Protein Folding Simulation and Virtual Screening of Dual Specificity Phosphatase in Parallel

Charles Xue
Professors Jason Haga and Susumu Date
Osaka University, Japan
August 20\textsuperscript{th}, 2010
Final Presentation
Table of Contents

• Summary of Project
  – Docking
  – Modelling

• Methods
  – Dock methods (brief)
  – Modeller methods

• Data/Results
  – Consensus ranking for protein 1M3G
  – Folded model of 2NT2
  – Folded model of DUSP1

• Discussion
  - Significance of Results
  - Future Work

• Cultural

• Acknowledgements
DOCK6

- Program utilizes force field in order to simulate ligand-receptor binding to determine the best candidates for enzyme inhibitors.

- Goal of project is to screen over 20,000 compounds within the ZINC database against the enzyme in parallel on the PRAGMA grid.

- Tested with both grid-based score and Amber.
Project Summary Cont.

• Protein folding (*Ab initio*)
  – Utilizes computer algorithms to simulate natural forces in order to obtain a result.

• Protein folding (homology)
  – Utilizes similar known structures in order to predict the model of an amino acid sequence
Project Summary Cont.

MODELLER9v8

- Program utilizes homology modelling algorithms to determine best tertiary structure of an amino acid sequence.

- Requires proteins with known configurations and similar sequences, ideal for the dual specificity phosphatase family.

- Goal is to implement the entire folding process and loop refinement in parallel on the PRAGMA grid.
Methods/Procedures

DOCK6

- Run Dock6 utilizing the built in mpi’s and scripts written by Marshal Levesque.

- Separate the thousands of compounds into slices and run each slice independently on clusters in the grid.

- Compile final results and organize based on energy score.
Why Modeller9v8 – Alternatives?

- Homology modelling is fastest
- DSP family has a high level of structural similarity
- Modeller offers built in parallel support

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homology</td>
<td>Evolutionary information recognition</td>
<td>Webserver server</td>
</tr>
<tr>
<td>MgenTHREADER</td>
<td>Sequence profile and predicted secondary structure</td>
<td>Webserver main page</td>
</tr>
<tr>
<td>LOOPP</td>
<td>Multiple methods</td>
<td>Webserver server</td>
</tr>
<tr>
<td>MUSTER</td>
<td>Profile-profile alignment</td>
<td>Webserver server</td>
</tr>
</tbody>
</table>

3. Ab initio structure prediction

<table>
<thead>
<tr>
<th>Name</th>
<th>Method</th>
<th>Description</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-TASSER</td>
<td>Combination of ab initio folding and threading methods</td>
<td>Structural and function predictions</td>
<td>Main page</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>Rosetta homology modeling and ab initio fragment assembly with CIPsys domain prediction</td>
<td>Webserver server</td>
<td>No</td>
</tr>
<tr>
<td>Bhageerath</td>
<td>A computational protocol for modeling and predicting protein structures at the atomic level</td>
<td>Webserver Server</td>
<td>No</td>
</tr>
</tbody>
</table>

4. Secondary structure prediction

<table>
<thead>
<tr>
<th>Name</th>
<th>Method</th>
<th>Description</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>NetSurfP</td>
<td>Profile-based neural network</td>
<td>Webserver server</td>
<td>No</td>
</tr>
<tr>
<td>GOR</td>
<td>Information theory/Bayesian inference</td>
<td>Many implementations</td>
<td>Basic GOR GOR V</td>
</tr>
<tr>
<td>PrediProtein</td>
<td>Profile-based neural network</td>
<td>Webserver server</td>
<td>No</td>
</tr>
</tbody>
</table>

5. Transmembrane helix and signal peptide prediction

<table>
<thead>
<tr>
<th>Name</th>
<th>Method</th>
<th>Description</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMTOP</td>
<td>Hidden Markov Model</td>
<td>Webserver/standalone main page</td>
<td>No</td>
</tr>
<tr>
<td>MEMSAT</td>
<td>Neural networks and SVMs</td>
<td>Webserver/standalone main page</td>
<td>No</td>
</tr>
<tr>
<td>PSIpred</td>
<td>Multiple alignment-based neural network system</td>
<td>Webserver/standalone server</td>
<td>No</td>
</tr>
</tbody>
</table>
Methods/Procedures Cont.

