all pairs of conformers that do not contain the same structure
- Run DIFGAPE using GA search on the conformer libraries and associated pair-wise alignments
- Compare the DIFGAPE overlay with the crystallographically observed overlay. If the survival rate $\geq 0.5$ and the RMSD $< 2\AA$ then the experiment is considered a success
## Targets and Datasets

<table>
<thead>
<tr>
<th>Protein target</th>
<th>Protein Family</th>
<th>Subset</th>
<th>No. of Ligands</th>
<th>MACCS_Avg</th>
<th>MW_Avg</th>
<th>b_rotN_Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2</td>
<td>Transferase (Kinase)</td>
<td>CDK2_Focused</td>
<td>9</td>
<td>0.49</td>
<td>352</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CDK2_Diverse</td>
<td>10</td>
<td>0.39</td>
<td>322</td>
<td>4.5</td>
</tr>
<tr>
<td>DHFR</td>
<td>Oxidoreductase</td>
<td>DHFR</td>
<td>12</td>
<td>0.60</td>
<td>394</td>
<td>10.7</td>
</tr>
<tr>
<td>Elastase</td>
<td>Hydrolase (serine Proteinase)</td>
<td>Elastase</td>
<td>5</td>
<td>0.70</td>
<td>468</td>
<td>16.6</td>
</tr>
<tr>
<td>ESR1</td>
<td>Hormone Receptor</td>
<td>ESR1</td>
<td>13</td>
<td>0.43</td>
<td>358</td>
<td>6.9</td>
</tr>
<tr>
<td>FXa</td>
<td>Hydrolase</td>
<td>FXa_Focused</td>
<td>11</td>
<td>0.83</td>
<td>541</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FXa_Diverse</td>
<td>8</td>
<td>0.51</td>
<td>482</td>
<td>9.9</td>
</tr>
<tr>
<td>HIV1 protease</td>
<td>Hydrolase (Acid Protease)</td>
<td>HIV_Div</td>
<td>13</td>
<td>0.52</td>
<td>612</td>
<td>24.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HIV_Div_MW</td>
<td>8</td>
<td>0.48</td>
<td>519</td>
<td>19.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HIV_Div_MW_RB</td>
<td>4</td>
<td>0.47</td>
<td>486</td>
<td>14.0</td>
</tr>
<tr>
<td>p38</td>
<td>Transferase (Kinase)</td>
<td>p38</td>
<td>12</td>
<td>0.47</td>
<td>377</td>
<td>4.5</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>Virus</td>
<td>Rhinovirus</td>
<td>8</td>
<td>0.80</td>
<td>350</td>
<td>11.1</td>
</tr>
<tr>
<td>Trypsin</td>
<td>Hydrolase</td>
<td>Trypsin</td>
<td>7</td>
<td>0.51</td>
<td>174</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Results Using Crystal Structure Bound Conformations

<table>
<thead>
<tr>
<th>DataSet</th>
<th>RMSD(Å)</th>
<th>No. of Input Ligands</th>
<th>No. of Output Ligands</th>
<th>Survival Rate</th>
<th>Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2_Focused</td>
<td>0.58</td>
<td>9</td>
<td>5</td>
<td>0.56</td>
<td>yes</td>
</tr>
<tr>
<td>CDK2_Diverse</td>
<td>2.81</td>
<td>10</td>
<td>3</td>
<td>0.30</td>
<td>no</td>
</tr>
<tr>
<td>DHFR</td>
<td>0.46</td>
<td>12</td>
<td>7</td>
<td>0.58</td>
<td>yes</td>
</tr>
<tr>
<td>Elastase</td>
<td>1.49</td>
<td>5</td>
<td>4</td>
<td>0.8</td>
<td>yes</td>
</tr>
<tr>
<td>ESR1</td>
<td>0.34</td>
<td>13</td>
<td>7</td>
<td>0.54</td>
<td>yes</td>
</tr>
<tr>
<td>FXa_Focused</td>
<td>0.39</td>
<td>11</td>
<td>10</td>
<td>0.91</td>
<td>yes</td>
</tr>
<tr>
<td>FXa_Diverse</td>
<td>1.04</td>
<td>8</td>
<td>6</td>
<td>0.75</td>
<td>yes</td>
</tr>
<tr>
<td>HIV_Div_MW_RB</td>
<td>0.3</td>
<td>4</td>
<td>2</td>
<td>0.50</td>
<td>yes</td>
</tr>
<tr>
<td>HIV_Div_MW</td>
<td>0.67</td>
<td>8</td>
<td>3</td>
<td>0.38</td>
<td>no</td>
</tr>
<tr>
<td>HIV_Div</td>
<td>2.81</td>
<td>13</td>
<td>4</td>
<td>0.31</td>
<td>no</td>
</tr>
<tr>
<td>P38</td>
<td>1.04</td>
<td>12</td>
<td>4</td>
<td>0.33</td>
<td>no</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>4.56</td>
<td>8</td>
<td>6</td>
<td>0.75</td>
<td>no</td>
</tr>
<tr>
<td>Trypsin</td>
<td>1.12</td>
<td>7</td>
<td>5</td>
<td>0.71</td>
<td>yes</td>
</tr>
</tbody>
</table>
Rhinovirus

