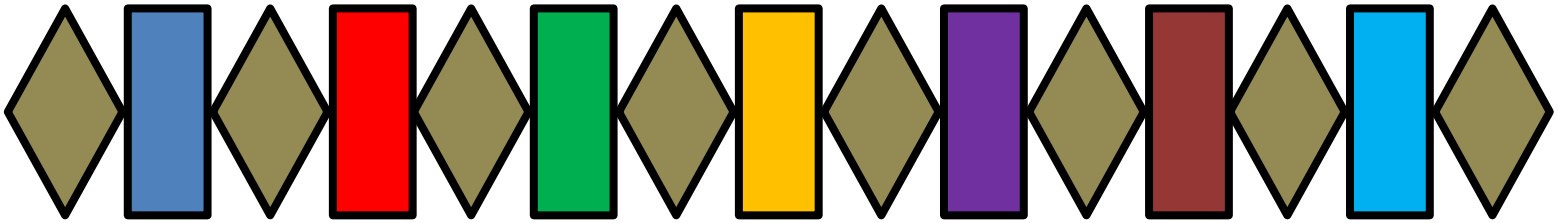
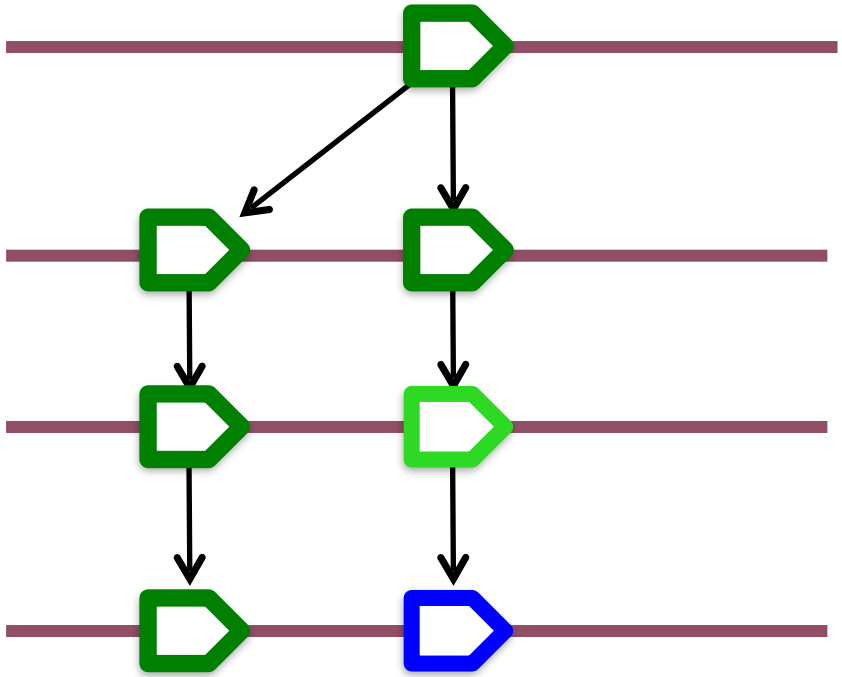


BIOL/CS 123B
Midterm 2 Review



Spring 2021

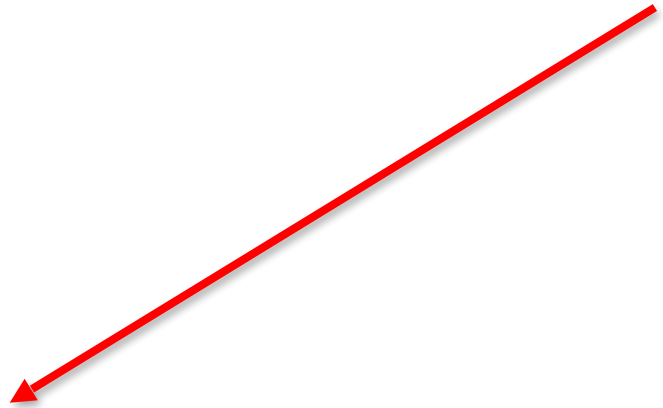
Philip Heller

NASA GeneLab Boot Camp and Internships

- 6 internships (lightly paid) … must attend Boot Camp
- 24 extra spaces in Boot Camp, no internship
- Boot Camp: June 7 - June 11
- Internship: June 14 - Aug 18
- Apply at

  https://nams.usra.edu/programs/education/student-r-d-opportunities/

Student R&D Opportunities



Apply Now

Biosciences

Evaluation and Development of NASA GeneLab Data Processing Pipelines

Overview

Up to 6 interns will be selected for this internship and will be split into 3 groups. Each group will work on one of the following 3 projects. In your cover letter, please rank the 3 projects listed below from most to least interesting to you, why you are interested in this internship, what you hope to gain from the internship, and what makes you qualified for this internship. Please provide links to a cover letter, your CV and an unofficial transcript. Provide a link to your unofficial transcript in the last question (i.e. additional materials).

Projects

After you click “Apply Now”…

- Q3 (Nationality) – US citizenship is not required
- Q4 (Internship period) - June 14 - Aug 18
- Q5 (Technical area) - Biosciences
- Q6 (Capstone) - No
- Q7 (How did you hear about) - “SJSU Bioinformatics program”
- Q8 (Cover letter)
  - 100% written by you
  - Include Bioinformatics Minor status (declared, plan to declare, won’t declare)
  - Include SJSU i.d.
  - State if you are applying for boot camp + internship, or just boot camp
- Q9 (C.V.) - CV = “Curriculum Vitae” = resume
- Q10 (Additional materials) – Unofficial transcript from MySJSU

Recommendation email

- From a faculty member, but not your 123A or 123B instructor

Instructions to your recommender (applying for boot camp only)

Please write a short paragraph (50-100 words) explaining why you recommend this student for a 1-week boot camp to learn about NASA's GeneLab RNA-seq analysis pipeline. State the strength of your recommendation on a scale of 1 (weakest) to 10 (strongest). Email your recommendation to philip.heller@sjsu.edu by the end of day on Friday April 16.

Instructions to your recommender (applying for boot camp and internship)

Please write a short paragraph (100-200 words) explaining why you recommend this student for a summer internship at NASA's GeneLab, analyzing gene expression data from Space Shuttle missions and the International Space Station. State the strength of your recommendation on a scale of 1 (weakest) to 10 (strongest). Email your recommendation to philip.heller@sjsu.edu by the end of day on Friday April 16.

Application Deadline

- End of day Friday April 16
- Be time zone aware
- Late or incomplete applications will not be accepted for any reason

Biol/CS 123B Spring 2021
Midterm 2

Rules:

- Edit this doc. Use blue text for your answers.
- You may refer to your notes, homework, labs, and slides.
- You may not use the web except as directed by Question 8.
- You may not communicate with anyone.
- For #1 - #5, your answers must be entirely in your own words. Using someone else's words is plagiarism.
- For #6, #7, and #8 you must show all your work.
- Don't add extraneous information to your answers. Just answer the questions. Points will be deducted for extraneous information.
- You may only submit once.
- Edit this doc. Upload to "Midterm2" on Canvas by 10 minutes after the end of class.
- You must attend the zoom session with your camera on, until you submit. If you have a question, unmute and ask to go to a breakout room.

Projects are due soon

- Sunday April 18, just before noon … report and presentation, 2 separate uploads.
- Please put your name, i.d., and major or field on your report and on the 1st slide of your presentation.
- No last-minute crises! If you need help, come to office hours. Weekend emails about projects won't get a reply until after the weekend.

Review of Big-Oh

- ***Not*** the official CS146 definition
- A way to compare execution time of algorithms
  - Not particular programs which implement algorithms
  - Independent of implementation, independent of computer
- "An algorithm is O(n^2)" roughly means that execution time is proportional to n^2
  - n = input size
  - time α n^2 ➔ time = k * n^2
  - k varies across implementations and computers

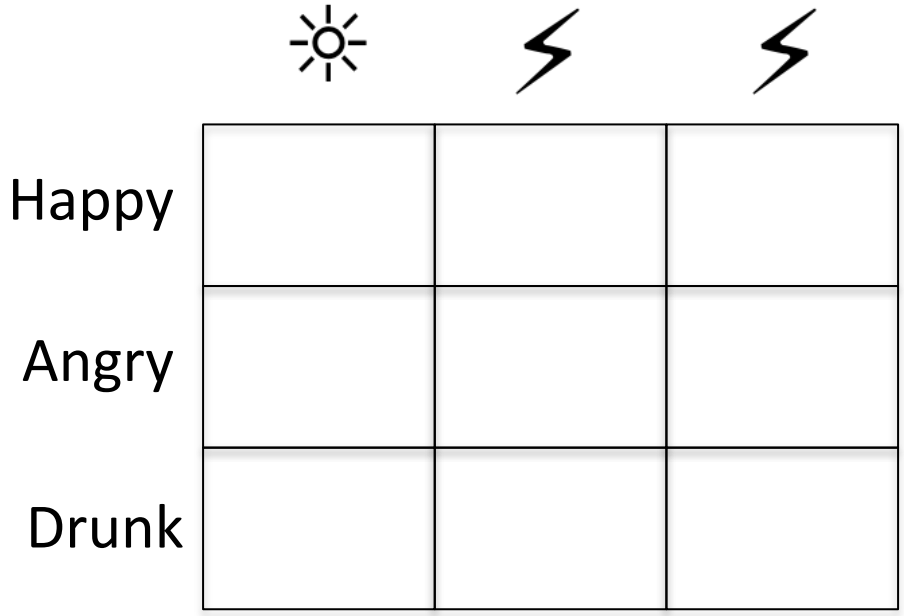
Using Big-O

- Your data might be too big to analyze in reasonable time.
- Do an experiment:
  - Use a very small data set (n << actual n).
  - Run the algorithm and measure t (execution time, hopefully reasonable).
  - Now you know t_{little} and n_{little} in the Big-O formula. Solve for k.
- Compute t for the full data set:
  - You know k and n_{big} in the Big-O formula. Solve for t_{big}.

Viterbi Algorithm

Example: Find the Viterbi (most likely) path through the Thor HMM that generates Sun/Thunder/Thunder.

