Computational Genomics

BIOL 7210 A Spring 2020

Instructor: Prof. King Jordan
TA: Shashwat Deepali Nagar

Office hours by appointment

king.jordan@biology.gatech.edu
EBB 2109

shashwat@gatech.edu
EBB 2200

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 (actionable) 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 webservers that allow for automatic analysis of data
  8. Create databases that allow for searching and dissemination of genome annotations

Therefore: Computation is more critical to genomics than experimentation!

Reality-based course

- 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 from Centers for Disease Control and Prevention (CDC) foodborne illness surveillance and outbreak investigations
- *Pinpoint the source of the outbreak – the species/strains involved and their virulence profiles*
- Finishing with a publicly available webserver and predictive tool that can automate your most important analysis steps

Reality-based course

- This course is probably 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 Illumina sequence data and use the latest in analytical (computational) tools
- We will provide instruction on the fundamentals of the specific applications that you will use – e.g. genome assembly, gene prediction etc.
- But there is no way to 'spoon-feed' this kind of knowledge and experience to students ('sage on the stage' will not work here) … the technology changes every year
- 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

Course Structure

- The class will be divided into **three teams** and each team will be sub-divided into **five working groups**
- Each team will be responsible for conducting an outbreak investigation based on the analysis of genome sequence data from 50 isolates
- Outbreak investigation will entail identifying the species/strain of bacteria that cause the outbreak, pinpointing the origin and source of the outbreak, characterizing the functional profile of the virulent isolates (i.e. virulence factors and antimicrobial profile), making specific recommendations about outbreak response and treatment – much more information will be provided on this later
- Each team will be made up of five working groups: (1) **genome assembly**, (2) **gene prediction**, (3) **functional annotation**, (4) **comparative genomics**, (5) **predictive webserver**
- Each student will be a member of two working groups
- All students will be responsible for completing exercises corresponding to the tasks from the individual working groups

Course Teams & Groups

Team 1	Team 2	Team 3
1. Ahmad, Maria	1. Astore, Courtney A	1. Cheng, Shen-Yi
2. Anis, Hira	2. Hamilton, James Matthew	2. Gupta, Sonali
3. Cheon, Hyeonjeong	3. Hang, Xin	3. Kristof, Andrew Joseph
4. Gan, Shuheng	4. Hazra, Ujani	4. Kundnani, Deepali L
5. Gerhardt, Kenji Allen	5. Lacek, Kristine Anna	5. Maddala, Aparna
6. Kesar, Devishi	6. Lee, Kara Keun	6. Melkote Sujay, Ahish
7. Mora, Laura Maria	7. Lin, Shuting	7. Misra, Pallavi
8. Mulligan, Jessica R	8. Nikitina, Arina Antonovna	8. Rozanski, Allison Nicole
9. Narayanan, Priya	9. Oh, Sooyoun	9. Singu, Swetha Gowri
10. Patrick, Heather	10. Parekh, Paarth Jatin	10. Ulukaya, Gulay Bengu
11. Pfennig, Aaron Ruben	11. Sharma, Rhiya	11. Yang, Ruize
12. Vegesna, Manasa	12. Sharma, Shivam	12. Zhang, Yuhua
13. Xiao, Yiqiong	13. Temples, Danielle Alexa	13. Zhou, Jie
14. Zheng, Winnie	14. Wang, Hanchen	

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 Predictive Webserver

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 have 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 – but students will also want to join groups that provide an opportunity to learn new skills (2 groups per student)

