Syllabus: Computational Biology (BIOL F692)

August 31, 2011

1 Instructor contact information

Naoki Takebayashi, WRRB 226, 474-1178
e-mail: ntakebayashi@alaska.edu
Office hour: any time but you can drop me an e-mail before you come.
E-mail is the best way to contact me.

2 Meeting time and place

Irving I 303
TR, 9:45-11:15AM
3 credits

3 Course Description

Computation has been used in biology since 1960's. In the recent years, computational biology has been moved into the central domain of biological science. The boom in computational biology is motivated by both availability of enormous data set (e.g., in bioinformatics) and complexity of biological systems (e.g., in simulational studies). Programming skill is essential in computational biology, but it may not be readily accessible to most biologists without guidance. Yet, practical programming skills needed for biological problems are relatively simple. It can be learnt and applied for their own biological research without formal computer science courses. The course will expose students to the first-hand experience of programming, specifically tailored for biological applications. The goal of the course is that students without any previous programming experiences become able to apply the programming skills to solving their daily biological problems.

This course is motivated by my personal experiences helping my friends, who were painfully analyzing huge data sets by hand (with spreadsheet programs) or whose brilliant idea could not be tested, because they did not know how to do computer simulations.

First, we will cover basic unix environment. Then, we will learn higher-level languages, such as Perl and R, which are useful for large-scale data analysis and visualization. In each section, we will begin with biological questions, and then we will investigate how to approach the problem. The underlying theory or statistical techniques will be discussed, and programming techniques and algorithms will be explained. We will employ several programming languages (perl, R), which have strengths and weaknesses and complement each other.

The students should have elementary knowledge of computers (e.g., how to use keyboard, mouse, etc), but are not expected to know how to program or work within unix computer environment. During the class, we will meet in a computer lab, and access unix server, but students are encouraged to use their own computer.

4 LSI login server

We use the login cluster, maintained by UA Life Science Informatics Core. To access it, you use the following command:

```
ssh -p 55000 -Y username@tuxedo.inbre.alaska.edu
```

5 Course home page

Lecture materials can be found in:

```
http://raven.iab.alaska.edu/ ntakebay/teaching/programming/index.html
```

6 Approximate schedules

Week 1. (Sept 6-8) Unix environments 1:
- Unix basic commands, text-editor
- Practice CLI with Naoki's Perl script

** CLI: p.1-43, emphasis on Chapter 5
** CLI: p.47-54 (permission and ls), p.78-81 (redirection)

Week 2. Introduction to Perl 1, Part 1
- Quick Elements of Perl Programming
- Variable types, Array, Hash

Quiz about CLI on Thr.

** PERL: Ch. 2 (Scalar data)
** PERL: Ch. 3 (Array and list data)

Week 3. Introduction to Perl 1, Part 2
- Flow Control
- File I/O
- Subroutines

** PERL: Ch.6 (hashes, Ch.5 in PDF)
** PERL: Ch.5 (Basic Input Output, Ch.6 in PDF) & Ch.4 (subroutines, Ch.8 in PDF)

Week 4. Introduction to Perl 2 - text handling
- regular expression

** PERL: Ch.7-8 (Regular expression, Ch.7 in PDF)

Week 5. Introduction to Perl 3
- Functions for Array and scalar operations

** PERL: Ch.9 (Processing Text, part of Ch.7 in PDF)
** PERL: Ch.10 (control structure, Ch.4 in PDF)

Week 6. Application of perl to DNA sequence analysis
- Bioperl
- interface to genbank

Week 7. Application of perl to DNA sequence analysis
- sequence manipulation
- command line blast
- 2nd gen sequencing data manipulation

Week 8. Probability theory
- basics
- Patterns in DNA

Week 9. Visualization and statistics with R 1
- Elements of R
- Graphics

Week 10. Visualization and statistics with R 2
- Basic statistics (e.g. linear models)
- Simple simulations with R

Week 11. Approximate Bayesian Computation
- Bayesian thinking
- Computational approach

Week 12. ABC in phylogeography
- coalescent simulation (ms)
- msbayes
- Demographic model
- Thanksgiving break (no class on Thr)

