SPECIAL TOPICS IN BIOCHEMISTRY BCHM 8995

INTRODUCTION TO APPLIED BIOINFORMATICS

FALL SEMESTER 2008

Professor: Ricardo González Méndez, PhD

Professor of Radiological Sciences

758-2525 ext. 2121

rgonzalez@rcm.upr.edu

Units: 3 credits

Day/Time: Tuesdays 9:00 -11:30 am.

Description: This course is designed to provide an understanding of important topics in applied bioinformatics and computational biology. Students will be able to learn about problems involved in the analysis of biological data such as DNA/protein sequences and protein structures. The course is intended to provide a good understanding of the commonly used algorithms in the analysis of genomic data, hands-on experiences with accessing and using relevant databases, the use of advanced computer programs for sequence analysis, protein sequence phylogenetic analysis, visualization of results, and a basic overview of genome analysis. The course is intended to be flexible and adaptable to students’ needs.

Objectives: To provide students with a thorough understanding of current computational techniques available to biomedical scientists for processing, analyzing, and understanding DNA and protein sequence and structural analysis. This will be accomplished through lectures, coupled with lab sessions where students will interact with the resources discussed in class. Student comprehension of computational biology concepts and utilization of computational resources will be assessed through individual projects which will evolve throughout the semester. In addition, students will be exposed to a variety of research avenues available to the computational biomedical scientist.

Prerequisites: Basic Biochemistry and Molecular Biology and undergraduate math.

Students: 10 students maximum

Textbook: Bioinformatics: Sequence and Genome Analysis, 2nd Edition

David W. Mount

CSHL Press, Cold Spring Harbor, New York, 2004

Supplemental readings (will be provided by the Professor as needed):

Bioinformatics for Dummies 2nd edition – J-M Claverie, C Notredame - 2007

Original papers and reviews from the scientific literature.

Additional Resources for computing (available from the Professor):

Learning the Unix Operating System, 5th Edition: A Concise Guide for the New User (2001) *Jerry Peek, Grace Todino, & John Strang*, O'Reilly & Associates, Inc. Sebastopol, California.

Developing Bioinformatics Computer Skills (2001) *Cynthia Gibas & Per Jambeck*, O'Reilly & Associates, Inc. Sebastopol, California.

Grading: Final grades will consist of attendance (10%), homework problems (40%), and an individual project based on a student-selected family of sequences (50%). The project grade will consist of two progressive versions due during the course of the semester (10% for each part), and a brief (15-20 min) oral presentation (10%) of the final report (20%) to the class during the last week of the course. A breakdown of the grading is provided below.

|  |  |  |
| --- | --- | --- |
| Activity | Percent | Points |
| Attendance | 10% | 50 |
| Homework | 40% | 200 |
| Course Project parts 1-2 | 20% | 100 |
| Project Presentation and final report | 30% | 150 |
| Total | 100% | 500 |

# Instructional Strategies: We will use lectures, and remote, interactive videoconferencing to present the theory. We will laboratories and homework exercises to provide an active learning experience in the application of the theory. Finally, we will use a student selected project as an active learning experience that is directly applicable to his research career as a scientist including audiovisual presentation and written reports to demonstrate the skill set and expertise acquired throughout the course.

# BCHM 8995 Course Project

A significant part of your grade in Introduction to Applied Bioinformatics will be based on a four-part project. The project will mirror our class discussions, and involves an extensive analysis of a family of protein sequences. The sequence analysis portion of your project should be integrated with the biological, phylogenetic, and structural information you obtain about your family. The final goal will be to have a report that would be useful in designing laboratory experiments to further characterize and understand the protein family.

Parts of the project will require the use of the GeneDoc multiple sequence alignment editor. This program is available at no cost from <http://www.nrbsc.org/gfx/genedoc/index.html>. GeneDoc runs on PC’s and laptops using the Windows XP Operating system. You should obtain the program and work through the tutorial material that comes with it in order to become familiar with its use.

**Part I**

**A - Goal**: Submit three candidate families for your course project.

**Assignment**: Search for and evaluate a few protein families to use for the project. You will answer the following questions:

1. Approximately how many different sequences are available in this family?
2. What is the approximate sequence length of members of this family?
3. What is the average sequence identity between members of this family?
4. Is there a known three dimensional structure for at least one member of the family (supply the PDB code)?

**Required Report Elements**: Pick three sequence families/domains for which at least one three dimensional structure is known and write a 2-3 page (double-spaced, typed) report on each of the three families/domains. Each report should contain answers to the four questions above, as well as a brief description of the protein's biological function. Rank the three families/domains in order of your interest.

**B - Goal**: Collect an initial set of sequences that belong to your protein family, generate a multiple sequence alignment, and provide a GeneDoc representation of your results.

