University of Colorado at Colorado Springs

CS 582: Bioinformatics

Project Ideas

You are required to write a project proposal that is brief, say no more than 2 or 3 pages, single spaced, using letters that are about the size you see in this document. The proposal should contain the following sections: Introduction, Problem Statement, Related Research, Proposed Approach, Proposed Schedule, References. Each proposal must show that you have read 5-10 relevant research papers or documents, or books. Use the Web, and the print and electronic resources in our library. Because, we are novices in biology, I know it is going to take quite a bit of work to read relevant material and filter out the parts that we need.

The project ideas given below are preliminary. You can refine them any way you want. Several people can work on one topic as a group. If so, each person must focus on a specific aspect of the problem, and the software generated must work together. Feel free to contact anyone anywhere in the world working on a topic you choose with regards to research papers, data, and advice. There is a lot of data available on the Web, but one needs to know where the data is, and then be able to interpret and use it. Also, our text books have lots of chapters that we will not be able to cover or cover cursorily. You can research into any of the topics discussed in our text, read related papers, get relevant data, and perform experiments.

What I am going to look for in judging your project is the amount of work you have put in and the status of the working aspects of the project. I expect everyone to put in sustained work over the semester and keep me informed of what you have accomplished. Come and show me what you have done during my office hours every week or two.

1. **Working with our Biology Department** Talk to professors in our Biology Department, professors such as Melamede, Newell, Mattoon, Berry-Lowe, and Broker, etc., and see if they have any problems they are interested in. Work with students or professors in Biology.

2. **Coiled-Coil Protein DB**: Develop a database based on David Brinkman's MS thesis from last year. He worked on so-called coiled coil proteins and developed a set of tools to analyze them. We need to enshrine all in a database and update this database on a regular basis.

3. **Genome database for a simple organism** Understand the genome sequence of a simple organism such as *E. coli* and a few other simple species and develop a database that can be queried in various ways by experts and non-experts.

4. **Simulation of life's origin** Consider the 20 or so chemical elements that play a role in life. These were the chemical elements that were present in the early stages of earth's creation and development of life on earth. Research into one or more theories of how life may have evolved in earth over 3 billion years or so. Develop a simulation model to show the creation of life on earth from the simple chemical elements. Verify the model with respect to the paper(s) you read.

5. **Whole Genome Comparison** Research into how whole genomes can be compared. Compare genomes of 2-20 bacteriophages or viruses using whole genome comparison techniques.

6. **Whole Genome Comparison** Collect as much data as possible about all types of genomes that have been sequenced: size, number of genes, frequency of A and C, etc. Now, make a database that can be queried. Provide some whole genome comparison tools. E.g., we can try to build a phylogenetic tree based on simple things such as the size of a genome, the number of genes, the amount coding/genic areas in the genome, C-value, etc. Or, determine based on some simple criteria the distance between two genomes, etc.

7. **Mitochondrial DNA**: Find on the Web where it's possible to obtain a large number of mitochondrial DNA sequence data. Perform comparison experiments on the DNA, e.g., for finding the Mitochondrial Eve.

8. **Ion Channels**: Study/research into ion channels and make some simplistic software models.

9. **BioPerl** BioPerl is a set of Perl modules for various bioinformatics tasks. It is a open-source project. Go to a site such as *www.bioperl.org* and see how you can help write one or more modules and contribute it to the BioPerl project.

10. **Graphics and Visualization Tools** Reseach into protein visualization programs available on the Web. Download a few, read about it, and implement one of your own.

11. **Taxonomic Database** Create a taxonomic database like the one is `www.sp2000.org`. Design the database considering the questions people may ask. Use MySql or PostgresQL and make it work on one of our Linux systems. Link it from the top page of the Bioinformatics Web site. Allow a Web-based interface to populate the database. Think of quality control issues when anyone is allowed to enter data into it. Show graphics for as many species as possible. Show trees by creating them on the fly. Link to relevant sites. Come up with other ideas to make it worthwhile to come to the site.

12. **Genome Projects** There are about 100 fully mapped genomes and all these are available on the Web. There is work in progress on about 800 genomes according to what I have read. Create a Web portal to these genome sites and link it from the Bioinformatics site at UCCS. Think of what should go into such a portal. For example, one could crawl the Web automatically, update data about statistics and other relevant material, say on a daily or weekly basis. One can provide meta-search abilities of various kinds. One can provide illustrations of various kinds. One can link or write programs for comparative genome searches. One can provide links to taxonomic databases. Etc.

13. **Modeling Cell Division** Develop a class structure for a cell. Develop methods that depict cell division in detail. Provide graphical simulation. Go into as much details as possible. This project should be done in coordination with a few MS students who are working on a similar project already. Use and XML-base language to write your model first.

14. **Modeling Viruses or Bacteria** Develop a model, in an XML-based language for viruses or bacteria. Read about how they work, how they attack their prey, multiply, etc. Read about the genetic issues. Model all this and then program it.

