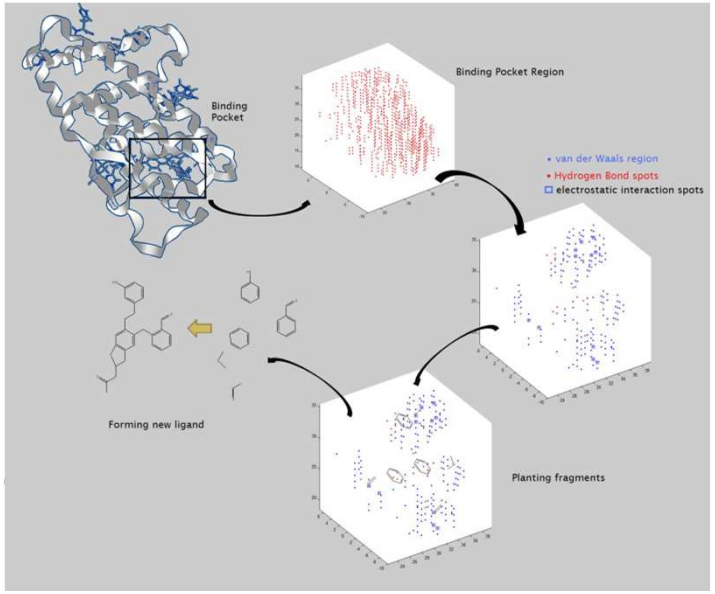
Homology Modeling and Virtual Screening of Dual Specificity Phosphatase Inhibitors

Daniel Li, Brian Tsui
August 25, 2011
PRIME 2011
Osaka University, Cybermedia Center
University of California San Diego, Department of Bioengineering

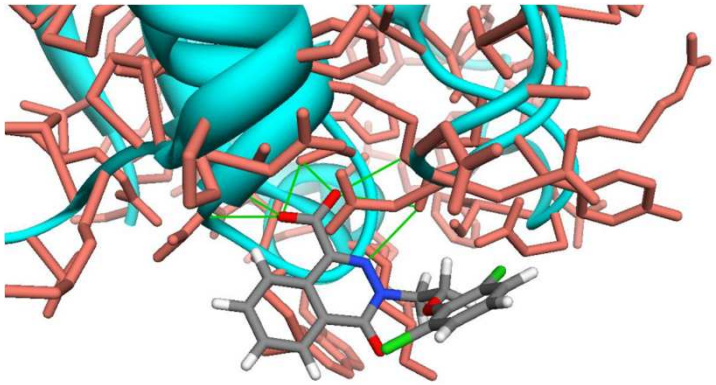
Background: Drug Design

- Drug design is based on how well a ligand (binding material) binds to a receptor (binding target).
- Structure based drug design: experimental structure of a target is predetermined and ligands' binding affinities are calculated



Background: Dock Virtual Screening

- Large scale wet lab testing of ligands is inefficient
- Virtual screening quickly determines approximate energy functions of ligand-receptor interactions



- Simulated docking experiments are much faster and allows scientists to choose only the top results to be tested in wet lab

Background: Homology Modeling

- Construction of a protein model from an amino acid sequence using experimentally determined structures of closely related proteins
- Main errors include sequence alignment and low sequence identity

