

# CS612 Homework Assignment 3

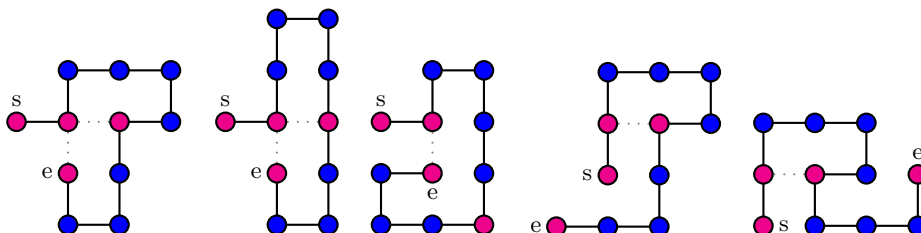
Due Thu. March 31, 2016

1. **HP lattice model:** Given the following two sequences:

- S1 = HHPPPPHPPPH
- S2 = HHPHPPHPHPH

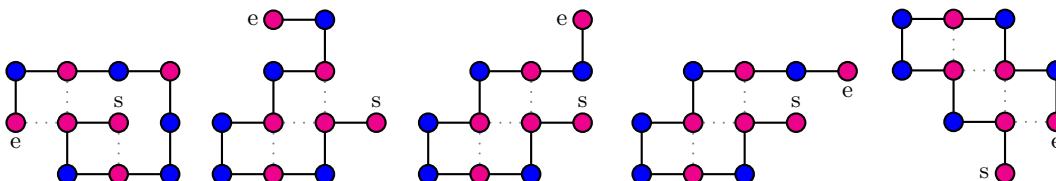
- (a) Find five possible high-scoring self-avoiding 2-D grid arrangements for S1 and calculate each one's "potential energy" according to the model discussed in class (award one point for each pair of H-H points that are one grid point apart (not diagonal), but not adjacent on the sequence. For more info look at the class notes – I'm referring to the white dashed lines. Attach the five arrangement (indicate which one is the start and which one is the end). Color P in blue and H in red or magenta, like in class. Notice that you don't have to find the absolute globally best structure, just arrangements that have a good score.

Here are five examples. I could not find any conformation with  $> 2$  H-H interactions. The values (left to right) are 2, 2, 1, 1, 1.

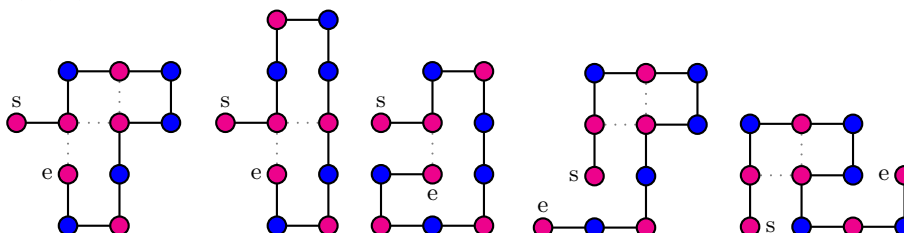


- (b) Find five possible high-scoring self-avoiding 2-D grid arrangements for S2 and calculate their score as in (a).

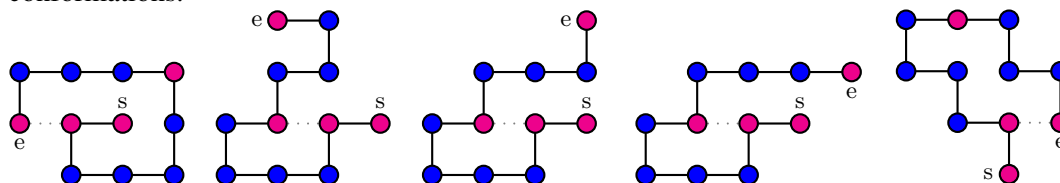
Here are five examples. The best scoring example had 4 H-H interactions. The scores are: 3, 3, 3, 3, 4.



- (c) For the same grid arrangements as in (a), "thread" S2. That is, use the same arrangements as in (a) above but the amino acid sequence in S2. Calculate the score again. The scores are, left to right: 3, 2, 1, 2, 2.



- (d) For the same grid arrangements as in (b), “thread” S1. That is, use the same arrangements as in (a) above but the amino acid sequence in S1. Calculate the score again. The scores here are 1 for all the conformations.



- (e) Explain the observed differences briefly.

**Answer:** The main observations are that: 1) S2 has more hydrophobic amino acids, so its scores generally tend to be higher. Also, the conformations that are suitable for S1 are not the ideal for S2 and vice versa.

2. **Hands-on homology modeling exercise using a template:** In this assignment you will perform homology modeling using the SwissModel server.

As a nice and easy example, let us start by modeling a short protein from a family called Crambin. Go to the SwissModel server at: <http://swissmodel.expasy.org/workspace>. There are three ways of doing homology modeling using the SWISS-MODEL server – a fully automated approach (called Automated mode), Alignment mode which allows you to submit a multiple sequence alignment of your target with one or more templates and a Project mode, which allows you to manually optimize your alignment. We will use the automated.