Course Teams & Groups

	Team 1	Team 2	Team 3
Genome Assembly	**Kesar, Devishi** Mora, Laura Maria Cheon, Hyeonjeong Mulligan, Jessica R Patrick, Heather Xiao, Yiqiong	**Lacek, Kristine Anna** Sharma, Shivam Hamilton, James Matthew Lin, Shuting Sharma, Rhiya Wang, Hanchen	**Yang, Ruize** Kundnani, Deepali L Kristof, Andrew Joseph Maddala, Aparna Singu, Swetha Gowri
Gene Prediction	**Ahmad, Maria** Pfennig, Aaron Ruben Anis, Hira Mulligan, Jessica R Narayanan, Priya Zheng, Winnie	**Hang, Xin** Temples, Danielle Alexa Astore, Courtney A Hamilton, James Matthew Parekh, Paarth Jatin	**Misra, Pallavi** Yang, Ruize Cheng, Shen-Yi Melkote Sujay, Ahish Zhou, Jie
Functional Annotation	**Gerhardt, Kenji Allen** Ahmad, Maria Cheon, Hyeonjeong Gan, Shuheng Vegesna, Manasa Zheng, Winnie	**Temples, Danielle Alexa** Lee, Kara Keun Hazra, Ujani Lin, Shuting Oh, Sooyoun Sharma, Rhiya	**Rozanski, Allison Nicole** Misra, Pallavi Cheng, Shen-Yi Kristof, Andrew Joseph Ulukaya, Gulay Bengu
Comparative Genomics	**Mora, Laura Maria** Gerhardt, Kenji Allen Anis, Hira Patrick, Heather Vegesna, Manasa	**Lee, Kara Keun** Lacek, Kristine Anna Astore, Courtney A Hazra, Ujani Nikitina, Arina Antonovna	**Kundnani, Deepali L** Gupta, Sonali Singu, Swetha Gowri Ulukaya, Gulay Bengu Zhang, Yuhua Zhou, Jie
Predictive Web Server	**Pfennig, Aaron Ruben** Kesar, Devishi Gan, Shuheng Narayanan, Priya Xiao, Yiqiong	**Sharma, Shivam** Hang, Xin Nikitina, Arina Antonovna Oh, Sooyoun Parekh, Paarth Jatin Wang, Hanchen	**Gupta, Sonali** Rozanski, Allison Nicole Maddala, Aparna Melkote Sujay, Ahish Zhang, Yuhua

Course Grading

Points	Percent	Evaluation
100	10%	Class attendance & participation
200	20%	Exercise sessions (4 x 50 points each)
200	20%	Class presentations (4 x 50 points each)
400	40%	Group results (2 x 200 points each)
100	10%	Documentation (2 x 50 points each)

Grading Rubrics & Expectations – Class participation

- Class attendance and participation are **mandatory**
- Lecture and discussion participation
  - asking and answering questions
  - offering ideas and opinions
- Group presentation participation
  - asking questions during others' presentations
  - engaging the audience during your own presentations
  - connecting their presentations to previous class discussions
- Exercise session participation
  - Attendance at and participation in exercise sessions

Grading Rubrics & Expectations – Exercise sessions

- Shashwat will run exercise sessions for the first four group activities: genome assembly, gene prediction, functional annotation, and comparative genomics
- Exercise sessions will provide the fundamental practical knowledge for how to run applications corresponding to each group activity – you can think of this as a starting point or the minimum amount of knowledge required to complete each task
- He will demo the tools and then provide a series of tasks/questions that build on the demo
- Students will be required to submit the results of analyses based on the demo session – the exact format and requirements will be specified for each demo

Grading Rubrics & Expectations – Class presentations

- Each working group will conduct two class presentations of 20 minutes

- Presentation #1 – Background & Strategy
  - background and overview of area, core concepts
  - description of possible tools/algorithms that will be used – qualitative comparison
  - proposed approach and analysis workflow
  - description of task delegation – who does what

- Presentation #2 – Final Results
  - quantitative comparison of results of different tools/algorithms tested
  - justify choice of final tools and approach used
  - information on how tools were combined and results merged
  - final results and deliverable for next group

Grading Rubrics & Expectations – Group results

- Group results are the single most important metric for success (40% of your grade)

- Genome assembly
  - staring from FASTQ files – perform sequence read quality control (QC)
  - perform genome assembly using multiple tools
  - evaluate the performance of multiple tools (assembly metrics)
  - combine tools and merge assemblies as needed
  - **deliverable #1** – QC and assembly pipeline on GitHub (more from Shashwat)
  - **deliverable #2** – FASTA files with assembled contigs to gene prediction group

Grading Rubrics & Expectations – Group results

- Group results are the single most important metric for success (40% of your grade)