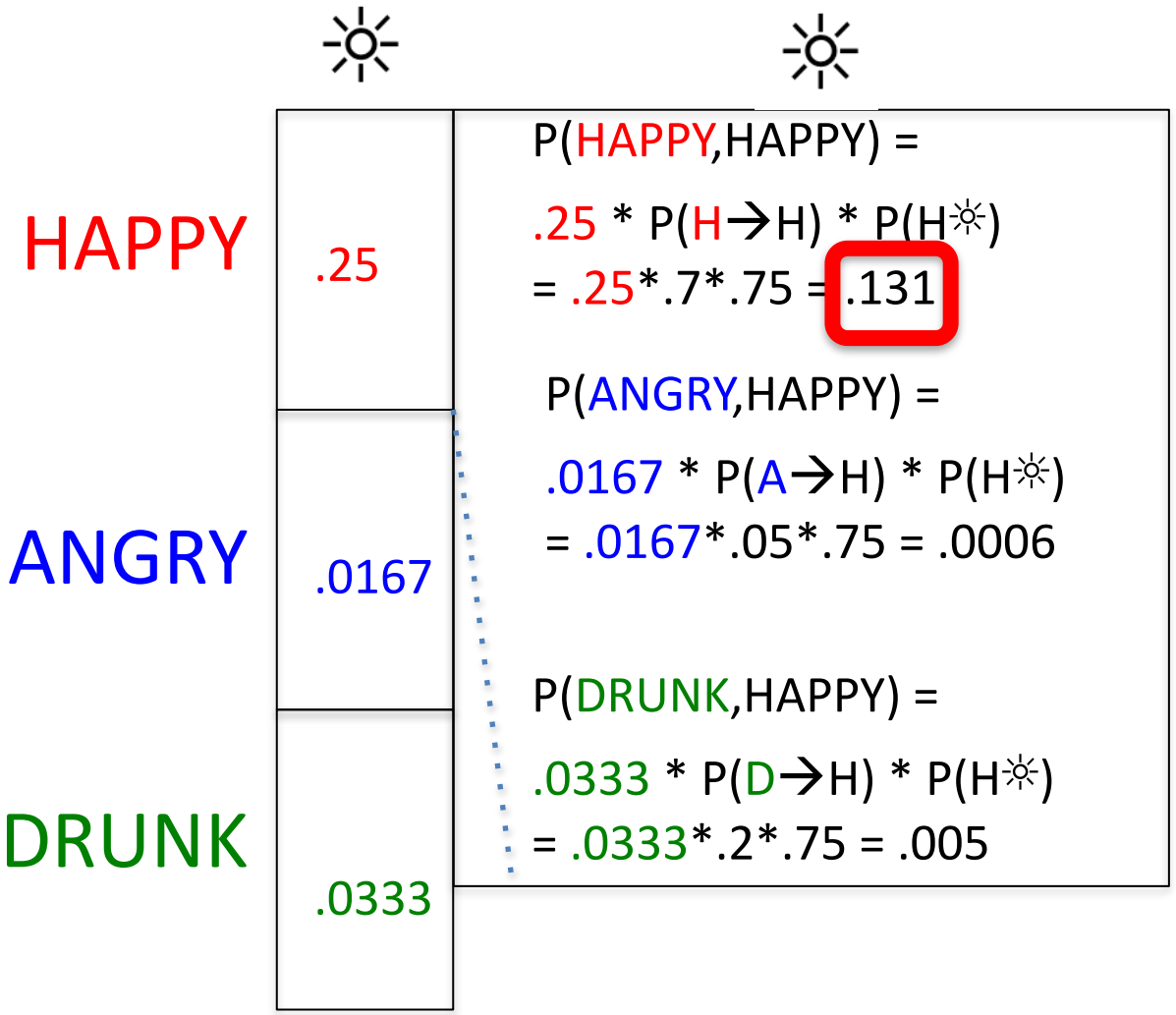
Step 1: Draw a grid. 1 row for each state of the HMM, 1 col for each member of the observation.



Viterbi Algorithm, first column:

	☼	⚡	⚡
Happy	P(start = State for row) * P(Emit weather for col from state for row)		
Angry	P(start = State for row) * P(Emit weather for col from state for row)		
Drunk	P(start = State for row) * P(Emit weather for col from state for row)		

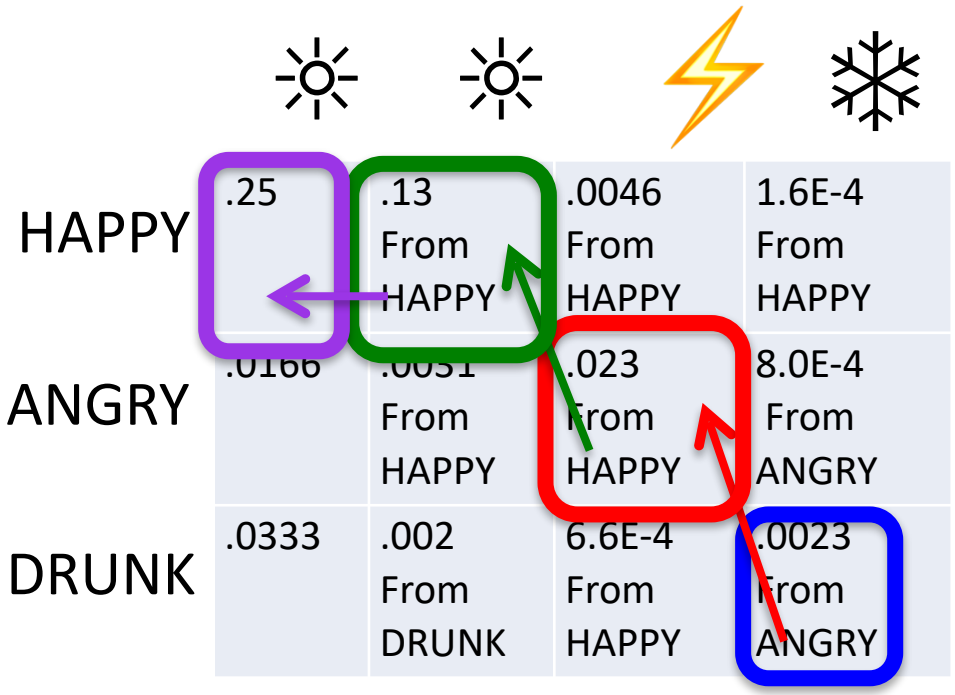
Step 3: Fill in 2nd column, HAPPY cell



Max of the 3 probs is for path = HAPPY,HAPPY

→ This is the most probable path that emits ☼☼ and ends at HAPPY. Its probability is .131

Retain .131 and remember that state from prev col was HAPPY



HAPPY	.25	.13 From HAPPY	.0046 From HAPPY	1.6E-4 From HAPPY
ANGRY	.0166	.0031 From HAPPY	.023 From HAPPY	8.0E-4 From ANGRY
DRUNK	.0333	.002 From DRUNK	6.6E-4 From HAPPY	.0023 From ANGRY

The rest of the Viterbi path is found by tracing back along the "From" states

Viterbi path = HAPPY HAPPY ANGRY DRUNK

5x10^-324: A special number

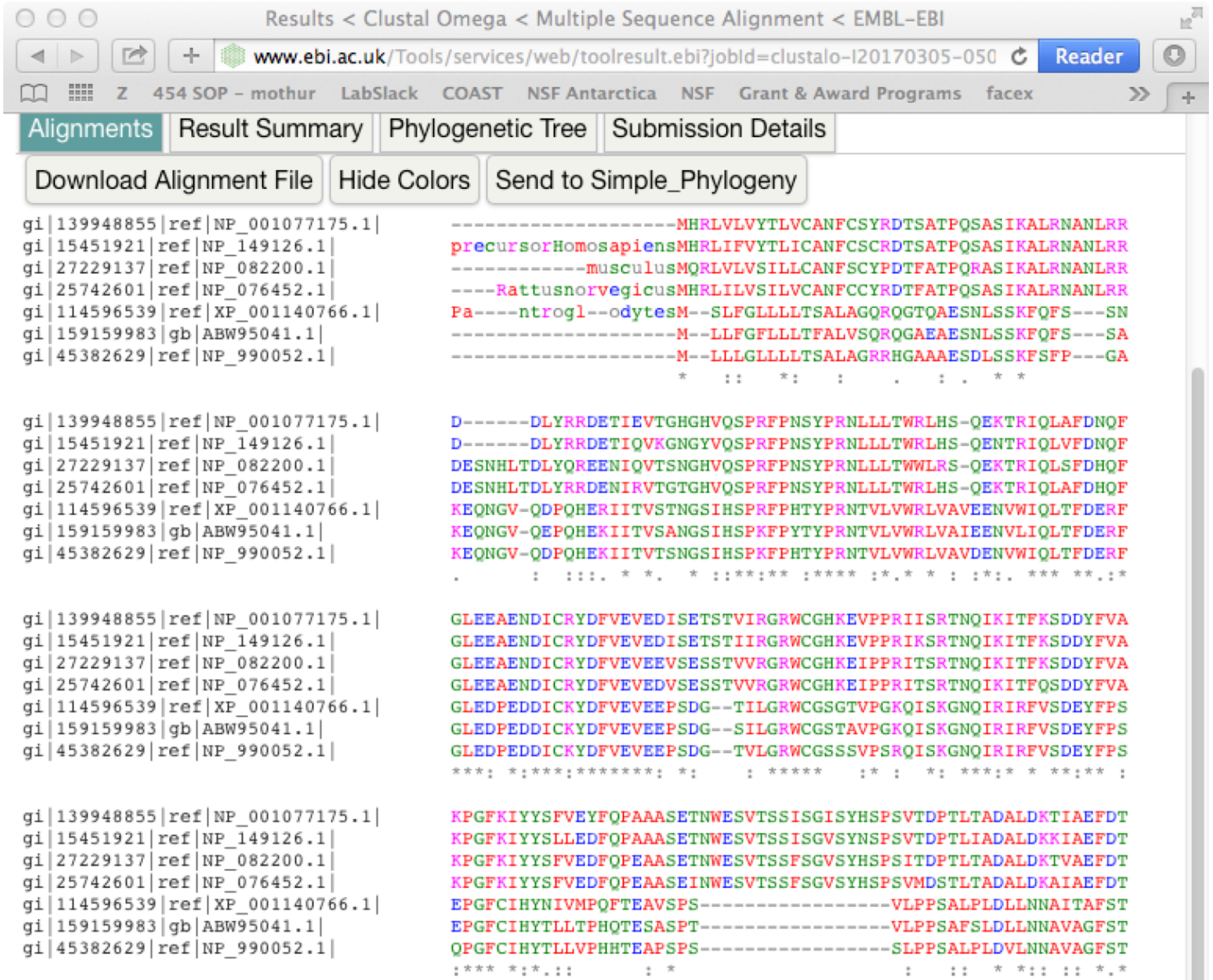
- The smallest fraction that most computers can represent
- If you multiply this by *any fraction*, the result will be rounded down to zero
- Viterbi or Forward score(any long enough sequence | any realistic HMM) will be wrongly reported as zero
- There's an easy fix for Viterbi
- ***No fix is possible for FA (or BA)***

What we want, what we settle for

- What we want
  - Given a sequence and some HMMs, compare probabilities that the HMMs emitted the sequence
  - Requires Forward Algorithm
- What we settle for (when sequence is too long)
  - Given a sequence and some HMMs, compare Viterbi probabilities for the sequence from each HMM
  - Use Viterbi probability (probability of best path) as a "proxy" for sum of probabilities of all paths

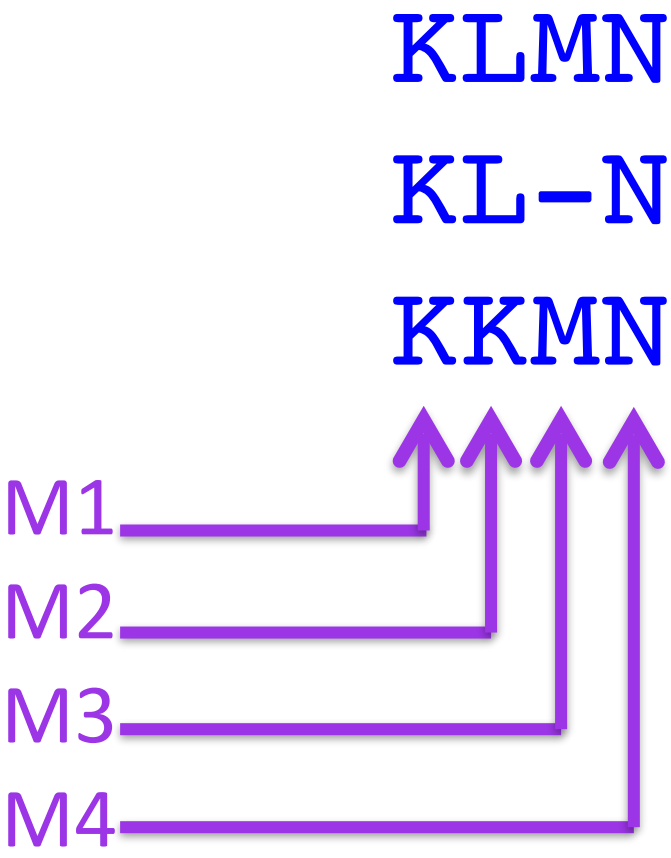
Step 1: Collect a training set (search GenBank)