Background: The Dual Specificity Phosphatase (DSP) family

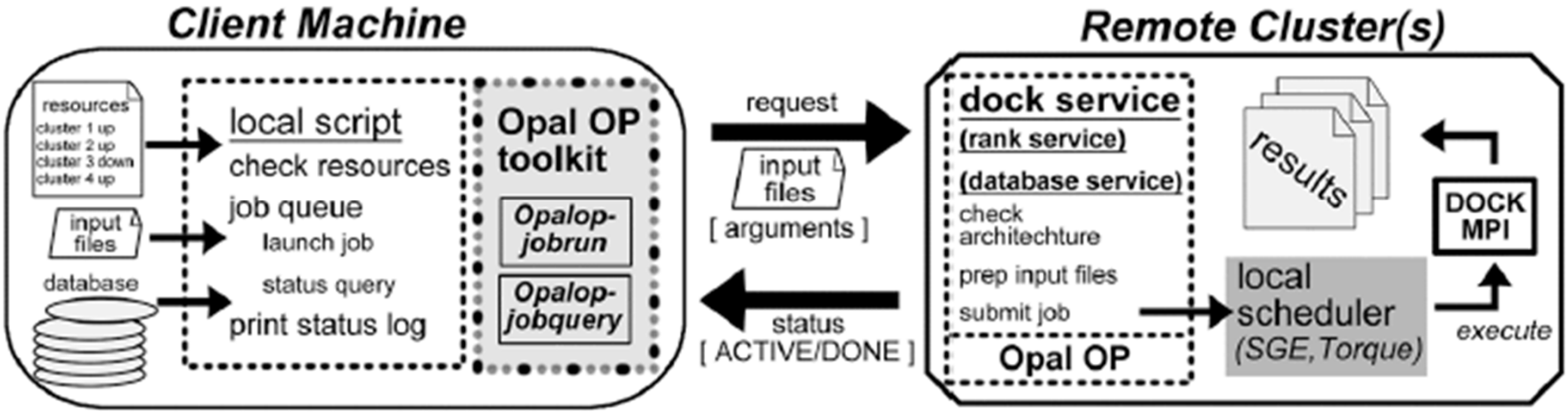
- Defined by its ability to catalyze the removal of two covalently attached phosphate groups from tyrosine and serine/threonine residues on the same substrate
- Part of the Protein Tyrosine Phosphatase superfamily, which is characterized by a highly conserved PTP loop sequence (HCXXXXXR)

Project Goals

- Create detailed, complete datasets by combining both structural and binding affinity data with a series of ligands (from screenings of well-defined structures) for DSPs that lack crystal structures.
- High quality data for a representative set of protein targets of therapeutic significance will provide a benchmarking set that can assist in improving current molecular computational models, methods and software.
- Continue Docking project started by previous PRIME students.

Materials and Methods (DOCK)

- The DOCK process is divided into two phases (grid based score and AMBER score)

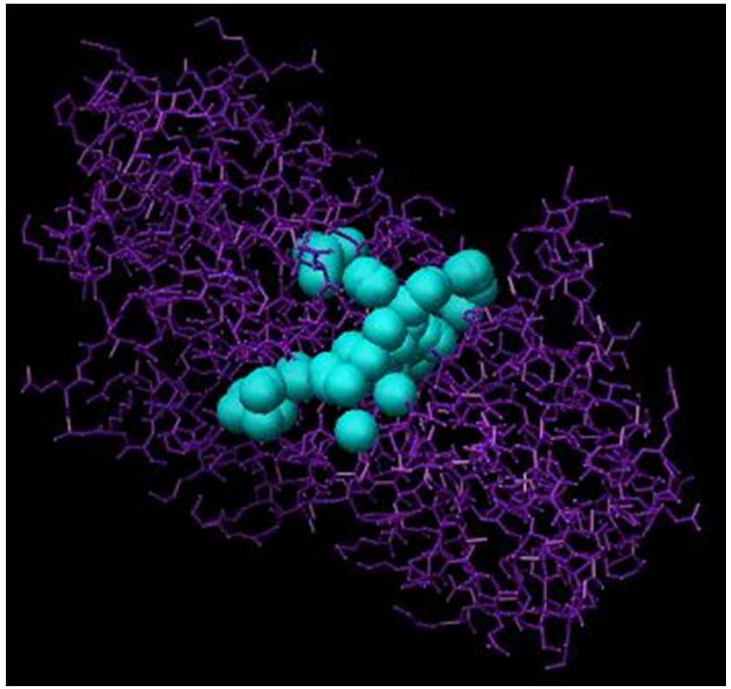


Materials and Methods (DOCK)

- Use slice_distribute.pl to send 176 slices to remote clusters
  - Each slice contains many different ligands
- Use bigrun.pl through opal-op to submit the DOCK job on the cluster

Materials and Methods (DOCK)

- X-ray crystal structure of a drug/receptor complex is obtained and the active site is identified.
- Points within this site, known as “spheres,” are used to define the volume or space within the active site pocket where the drug binds.
- Sphere centers are matched with ligand (drug) atoms to generate thousands of orientations of the ligand in the active site within the program DOCK.



- This simulation is run in parallel to greatly reduce computation time.

Materials and Methods (DOCK)

- Use slice_redistribute.pl to submit slices for phase two of docking
- Use bigrun_amber.pl and opal-op to submit the job to the cluster
  - AMBER scoring requires separate preparation files that are generated on the frontend node

Materials and Methods (DOCK)

- Because each ligand for Amber must be prepared separately with another DOCK accessory program, it requires significant computation time compared to grid based scoring
- AMBER allows the receptor to be moved around and takes into account more energy potentials.