**Note:** SwissProt contains a rather extensive help section, please take a look if you need a clarification. I suggest you create a login with them, so that your jobs will be saved on the server for about a week.

On the top window, copy and paste the following sequence:

SVCCPSLVARTNYNVCRLPGTEAALCATFTGCIIPGATCGGDYAN

Click on “search for templates” and wait. It should take a couple of minutes. When you are done, you can access the templates by clicking on “templates”. The page contains information about each template – its PDB code, its coverage of the query sequence, its sequence identity etc.

- a. How many templates did you get? What is the PDB code, coverage and identity of the top template?

**Answer:** Using the default settings, eight templates should be obtained. The top template has a PDB code of 2eyb.1.A (2eyb, chain A). The coverage is range 1–46 (the whole sequence) with 80.43% identity. The full list contains 32 extra templates not mentioned in the table, so I accepted both 8 and 40 as correct answers.

Now, check the boxes near the top three templates and click “build model” on the top right. The server will now try to build three models, based on the three templates. Again, you will probably have to wait a while.

- b. What is the QMean score of each of the models? (this is a score that combines various aspects of interactions in the molecule. This statistical score is logarithmic and the lower, the worse.

**Answer:** The QMean scores are -1.66, -0.19 and -1.74, respectively.

- c. What is the overall score, GMQE (Global Model Quality Estimation) of each one of the models?

**Answer:** The GMQE scores are 0.93, 0.96 and 0.93, respectively.

The scoring functions combine the energy terms used for the modeling, which are shown on the model window on a blue (good) to red (bad) scale. Notice the three superimposed models on the right window. You will notice that the protein is depicted on a blue to orange scale which correspond to model quality. As you may tell, the orange parts correspond mostly to loop regions.

At the top of the page you will see an icon of a page and if you hover the mouse over it, it will say "one page project report". Please attach it to your submission.

- d. Now measure the RMSD of each model to its respective template. The easiest way IMO is to save the three models to your computer by clicking on the button near the model and save as pdb. The names of your file will be 01.pdb, 02.pdb and 03.pdb. For each model do the following: Upload it to Chimera. Then click File→Fetch by ID and put the PDB ID of the template. Notice that for templates 1 and 2, being NMR files, you will get multiple structures (remember our earlier discussion about NMR structures). Then go to Tools → Structure Comparison → Match Maker to determine the optimal structural alignment between the template and model. Select your model file as reference and the first NMR model (if relevant), marked by (#1.1). Click "OK". Then go to Tools → Structure Comparison → Match align to obtain a position by position sequence alignment. Select only your model and the first template again. A sequence alignment shows up. Finally, open Favorites → Reply Log. The line "overall RMSD" is a position-by-position RMSD, based on the alignment. This is what I'm looking for. Do it for all three models.

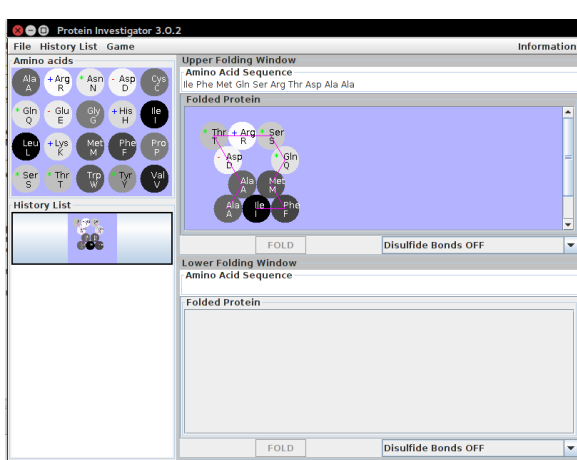
**Important!** Notice that even though the models have the same coverage, the same sequence identity and they are overall quite good, model three beats the other two in all the categories. The reason is that the first two models were generated by NMR and the third was generated by X-ray and is very high-resolution. NMR structures tend to be more fuzzy, and in the case of homology modeling selecting a good quality, high-res structure can be quite important. While all three models came out OK in this case, it can be quite critical in other cases (but as of now I could not find a better example that still works. The one from two years ago was more drastic but no longer relevant – better structures came out). If at all possible – aim for X-ray structures when you create your models, and try not to go below a 2Å resolution.

