Computational Genomics

BIOL 8803 A     Spring 2009

I. King Jordan & Andrew Conley

Office hours by appointment:
king.jordan@biology.gatech.edu
404-385-2224
Cherry Emerson 215

aconley@gatech.edu
404-385-1264
Cherry Emerson 217
Genomics & Computation

Genomics involves the characterization & study of complete genomes.

Genomics = experimentation + computation.

Computers needed to handle large data sets (obvious, perhaps trivial).

Computers needed to convert information into knowledge.

Genome sequencing efforts (along with functional genomics efforts) yield information alone.

Computational tools must be applied to bring light to that information.
Experimentation vs. Computation

Experimentation:
1. Extract DNA from biological sample
2. Produce (characterize) sequence from extracted DNA

Computation:
1. Interpret (read) results from sequencing reactions
2. Output experimental results in human/computer readable format
3. Assemble sequence fragments into contiguous sequences (contigs)
4. Find (predict) gene locations in raw sequence (exon/intron boundaries)
5. Annotate (predict) the function of the genes
6. Compare genome sequences within and between species
7. Create databases that allow for searching and dissemination of genome annotations

Therefore:
Computation is more critical to genomics than experimentation!
In this class, you the students will complete all of the computational phases of a complete (microbial) genome project.

Starting with unassembled genome sequence data

*Neisseria meningitidis* provided by the Centers for Disease Control (CDC)

Finishing with a publicly available genome sequence browser

This course is unlike any course you have had before

This course is entirely practical

This course is centered on work and results

This course is real – you will be solving an actual problem with real data
Why run a course like this?

This course meets a specific need for more practical training that has been articulated by Bioinformatics students and faculty.

Real world training on the most up-to-date technological platforms – e.g. we will analyze 454 sequence data and use the latest in analytical (computational) tools.

There is no way to ‘spoon-feed’ this kind of knowledge and experience to students (‘sage on the stage’ will not work here).

The only way to relate these skills is to have you do them yourselves – this is the ultimate ‘active learning’ course.

The burden of making this course successful will be placed squarely on the students.
The corporate model

In order to facilitate this novel pedagogical model, we will be adopting a corporate model for the course

Chief Executive Officer (CEO) – King Jordan

Chief Operating Officer (COO) – Andrew Conley

Chief Information Officer (CIO) – Troy Hilley

Share holders – Leonard Mayer and the Meningitis lab at the CDC

Management & Employees – you the students

Consultants – expert guest lecturers
Corporate responsibilities

CEO – establish plan of attack, assemble team, provide resources, delegate activities

COO – oversee and inform all use of technology, liaison between employees (i.e. students) and CEO and CIO, runs software

CIO – maintain and run hardware, install software, trouble-shoot

Share holders – set up problem and provide raw data

Consultants – provide employees with expert guidance on the use of technology and analysis of data
Guest lecturers (consultants) [i]

Leonard Mayer - CDC Meningitis Laboratory - *Neisseria meningitidis*

Lee Katz - Georgia Tech Bioinformatics - Computational genomics of *Neisseria meningitidis*

Andrey Kislyuk - Georgia Tech Bioinformatics - Genome Assembly

Scott Sammons - CDC Bioinformatics Core Facility - Bioinformatics of 454 Genome Sequencing
Guest lecturers (consultants) [ii]

Mark Borodovsky - Georgia Tech BME & CSE - Gene Prediction

Dhwani Govil - CDC Bioinformatics Core Facility - CDC Functional Annotation

Andrew Conley - Georgia Tech Bioinformatics - Generic Model Organism Database Software Platform
Employee (student) responsibilities

Technology acquisition – learn relevant approaches and tools including the underlying theory

Choice of appropriate technology – evaluate the performance of different tools, choose the best tool(s) for the job

Explanation of technology acquisition and choice – clearly relate to your peers why you made the choices you did, relative assessment of performance should be used here, demo showing preliminary results, if complementary approach needed then explain

Knowledge distribution – for each group, ensure that classmates from other groups also acquire knowledge and experience in your domain of expertise

Perform analysis – do the actual analysis your group is charged with, report the results to the class in a lecture and on the Wiki, get the results into the genome browser, iterate as needed
Benchmarks for success

1. Actively engage in classroom discussions and lab work

2. Demonstrate that your group understands the theory and the state-of-the art for your specific analytical phase (Group Presentation I)

3. Clearly justify your choice of the tool(s) to be used for your analytical phase, demonstrate comparative performance (Group Presentation II)

4. Do analysis, produce & document results, present results and integrate into genome browser

5. Work closely as needed to help other groups succeed in their phases (see ‘Virtual’ Distributed Computing) and to help other groups acquire knowledge and experience in your domain
Group activities

Students will break into 5 groups, each of which will be charged with completing one specific computational phase of the project

1. Genome Assembly
2. Gene Prediction
3. Gene Functional Annotation
4. Comparative Genomics
5. Production of a Genome Browser
Group composition

Bioinformatics students have varying backgrounds and skill sets

E.g. Some of you come from math/physics, some may be biologists, others may be programmers (of course the ideal student will have a combination of these skills)

