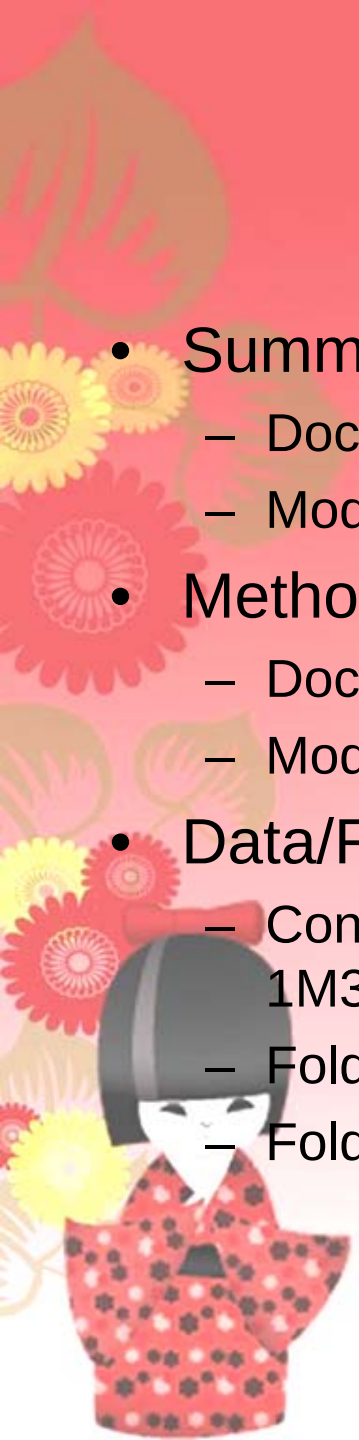
Protein Folding Simulation and Virtual Screening of Dual Specificity Phosphatase in Parallel

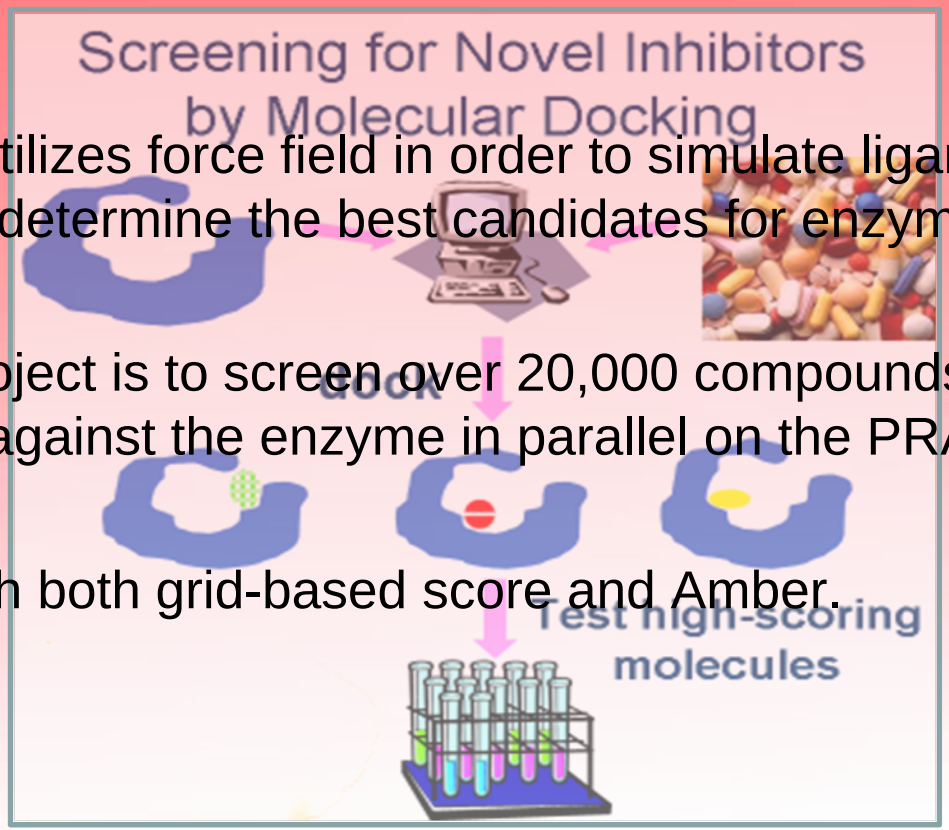
Charles Xue
Professors Jason Haga and Susumu Date
Osaka University, Japan
August 20th, 2010
Final Presentation

