CS483 Project Requirements

Overview

Here are some suggestions for your final project in structural bioinformatics. This will involve software development, project write up and presentation of the application along with background information to motivate the application. The main emphasis of the course is the development and implementation of algorithms related to one or more of the following:

- Accumulation of useful protein statistics
- Searching for examples of structural motifs, for example, the catalytic triad[1] or metal binding sites (see the paper by Patel et al.[2]).
- Visualization of some phenomenon related to molecular structure
- Facilitating changes to molecular structure

The topic may be related to one or more research papers, for example, the paper on backrub motions[3] to be described later.

General guidelines for content

As stated earlier, the topic must deal with the 3D structure of biomolecules (tertiary or quaternary – Please: no secondary structure predictors).

It is expected that the topic will likely be related to some algorithm or procedure that serves some practical purpose related to the 3D structure or provides some insight or analysis about a structure. Note that such a study need not involve the entire protein. Scientists studying proteins will often concentrate most of their efforts on a protein domain or the binding pocket of the protein.

Deliverables

All projects are to be handed in on the day of the presentations (to be held at the time of the final exam currently scheduled to be: MC4040 Saturday 9 a.m. April 11).

Your project write up should be about 15 pages, double spaced, say 4000 words in length. A presentation should be, at most, 20 minutes in length followed by 5 minutes for questions.

As part of Assignment 4, you will be required is to generate a one or two page description of the project so that I can offer suggestions and so that I can tell you whether or not the topic is acceptable.

Your project proposal should include the following:

- Clearly state the problem being studied with an emphasis on the structural aspects.
- Provide motivation. Why is this topic important?
- What are the algorithms involved in this study?

[1] http://en.wikipedia.org/wiki/Catalytic_triad

[2] K. Patel, A. Kumar, and S. Durani. Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. *Biochimica et Biophysica Acta* **1774** (2007), 1247-1253.

[3] I.W. Davis, W. B. Arendall, D.C. Richardson, and J.S. Richardson. The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, **14**(2006), 265-274.

- State what the deliverables will be (anticipated functionality of the application and nature of the user interface – this should be a Graphical User Interface if possible).
- Be sure to include a set of references to papers that you will be using.

References and leads for project ideas

The following list is an assortment of web pages that may provide some starting points for you when you search for a topic:

1. 3Dsig: A fine selection of current research topics can be found in the archived proceedings of previous 3Dsig meetings. Here is a list of URLs for recent proceedings:
   http://bcb.med.usherbrooke.ca/3dsig12/files/proceedings/proceedings_2009.pdf
   http://bcb.med.usherbrooke.ca/3dsig12/files/proceedings/proceedings_2010.pdf
   http://bcb.med.usherbrooke.ca/3dsig12/files/proceedings/proceedings_2011.pdf
   http://bcb.med.usherbrooke.ca/3dsig12/files/3DSIG_2012_Booklet.pdf
   http://bcb.med.usherbrooke.ca/3dsig13/files/3Dsig2013_Booklet.pdf
   http://bcb.med.usherbrooke.ca/3dsig14/files/3Dsig2014_Proceedings.pdf
   You should review some of the abstracts in the proceedings. If you find a topic of interest you can usually do a Google search to get more information, such as a complete paper on the research topic.

2. Molecule of the month: http://www.rcsb.org/pdb/101/motm_archive.do This is an excellent starting point for various studies in proteins. Goodsell usually relates function to structure and you can read the research papers associated with the protein to get more information. Recall that every protein in the PDB has at least one reference to a research paper dealing with that protein.

3. For an excellent starting point when you need information about a specific protein, try: www.sbkb.org/.

4. Chimera animations:
   http://www.cgl.ucsf.edu/chimera/animations/animations.html

5. Here is a list of miscellaneous papers that may also be of interest:

   http://www.proteomesci.com/content/pdf/1477-5956-10-S1-S1.pdf
   http://www.biomedcentral.com/1471-2105/13/286
   http://www.mpibpc.mpg.de/275448/Hub_PCB_2009.pdf

6. Structural Biology software:

   http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3640466/pdf/d-69-00701.pdf
   http://bioinformatics.oxfordjournals.org/content/27/5/723.full.pdf+html
   http://arxiv.org/ftp/arxiv/papers/1009/1009.4801.pdf
   http://bioinformatics.oxfordjournals.org/content/early/2014/02/05/bioinformatics.btu020.abstract

The Contents of the Project Report

The following points are suggestions for the content of your final report:

1. Introduction and motivation

a. Clearly state the problem or topic.

b. Why is the problem or topic important?

c. In the study of this problem, what are the techniques or algorithms used by various researchers? It will be important for you to organize this material in a way that makes it understandable and clear.

2. Algorithm to be studied

Clearly describe the algorithm being implemented along with any suspected limitations or trade-offs.

3. Critical evaluation

a. How well does the technique or algorithm achieve its objectives?

b. What are the strengths and weaknesses of the various approaches?

c. What problems or issues remain to be studied?

4. Conclusion

a. What can we learn from this algorithm?

b. Can you suggest any new avenues of investigation?

