Pedagogy of Data Science within other Disciplines

Eric Miller, Assistant Professor
Department of Biology

HAVERFORD
COLLEGE

Pedagogy of Data Science ~~within other Disciplines~~ –in Biology

Eric Miller, Assistant Professor
Department of Biology

HAVERFORD
COLLEGE

Data Science + Biology

- How do we introduce/demonstrate the intersection of data science with discipline knowledge to all students?

- How do we prepare a subset of students for graduate school in topics around this intersection?

- How can I prepare students with the framework / skills to help with my lab's bioinformatic research?

Data Science-y Bio Courses

- Advanced Lab in Bioinformatics

  aka Bioinformatics Superlab (2021, 2022)

- Programing in Biology: Bio104 (Sp2024)

  + other 104 courses: astrophysics, chemistry, linguistics, physics

- Biostatistics (Fa2023 + Future curriculum)

Superlabs @ Haverford

- Discovery-based learning; novel research for 7-14 weeks
- Give students the science background + general question
- Student research in 7-20 parallel groups

Superlabs @ Haverford

- Discovery-based learning; novel research for 7-14 weeks
- Give students the science background + general question
- Student research in 7-20 parallel groups
- Develop hypothesis, design experiment, complete experiment, analysis, communicate findings
- Students do this 2-4 times in their 3rd year
- Resulting data is publishable (initial studies, prelim. grant data)

Superlabs @ Haverford

- 50+ years at Haverford Biology; offshoots of:
  - Chemistry Superlab
  - Biochemistry Superlab
  - Neurobiology Superlab
  - Bioinformatics Superlab
- High level of departmental support

Bioinformatics Superlab

- Goals
  - Working with 'big data' — but without coding
  - Using Galaxy (genetic analyses) and RStudio (graphing and stats)

Bioinformatics Superlab

Bowtie2 - map reads against reference genome (Galaxy Version 2.4.2+galaxy0)

Favorite | Options

Is this single or paired library

Single-end

FASTA/Q file

51: ClustalW on data 49: fasta

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

Yes | No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes | No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome

No options available

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

History

search datasets

BLASTing

31 shown, 27 deleted

429.61 MB

51: ClustalW on data 49: fasta

50: MAFFT on data 49

49: luxSUnique_Outgroups.faa

48: tblastn luxS_typical vs '5SpeciesOutgroup'

47: tblastn luxS_typical vs '5SpeciesOutgroup'

46: tblastn luxS_typical vs '5SpeciesOutgroup'

43: luxS_typical

42: Join neighbors on data 41: Calculated distances

21 lines

format: **nhx**, database: **?**

0.00%
0.00%
0.01%

Bioinformatics Superlab

- Goals
  - Working with 'big data' — but without coding
  - Using Galaxy (genetic analyses) and RStudio (graphing and stats)
- All parts of 'sciencing':

  Observation / prediction; generate testable hypothesis; design + execute experiment; results + interpretation; communication and 'next steps'

Bioinformatics Superlab

Dates	Monday (1.5 hours)	Tuesday (3 hours)	Homework for Thursday	Thursday (3 hours)	Homework for Monday
January 18 - 21	No Class	Course dynamics and syllabus. Talk about lab notebooks. Present research question and Streptococcus. Discuss quorum sensing and HGT. Bioinformatics basics: Fears / anxiety about this class. What is data? Text vs. binary. Sublime text to look at data. Excel to work with data.	RStudio reading (Chapter 2 -The Very Basics, 2.4,2.5 optional based on interest only) https://rstudio-education.github.io/hopr/basics.html	Working with RStudio: basics + getting data into RStudio (Tutorial on Moodle)	Read parts of a QS paper for discussion on Monday. Submit a question about paper by 10pm on Sunday.
January 24 - 28	Discussion of QS paper	Introduction of scientific writing assignment ~~Working with RStudio: Tutorial on Graphing~~ Lecture: Statistics in R, Work on independent Graphs	Create a novel graph using S. suis metadata. Be prepared to give a 1 minute presentation of your graph to the class, or no graph but an explanation of what was challenging in your attempt to make a graph. Submit graph (or not graph and explanation of what was challenging) on Moodle by 10am Thursday.	Discussion of graphs — Break into 2 groups of 7 to do this? ~~Lecture: Statistics in R~~ Start Statistics Tutorial	Submit a question on statistics in R by 10am on Monday Work on outline of scientific writing assignment
January 31 - Feb. 4	Questions on statistics in R Lecture on Next Generation Sequencing and mapping to reference genomes and variant calling.	Finish statistics tutorial	Submit question that you have on mapping reads (from the tutorial or more of a conceptual question) to ask on Thursday.	Download data, then discuss Next Generation Sequencing questions. Working with Galaxy: basics + tutorial on short read mapping RStudio part of mapping tutorial.	~~Finish creating a graph of your read sets (from the RStudio part of the tutorial), and be prepared to share the graph and your interpretation on Monday.~~ ~~Submit graph by 10am on Monday.~~
Februrary 7 - 11	~~Review of graphs~~	RStudio part of mapping tutorial (if	Check in with Galaxy; Graph your results	Review/discuss graphs ~~In-class exercise of which assemblies to keep for our dataset.~~ Presentation of quoroum sensing diversity challenge — students form 'Supergroups' to look at diversity across 4 QS systems. In-class reading of group papers.	Outline of scientific writing assignment due
				Student group time to work on project	As a group, ~~prepare a graph on your data~~ Summary (table) of diversity (signal) for Tuesday.