Step 2: Align (e.g. ClustalΩ)

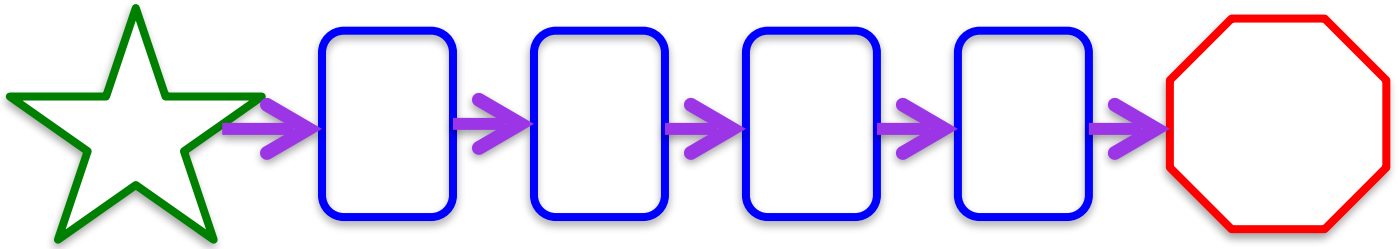


Step 3: Look for conserved domains

Each alignment column becomes a state



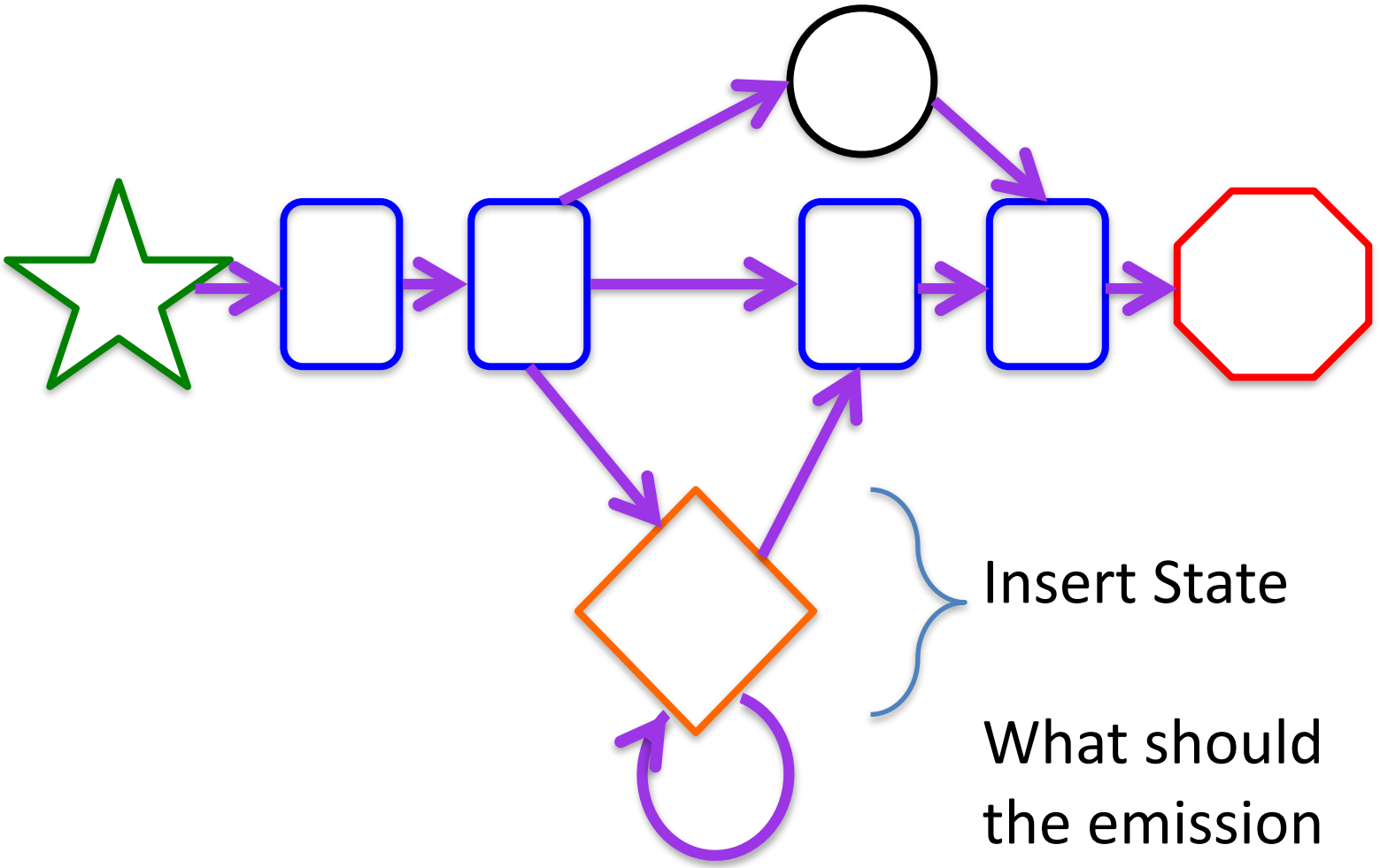
Too Simple: 3 problems



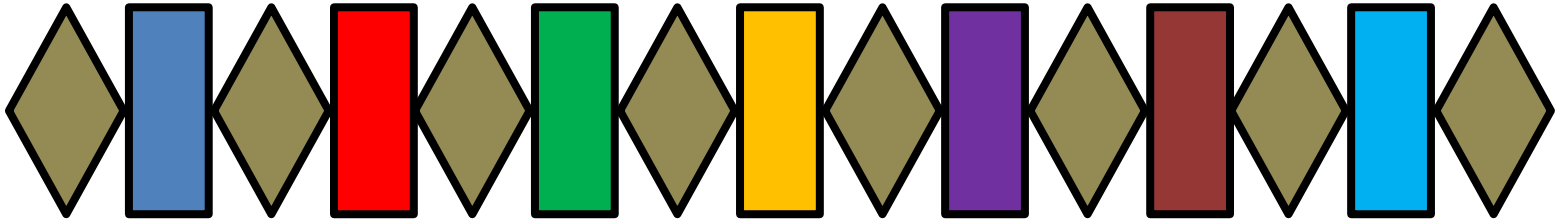
- P(Any sequence containing aas not represented in the positive training set) = 0
- Indel isn't really an emission in the same sense as the amino acids
- This HMM can only handle sequences of length=4
  - P(Any sequence of any other length) = 0

Problem: HMM can only handle sequences of same length as alignment of positive training set

- Delete states take care of shorter sequences
- Another special state takes care of longer sequences:
  - Insert States
  - Usually drawn as diamonds between/below the match states


Insert State
What should the emission probabilities be?

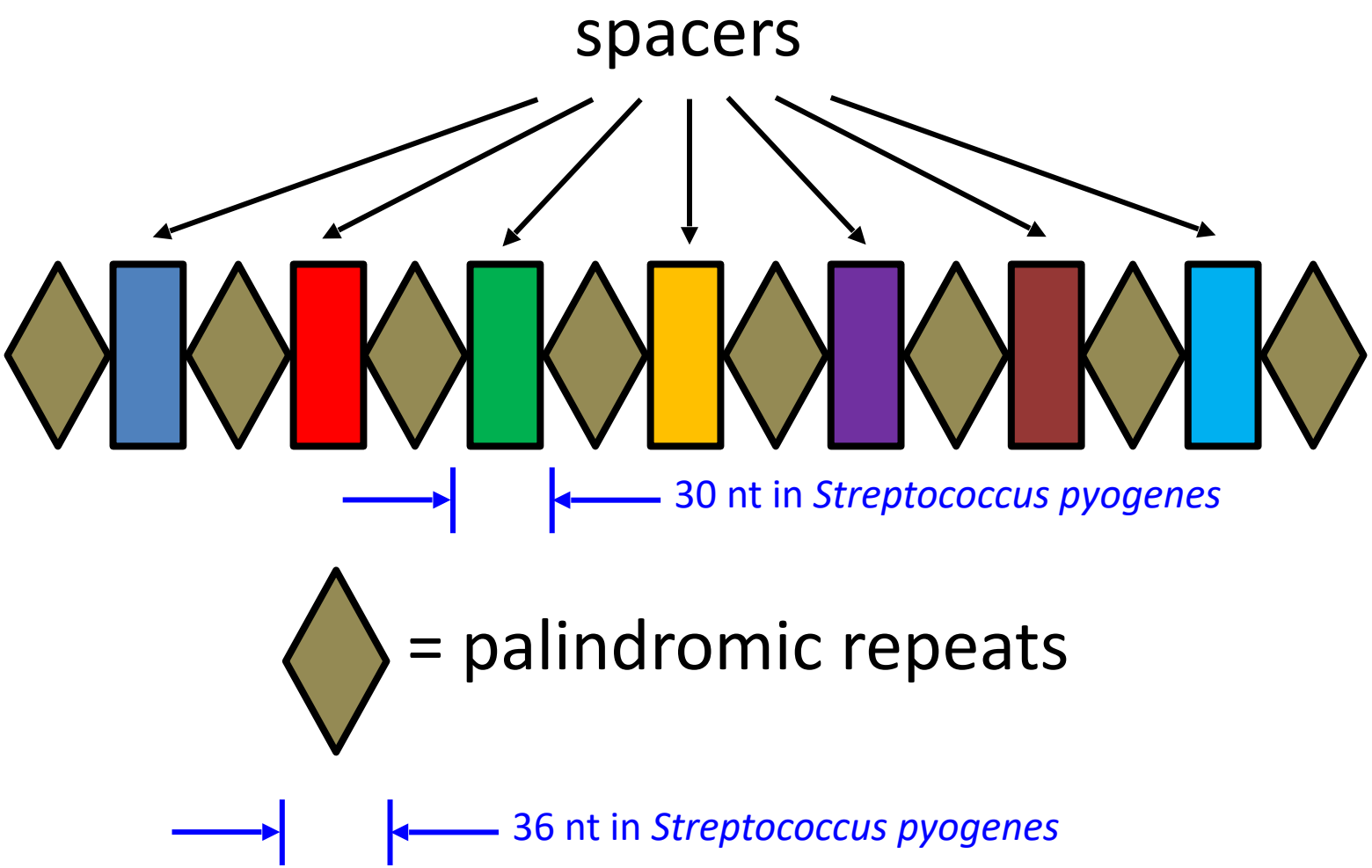
CRISPR: An ancient immune system drives new biotech