Table of Contents



Project Summary

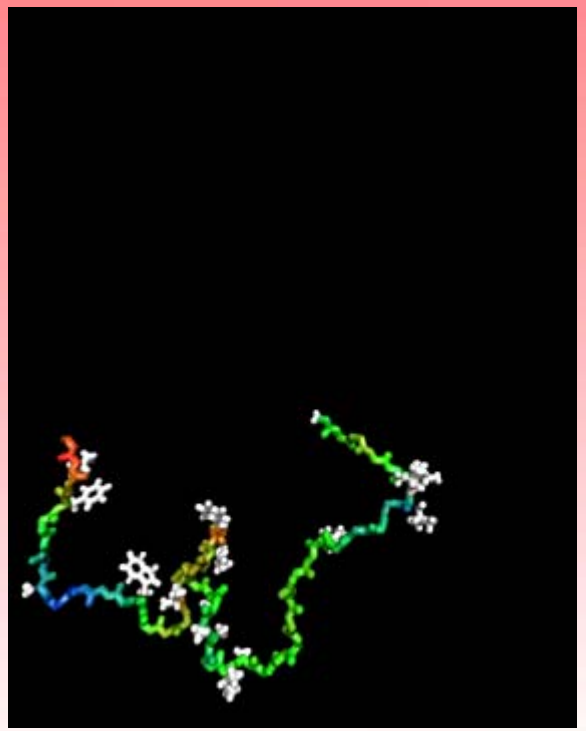
DOCK6



- 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.

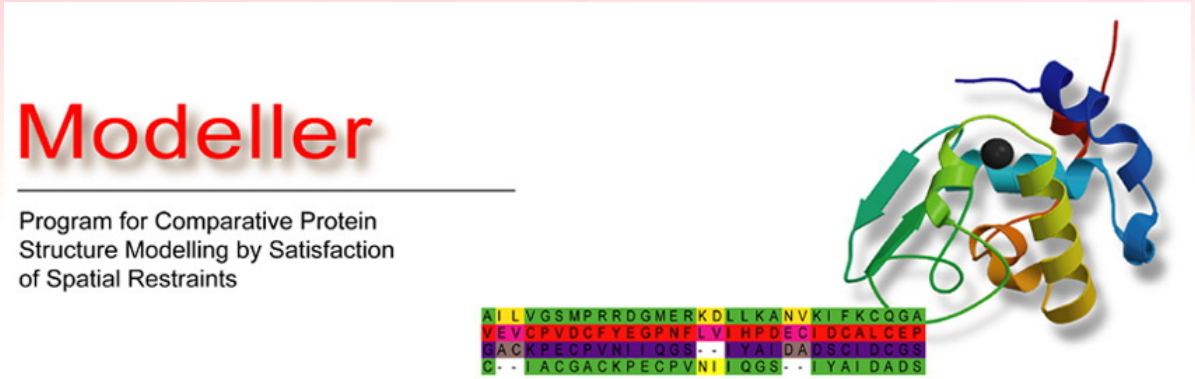


- Protein folding (homology)
  - Utilizes similar known structures in order to predict the model of an amino acid sequence

Project Summary Cont.

MODELLER9v8

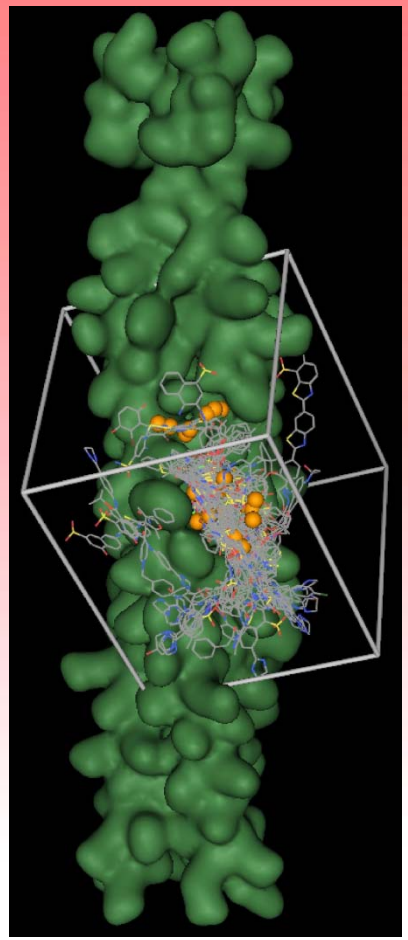
- Program utilizes homology modelling algorithms to determine best tertiary structure of an amino acid sequence.
- Requires proteins with known configurations and similar sequences, ideal for the dual specificity phosphatase family.
- Goal is to implement the entire folding process and loop refinement 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.