**Answer:** The RMSD values are 0.067, 0.064 and 0.092Å, respectively.

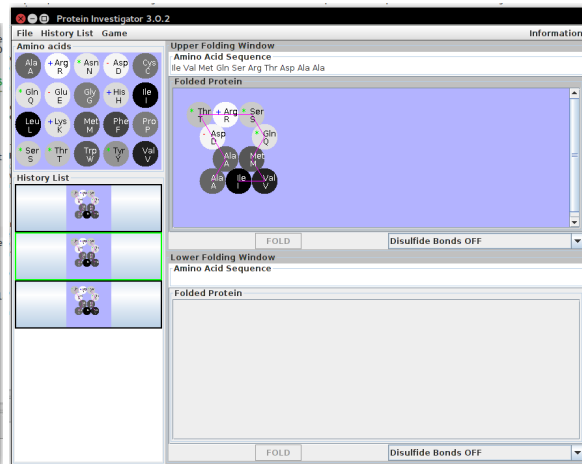
3. **Basic protein folding exercise:** Use Protein Investigator (the software we demonstrated in class) at <http://intro.bio.umb.edu/MOOC/jsPI/JsPI.html>. It requires the Java running environment to run. On the upper folding window type the following sequence: IFMQSRTDAA (Ile-Phe-Met-Gln-Ser-Arg-Thr-Asp-Ala-Ala). Type "Fold" and see the shape of the folded protein. The energy function is based on hydrophobic contacts, ionic interactions (opposite charges attract, similar charges repel each other), and hydrogen bonds between polar amino acids. For the classification of hydrophobic, charged and polar amino acids see class notes.
  - (a) Create a mutant protein by changing **one** amino acid in the sequence above, such that the mutation has no effect on the shape of the mutant protein. Explain. Attach a screenshot of the resulting protein.
  - (b) Create a mutant protein by changing **one** amino acid in the sequence above, such that the mutation has a large effect on the shape of the mutant protein. Explain. Attach a screenshot of the resulting protein.
  - (c) Design a protein of at least 8 amino acids such that a salt bridge (an ionic interaction between charged amino acids) is critical to its shape. Explain and attach a screenshot.

**Answer:** See the figures below. This is just a suggestion. I accepted any reasonable answer. (a) The original sequence. (b) The sequence after mutating Phe (the 2nd amino acid) to Val (Valine). They are

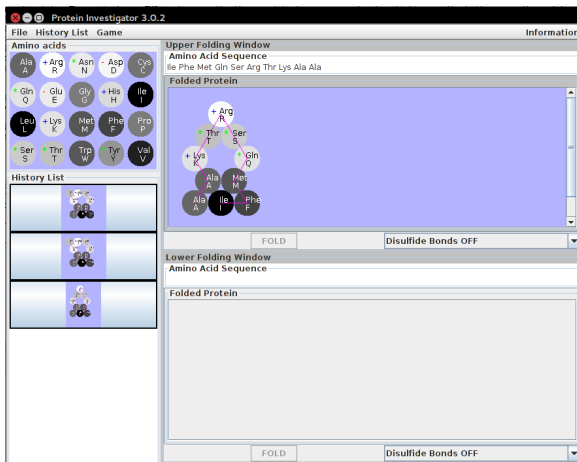
both hydrophobic and the shape doesn't change. (c) When Asp (8th amino acid) is changed into Lys, it is a change from a negative to a positive amino acid. This causes a significant change in the structure. (d) The sequence contains several oppositely charged amino acids that create several salt bridges.



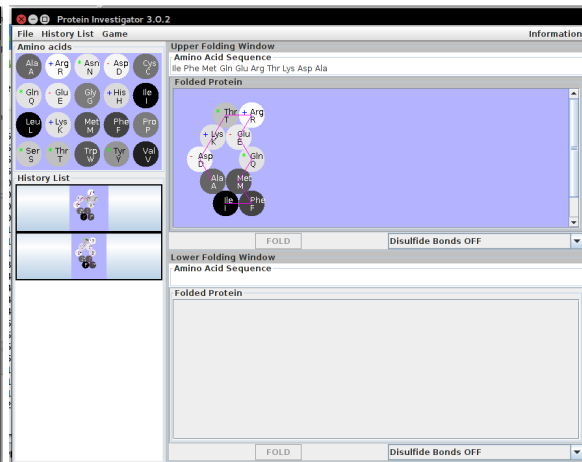
(a)



(b)



(c)



(d)

4. Given the following point sets:

Point Set A:

0.9003	-0.3258	-0.2888
-0.5377	0.2196	-0.8140
0.2137	0.8614	-0.4608
-0.0280	-0.0740	-0.9969
0.7826	0.2782	0.5569
0.5242	-0.7065	0.4755
-0.0871	0.9154	-0.3929
-0.9630	0.2336	-0.1344
0.6428	-0.6475	0.4094
-0.1106	0.7801	-0.6158

Point Set B:

-0.8842	0.4649	0.0448
-0.2943	-0.0193	-0.9555
0.6263	-0.7336	0.2636
-0.9803	0.1798	-0.0821
-0.7222	-0.6759	0.1467
-0.5945	-0.7013	0.3934
-0.6026	0.4536	-0.6566
0.2076	-0.9660	-0.1540
-0.4556	0.2610	0.8511
-0.6024	-0.3751	-0.7046

(text versions available as set1.txt and set2.txt, enclosed).

- (a) Determine the RMSD between the two point sets.
- (b) Determine the optimal RMSD between the point sets given that they are allowed to translate but not rotate. It can be shown that the optimal RMSD is obtained when the two point sets are translated so that their centroids are at the same point. The centroid or center of mass of a set of points is a point whose x,y,z coordinates are the average of the x,y,z coordinates of the point set, respectively. You can use any of C/C++, Java, Matlab, R, Python or any other acceptable programming language. Please attach your source code.

**Answer:** See R and Matlab codes as separate files. The RMSD for the original point set is approximately 1.44Å and for the translated set approximately 1.27Å.