5. References

What to hand in for the project proposal:
Give the project a title and briefly describe the topic that will be studied (in the form of an abstract). State the goals of the project and what you hope to accomplish. The proposal should specify the problem and the algorithm to be implemented. Provide a list of references. The next page contains a marking template that will be used for the final evaluation of the project.

Project Notes:
I would highly recommend that students do their project using a grammar checker such as that available with Microsoft Word. Marks will be deducted for extensive bad grammar that is so bad that it could easily have been detected by Microsoft Word.

Project Marks

Technical Style of the report:

Grammar and Spelling: ___________ (/8)

Organization: ___________ (/6)

Clarity: ___________ (/6)

Content:

Introduction and motivation: ___________ (/5)

Background material: ___________ (/10)

The algorithm: implementation and testing: ___________ (/25)

Your critical evaluation or your software: ___________ (/10)

Discussion/Conclusion: ___________ (/5)

References ___________ (/5)

Presentation style:

Grammar and Spelling: ___________ (/4)

Organization: ___________ (/4)

Clarity (speaking & pacing): ___________ (/4)

Design of slides (font size, pictures, examples): ___________ (/4)

Interaction with audience: (enthusiasm, handling questions) ___________ (/4)

Total: __________ (/100)

Project Suggestions:

The following examples represent some project "sketches" that could be filled out as a project proposal.

P1: Backrub Movies
Read the paper by Davis et al.:

I.W. Davis, W. B. Arendall, D.C. Richardson, and J.S. Richardson. The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, **14**(2006), 265-274.

This project would involve the implementation of a GUI (Graphical User Interface) that initiates the following functionality:

- The user will fetch an arbitrary PDB file or one of the files suggested by the research paper.
- The application will locate the portion of the protein backbone that is involved with the type of backrub motion described in the research paper.
- The application will determine the start and end conformations associated with the backrub motion and then use a "morphing" algorithm to derive intermediate conformations for the motion.
- These are to be assembled into a Chimera movie that can repetitively illustrate the backrub motion. It is important to see how the sidechains move in conjunction with the backbone motion so that there is no steric collision.
- In addition to the movie functionality the application should measure the amount of motion involved with the objective of correlating the motion with the type of residues involved.

P2: NMR loops and the TSP
This project would involve the implementation of a GUI (Graphical User Interface) that initiates the following functionality:

- The user will fetch an arbitrary PDB file containing protein conformations determined by NMR. This will typically involve several sub-models (usually 20 conformations and sometimes more).
- The application should give the user the option of selecting an internal loop from the proteins secondary structures ("internal loop" means that the residues form a coil that is flanked on both ends with a either a helix or strand, i.e. not a coil at either end of the protein).
- The loop should be extended to include four additional residues, namely the two non-coil residues at both ends of the loop.
- The sub-models should be moved and rotated in the display so that the four alpha carbon atoms of these four residues are brought into overlap. Use the `Overlap` class in the StructBio package.
- The idea is to generate a movie that uses the loop conformations from the NMR data to act as successive frames in the movie. However, we want the movie to play the frames in an order that gives the smoothest perceived motion. To do this you should calculate the "distance" between each possible pair of loops. These distances can be used as input data for a heuristic that derives a "solution" for the Travelling Salesperson Problem. The salesperson tour will determine the sequence order of the conformations making up the movie. A reasonable heuristic strategy is to subject an initial tour (perhaps determined by nearest neighbour choices) with a sequence of "2-opt" moves. There are several Internet sources that can explain 2-opt

including YouTube videos showing how it systematically improves an initial tour generated by a naïve algorithm.

- The application should give the user various choices for the distance calculation between a pair of loops. Here are the possibilities:
    - Calculate the RMSD for the two loops using only the alpha carbon atoms.
    - Calculate the RMSD for the two loops using all the atoms.
    - Scan each loop and generate a P, M, or T label for each residue in the loop. The label will specify the chi1 setting of the residue. The "distance" between two loops would involve counting the positions in the PMT string where the labels are different.
    - Scan each loop and generate a PMT "word" for each residue in the loop. A PMT word will have as many characters as there are chi angles in the residue under consideration. The distance between two loops would involve a calculation similar to that used in the last point but the labels associated with chi2 are only involved if the labels for chi1 are the same. In a similar fashion, chi2 labels (if there are any) would only be used if the labels for chi1 and chi2 are the same, etc. The distance increment due to chi2 differences would be weighted less (say ½ of that used for chi1). Similarly, the distance contribution corresponding to chi3 differences would be ¼, with a similar pattern for chi4 if it is involved.
- The real objective of the application is to correlate backbone motion with changes in the rotameric settings of the side chains. This is why the PMT distance calculations would be more important. We want to get a better understanding about the movement of the backbone when the side chains change there rotameric settings.