Methods/Procedures Cont.

Why Modeller9v8 – Alternatives?

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

Bioingbu	Evolutionary information recognition	Webserver	server	No
mGenTHREADER/GenTHREADER	Sequence profile and predicted secondary structure	Webserver	main page	No
LOOPP	Multiple methods	Webserver	server	No
MUSTER	profile-profile alignment	Webserver	server	No
3. *Ab initio* structure prediction				
Name	**Method**	**Description**	**Link**	
I-TASSER	Combination of ab initio folding and threading methods	Structural and function predictions	main page	No
ROBETTA	Rosetta homology modeling and ab initio fragment assembly with Ginzu domain prediction	Webserver	server	No
Bhageerath	A computational protocol for modeling and predicting protein structures at the atomic level.	Webserver	Server	No
Abalone	Molecular Dynamics folding	Program	Example	Yes
LeanMD	Parallel Protein Folding on PetaFLOP Machines	Program	Download	Yes
NAMD	**Parallel Molecular Dynamics**	**Program**	**Download**	**Yes**
ProFoGa	Open Source Folding Simulator	Program	Download	Yes
ProFaSi	Protein Folding and Aggregation Simulator	Program	Home Page	Yes
4. Secondary structure prediction				
Name	**Method**	**Description**	**Link**	
NetSurfP	Profile-based neural network	Webserver	server	No
GOR	Information theory/Bayesian inferen	Many implementations	Basic GOR GOR V	No
Jpred	Neural network assignment	Webserver	server	No
Meta-PP	Consensus prediction of other servers	Webserver	main page	No
PREDATOR	Knowledge-based database comparison	Webserver	server	No
PredictProtein	Profile-based neural network	Webserver	server	No
PSIPRED	two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST	Webserver	server	No
YASSPP	Cascaded SVM-based predictor using PSI-BLAST profiles	Webserver	server	No
5. Transmembrane helix and signal peptide prediction				
Name	**Method**	**Description**	**Link**	
HMMTOP	Hidden Markov Model	Webserver/standalone	main page	No
MEMSAT	Neural networks and SVMs	Webserver/standalone	main page	No
PDHtm in PredictProtein	Multiple alignment-based neural network system	Webserver/standalone	server	No

Methods/Procedures Cont.

MODELLER9v8

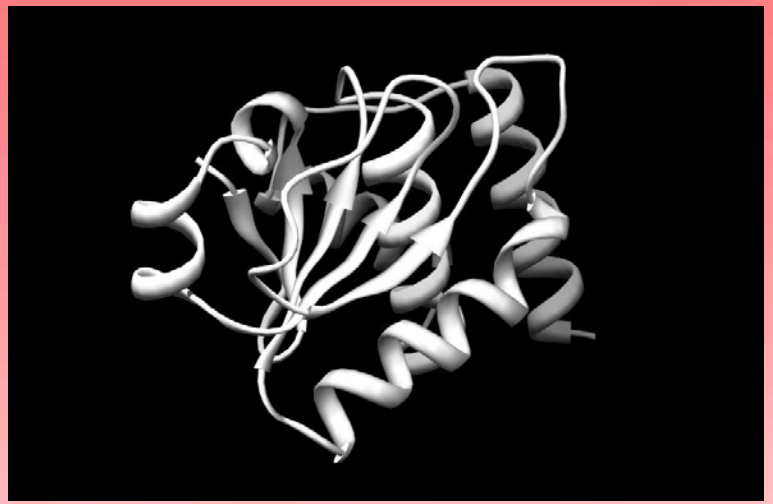
(Fit_distribute.pl/modrun.pl)