Materials and Method (DOCK)

- Use rerank.pl to rank all the binding energy of the ligands
- Weight AMBER and grid score equally to determine list of ligands that have the best binding ability to the protein
- The above procedure is repeated for all the proteins in the DSP family

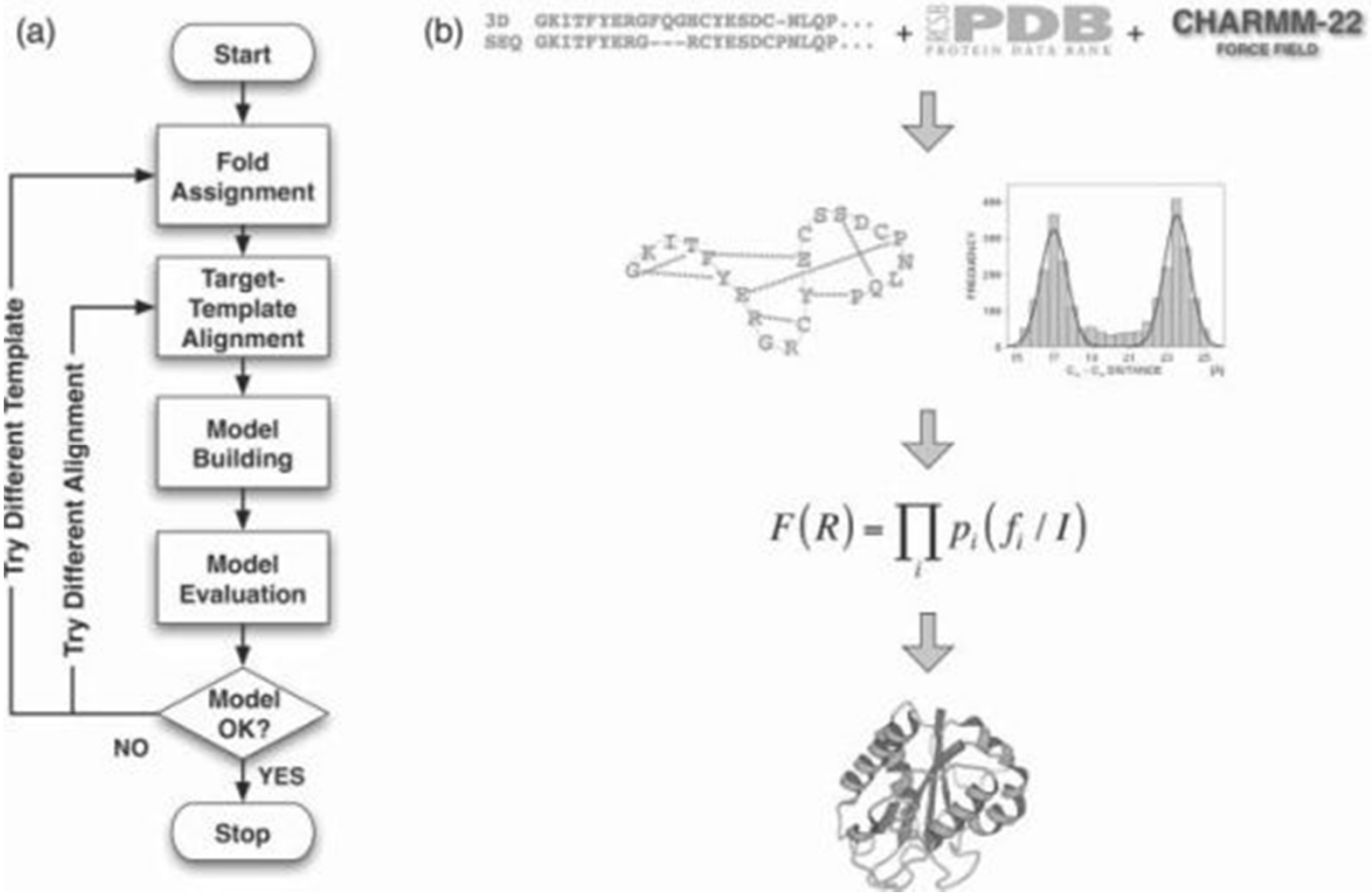
Materials and Methods (MODELLER)

- Use fit_distribute.pl to submit slices to remote clusters
  - Each slice contains information to tell the cluster to generate 600 models
- Use opal-op and modrun.pl to submit the MODELLER job in parallel
- The number of slices depends on the number of templates