- Highly structured classes
  - Lectures
  - Primary literature
  - Skill-based tutorials
  - Informal peer reviews
  - Emphasis on students presentation of their work, early and often
- Balanced with flexibility for when research goes sideways

Bioinformatics Superlab

- Datasets:
    - Speciation and hybridization of 100's of *Saccharomyces* yeast genomes
    - Searching 22,000 *S. pneumoniae* bacterial genomes for variation in signaling peptides; next steps with this variation

Bioinformatics Superlab

- What's good:
  - Increased subject knowledge, broad-scale bioinformatics knowledge
  - Greatly improved hypothesis testing, scientific communication
- Unexpected difficulties
  - Technical issues (dataset, BLAST algorithm not working)
  - Need to spend time on basic computing
  - Not enough time for 'proper' discussion of statistics
  - Harder to show students why an approach is suboptimal

Programming in ____

Programming in ____

- 2 pathways for basic programming CS105 + CS106

  ___104 + CS107

- Basics of: types, data structures, OOP, data management + access

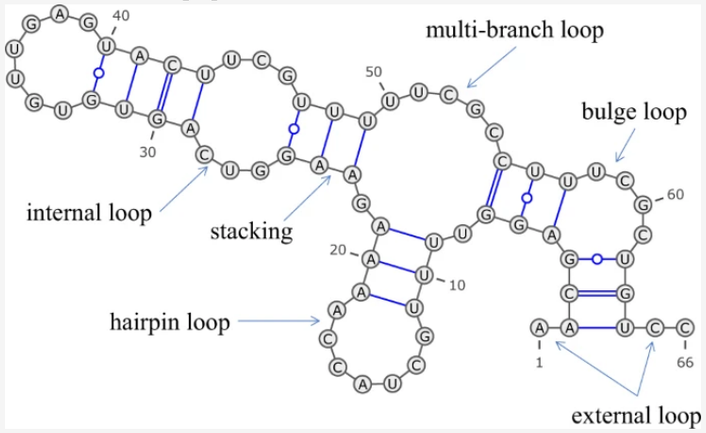
- Co-taught within department and CS faculty (for labs)