(getbest.pl/getbestlocal.pl)

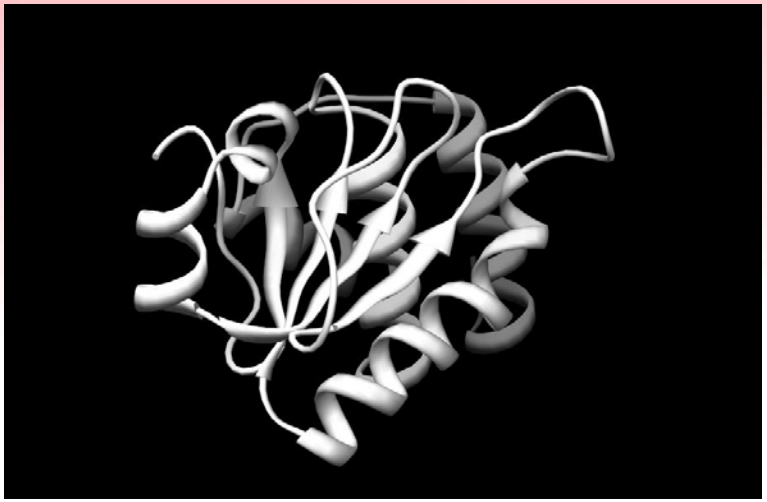
- Develop script for Modeller to submit jobs with sge based on its built in task-based interface.

- 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.



Simulated image of SSH2 protein in chimera



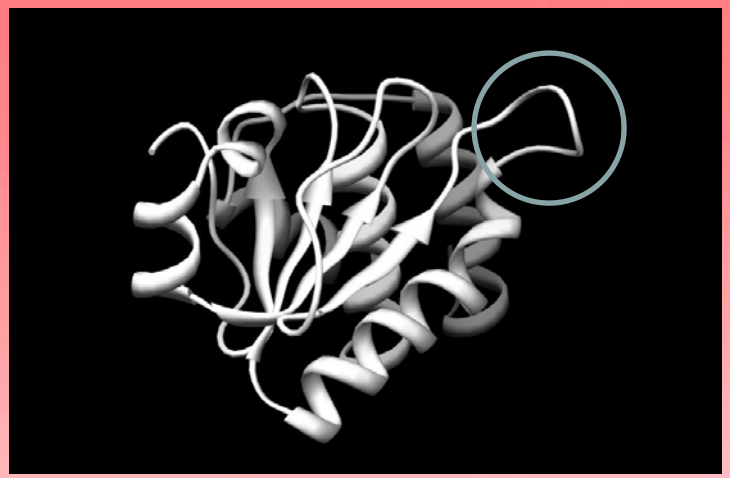
Generated model before loop refinement

Methods/Procedures Cont.