Materials and Methods (MODELLER)

- After the job is done, MODELLER gives three scores: GA341, molpdf, and DOPE
- The models are ranked by an aggregate score based on molpdf and DOPE
- The best model is chosen for loop refining and energy minimization
- The model is then used as a protein for the DOCK process

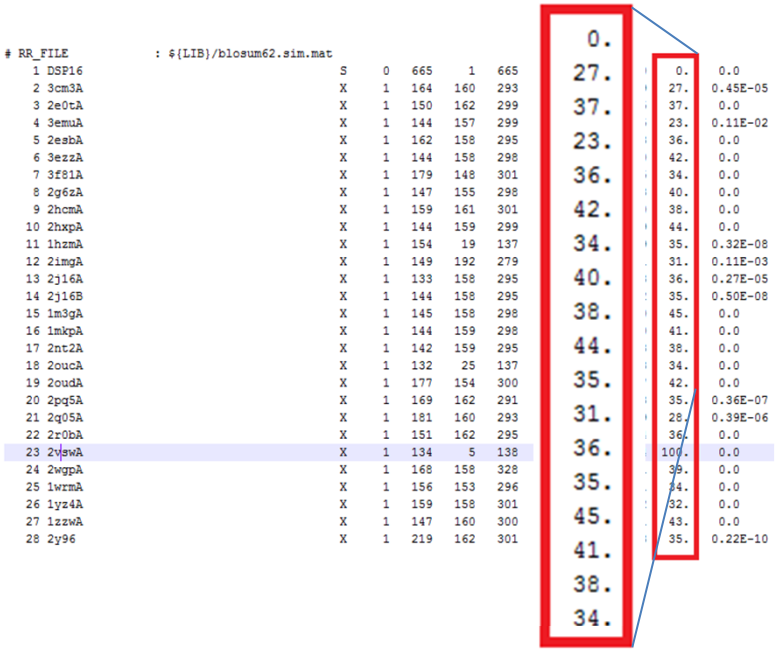
Materials and Methods (MODELLER)



Materials and Methods (MODELLER)

1. Fold assignment

- Swiftmodeller applies a protein blast and selects the top templates with 40% sequence identity as cutoff. Templates are chosen based on their homology to each other seen in the generated output dendrogram. Distances in the dendrogram are relative to the percentage similarity in the residues.



```
# RR_FILE          : ${LIB}/blosum62.sim.mat
   1 DSP16                          S    0   665     1   665      0.     0.0
   2 3cm3A                          X    1   164   160   293     27.   0.45E-05
   3 2e0tA                          X    1   150   162   299     37.     0.0
   4 3emuA                          X    1   144   157   299     23.   0.11E-02
   5 2esbA                          X    1   162   158   295     36.     0.0
   6 3ezzA                          X    1   144   158   298     42.     0.0
   7 3f81A                          X    1   179   148   301     34.     0.0
   8 2g6zA                          X    1   147   155   298     40.     0.0
   9 2hcmA                          X    1   159   161   301     38.     0.0
  10 2hxpA                          X    1   144   159   299     44.     0.0
  11 1hzmA                          X    1   154    19   137     35.   0.32E-08
  12 2imgA                          X    1   149   192   279     31.   0.11E-03
  13 2j16A                          X    1   133   158   295     36.   0.27E-05
  14 2j16B                          X    1   144   158   295     35.   0.50E-08
  15 1m3gA                          X    1   145   158   298     45.     0.0
  16 1mkpA                          X    1   144   159   298     41.     0.0
  17 2nt2A                          X    1   142   159   295     38.     0.0
  18 2oucA                          X    1   132    25   137     34.     0.0
  19 2oudA                          X    1   177   154   300     42.     0.0
  20 2pq5A                          X    1   169   162   291     35.   0.36E-07
  21 2q05A                          X    1   181   160   293     28.   0.39E-06
  22 2r0bA                          X    1   151   162   295     36.     0.0
  23 2vswA                          X    1   134     5   138    100.     0.0
  24 2wgpA                          X    1   168   158   328     39.     0.0
  25 1wrmA                          X    1   156   153   296     34.     0.0
  26 1yz4A                          X    1   159   158   301     32.     0.0
  27 1zzwA                          X    1   147   160   300     43.     0.0
  28 2y96                           X    1   219   162   301     35.   0.22E-10

   0.
  27.
  37.
  23.
  36.
  42.
  34.
  40.
  38.
  44.
  35.
  31.
  36.
  35.
  45.
  41.
  38.
  34.
```

```
Weighted pair-group average clustering based on a distance matrix:

                                                .----------------------- 1b8pA @1.9    39.0000
                                                |
                                        .---------------------------------- 1bdmA @1.8    50.5000
                                        |
                                .----------------------------------------- 5mdhA @2.4    55.3750
                                |
                                |                                  .--- 1civA @2.8    13.0000
                                |                                  |
        .-------------------------------------------------------------- 7mdhA @2.4    83.2500
        |
      .-------------------------------------------------------------- 1smkA @2.5

     +----+----+----+----+----+----+----+----+----+----+----+----+----+
   86.0600   73.4150   60.7700   48.1250   35.4800   22.8350   10.1900
     79.7375   67.0925   54.4475   41.8025   29.1575   16.5125
```

