Navigating high-throughput docking results

Keana Scott
Noel Southall, Trung Nguyen
Lalit Verma, Boris Fain, Ajay

Celera Genomics

228th ACS National Meeting, Philadelphia, PA
August 25, 2004
High-throughput docking

• Results are often subjected to strict filters
  - based on multiple scoring functions
  - applied in linear fashion to reduce the docked poses to a manageable number for visual inspection

• Don’t throw out the baby with the bathwater

• Leverage modeler’s intuition
  - Don’t make harsh assumptions about where to look
  - Let modelers navigate the result space
  - Identify diverse and interesting candidate hypotheses
In-house pipeline

- High-throughput docking issues
  - Very large data set
    - 10Ks library cmpds x 100s conformers ➞ 1Ms poses
  - Scripts
    - To parse outputs
    - To keep track of multiple run data

But what do you do with millions of poses?
Filters

- 3rd party and custom score functions
- Ligand specific interactions:

Going beyond ligand specific interactions

- Other possible interaction points on protein – other binding modes
- Cavity volume filled
- Poisson-Boltzmann electrostatics treatment
Tried and true(?) approach

> 1M structs

Scoring functions

Keep top 5%

Filter 1

Filter 2

Filter n

~ 100’s structures

Hopefully, with minimal loss of true hits
Are we seeing all the interesting things to see?

- We should be asking
  - What about others - false negatives?
  - How are these regions related?
  - What can we glean from various inter-relationships?

- Don’t make cuts based on pre-conceived ideas

- Figure out how best to explore, collect, and organize patterns from very large result sets
Good analysis tool should provide

- Access to complete picture of the pose space
  - No up-front elimination of results
- Means to explore patterns in the data
  - Easy manipulation of results
    - sorting
    - algebraic and logical manipulations
    - comparison
    - grouping
    - chemistry aware tools
  - Easy sub-setting and persistence of analysis

Let scientists play with the data.
Leverage modeler’s intuition.
Every scientist’s favorite tool

• **Spreadsheet**
  - Move columns around
  - Sort values
  - Perform basic statistical and algebraic manipulations
  - Plot

• But spreadsheets usually lack:
  - Very large data support
  - Visualization
  - Chemistry awareness (unless using plug-in modules)

• And these issues are not limited to docking
SEURAT data browser

Usual spreadsheet features, plus...

- Large data support with Oracle backend
  - No row/column limitation
  - Quick data manipulation
  - Access to docking information
    - Result summary and statistics
    - User annotation
    - Setup parameters
- Visualization with PyMOL
- Chemistry awareness
SEURAT data browser

Usual spreadsheet features, *plus*...

- Large data support with Oracle backend
- Visualization with PyMOL
  - Fully functional PyMOL
    - Many thanks to Warren DeLano
    - Protein visualization
    - Programmatically controllable
    - Freely available
  - Java wrapper around PyMOL
  - Synchronization between PyMOL and Data Browser
- Chemistry awareness
SEURAT data browser

Usual spreadsheet features, plus…

- Large data support with Oracle backend
- Visualization with PyMOL
- Chemistry awareness
  - Ordering/comparison by structural similarity
  - Sub-structure search
  - Handling of non-scalar data
    - > 0.5
    - 10 – 55
Data browser overview

fully functional PyMOL

perform logical operations on columns

select a docking run to analyze

visual inspection results from PyMOL

selected poses are displayed in PyMOL
Data browser
Data browser
Mode generation

mode 2: poses that form polar interaction 1 and occupy hydrophobic pocket
Data browser

\[ \sum \text{mode1}: k > 0 \text{ and } g > 0 \text{ and } g < 4.0 \text{ and } l > 0 \]

<table>
<thead>
<tr>
<th>SCORE_1</th>
<th>SCORE_2</th>
<th>HEURID_1</th>
<th>HEURID_2</th>
<th>HEURID_3</th>
<th>HEURID_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.5</td>
<td>3.546</td>
<td>514.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.17</td>
<td>3.816</td>
<td>470.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.15</td>
<td>3.724</td>
<td>490.708</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.50</td>
<td>3.73</td>
<td>480.048</td>
<td>2.847</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.58</td>
<td>5.453</td>
<td>510.151</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.82</td>
<td>4.031</td>
<td>501.857</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.48</td>
<td>4.25</td>
<td>435.021</td>
<td>4.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.48</td>
<td>2.251</td>
<td>457.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.48</td>
<td>2.25</td>
<td>435.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.42</td>
<td>1.406</td>
<td>434.021</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data browser

![Data browser interface](image)

- **Dock Run:** 4c0x fixed with chemscore
- **Description:** chemscore
- **Name:** 4c0x fixed with chemscore
- **Parameters:** chemscore
- **Project Name:** Project 2
- **Protein ID:** 109
- **Protein Name:** 4c0x_protein_clean.mol2

**New model column** highlighted in red.
Data browser

![Data browser interface with 5 binding mode hypotheses highlighted.](image)
poses we want to visually inspect
Visualization with PyMOL
## Conclusions

<table>
<thead>
<tr>
<th>problem</th>
<th>Identify diverse and interesting set of candidate molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>approach</td>
<td>Take advantage of scientists’ intuition - navigate the pose landscape and explore, identify, collect, and organize patterns and generate binding mode hypotheses</td>
</tr>
<tr>
<td>results</td>
<td>several modes of interactions not anticipated from the crystal structure</td>
</tr>
<tr>
<td>lessons learned</td>
<td>Ability to explore, collect, and organize hypotheses in the result space is crucial in discovering interesting and unusual binding modes</td>
</tr>
</tbody>
</table>
Acknowledgments

Celera:

Anand Basu
Paul Sprengeler  Ken Brameld  Richard Xie
Ryan Weekley  John Eksterowicz
Jim Janc

IBM:

Lawrence Hannon