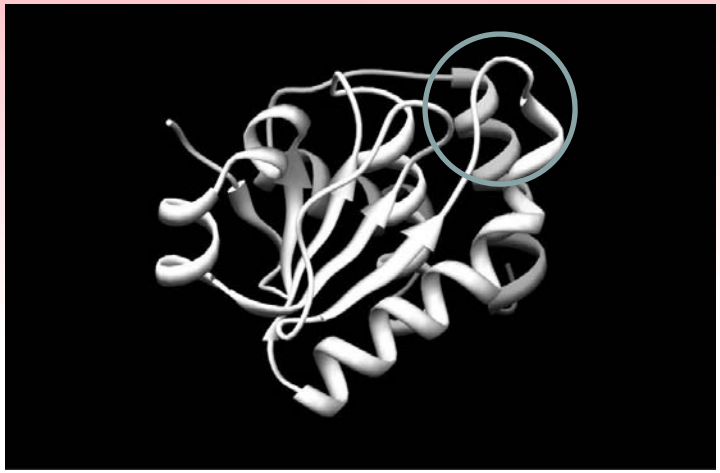
MODELLER9v8

(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

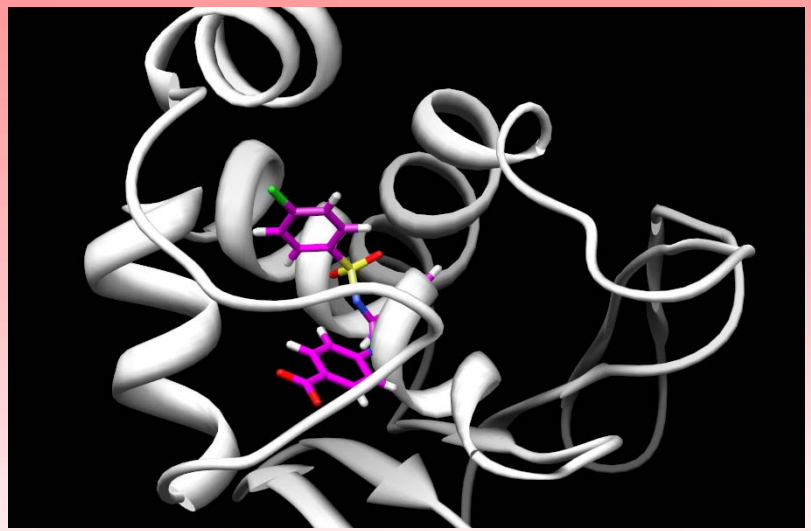


Protein after the loop refinement, the loop has been re-simulated to better resemble actuality

Results

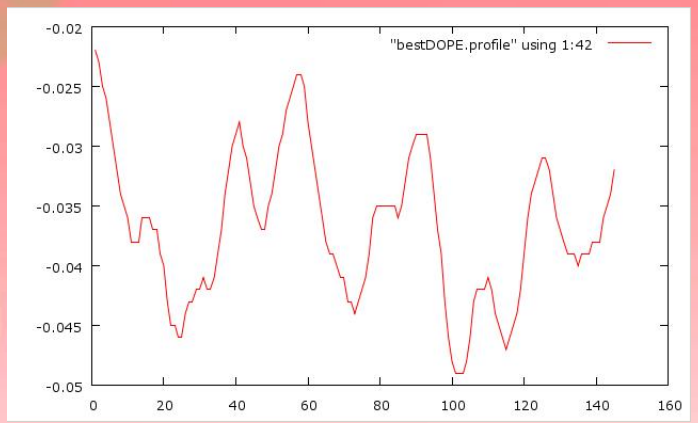
DOCK6 (DSP2 aka 1M3G)

1.	ZINC02352689	6	3	3	-93.692276	-142535802880.000000
2.	ZINC06645917	48	43	5	-46.682549	-41836152.000000
3.	ZINC06645918	66	65	1	-45.843616	-510907777024.000000
4.	ZINC06645916	331	327	4	-42.563263	-3793709312.000000
5.	ZINC05047674	596	359	237	-42.357357	-37.623665
6.	ZINC06645550	606	228	378	-43.194515	-34.503338
7.	ZINC02649005	666	312	354	-42.631413	-35.088120
8.	ZINC02921257	835	426	409	-42.014664	-34.038483
9.	ZINC03457610	848	293	555	-42.763611	-32.087337
10.	ZINC01052992	945	351	594	-42.402878	-31.573633

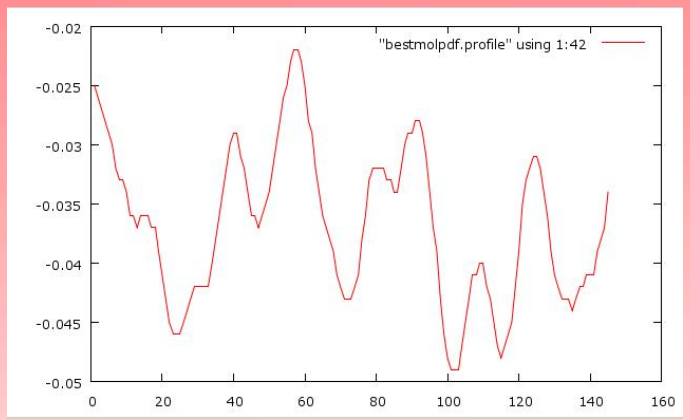


Results Cont.