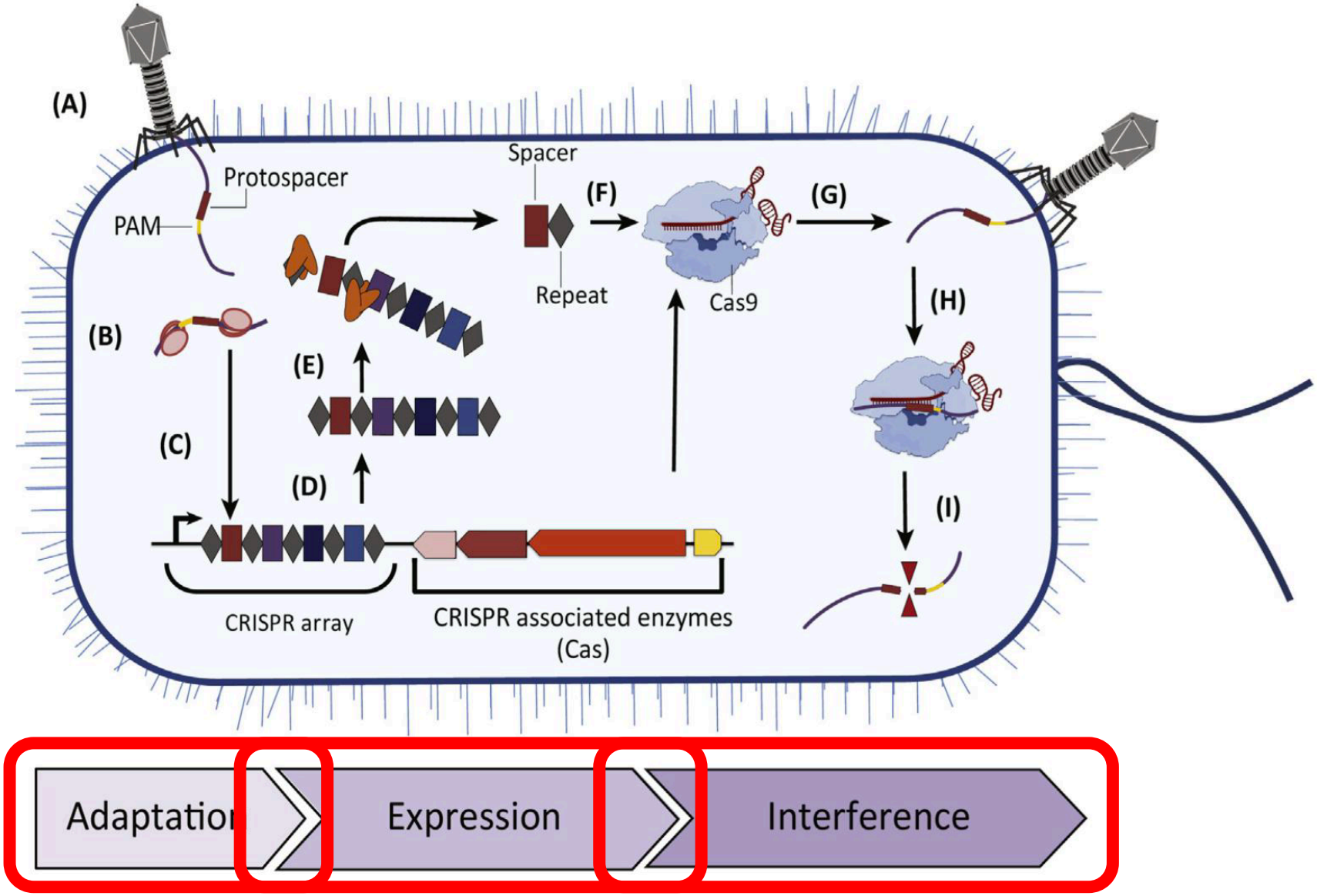
Much material adapted from

"Advanced Bioinformatics for Biotechnology"

Structure of a prokaryotic CRISPR



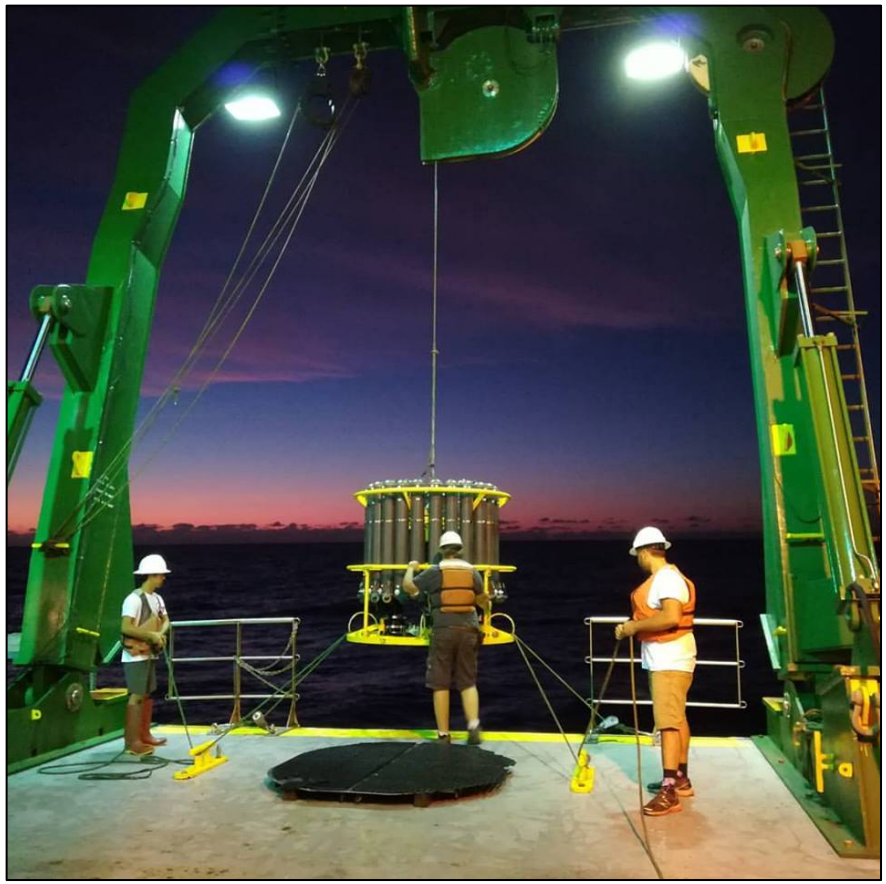
The CRISPR-Cas System: 3 Stages



The CRISPR-Cas System: 3 Stages

Let’s break down these steps…

BIO/CS 123B
Part 6: Metagenomics and UCYN-A



Spring 2021

Phil Heller

Questions that a metagenomic study can answer

- Taxonomic identification: What species are present in the community, and in what proportions?
- Functional identification: What is the *genetic potential* of the community?
  - What genes are present, regardless of what species own them
  - E.g. presence of *nifH* means the community might fix nitrogen … probably does, but not definitively proved

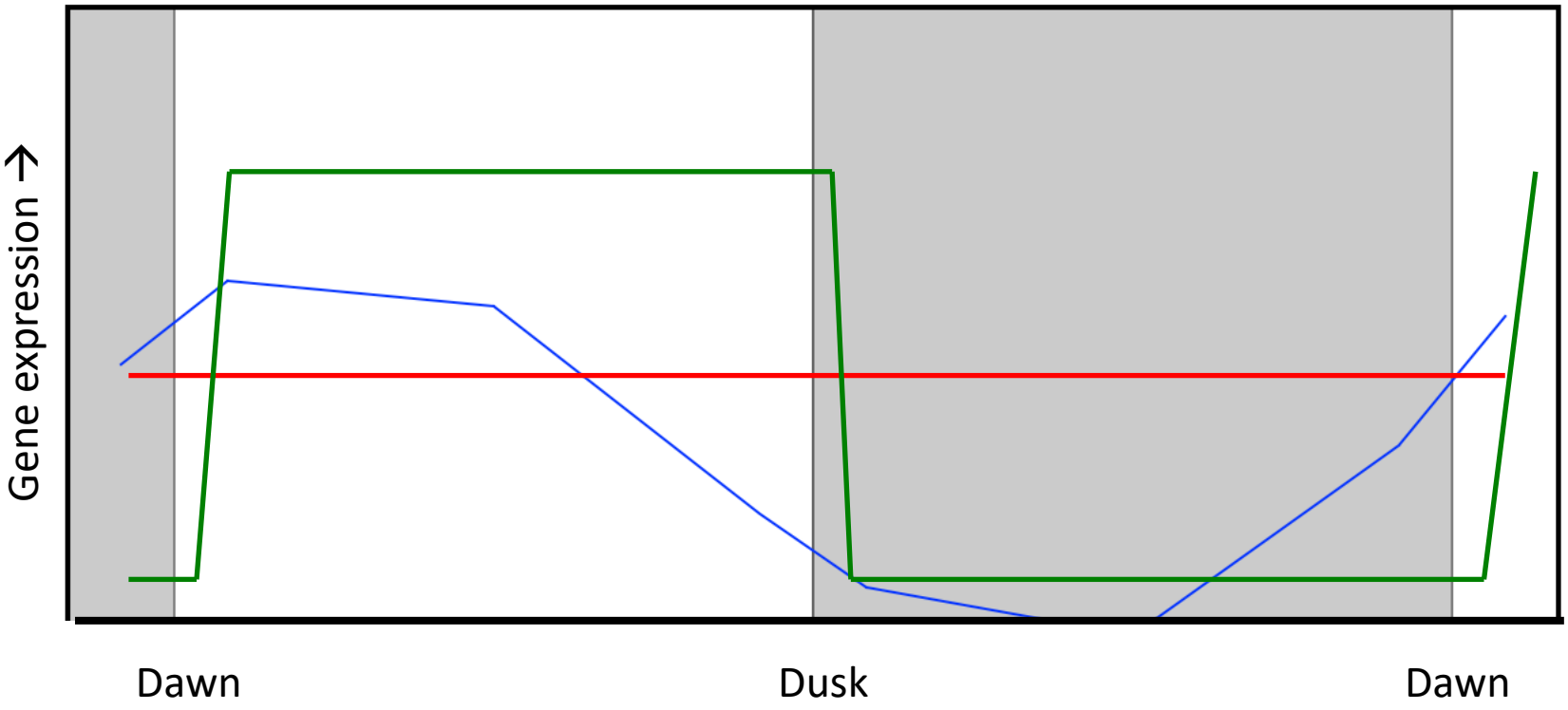
Metatranscriptomics

- Better than genetic potential, for a price
- What genes were actually being expressed at the moment you sampled?
- Expression volume can fluctuate over a 24-hour cycle: "diel" expression → have to sample n times over 24 hours
  - Example: photosynthesis genes: half-life < 12 hours
- Technology:
  - RNA-Seq
  - cDNA ("*complementary* DNA")

The kinds of thing we can learn from metatranscriptomics

- Cyanobacteria = the phylum of photosynthetic bacteria.
- Ecologically very important to marine ecologists.
- Photosystem II genes (which code for light-harvesting proteins) are costly and don't last very long.
- What strategy do cyanobacteria use to optimize resource use?