Materials and Methods (MODELLER)

2. Target-template sequence alignment

- Noniterative salign is run prior to multiple template alignment and the salign files are checked for quality score. If quality score < 70% then iterative salign is performed.
- Iterative performs many multiple alignments considering only local and global atom distances until scores reach consensus while noniterative alignment takes into more chemical factors of the residue.

Materials and Methods (MODELLER)

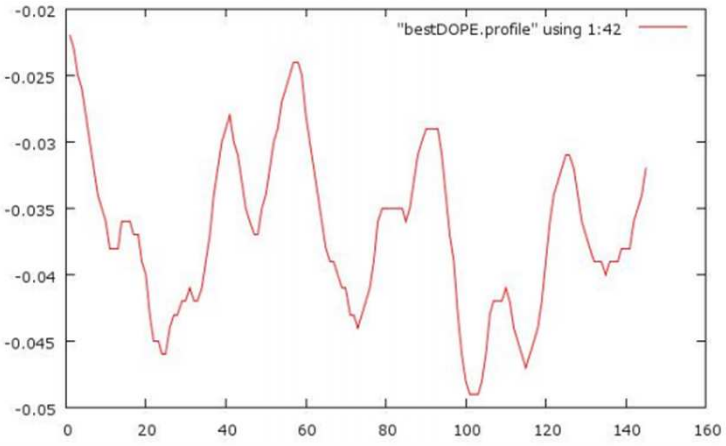
3. Run 600 modeling experiments of the protein

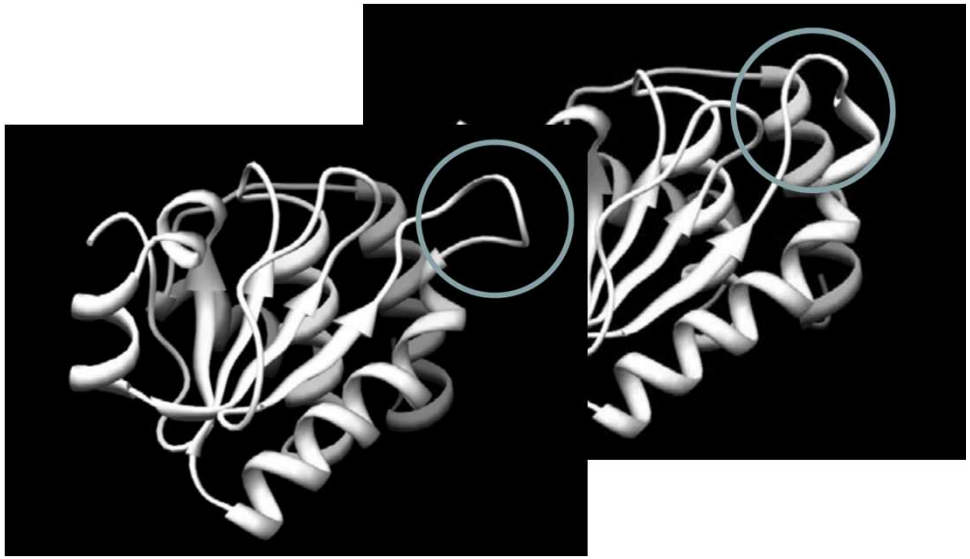
- To evaluate the folding quality, a GA341 score is taken. This score takes into account the sequence identity, compactness of the protein, and the combined potential energy of the model. Scores are filtered with cutoff at 0.7 out of 1. If number of samples after filtering is not enough then 3000 samples are rerun to obtain a more convergent GA341 score among the samples.
    - Create aggregate score of DOPE + molpdf (equally weighted) and compare based on lowest total value. Use normalized dope score to evaluate models between different alignments.