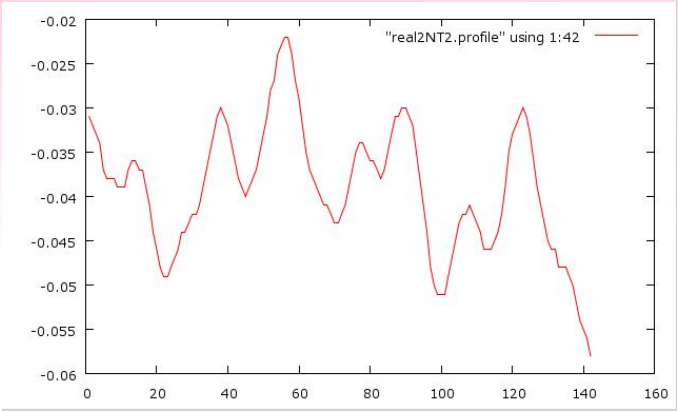
MODELLER9v8 (SSH2 aka 2NT2)



Plot of molecule before loop refinement



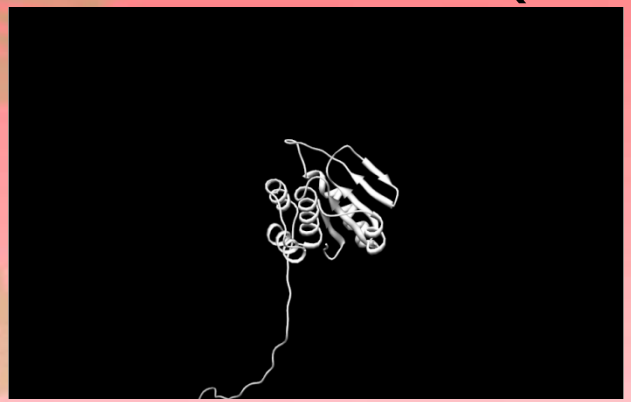
Plot of molecule after loop refinement



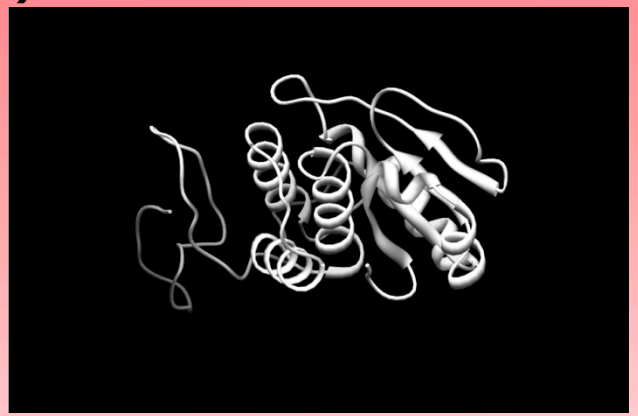
Plot of actual SSH2 protein

Results Cont.

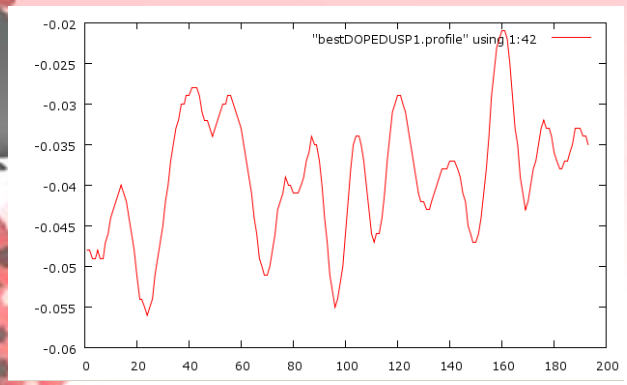
MODELLER9v8 (DUSP1)