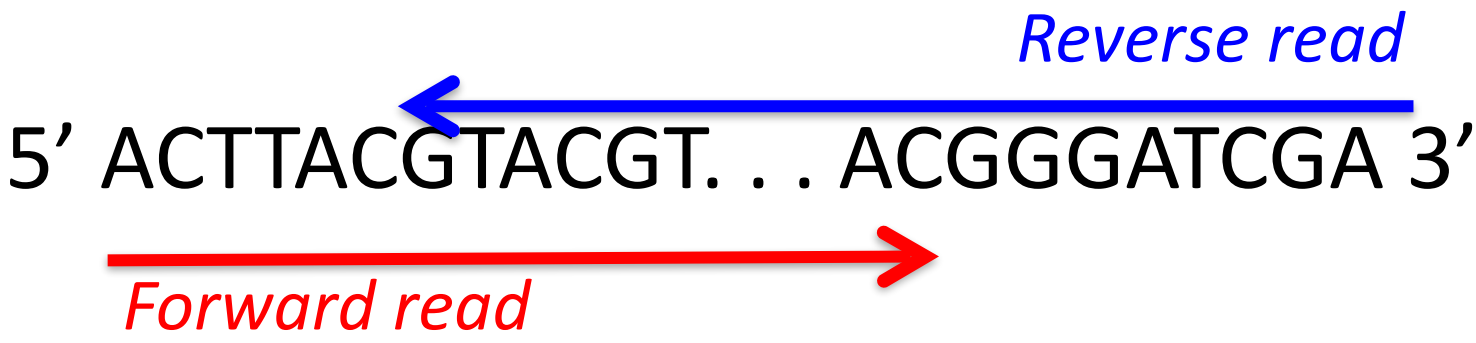
Possible strategies for photosystem gene expression timing



- Worst: constant expression
- Better: trigger by light level
- Best: just-in-time manufacturing

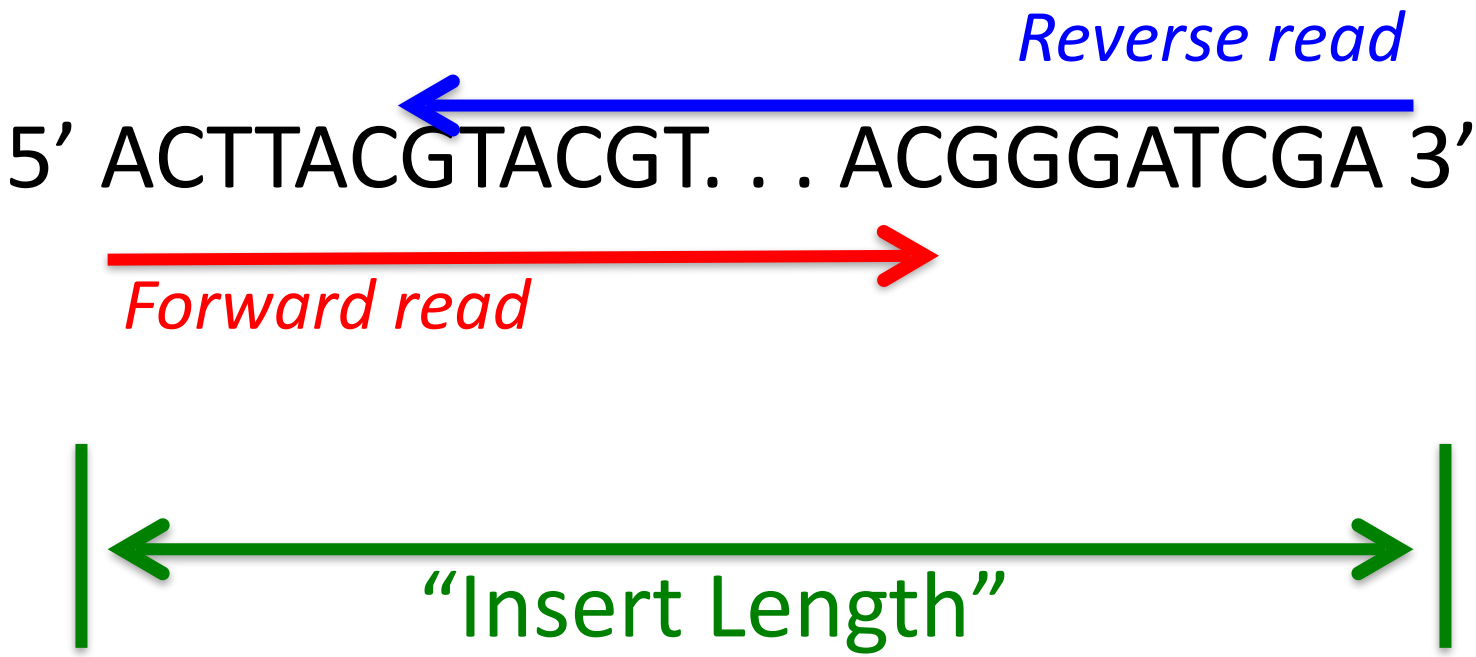
Step 2: Sequence

- Result is fastq file or 2 files
- Paired-end sequencing
  - Next-Gen sequencing (pyrosequencing) can only reliably read 200-800 bases from 5’ of a fragment
  - Best quality is near 5’ end of fragment, gets progressively worse in 3’ direction
  - Therefore quality near 3’ can be pretty bad
  - So sequence from both ends



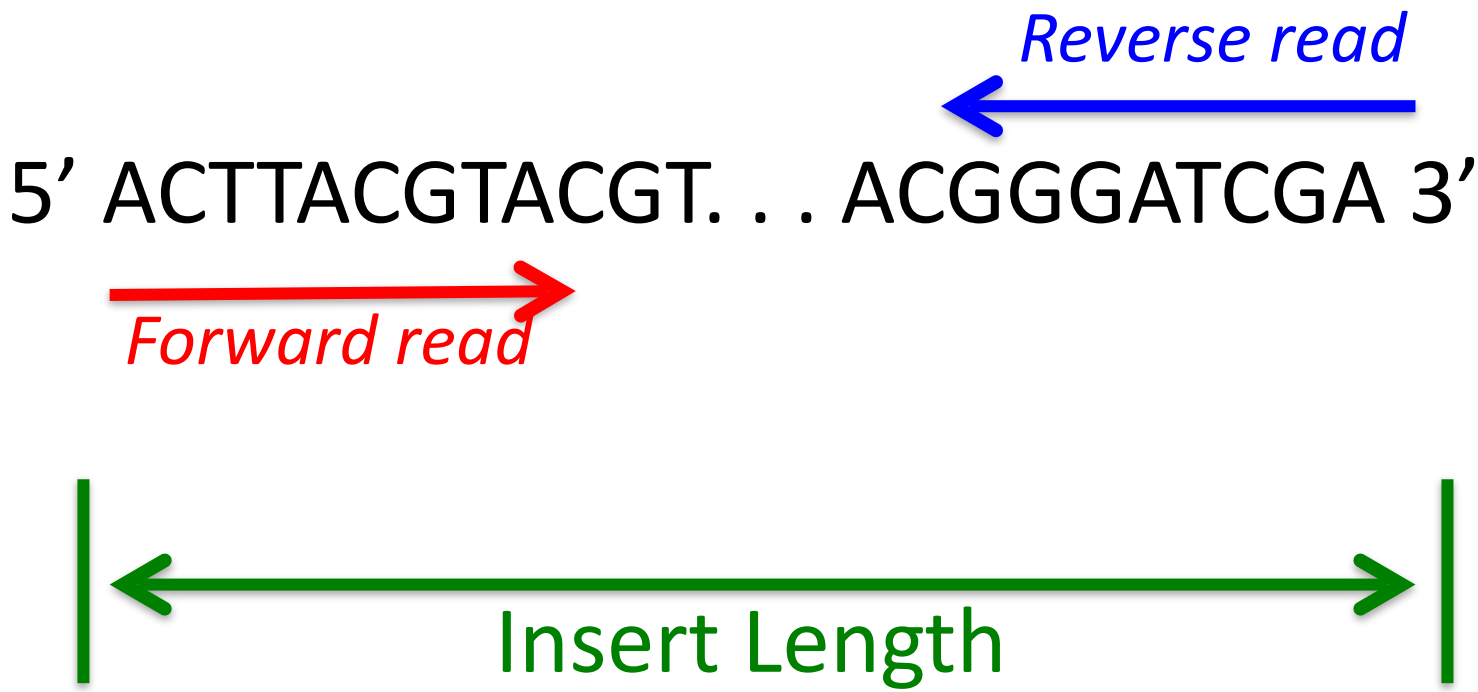
Paired-end sequencing: bonus information