15. **Aspects of the Immune System** There are a few MS students who have started working with on modeling aspects of the immune system. They are working on a gross model, a cellular level model, a genetic level model. There will be XML-language based models and also graphical models that will accompany the simulations. You can talk to me or some of the students if you want to participate in a certain aspect of this project.

16. **Disease Database** Design and build a database of various diseases for humans and other organisms. Provide descriptions of diseases. Research into the genetic causes of diseases. Show the genetic causes in the database or provide links to appropriate places. Allow searches of various kinds. Write in such a way that newcomers can understand, but at the same time, there is sufficient and useful information. Allow outsiders to add information, but be mindful of quality control issues. Link it to the Bioinformatics Web site at UCCS. Look at the *OMIM* database at www.ncbi.nlm.nih.gov for a similar but very detailed database.

17. **Immunogenetic Database** Look at the Web site http://imgt.cnusc.fr:8104/. It provides data regarding how genetics and the immune system are related in a lot of detail. See if you can design and create a database like this at UCCS. Make it at a level that even newcomers understand. Allow searches and other facilities.

18. **Clustering Algorithms and Gene Array Expression Data** I have access to some gene array expression data and there are tons of gene array expression data on the Web, although in different formats. There is an MS student working on her project implementing a few clustering algorithms. You can research into clustering algorithms of various kinds and work with either data I have obtained from the UCHSC or data you collect from the Web. Come up with a new clustering algorithm and implement it, or implement several clustering algorithms and compare results.

19. **Phylogenetic Trees** Research into various techniques for phylogenetic database creation. Look at our texts and any other books and Web sites. Implement a few of the algorithms on realistic data you collect from the Web.

20. **Protein Structure Prediction** Research into various algorithms for protein structure prediction. Look at our texts, any other books and Web sites. Implement one or more algorithms. Test on realistic data you collect from the Web.

21. **Gene Prediction** Gene prediction consists of identifying regions of genomic DNA that encode proteins. Some of the existing models that identify and distinguish coding regions from non-coding regions are based on: Hidden Markov Models, Neural Network, Probabilistic Models, Linear Discriminant Analysis, Decision Tree Classification, and Quadratic Discriminant Analysis. Choose one or more of these techniques and implement a prediction (i.e., search algorithm) that will be able to search a given database for genes that do code for proteins.

22. **Drug Discovery**: This is a very hot area now. It involves modeling molecules, making databases of molecular structures, and then doing pattern matching in 3-D. Research into academics and companies that are doing work in drug discovery. Search using search engines. Look at `www.researchindex.com` for papers. Read them, and implement one or more algorithms, or come up with algorithms of your own.

23. **Distributed Computing** Many of the algorithms in bioinformatics deal with huge amounts of data. Research into parallelizing one or more algorithms, say sequence alignment, similarity searches, or clustering.

24. **Genetic Networks** Research into how one can build genetic networks for regulation. Look into techniques, and algorithms, and implement one or more.

25. **Cancer and Bioinformatics** Genes play a big role in cancer. There is a large amount of study linking genetic disorders to cancers of various types. Research into this area and see how one can make computer models based on this research for predicting cancer. For example, one can use gene expression data for cancer diagnosis. I have a paper on this topic.

26. **Protein problems**: The book *Bioinformatics* by Baldi and Brunak lists a lot of problems and issues in regards to proteins in Section 1.5, page 43, of their book. See if any problem catches your fancy. Read the papers listed. Contact the authors if you need to.

27. **Karp's paper** Read Karp's paper on *Mathematical Challenges in Bioinformatics* and focus on any problem he refers to. Research into the problem and implement one or more algorithms.

28. **Looking at workshop and/or conference sites** Look at a Government Web site such as http://www.ornl.gov/hgmis/publicat/02santa/bioinform.html that contains research abstracts from the DOE Genome Contractor-Grantee Workshop IX, or similar sites. Look at the Pacific Symposium on Biocomputing at http://psb.stanford.edu/psb01/ and other such sites. Read the abstracts and see if there is anything you like. Get the paper and read it. Contact the author. Implement the ideas mentioned in the paper.

29. **Researchindex.com** This Web site contains a large collection of computer science research papers. Go to this Web site and search for topics in bioinformatics. Find a few recent papers, read them, and implement one or more algorithms in the papers. Compare results.

30. **Chapter 12 of Pevzner's Book** There is a very large collection of problems for projects in Chapter 12 of Pavel Pevzner's book on all sorts of problems in Computational Biology. Please go through the list and see if there is anything of interest.

For each project, please make a Web page on the *dirac.uccs.edu* machine. Please post updates and documents on the site. It would be good to have PDF documents instead of Word documents. I will link the pages from the Bioinformatics Web page.
The final report for your project should have the following sections.

1. Title and author information
2. Background and literature search
3. A description of the data you use
4. A description of the methods, algorithms, etc., you use.
5. Results you have obtained.

6. Conclusions you have made based on the results.
7. Possible directions of future research.