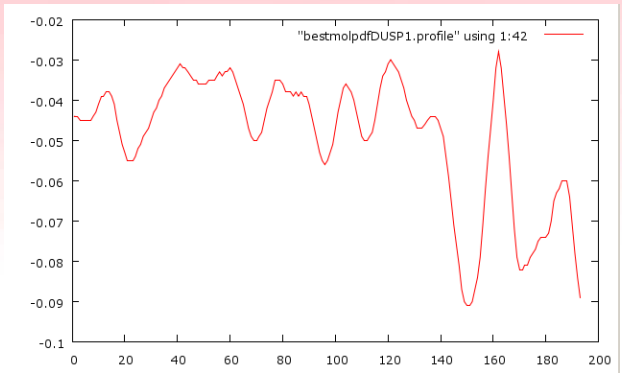


Lowest DOPE score before loop refinement



Lowest DOPE score after loop refinement





Discussion

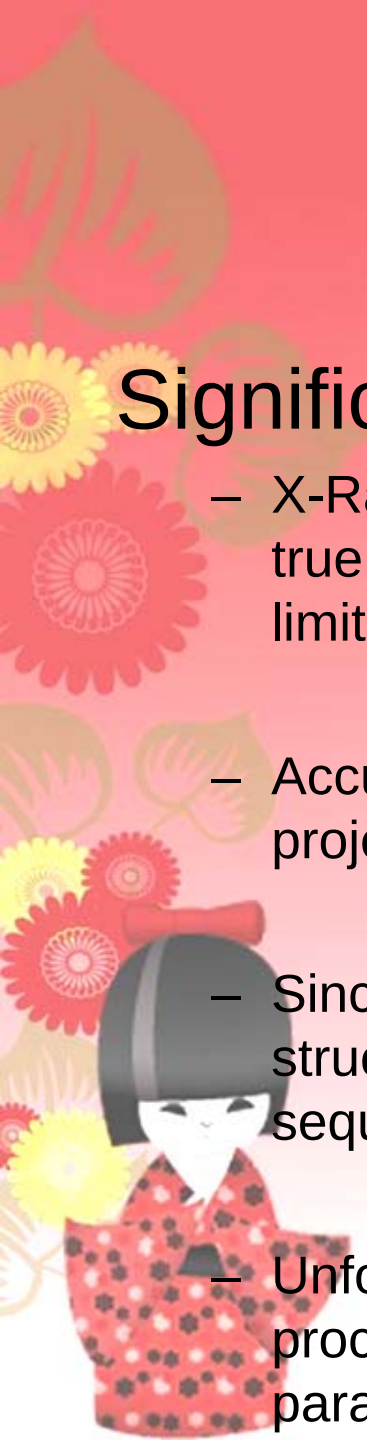
Significance of Project (DOCK):

- 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):

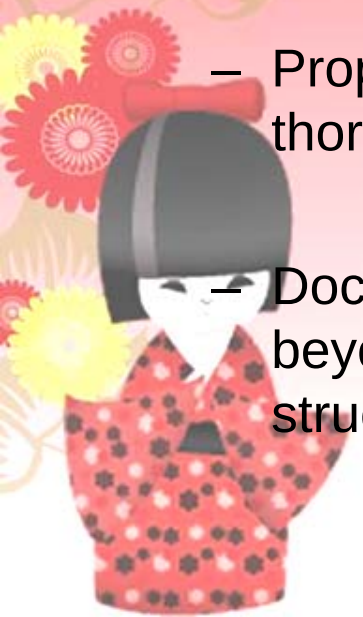
- X-Ray crystallography and NMR spectroscopy to determine the true structures of proteins at a suitable resolution is extremely limited by supply.
- Accurate folded proteins *in silico* is of high demand and many projects started for that purpose (i.e. FoldingAtHome).
- Since structure of protein determines 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.

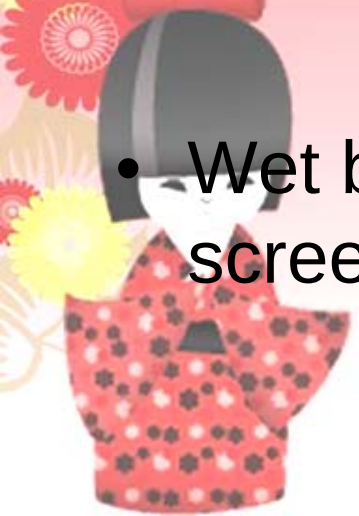
Significance of Project (DOCK with MODELLER:

- Many members of the DSP family (of which SSH belongs to) has 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.)



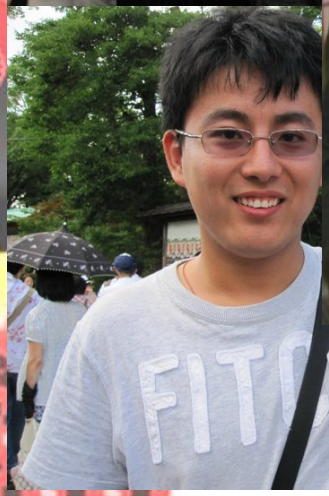
Future Work

- 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!