Materials and Methods (MODELLER)

4. Optimize the protein using energy minimization methods and loop refinement.

- Use the Molprobity server and PROCHECK to observe how much the protein has been optimized. Models will have hydrogens added, residues flipped to minimize clash. Energy minimization of the structure while holding active size (hcx5r(s/t)) is then performed. Energy profiles will be read and loop refinement performed.





Materials and Methods (MODELLER)

- Loop Refinement
- The pdb file name and the regions to be loop refined are stored in the file “loop_pieces”
- The PDB name of the protein and how many models to be generated are stored in “important_info”

Materials and Methods (MODELLER)

- “loop_distribute.pl” designates which protein sequence to be refined and number of refinements per slice of data.
- The input file loop-task.py.generic.original is converted to loop-task.py.generic
  - Inserts the pdb file name and the PDB code into the script
- Creates a slice_array file that contains information on how many slices to divide the job into

Materials and Methods (MODELLER)

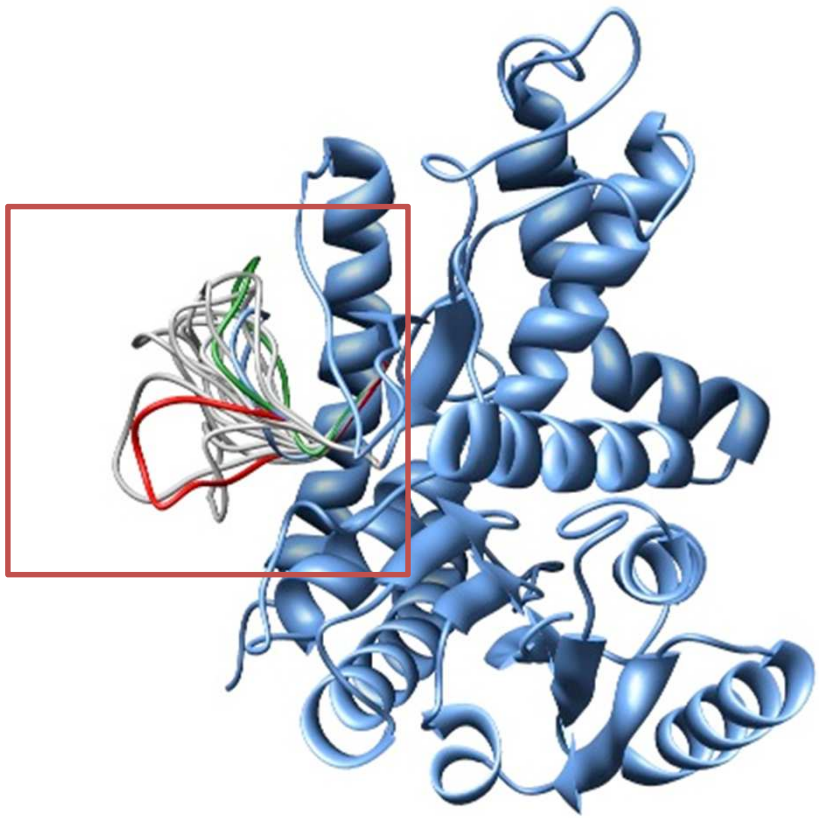
- “looprun.pl” uses input from loop-task.py.generic to generate loop-task.py. This input file specifies:
  - The region where the protein should be loop refined
  - Instructs MODELLER begin loop refinement in parallel