- Gene prediction
  - starting from assembled genomes – predict genes (features) using multiple tools
  - compare the results of multiple tools
  - validate ab initio predictions with homology
  - combine tools and merge predictions as needed
  - coherent gene naming scheme
  - provide confidence levels for gene predictions
  - **deliverable #1** – gene prediction pipeline on GitHub (more from Shashwat)
  - **deliverable #2** – gff files for functional annotation group
  - **deliverable #3** – FASTA files for gene nucleotide and protein sequences

Grading Rubrics & Expectations – Group results

- Group results are the single most important metric for success (40% of your grade)

- Functional annotation
  - starting from gene (feature) predictions – predict various aspects of gene (protein) function
  - perform both *ab initio* and homology-based prediction as appropriate
  - aspects of function to predict – biochemical activity, molecular function, (sub)cellular localization, domain and motif composition, higher level features such as protein families or operons, enzymatic activity, virulence factors etc (note that this list is not exhaustive)
  - **deliverable #1** – functional annotation pipeline on GitHub (more from Shashwat)
  - **deliverable #2** – gff files for comparative genomics group
  - **deliverable #3** – annotated FASTA files for gene nucleotide and protein sequences

Grading Rubrics & Expectations – Group results

- Group results are the single most important metric for success (40% of your grade)

- Comparative genomics
  - combining data from all three previous groups – compare genome sequences in order to perform outbreak analysis (much more on this later)
  - what is the identity of the species/strains that cause the outbreak?
  - how are the isolates related to each other? how do they differ?
  - which isolates correspond to outbreak versus sporadic strains?
  - what are the virulence and antibiotic resistance profiles of the outbreak isolates?
  - what is the recommended outbreak response and treatment?
  - **deliverable #1** – comparative genomics pipeline on GitHub (more from Shashwat)
  - **deliverable #2** – specific public health recommendations to CDC

Grading Rubrics & Expectations – Group results

- Group results are the single most important metric for success (40% of your grade)

- Predictive webserver
  - combining data from all four previous groups – create a predictive webserver that allows for automated analyses of your species of interest
  - possible functional utility of the predictive webserver include (but may not be limited to)
    - distance from the closest isolate in the database
    - assigning whether the uploaded isolate looks like outbreak or sporadic strain(s)
    - visualization of the distance between your isolate and database isolate as a phylogenetic tree and/or heatmap
    - virulence factor and antimicrobial resistance profiling of your isolates
  - **deliverable** – fully functional online predictive webserver

Grading Rubrics & Expectations – Documentation

- Working groups are expected to thoroughly document their results on both the Wiki and GitHub pages

- The gold standard for documentation is reproducibility – there should be sufficient detail such that an independent group working in a different part of the world could reproduce your results

- One of the best resources for documentation is previous classes Wiki pages
  http://www.compgenomics2017.biology.gatech.edu/
  http://www.compgenomics2018.biosci.gatech.edu/
  http://www.compgenomics2019.biosci.gatech.edu/

- Shashwat will provide substantial detail on the documentation requirements for both the Wiki and GitHub pages in subsequent classes

Additional 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

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

4. **Do analysis, produce & document results, present results, and integrate into webserver**

5. Work closely as needed to help other groups succeed in their phases and to help other groups acquire knowledge and experience in your domain

6. Innovation is key – you must show innovations & improvements over previous years classes

No free lunch

- 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
- Group questionnaires during and after semester to evaluate individual effort
- Collegiality and respect are essential and mandatory
- 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 TA (Shashwat) as to how best resolve issue
    3. If steps 1 and then 2 fail, consult with me and I will address the issue

Team contracts

- Team contracts will help you set expectations about frequency and quality of work, along with ground rules for transgressions.

- Before you begin working together, discuss
  - Expectations
    - Frequency of meetings, frequency of communication, roles, division of labor
  - Conflict resolution policies
  - Consequences
    - How would non-performance be addressed?
    - How would the grade be divided (even | penalization based on absence or non-performance | something else)