- The “insert length” is known, and roughly constant for all reads

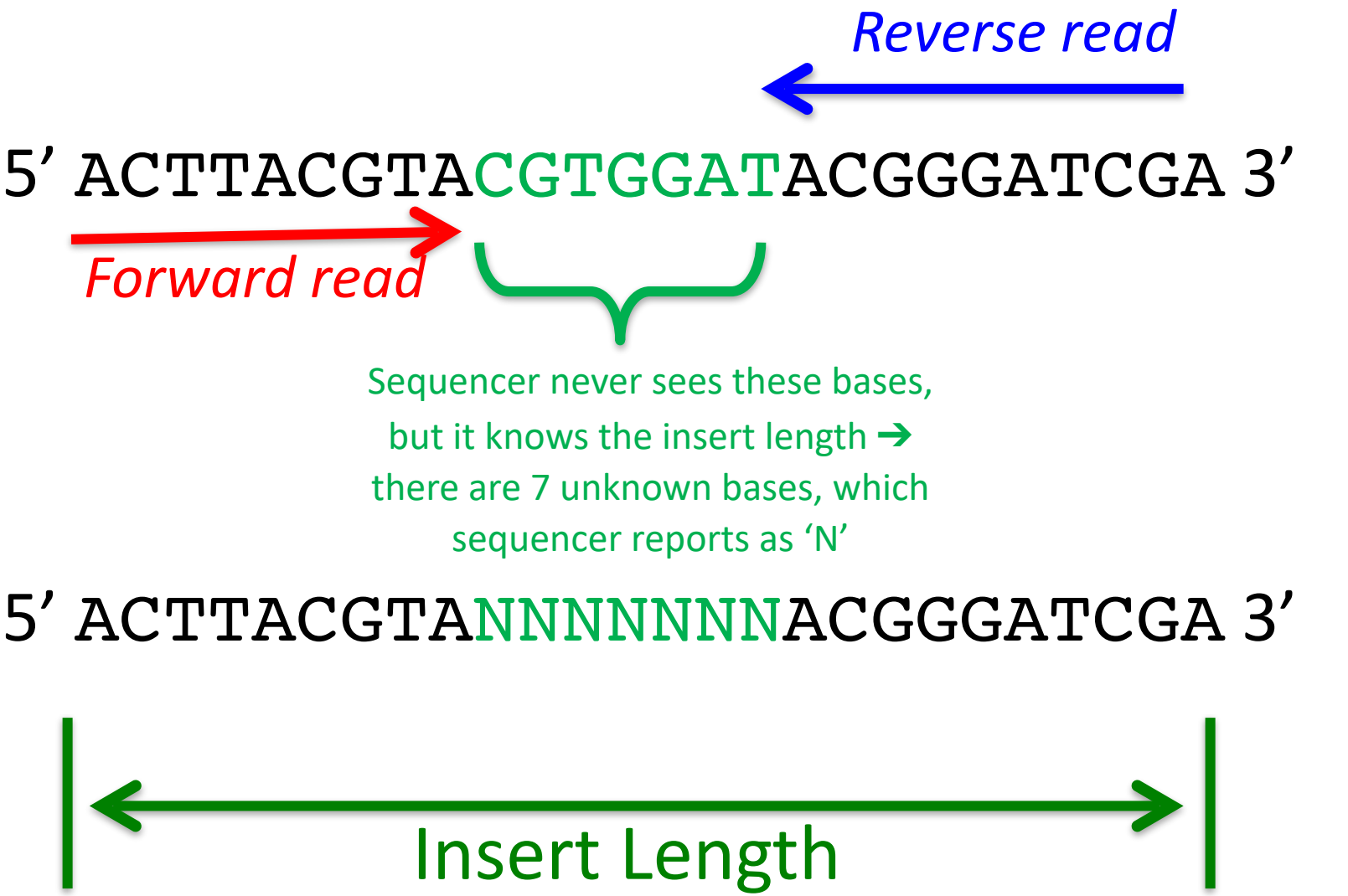


Paired-end non-overlapping reads

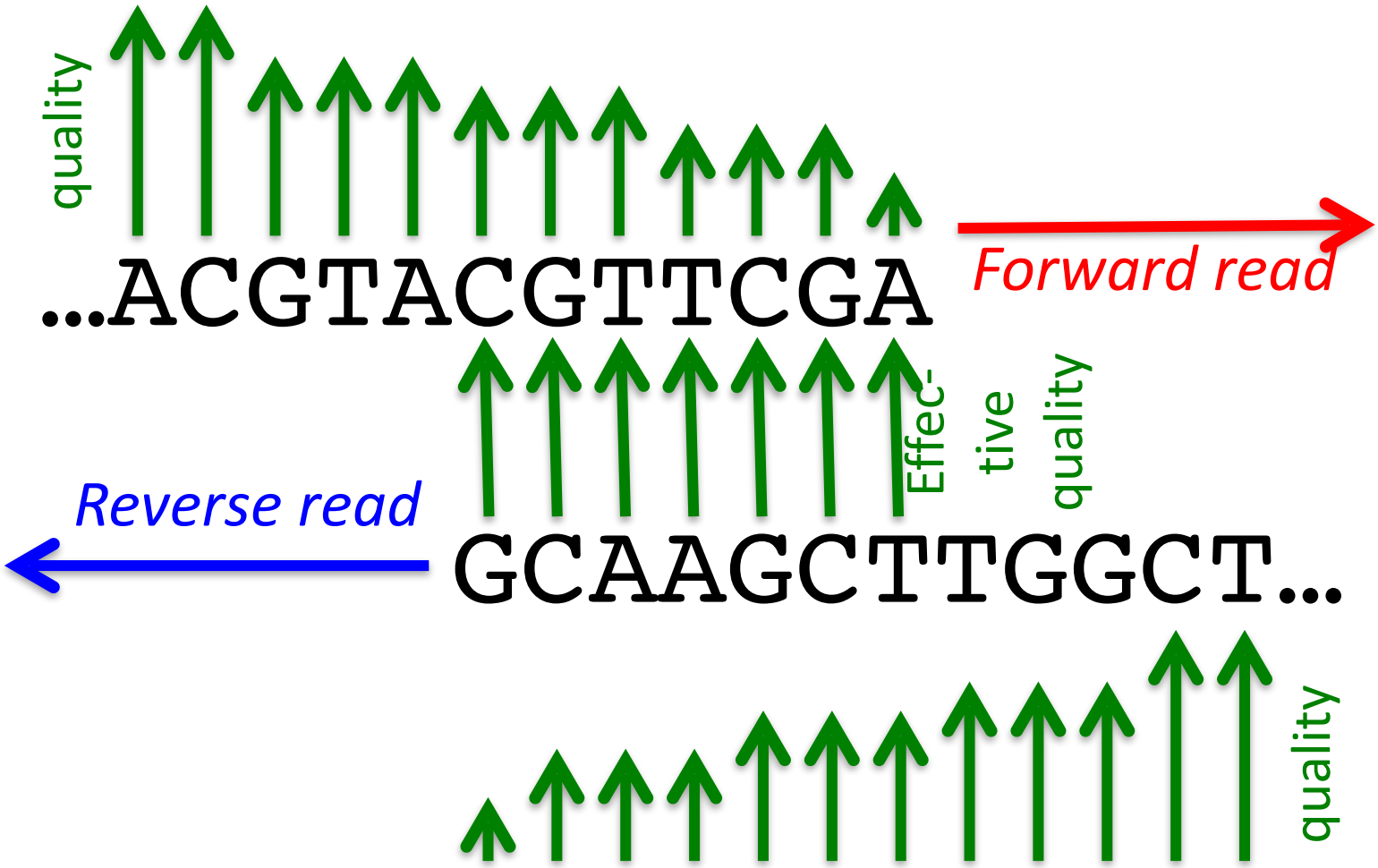
- If insert length > forward read length + reverse read length
- Relative positions of the pair of reads is valuable information for assemblers



Paired-end non-overlapping reads



Paired-end overlapping reads: bonus quality where you need it most



Paired-end overlapping reads

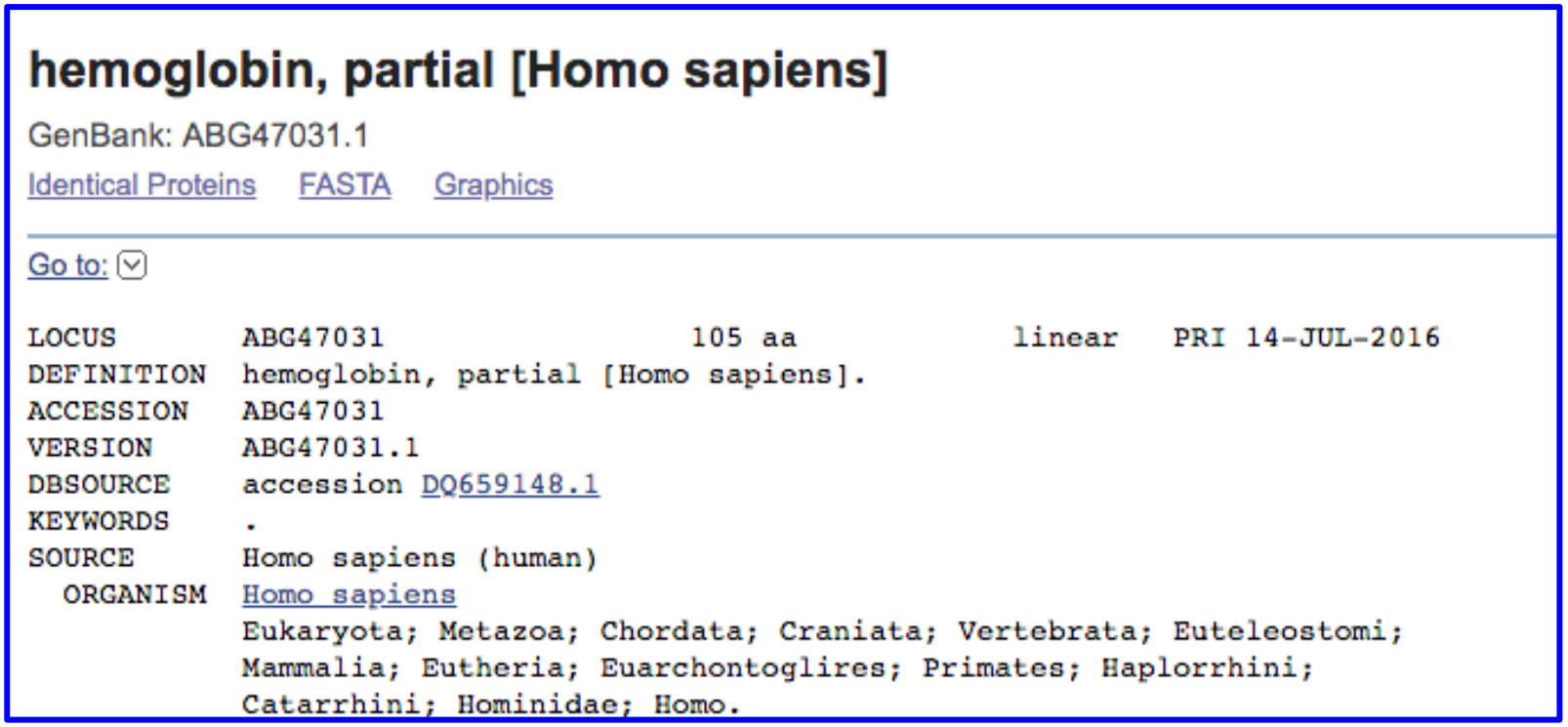
- 3’ ends of reads have poorest quality
- But they overlap, so it’s meaningful if they agree
- If 2 unreliable witnesses independently report the same event, P(event really happened) is high

Step 5: Identify

- Identification approaches are alignment-based
  - Blastn each read against a “reference database”
    - Usually GenBank
    - Custom vouchered reference database

Blasting against GenBank

- GenBank contains organism and function annotations



hemoglobin, partial [Homo sapiens]

GenBank: ABG47031.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       ABG47031                 105 aa            linear   PRI 14-JUL-2016
DEFINITION  hemoglobin, partial [Homo sapiens].
ACCESSION   ABG47031
VERSION     ABG47031.1
DBSOURCE    accession DQ659148.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

A sharper tool than GenBank: Custom voucher-based databases

- Expensive.
- Best bioinformatics practices are not yet developed.

Voucher-based studies

- An expert identifies an organism
- Extract tissue from the organism
- Sequence some of the tissue
- Cold-store remaining tissue: the “voucher”
  - In case of controversy, the voucher *vouches for* the identification
  - “You say that’s vampire squid DNA? Prove it!”
  - Can’t do that with GenBank records

Aquarium of the Pacific

Yes, vampire squid is a thing.

Voucher-based studies

- Sample an environment
- Compare sampled sequences against vouchered database
  - Blast reads against custom database
  - Usually E-value of best hit is << (much much better than) other hits
  - Custom database, so need to develop new intuitions about range of E-values that mean strong hits
- Advantages over GenBank:
  - *Much* higher confidence in identity of subjects
  - Faster blast
- Disadvantages:
  - You can't identify anything that isn't in your database

Cryptic species

- "Cryptic" means can't be visually distinguished from a different species.
- Invasive species are often hard to distinguish from natives.
- Example: mussels

Mytilus trossulus
Native to California coast

Mytilus galloprovincialis
Invaded southern California
? 19th century ? Early 20th ?