Week 13. Topics in Bioinformatics
- Applications used in Bioinformatics

Week 14. Project Presentation
* Presentation by grad students
* Presentation by undergrad is optional

** Reading Assingment Abbreviation:

CLI: Joe Barr 2007 CLI for Noobies: A Primer on the Linux Command Line. Prentics Hall
PERL: Schwartz, R. L. and T. Christiansen. 2001. Learning Perl. O'Reilly and Associates.
The current one is the 6th edition. PDF is based on the 4th edition, the chapters, so the corresponding chapters of PDF are indicated in the parentheses.

6.1 Important dates

Week 4.

** Student presentation of project ideas

Week 14.

** Presentation of Project.

7 Course readings/materials

Textbook:

- Schwartz, R. L. and T. Christiansen. 2001. Learning Perl. O'Reilly and Associates.
- Barr, J. 2007 CLI for Noobies: A Primer on the Linux Command Line. Prentics Hall
- Dalgaard, P. 2002 Introductory Statistics with R. Springer-Verlag, New York (recommended, not required)

8 Course goals

Students will learn basic programming skills useful for biological problems. After the completion of the course, students should understand how to abstract biological phenomenon and should feel comfortable in developing computer simulation, or make programs for biological data analyses.

9 Instructional methods

Students will learn through lecture, reading, and group discussion.

10 Course policies

You are expected to attend lectures and participate in discussion. You are expected to arrive at lecture on time.

11 Requirements

All students will be required to do readings and homework assignments. I encourage students to work on the homework assignments together. You are likely to "feel" the real meanings of concepts or techniques by exchanging different ways of interpreting them with your colleagues. Since practical skills are aquired only by doing them by themselves, there will be homeworks throughout the semester, and approximately 1/3 of grades comes from the homework assignments.

Additionally, part of the grade is based on your programming project. I encourage you to choose a topic which is related to your own research.

12 Evaluation/Grading

Student performance will be evaluated with the following factors
30% assignments
10% proposal (due week 4)
20% presentation (due final week)
30% final manuscripts (due final week)
10% participation to group discussion

Assignments: Majority of learning will come from homework assignments. I will assign homeworks after each lecture. The majority of homeworks is application of the concept from the lecture to solve some small programming problems. Each homework may take from 10 min. to 3 hours depending on the complexity. Although the total number of homeworks is not pre-determined, I expect that there will be at least $h = 15$ homeworks. Completion of each homework earns $30 / h$ point (approximately 2 points per homework).

Final project: One of the most important part of the course is working on a self-motivated programming project. The project should be related to your own interests: ideally a publishable results, or developing tools useful for your research. The students should demonstrate the competency in programing. Within the first 4 weeks, the students will hand in a brief (1-3 pages) proposal, outlining the biological problems

you are going to tackle. Before the proposals, students are encouraged to come to talk to me about your ideas of projects. Students will present the results during the final week. Then the manuscripts is due in the final week. I expect the manuscripts to be full-blown, publication quality (15-25 pages).

Group discussion: The class will be interactive, and there will be oppotunities for students to discuss pros and cons of different approaches to solve biological problems of algorithms. I expect that all students will participate in the discussion. The volume of comments is irrelevant, and thoughtful, constructive comments are evaluated high. Maximum of 10 points.

Final grades: A: $\geq$ 90 points, B: $\geq$ 80 pts, C: $\geq$ 70 pts, D: $\geq$ 60 pts, F: $<$ 60 pts.

Students are required to obtain $\geq$ 60 points to pass the class. Students are not evaluated relative to the class mates; they are not enemies, but they are your colleagues. I will not modify this absolute scale (i.e. no curving).

13 Support Services

If you require more assistance than can be provided in class, lab and office hours, you may want to contact Student Support Services (http://www.uaf.edu/sssp/).

14 Disability Services

If you have a disability, or think you may have a disability, please contact the Office of Disabilities Services (203 WHIT, 474-7043). We will work with this office to provide reasonable and appropriate accommodation to students with disabilities.