**Assignment**: For your approved family/domain, identify related sequences through searches of sequence and classification libraries. Use your results to create an initial multiple sequence alignment. Perform three database searches with the algorithm/program/website of your choice (Smith-Waterman, BLAST, iProClass, GenBank, UniProt, EBI, etc.). Each search should use a similarity matrix appropriate to a different degree of sequence divergence. The searches should use a complete protein sequence database (UniProt knowledge base; Swiss-Protein+TREMBL; or the NCBI RefSeq nr (non-redundant) database). Align sequences you have decided to pursue using ClustalW, and import this initial multiple sequence alignment into GeneDoc. Use GeneDoc to examine areas of amino acid identity, as well as areas of chemical and physiochemical conservation (see <http://prowl.rockefeller.edu/aainfo/contents.htm>). Examine the sequence database entries as well as the recent literature on this family for information about essential or critical residues for these proteins.

**Required Report Elements**: Supply a list of high scoring sequences from each database search. Indicate the program and scoring matrix used. Provide a list of the unique sequence identifiers you have chosen to pursue, with their species of origin indicated. Include a schematic view of the alignment in the report highlighted by conservation (4 levels), by chemical properties (2 levels), and by physiochemical properties. Discuss whether or not residues identified in he literature as known to be essential to the function of your protein, have been properly aligned between the sequences, as well as other inferences you make from the different highlighting modes. In addition, discuss whether any members do not seem to belong in this family.

**Part II**

**A - Goal**: (1) To improve the quality of the alignment, (2) use abstractions of your alignment to identify additional family members, and (3) create a phylogenetic tree based on your family members

**Assignment**: Edit the list of unaligned sequences used for part two of the project to remove duplicates, sequence fragments, and other undesirable sequences. Submit these sequences for a motif analysisi using MEME. Ask for MEME to return approximately one motif per 40 amino acids of your sequences length. Use your MEME results to:

1. Highlight motifs in GeneDoc to help you refine your initial alignment

After manually editing your GeneDoc alignment, create a PSSM and Hidden Markov Model based on your alignment and search the databases using PSI-BLAST or use the program MetaMEME for the HMM search to identify distantly related members of your protein family. After adding additional family members discovered by the three database searches above and optimizing the alignment, create a consensus phylogenetic tree using PHYLIP.

**Required Report Elements**: All report elements for this project must be contained within a single file (e.g., Microsoft Word), and accompanied by your latest alignment file. Your report should include the following:

1. A reminder of your protein's function and essential residues
2. For each motif discovered through MEME, report the "multilevel consensus sequence" and indicate the amino acid with the highest information content (see meme out file).
3. A list of your current family members in the same order as your GeneDoc alignment
4. A phylogenetic tree whose taxa labels indicate protein name and species name.
5. Use one of the group classification models to detect what residues are relevant to the subfamilies within your protein family alignment. If possible show this in your Genedoc visualization of the multiple sequence alignment.
6. For your discussion:
   1. Are the essential residues you reported in part 2 conserved across all family members? If not, discuss your concerns. Provide a graphic of this specific region with your essential residue(s) indicated. Are there any highly conserved areas not discussed in the literature?, etc.
   2. Relate the phylogenetic information you've discovered with your family. Do your sequences separate into obvious groups? Discuss why they may group this way. Are there any areas of conservation within a single group that are not shared with other family members?, etc.

**Final Presentation**

**Goal**: Create a final presentation on the analysis of your protein family and write a final report that collects all the elements of your presentation in detail.

**Required Presentation Elements**:

1. Your presentation should be in PowerPoint, and be approximately 15-20 min long.
2. A description of your proteins biological function
3. A graphic of at least one conserved motif in your multiple sequence alignment
4. At least one 3D structure clearly indicating the residue/motif you're discussing above
5. Other than the four required elements above, the theme of your presentation in up to you. I want you to tell us something interesting that you've learned about your family. Do your results suggest any future experiments that could be done? Anything in your MEME/phylogenetic tree/structure/homology modeling that proved enlightening? In particular, we’re looking for integration between the different methods (i.e. how your phylogenetic tree results helped to sort out your family members; did the structural visualization make sense of your conserved motifs; etc.) Are there key residues for the family or subfamilies in your alignment? Where are the key residues located structurally? What can you tell about protein function, residue conservation, and structure? How does this relate to subgroups or subfamilies?
6. In addition to your PowerPoint file hand in a copy of your final alignment file, meme output file, PDB file(s), and any VMD images you've generated.

**Final Report**

The Final Report should be no longer than 6-8 pages (typed double space) in addition to including relevant figures and diagrams from the presentation. It should provide a written summary of your presentation. It should follow the standard form a research paper including Intro, methods, results, discussion, conclusions and future directions, and bibliography.