Case Study: UCYN-A

Trichodesmium alone doesn't account for all marine nitrogen fixation

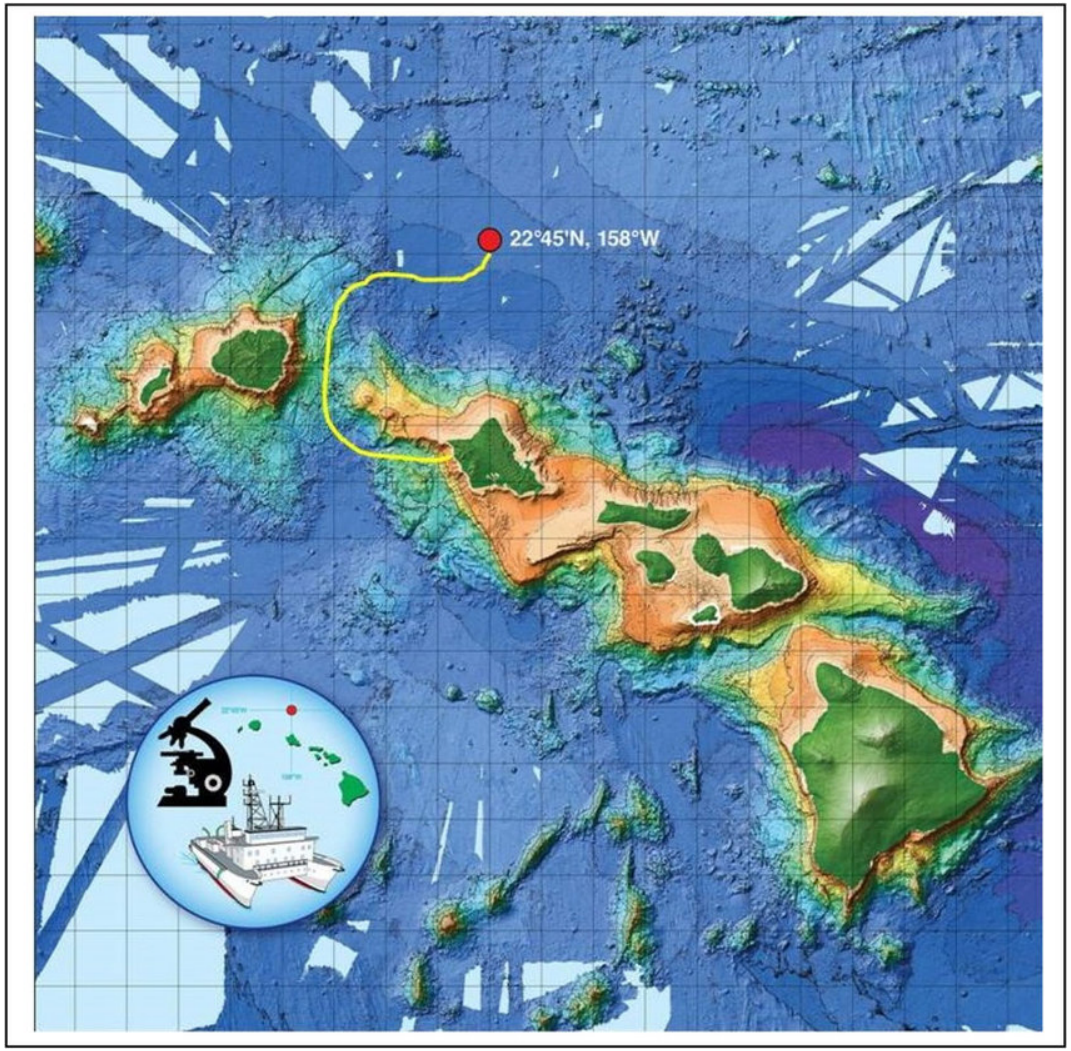
- "Direct estimates of N_2 fixation, largely of … *Trichodesmium*, can account for only a quarter to one-half of the geochemically derived rates of N_2 fixation in various ocean basins."

  - Mahaffey et al., 2005, Am.J. of Sci.

"Geochemically derived" = from computer models

- What else in the oceans is fixing 70 – 105 Tg of nitrogen annually?
  - ~= 1/10,000 of the atmosphere

A metagenomic search for marine nitrogen fixers at Station Aloha



A metagenomic search for marine nitrogen fixers at Station Aloha

- Sample the open ocean
- PCR with degenerate nifH primers
  - Wild-card search for nifH genes among all that DNA

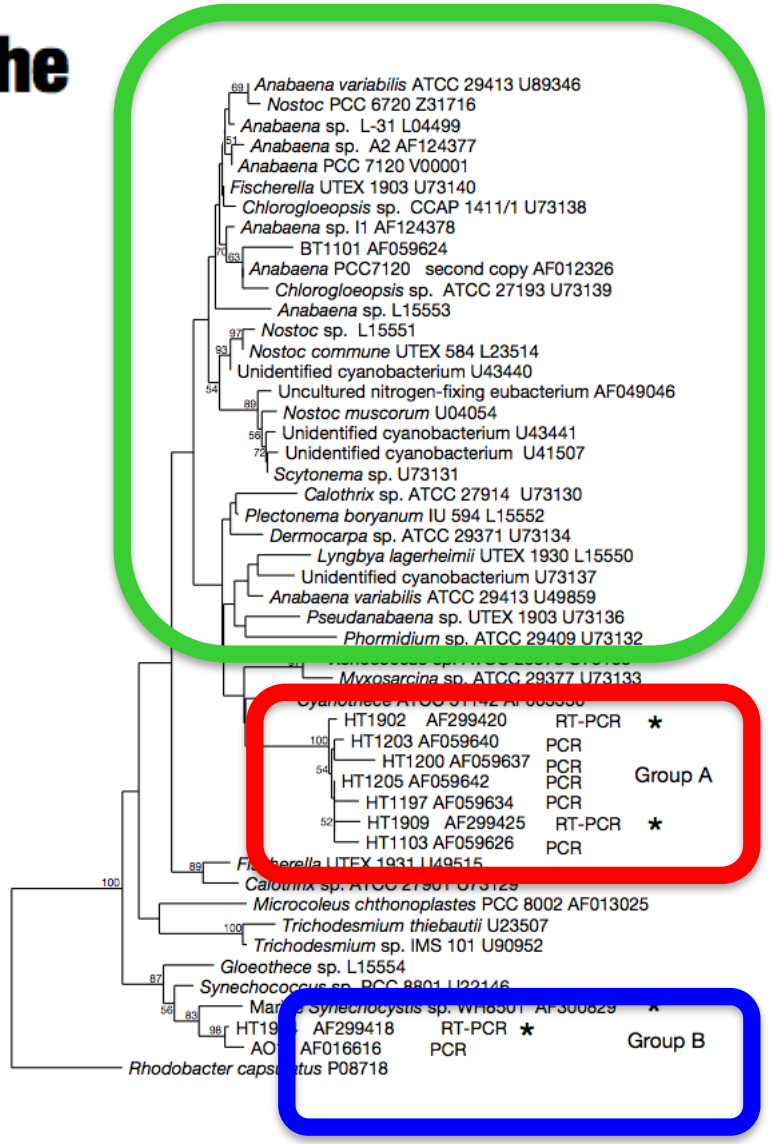
Discovery of unicellular marine nitrogen fixers

Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean

Jonathan P. Zehr*, John B. Waterbury†, Patricia J. Turner*, Joseph P. Montoya‡, Enoma Omoregie*, Grieg F. Steward*, Andrew Hansen§ & David M. Karl§

-- Nature, 2001

UCYN-A



Group B = *Crocosphaera watsonii*

123B S20 Module 7:
Future Directions of Bioinformatics

Remainder of the semester

3/22: Futures	3/24: Work on projects
3/29: Spring break	3/31: Spring break
4/5: ARBitrator lecture	4/7: ARBitrator lab
4/12: Midterm 2 review	4/14: Midterm 2
4/19: Project presentations	4/21: Project presentations
4/26: Project presentations	4/28: Project presentations
5/3: Project presentations	5/5: Project presentations
5/10: Final exam review	

Wed 5/19: Section 1 (9:00 AM lectures) final exam at 7:15 AM
Tues 5/25: Section 2 (10:30 AM lectures) final exam at 9:45 AM

Today's plan

- A little context
- Deep Learning
- Projects:
  - Poriferal vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

Adverb: Ad-hoc Viterbi

- 1990s: computers were just not all that
  - Slower
  - Less memory
  - Viterbi algorithm is O(nstates^2 * seqlen)
    - Time and memory
- Original bioinformatic HMMs were painstakingly designed and trained
  - Intended for heavy re-use
- 2020s: computers are all that
  - Your protein HMM app builds an HMM in ~1 sec
  - Single-use (“ad-hoc”) HMMs are a possibility

COI barcoding

- COI = Cytochrome C oxidase, subunit 1
- All animal species have it
- ”Barcode of animal life”
  - Unique across almost all animal species
- To identify a sample (e.g. blood, hair, tissue, ARMs plate)
  - Extract DNA
  - Amplify COI using primers
  - Sequence and blast
- The BOLD database
  - Vouchered
  - Stringent acceptance criteria

COI barcoding

- If the species of the tissue has previously been identified …
  - It's in your blastable database (BOLD, CO-ARBitrator, or GenBank)
  - You'll get a perfect hit
  - Or maybe slightly imperfect, due to amplification or sequencing error, but there's still exactly 1 lowest E-value hit
  - ➔ reliable identification
- If the species is previously unknown
  - It can't possibly be in your database
  - You can't possibly identify its species
  - Reasonable expectation: genus of best hit will be correct
  - But no!

COI barcoding and novel species

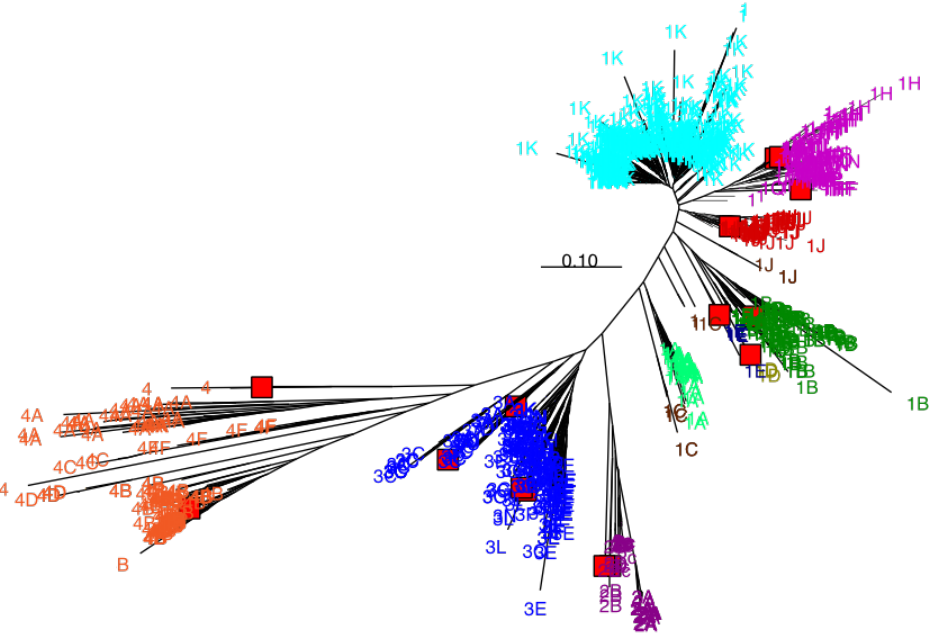
- The barcode concept was intended only for known species
- If query is novel, can we use the taxonomy of the best blast hit in some way?
  - P(correct phylum/class/order) ~= 100%
  - P(correct family) = 67%
  - P(correct genus) = really bad
- Many metagenomic/ARMS studies use COI identification
  - Sample contains both known and unknown species
  - Need an algorithm that can handle both cases

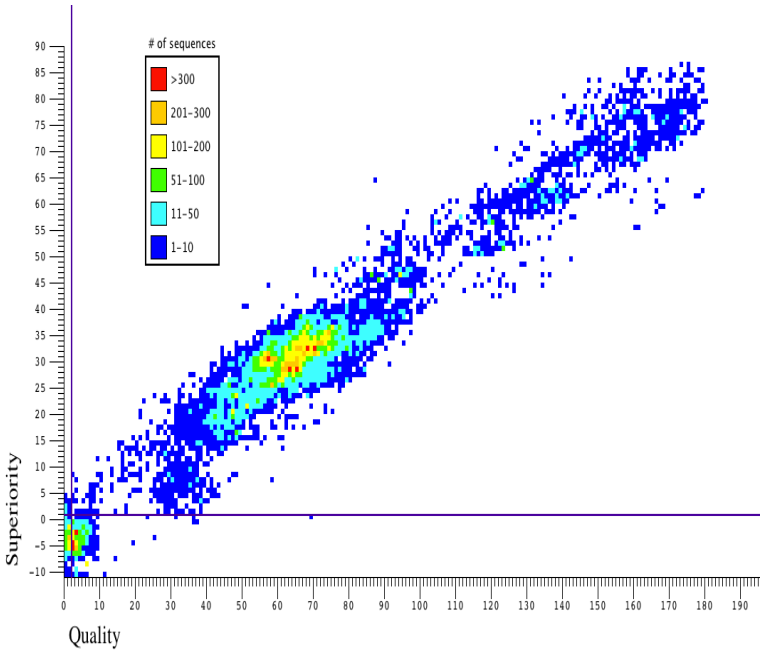
Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species

The Adverb Algorithm

- blastn your sequence against BOLD database
- Perfect or near-perfect hit?
  - ➔ previously known, accept species identification
- Imperfect hit?
  - Accept blast's order identification
  - Build an HMM for every family in the order
  - Compute Viterbi score of sequence on every family HMM
  - Highest Viterbi score indicated family identification
  - 90% accuracy
  - No reliable genus identification
- 18 months * 15 nodes * 28 cpus = 630 CPU-years

Bio/CS 123B Spring 2020
Module 8: Data Mining GenBank Conserved Domains and ARBitrator





Why don’t HMMs work well in identifying *NifH*?