P3: RIGS and surface colouration due to local closeness

This project deals with the discovery of protein binding sites that are candidates for allosteric drugs. In class we have discussed drugs that act as inhibitors by competitively binding to the active site of a protein to prevent an activity that is to be eliminated or diminished (think of a drug going into the binding site of HIV protease). An allosteric drug uses a different strategy: It binds to a nearby different binding site and causes a conformational change in the protein that distorts the binding site that is to be inhibited. This idea has been discussed in the following review paper:

I.N. Berezovsky. Thermodynamics of allostery paves a way to allosteric drugs. *Biochimica et Biophysica Acta*, **1834** (2013) 830 – 835.

This project would involve the implementation of a GUI (Graphical User Interface) that initiates the following functionality:

- The user will fetch an arbitrary PDB file.
- There should be the option of deleting chains that are considered to be not needed.
- Use the StructBio package to construct a residue interaction graph (RIG) that is to be displayed as a network graph (see Chapter 8 of the text). Nodes in the graph correspond to residues and will be displayed as small spheres that are placed at the centroid of a residue. Two nodes are connected by an edge if the corresponding residues are in *contact*. Two residues are in contact if any of their atoms are within 5 Angstroms of each other (use the `Shell` class from StructBio).
- For each node in the RIG you will calculate its *local closeness of degree m*. As in the paper use *m* = 4. Local closeness of degree *m* is defined by the formula:

$$C_m = \sum_{k=1}^{m} \frac{n_k}{k^2}$$

where n_k is the number of nodes whose shortest distance from the node under consideration is exactly *k* edges. The shortest distance can be determined by borrowing code from Script C_06 (structuralbioinformatics.com).

- Put a surface around the protein and use a strategy similar to that in Exercise 3 of Assignment 1 to give the surface a colouration that corresponds to the local closeness value. You should get a colouration that is similar to that seen in Fig. 1(b) of the Berezovsky paper.

Test your application on proteins with known binding sites to see if it can find them.

P4: Side chain packing atlas for Beta sheets.

Consider the data provided by the website: http://www.biochem.ucl.ac.uk/bsm/sidechains/

The atlas presents side chain packing distributions in terms of clusters. Typical clustering involves roughly 5 or 6 clusters (less than I would have guessed…).

The original side chain packing atlas was constructed in 1992 when there were far fewer proteins available. Interactive viewing of the distributions was done using RasMol a freely available application for molecular visualization. It was a first generation effort and does not have the display capabilities that Chimera currently has. This project would provide a new version of the atlas with a more sophisticated functionality provided by working with a larger input data set (consider the PISCES list used in Assignment 3) and your GUI designed to alter the display of information.

The atlas would provide distributions for all 400 possible pairings but to make the processing less onerous let us concentrate on pairings that have the first residue in a beta strand. The other residue would likely be on a strand as well but you should deal with any other scenario as well.

This project should involve a GUI and it is a somewhat more open-ended than the previously described projects.

The project should have three phases: data collection, data processing (clustering) and data display. The first two phases should not require a GUI.

Some suggestions:

- For data collection use the same PDB file list that you worked with in Assignment 3.
- For data processing you will have to decide on the clustering algorithm that should be used. You could try *K*-means with the RMSD calculation used to define distances. Selection of an appropriate *K* will be an issue here. The atlas is in the reference section of the Davis library and probably contains more details about the data processing done by the researchers.
- For the data display you should consider using classes from the StructBio package. I would suggest that the data be presented using a common frame of reference (similar to that done in Assignment 3 but displaying sidechains instead of lines representing hydrogen bonds).
- Consider a GUI that will give the user a choice of pairings (for example, GLU – SER) with the understanding that GLU will be in a strand. You might also wish to have different distributions for the various rotamers of the side chain. The display would then provide a representative for the clusters (as done in the website). You can skip the shadow effects on the walls of the 3D display octant (unless you have a clever way to do this[4]). You are encouraged to apply your own creativity in the design of the data display.

[4] Chimera can provide shadows (see the Effects tab of the Viewing dialog window) but I do not know if it can be used to cast shadows on a plane that you put into the scene.

P5: Structural motifs: data gathering, data processing and data display.

This project is similar to project P5 just described. The best example of such a project would be the collection, clustering and display of catalytic triads, more specifically, the SER-HIS-ASP catalytic triads. The following reference discusses the issues involved:

V. Gupta, N.A.U. Prakash, V. Lakshmi, R. Boopathy, J. Jeyakanthan, D. Velmurugan, and K. Sekar. Recognition of active and inactive catalytic triads: A template based approach. *International Journal of Biological Macromolecules*. **46**(2010) 317 – 323.

The paper implies that roughly 1 out of 12 proteins contain such a triad with the SER-HIS-ASP residues in close proximity to one another but the geometry of the triplet may indicate that catalytic activity is not likely. Consequently, the paper presents various distance constraints to eliminate the unwanted configurations.

As before, the application should display representatives of the triads along with data (bar charts) that characterize the geometry of the triad.

Use your imagination to come up with a display that helps to explain the functionality of the triad.

If you choose to go wtih one of these projects then send email to me as soon as possible. I want each student to have a different project. If two students want the same project it will be assigned to the student who has sent the email request first. A suggestion: instead of picking just one of the above projects, your request can be a list sorted in order of most preferred down to least preferred. In this way you are more likely to get your next choice if the first or second choice is already assigned.