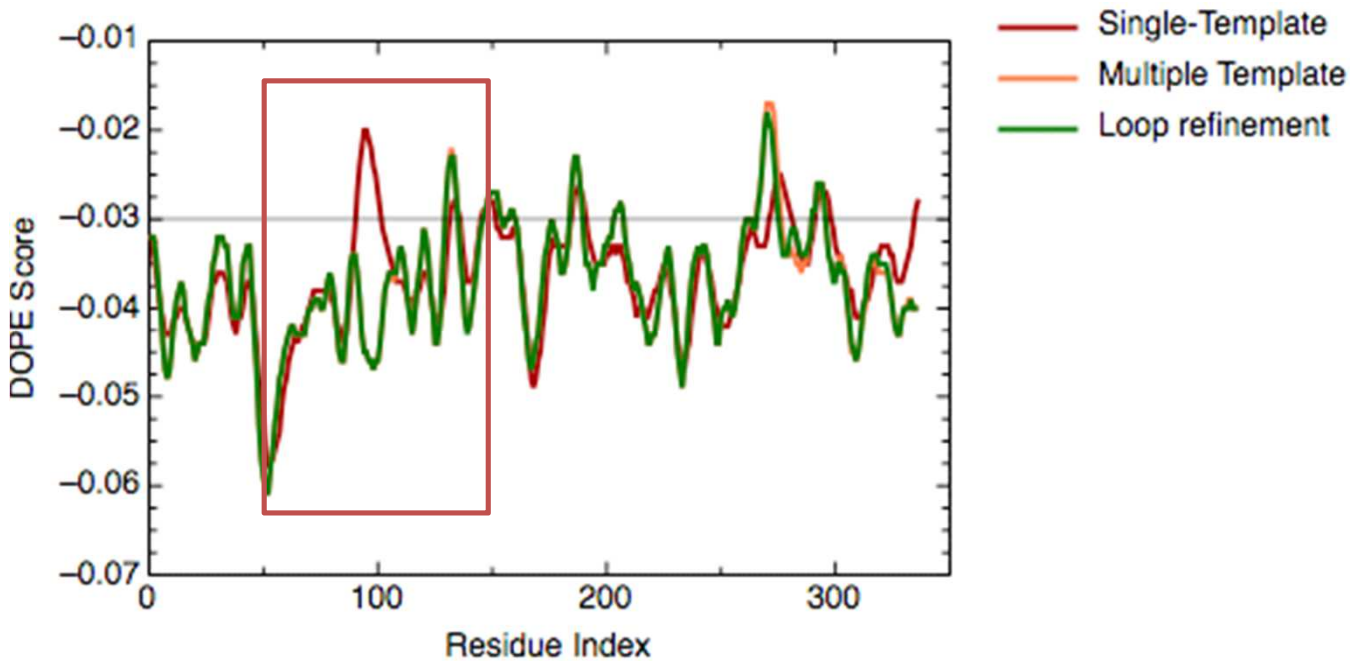
Materials and Methods (MODELLER)

- Loop-task.py generates output file loop-task.log
  - Lists the DOPE score of each of the different models after loop refinement
  - Choose the model with the lowest DOPE score

Materials and Methods (MODELLER)

- Loop Refinement
  - Allows minor modifications of loops, which are small regions of the protein without definite secondary structure
  - Often not modeled well
  - Loop refinement script uses a model generated from the earlier steps and tries to reduce the energy of the region by shifting the amino acids in the loop around
  - Generates even more models based off of the older model




−0.01
−0.02
−0.03
−0.04
−0.05
−0.06
−0.07
DOPE Score
0
100
200
300
Residue Index
Single-Template
Multiple Template
Loop refinement

Progress Week 8 and 9

- Successfully loop refined DUSP21
- Continued to perform AMBER screening on 3EZZ on rocks-200 and ocikbpra
- Modify loop refinement script so it can refine many areas at once

Before Minimization:

All-Atom Contacts	Clashscore, all atoms:	82.4	0th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	0.00%	Goal: <1%
	Ramachandran outliers	0.00%	Goal: <0.2%
	Ramachandran favored	97.34%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	2.51	47th percentile* (N=27675, 0Å - 99Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.53%	Goal: <0.1%

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.