Groups should be made of up of individuals with complementary skill sets:

Each group should one or more members who can program efficiently

Each group should have members who can work comfortably in the Unix/Linux command line environment

Each group should have members with biological training and perspective

Ideally, groups should have members with specific-skills relevant to each task – e.g. gene finding experience for gene prediction & database experience for the genome browser group
Group seed members
(please arrange to see me)

1. Genome Assembly – Daudi Jjingo
2. Gene Prediction – Jeffrey Martin
3. Functional Annotation – Pushkala Jayaraman
4. Comparative Genomics – Emily Rogers
5. Genome Browser – Viswateja Nelakuditi
# Group member questionnaire

<table>
<thead>
<tr>
<th>Programming</th>
<th>Unix/Linux</th>
<th>Biology</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daudi Jjingo</td>
<td>Jeffry Martin</td>
<td>Emily Rogers</td>
<td>Viswateja Nelakuditi</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daudi Jjingo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


No freeloaders

Active participation by all group members is required

Delegation of workload within groups will be entirely determined by the groups

Group members should invest substantial time and effort upfront to ensure optimal analytical design strategy and workflow

If problems arise in terms of effort distribution – i.e. if individual members are not contributing sufficiently – then there are 3 successive levels of control to address this:

1. Work to resolve issue within group (use peer pressure)
2. Consult with COO Andrew as to how best resolve issue
3. If steps 1 and then 2 fail, consult with me and I will address the issue
‘Virtual’ distributed computing

Distributed computing happens when tasks are split up and distributed among processors in order to speed up and facilitate their completion.

Students are strongly encouraged to perform ‘virtual’ distributed computing whereby tasks are split up and distributed among individuals.

Most students will be idle while each active group completes its specific task.

Active groups may choose to devise a working model that incorporates members of idle groups.

This will be left completely to the discretion and design of the groups (e.g. you may decide that it is more trouble than its worth)
Sharing knowledge across groups

Each group is ultimately responsible for one and only one phase of the genome project.

This means that for much of the course students will not be actively engaged in computational genomics problem solving.

How you choose to spend this time will determine to a great extent how much you get out of the course.

By no means should you wait until your part of the course to start working on your problem – research into your area and the tools available should begin right away.

In addition, groups will be responsible for sharing the knowledge they gain with members of other groups.

The process will also involve active learning and will take the form of in class laboratories and demonstrations that will be conducted by each group (more to follow on this).
Group presentations

Each group will be responsible for making a series of class presentations, labs and demos

1. Background & Strategy – explain theoretical background for what you will do, related state of the art, lay-out your general strategy

2. Tool Demo & Preliminary Results – explain and justify your selection of the best tool(s) to use to solve the problem, should include comparative results analysis, demo the tool in such a way that all students can use it, present some preliminary results

3. Results – give a detailed presentation of your final results, show carefully what was done and how final results were achieved, illustrate the kinds of problems that arose and how they were dealt with, results should be reproducible based documentation provided on Wiki page
Labs & Demos

In order to enable the sharing of knowledge and experience between groups, each group will be responsible for conducting one in class laboratory session and one in class demonstration.

The lab sessions will take place during the second group presentation.

In the lab sessions, each group will come up with a series of computational exercises that can be started in class and completed in or out of class.

Groups will be responsible for coming up with a rubric by which students from other groups work will be evaluated.

Group leading the demo will be responsible for assigning grades to each other group.

The demo sessions will take place during the third group presentation and may extend (as needed) into the fourth presentation.
Group evaluation & grading
(see syllabus for details)

1. All class members will be evaluated on their overall class participation including lab exercises – 12.5% of final grade

2. Group presentation I – 12.5% of final grade

3. Group presentation II & Lab – 12.5% of final grade

4. Group presentation III & Demo – 12.5% of final grade

5. Final Results and Documentation – 50% of final grade
Contingency plans

The coursework is inherently sequential & progressive

The successful completion of each phase of the analysis hinges upon the previous step

We will implement a series of contingency plans in the event that any given step in the analytical pipeline breaks down

E.g. if the assembly doesn’t work then we can provide an assembled genome, stripped of annotation, to the gene prediction group

Hopefully we will not have to resort to this
Computational resources

The School of Biology has provided a dedicated Linux server for this course – compgenomics.biology.gatech.edu

In addition, all lab computers are dual boot Linux and Windows (soon to be replaced with Macs running OSX)

We have installed a number of bioinformatics software packages on the server and on the lab computers – we can install more as needed

Andrew Conley will describe this resource and the other lab facilities shortly

All systems and install requests are to be made through Andrew Conley (NOT Troy Hilley)
Wiki Page

There is a course Wiki Page

http://www.compgenomics2009.biology.gatech.edu

Lectures, readings and all course information can be found on the page

Andrew will explain the page shortly

All protocols and results are to be carefully documented on the page

The first things to do today are:
1. Choose up teams (decide on group composition)
2. Build your personal Wiki page profiles
3. Log into the compgenomics server