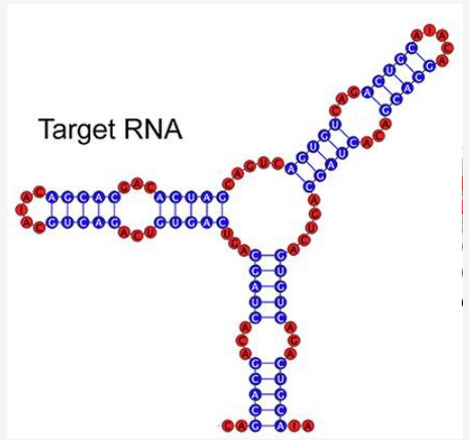
- Also as an intro to the topic

Programming in Biology

Genotype → Phenotype → Fitness

AACGAGGUUUU...UGUCC



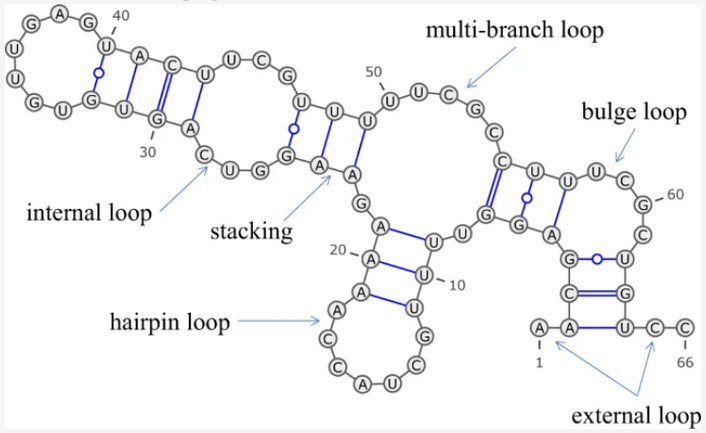


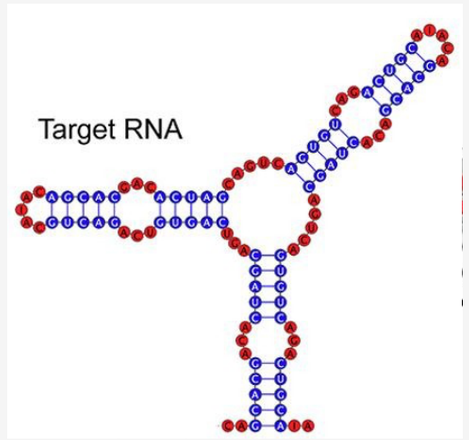
- Genotypes
- Phenotypes
- Differential fitnesses

Programming in Biology

Genotype → Phenotype → Fitness

AACGAGGUUUU...UGUCC





- Genotypes — Phenotypes — Differential fitnesses
- Genotypes mutating, — creating different phenotypes — and differential fitnesses

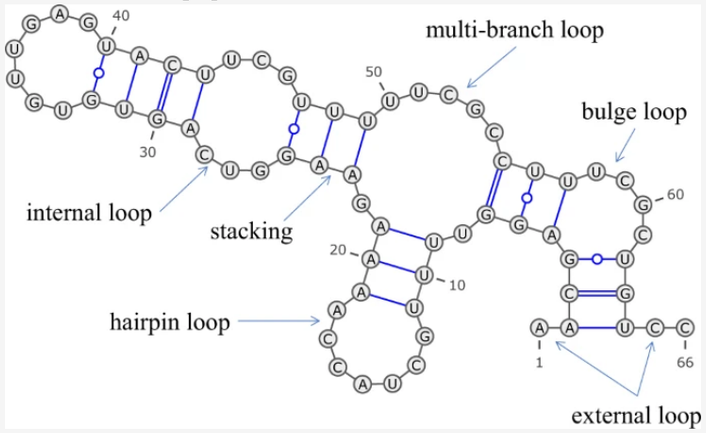
- Evolution of a population of RNA molecules towards a target shape

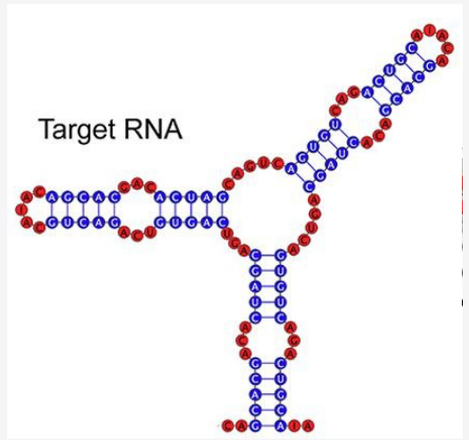
Influence of:

Programming in Biology

Genotype → Phenotype → Fitness

AACGAGGUUUU...UGUCC





- Genotypes — Phenotypes — Differential fitnesses
- Genotypes mutating, creating different phenotypes and differential fitnesses

- Evolution of a population of RNA molecules towards a target shape

Influence of:

- Population size
- Mutation rate
- Selection environment (simulated temperature)
- Ect.

Programming in Biology

Goals:

- Programing + documentation
- 4 forces of evolution + impact on asexual populations

Programming in Biology

Goals:

- Programing + documentation
- 4 forces of evolution + impact on asexual populations
- How computational biology can be used to test biological hypotheses
- Students start to think about the process, and their role, in science
- Help relieve CS faculty burden

Biostatistics

Biostatistics

Biostatistics

- 7-week elective (Fa2023)
- Needs to cover data-handling; RStudio; graphing; statistics
- Future: Semester-long course, required for major

Challenges

- Math anxiety / computing anxiety
- Starting point for computing skills
- Course scheduling
- Teaching capacity

Challenges

- Math anxiety / computing anxiety
- Starting point for computing skills
- Course scheduling

Fall Year1	Spring Year1	Fall Year2	
Language	Language	*Available*	Declare major
Writing	Writing*	*Available*	
Chemistry	Chemisty	Chemistry	
Available	Biology	Biology	

- Teaching capacity
  - Math / Stats department, CS department
  - Biology department: Who can teach biostats? Programming?

Eric Miller
Microbial Ecology and Evolution Laboratory
emiller3@haverford.edu

HAVERFORD
COLLEGE