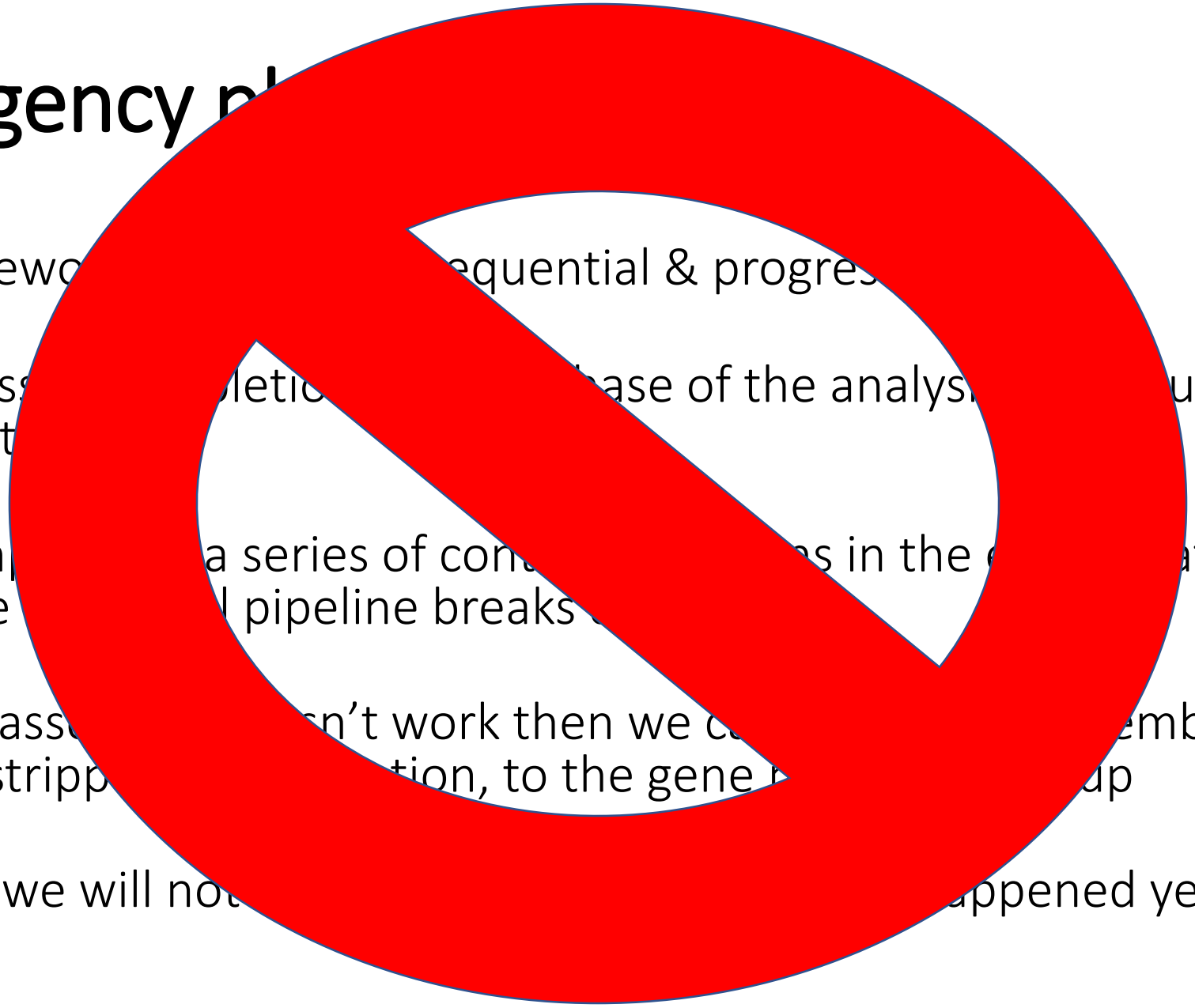
- For specific tasks, discuss
  - Deadlines
  - Person (/people) responsible for said task
  - Expected output files and location

Things to avoid in order to ensure success

1. Showing up Late to Class
2. Missing Class
3. Not Being Engaged in Class
4. Not Contributing to Group Efforts
5. Blind/Mis-informed Use of Tools
6. Copying From Previous Years Classes

Contingency pl

- The coursewo equential & progres
- The succes letic ase of the analys upon the previous st
- We will im a series of cont s in the t any given step in the l pipeline breaks
- E.g. if the ass n't work then we c embled genome, stripp tion, to the gene up
- Hopefully we will not appened yet)



Good luck!

Additional questions?