MODELLER9v8
(Fit_distribute.pl/modrun.pl)
(getbest.pl/getbestlocal.pl)
- Develop script for Modeller to submit jobs with sge based on its built in task-based interface.

- Split the models into slices and create an array which stores the data of each slice.

- Compile the final results, grabbing on the lowest DOPE and molpdf scores.
MODELLER9v8  
(loop_distribute.pl/looprun.pl)

• Develop script to refine individual loop segments from the result of fit_distribute.pl.

• Separate the various segments into slices and create an array to hold the data of each slice.

• Retrieve best model from each segment and utilize that model for the next one.

• Save only the final model with the lowest DOPE score.

Protein before the loop refinement, note the extended loop in the circled section

Protein after the loop refinement, the loop has been re-simulated to better resemble actuality
Results

DOCK6 (DSP2 aka 1M3G)

<table>
<thead>
<tr>
<th>Rank</th>
<th>ZINC ID</th>
<th>PubChem ID</th>
<th>ZINC ID</th>
<th>PubChem ID</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ZINC02352689</td>
<td>6 3 3</td>
<td>-93.692276</td>
<td>-142535802880.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ZINC06645917</td>
<td>48 43 5</td>
<td>-46.682549</td>
<td>-41836152.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ZINC06645918</td>
<td>66 65 1</td>
<td>-45.843516</td>
<td>-510907777024.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ZINC06645916</td>
<td>331 327 4</td>
<td>-42.563263</td>
<td>-3793709312.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ZINC05047674</td>
<td>596 359 237</td>
<td>-42.357357</td>
<td>-37.623665</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ZINC06645550</td>
<td>606 228 378</td>
<td>-43.194515</td>
<td>-34.503338</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ZINC02649005</td>
<td>666 312 354</td>
<td>-42.631413</td>
<td>-35.088120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>ZINC02921257</td>
<td>835 426 409</td>
<td>-42.014664</td>
<td>-34.038483</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ZINC03457610</td>
<td>848 293 555</td>
<td>-42.763611</td>
<td>-32.087373</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ZINC01052992</td>
<td>945 351 594</td>
<td>-42.402878</td>
<td>-31.573633</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results Cont.

MODELLER9v8 (SSH2 aka 2NT2)

Plot of molecule before loop refinement

Plot of molecule after loop refinement

Plot of actual SSH2 protein
Results Cont.

MODELLER9v8 (DUSP1)

Lowest DOPE score before loop refinement

Lowest DOPE score after loop refinement
Discussion

Significance of Project (DOCK):

– A viable in vivo inhibitor of a selected enzyme must not also inhibit enzymes of the same family.

– Testing in wet-bench conditions are costly and time consuming.

– Dock simplifies the situation by narrowing down the range of compounds to test for.

– Running parallel further speeds up the process to a suitable time range.
Discussion Cont.

Significance of Project (MODELLER):

- X-Ray crystallography and NMR spectroscopy to determine the true structures of proteins at a suitable resolution is extremely limited by supply.

- Accurate folded proteins *in silico* is of high demand and many projects started for that purpose (i.e. FoldingAtHome).

- Since structure of protein determines function, knowledge of the structure is far more valuable than simply knowledge of sequence.

- Unfortunately, it is very unforgiving of small deviations, causing processing to take a long time and be very precise, ideal for parallel computing.
Discussion Cont.

Significance of Project (DOCK with MODELLER):

- Many members of the DSP family (of which SSH belongs to) have not had a suitable structure determined.

- Proper screening requires that all members of the family be thoroughly screened.

Docking with a folded protein structure also has applications beyond that of the DSP family (other proteins with unknown structures, synthetic/altered proteins, etc.)
Future Work

• Continue to screen proteins of the DSP family, starting with the first simulated structure determined.

• Improve efficiency of program, utilizing hash as opposed to arrays.

• Wet bench work, to test the viability of the screened ligands and determine best inhibitor
Cultural Experiences Around Osaka
Cultural Experiences Around Kyoto
Cultural Experiences Other Places
Acknowledgements

• Thanks to all the labs:
  – Shimojo Lab
  – Takemura Lab
  – Fujiwara Lab

• Special thanks to:
  – Dr. Haga, Date-sensei, Ichikawa-sensei, Kokubo-San, and Matt for advice and help.
  – Marshall Levesque for his previous scripts.

• Finally, thanks to UCSD PRIME for a memorable trip!