- Unknown
- I don’t think anyone has looked into it
- I have my suspicions … paralogs?

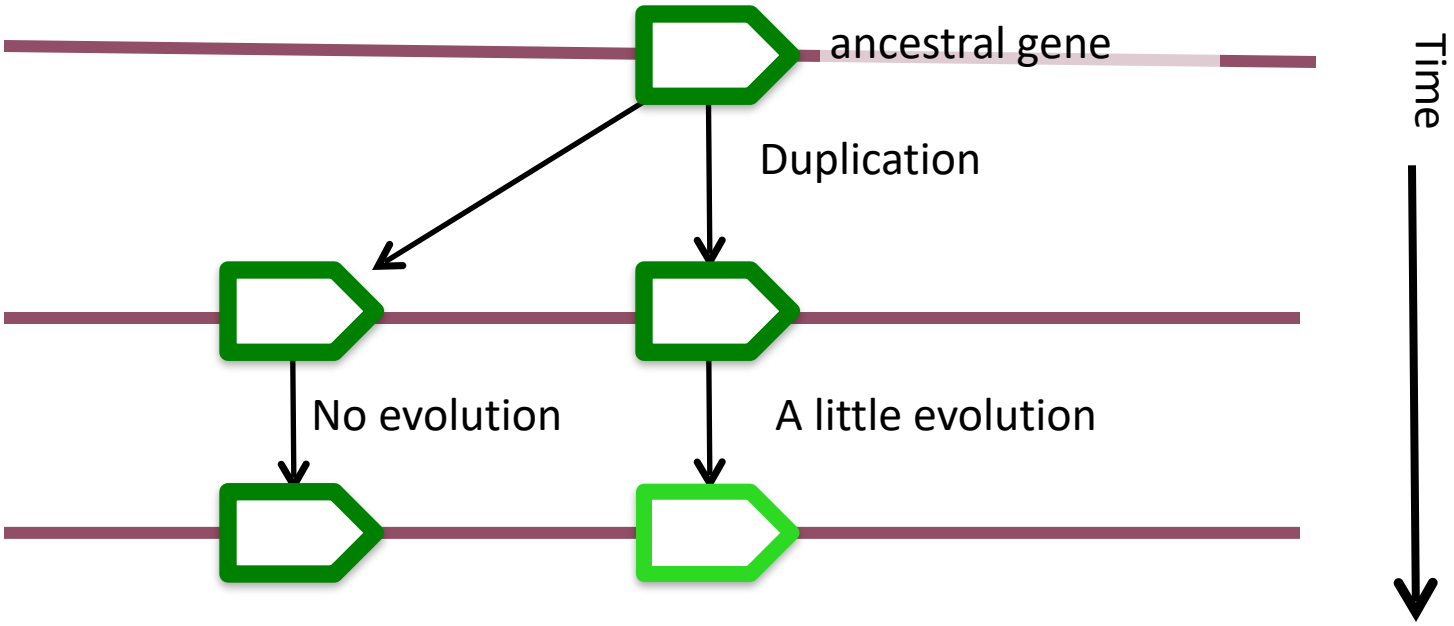
Kinds of Sequence Similarity

- Similarity
  - Coincidence
  - Homology
    - Orthology (orthologs): 2 different species
    - Paralogy (paralogs): 2 different genes in same genome

Origins of similarity

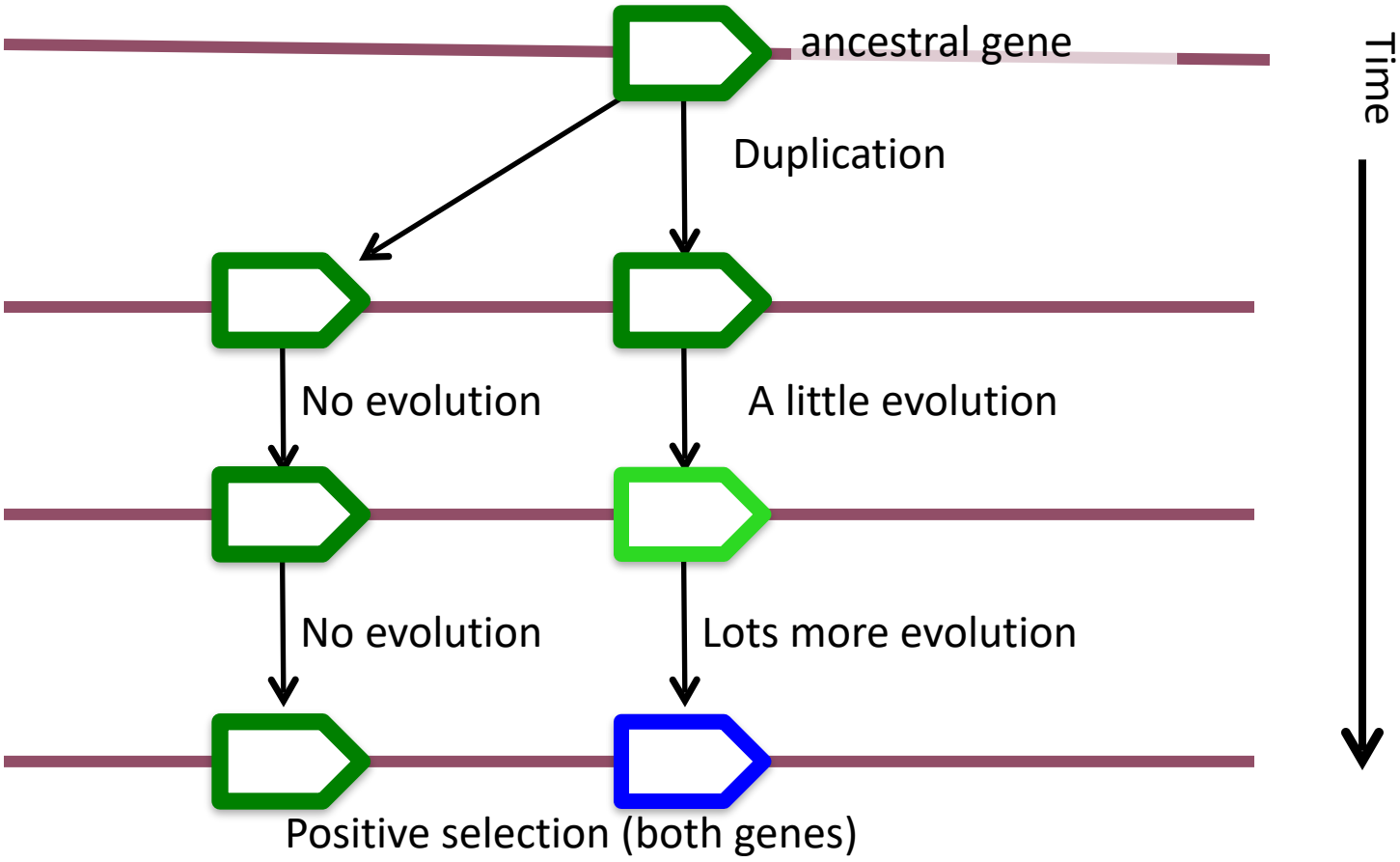
	Orthologs	Paralogs
Event type	Speciation	Gene duplication
Location of similar genes	2 different species	2 genes in same organism

Duplication

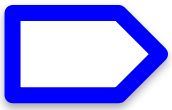


is stabilized by selection

Duplication: 1 possible destiny



has evolved to provide a new & beneficial function

 and  are *paralogs*:

- Sequences are similar but different
- Different function, but maybe related
- Common genome
- Related by descent from a common ancestor, via a duplication event

Some Classifier Terminology

- Sensitive = able to positively identify what you're looking for
  - Whether or not you mistakenly positively identify other stuff also
- Specificity = ability to avoid incorrect identifications
  - Whether or not you mistakenly overlook correct identifications

Example: panning for gold

Panning for gold

- Panning = searching “database” of river sediment
- Finding gold = a positive identification
- False positive error: you think it’s gold, but it’s not
- False negative error: you throw gold back in the river
- Sensitive = find all the gold in the river
  - Low false negative rate
- Specific = reject dirt, pyrite, rocks, etc.
  - Low false positive rate

Data mining GenBank for NifH

- True positive: a sequence that you correctly accept as NifH
- True negative: a sequence that you correctly reject as not NifH
- False positive error: you think it's NifH, but it's not
- False negative error: you reject NifH
- Sensitive = find all the NifH in GenBank
  - Low false negative rate
- Specific = reject NifD, NifK, all other not-nifH.
  - Low false positive rate

E-value definition

- Given:
  - A query sequence of length L
  - Which you blast against some database DB
  - And you get a hit with score S …
- The E-value of your hit is the probability of:
  - Blasting a random query of length L
  - Against database DB
  - With score >= S

Simplify the numbers: Superiority

- In phase 2, reject any sequence whose best rpsBLAST subject isn't the NifH CD
- Superiority only refers to sequences whose best rpsBLAST subject *is* the NifH CD
- Define Superiority =

log10(E-value of best negative hit)

Minus log10(E-value of best positive hit)

= by how many orders of magnitude (OOMs) is the best positive hit better than the best negative hit?

Superiority Example 1: In phase 2, rpsBLAST a sequence and get the following hits

Subject conserved domain	E-value of hit Subject C.D.
NifH conserved domain	1.0E-80
[illegible] domain	[illegible]
[illegible] domain	[illegible]

Superiority = log10(1.0E-60) – log10(1.0E-80)

= -60 – -80 = 20

Hit to NifH conserved domain is 20 O-O-Ms better than 2nd-best hit

Superiority Example 2: In phase 2, rpsBLAST a sequence and get the following hits

Subject conserved domain	E-value of hit Subject C.D.
NifH conserved domain	2.7E-35
NifH conserved domain	1.0E-32
Cellulose biosynthesis domain	2.7E-20

Superiority = log10(2.7E-20) – log10(2.7E-35)
= -20 – -35 = 15

Ignore the worse NifH conserved domain hit

Superiority Example 3: In phase 2, rpsBLAST a sequence and get the following hits

Subject conserved domain	E-value of hit Subject C.D.
NifH conserved domain	8.9E-61
Fer4_NifH family	1.0E-59
Cellulose biosynthesis domain	8.9E-55

Superiority = log10(8.9E-55) – log10(8.9E-61)
= -55 – -61 = 6

Ignore the family hit

Is there a good Superiority threshold that correctly classifies the positive and negative training sets?

- Yes!
- Threshold = 1 (for NifH)
- Tiny error rates