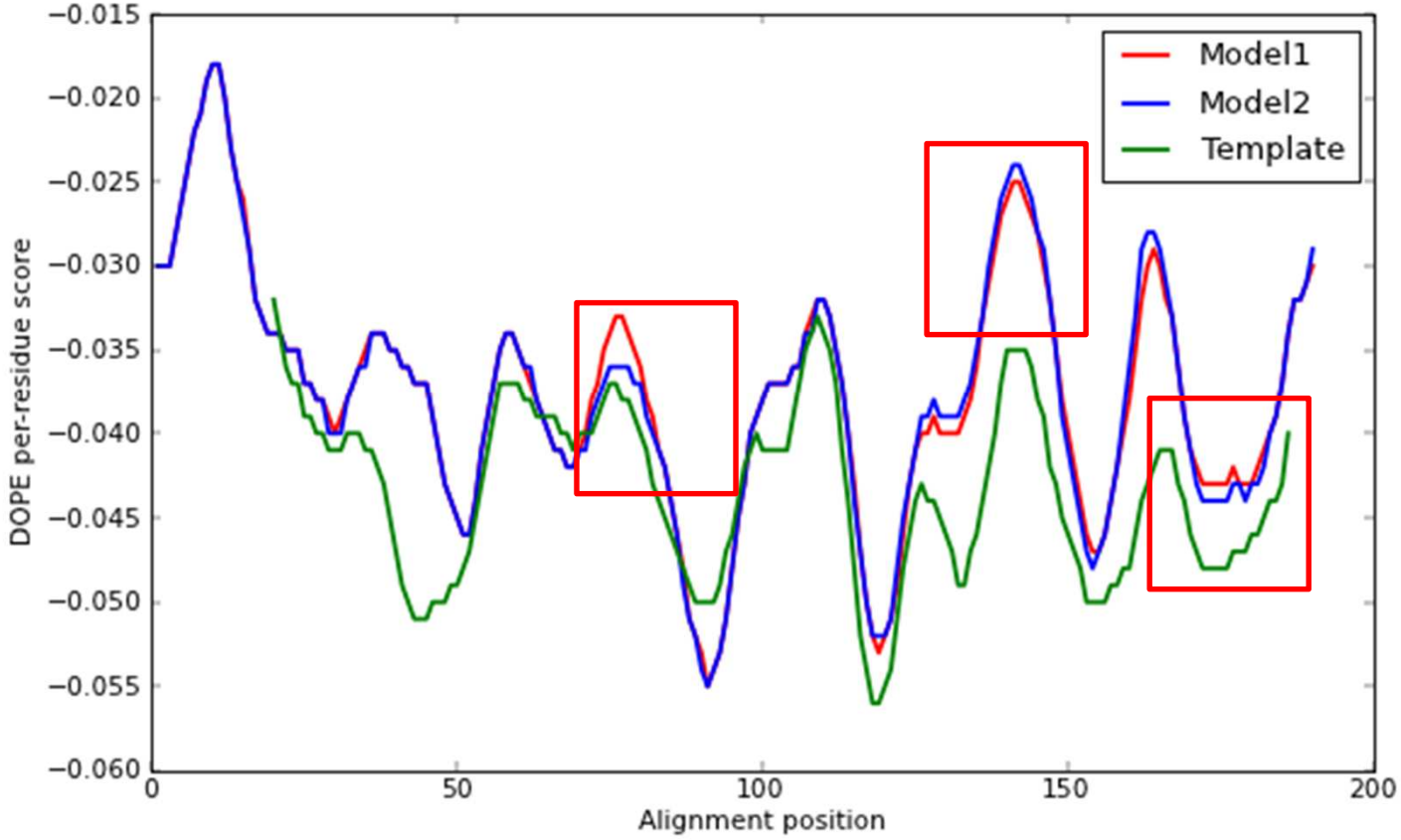
^ MolProbity score is defined as the following: 0.42574*log(1+clashscore) + 0.32996*log(1+max(0,pctRotOut-1)) + 0.24979*log(1+max(0,100-pctRamaFavored-2)) + 0.5

After Minimization:

All-Atom Contacts	Clashscore, all atoms:	3.3	97th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	0.00%	Goal: <1%
	Ramachandran outliers	0.00%	Goal: <0.2%
	Ramachandran favored	96.28%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	1.37	98th percentile* (N=27675, 0Å - 99Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.00%	Goal: <0.1%

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.

^ MolProbity score is defined as the following: 0.42574*log(1+clashscore) + 0.32996*log(1+max(0,pctRotOut-1)) + 0.24979*log(1+max(0,100-pctRamaFavored-2)) + 0.5



Regions modified: residues 73-78, 138-145, and 160-167

Residues 73-78

Residues 138-145

Residues 160-167

Future plans

- Perform Modeller experiments on proteins with known crystal structures.
  - Study common alignment problems.
  - Look for model patterns that occur using specific templates.
  - Learn to better interpret output scores and its relationship to model data.

Future Plans

- Dock the loop-refined and optimized DUSP21 to check that the docking results are similar to other DSPs
- Install Dock and MODELLER on the new cluster Milk
- Continue to run Dock for 2Y96 and 3EZZ

Acknowledgements

- Cybermedia Center, Osaka, Japan
  - Dr. Susume Date, Cybermedia Center, UCSD
  - Dr. Kohei Ichikawa, Cybermedia Center, UCSD
  - All the graduate students!
- UCSD, La Jolla, USA
  - Dr. Jason Haga
  - Dr. Peter Arzberger
  - Dr. Gabriele Wienhausen
  - Matt Mui, Charles Xue
  - General contributions from the NSF PRIME Undergraduate Research Program