2 of 6 ligands reversed by DIFGAPE

Crystal Structure

DIFGAPE using bound conformers
Crystal Structure

DIFGAPE using bound conformers
## Results Using Conformer Libraries

<table>
<thead>
<tr>
<th>DataSet</th>
<th>RMSD(Å)</th>
<th>No. of Input Ligands</th>
<th>No. of Output Ligands</th>
<th>Survival Rate</th>
<th>Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2_Focused</td>
<td>1.81</td>
<td>9</td>
<td>6</td>
<td>0.67</td>
<td>yes</td>
</tr>
<tr>
<td>DHFR</td>
<td>2.71</td>
<td>12</td>
<td>7</td>
<td>0.58</td>
<td>no</td>
</tr>
<tr>
<td>Elastase</td>
<td>3.23</td>
<td>5</td>
<td>5</td>
<td>1.00</td>
<td>no</td>
</tr>
<tr>
<td>ESR1</td>
<td>2.46</td>
<td>13</td>
<td>9</td>
<td>0.69</td>
<td>no</td>
</tr>
<tr>
<td>FXa_Focused</td>
<td>1.61</td>
<td>11</td>
<td>11</td>
<td>1.00</td>
<td>yes</td>
</tr>
<tr>
<td>FXa_Diverse</td>
<td>1.84</td>
<td>8</td>
<td>8</td>
<td>1.00</td>
<td>yes</td>
</tr>
<tr>
<td>HIV_Div_MW_RB</td>
<td>3.54</td>
<td>4</td>
<td>3</td>
<td>0.75</td>
<td>no</td>
</tr>
<tr>
<td>Trypsin</td>
<td>1.47</td>
<td>7</td>
<td>5</td>
<td>0.71</td>
<td>yes</td>
</tr>
</tbody>
</table>

P38 does as well as previously.
Survival rates and RMSD increased

• Improved survival rate:
  - Given a large number of conformations, likely to find a single conformation to fit in the overlay

• Worse RMSD
  - Due to improved survival rate. Extra conformations are likely to be wrong
  - Ligand binding mode may not be in conformational library.
  - Binding mode may be present, but GA does not find it
  - GA finds a false optimum, due to some artifact of the alignment and scoring functions
FXa

Crystal structure

DIFGAPE on Bound conformers

DIFGAPE on Conformer libraries

FXa
Focused

FXa
Diverse
Crystal structure:
Ligands fall into 2 classes

- DIFGAPE Fails to incorporate first class
- Three ligands 1SJO, 1XP1 and 1XP9 misaligned- OMEGA cannot create bioactive conformation
Other Systems

• DHFR
  - Works with 20 conformers/structure (1.3Å RMSD)
  - Lack of conformational energy term in scoring function may be an issue

• Elastase
  - Large peptidic compounds
  - Binding mode of 1ELC not in conformer library
  - 1ELB, 1ELC have unusual binding modes
  - 1ELC not included in bound conformers experiment
  - 1ELC and 1ELB present in solution, in incorrect conformations

• HIV_Div_MW_RB
  - Large peptidic compounds
  - Conformationally challenging
  - Binding modes not present in Omega conformations
Closest Omega Conformation to Crystal Structure

<table>
<thead>
<tr>
<th>RMSD Range (Å)</th>
<th>Omega MAXCONFS Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td>0-1</td>
<td>78</td>
</tr>
<tr>
<td>1-2</td>
<td>28</td>
</tr>
<tr>
<td>2-3</td>
<td>9</td>
</tr>
<tr>
<td>3-4</td>
<td>3</td>
</tr>
<tr>
<td>&gt;4</td>
<td>0</td>
</tr>
</tbody>
</table>

- ESR1: 1SJO, 1XPI and 1XP9
- Elastase: 1ELC (2.2Å)
- HIV_Div_MW_RB: all Omega conformations > 1Å from crystal structure
- 31% at MAXCONFS 50 have the closest OMEGA conformation > 1Å
Validation Summary

• 13 test systems
• 5 failed to create reasonable overlays from ligand binding modes
• 8 tested from 2D structures
• 3 failed due to OMEGA not including binding mode in conformer library (ESR1, Elastase, HIV_Div_MW_RB)
• Algorithm failed on DHFR
• 4 succeeded (FXa_Focused, FXa_Diverse, Trypsin and CDK2_Focused)
Conclusions

• Extremely difficult to predict binding modes from 2D structures without protein structure
• Partially successful in addressing this problem
• Final results dependent on conformer generator and alignment tool
• Algorithm improvements
  – Conformational energy term
  – Process multiple ROCS alignments per conformer pair
• Paper written, all test systems and results will be supplementary material
• Acknowledgements
  – Carleton Sage, Yinghong Gao
  – Paul Watson