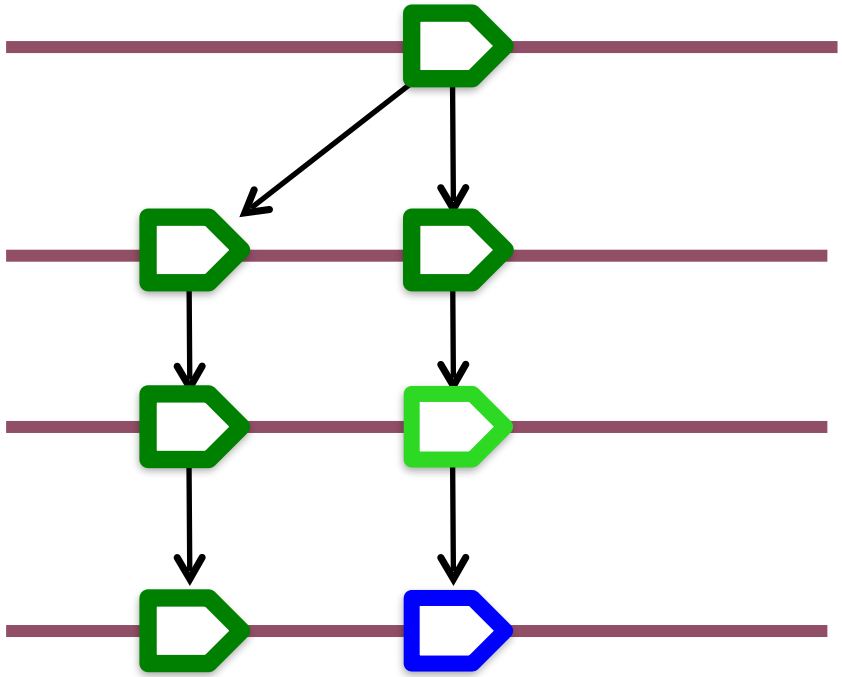
BIOL/CS 123B

Final Exam Review



Spring 2021

Philip Heller

End of semester business

- Canvas ***does not accurately compute overall grades!!***
  - It's not even close!
- If you want to compute your letter grade, use only your raw Canvas scores (MT1, MT2, Final Exam, Homeworks, Project Report). Apply the formula in the syllabus.
- If during the summer you think your letter grade is wrong, mail your computation to me.

Rules (9:00 section):

- You may refer to your notes, homework, labs, and slides.
- You may not use the web except as directed by a question.
- You may not communicate with anyone.
- Your answers must be entirely in your own words. Using someone else's words is plagiarism.
- You may only submit once.
- Edit this doc. Use blue for your answers. Upload to assignment "Final" on Canvas by 9:40 AM.
    - Unless you have a time accommodation. If you do, email your exam to philip.heller@sjsu.edu by 10:50 AM.
- You must attend the zoom session with your camera on. If you have a question, unmute and ask to go to a breakout room. If you need a biobreak, send "bathroom" as a private chat message and go; send "back" when you return.
- You may not discuss this exam with anyone except other students in this section until 2:00 PM on Wednesday May 25 (when the other section has finished their exam).

Very very strong suggestions:

- Don't add extraneous information to your answers.
- Work on question 1 first.

Rules (10:30 Section):

- You may refer to your notes, homework, labs, and slides.
- You may not use the web except as directed by a question.
- You may not communicate with anyone.
- Your answers must be entirely in your own words. Using someone else's words is plagiarism.
- You may only submit once.
- Edit this doc. Use blue for your answers. Upload to assignment "Final" on Canvas by 12:10 PM.
  - Unless you have a time accommodation. If you do, email your exam to philip.heller@sjsu.edu by 1:30 PM.
- You must attend the zoom session with your camera on. If you have a question, unmute and ask to go to a breakout room. If you need a biobreak, send "bathroom" as a private chat message and go; send "back" when you return.

Very very strong suggestions:

- Don't add extraneous information to your answers.
- Work on question 1 first.

The Big Ideas

- Technology
  - Amplification
  - Sequencing
- Classifiers
  - Hidden Markov Models
  - ARBitrator
  - Big-Oh
  - Terminology
- CRISPR
- Statistics
- Metagenomics

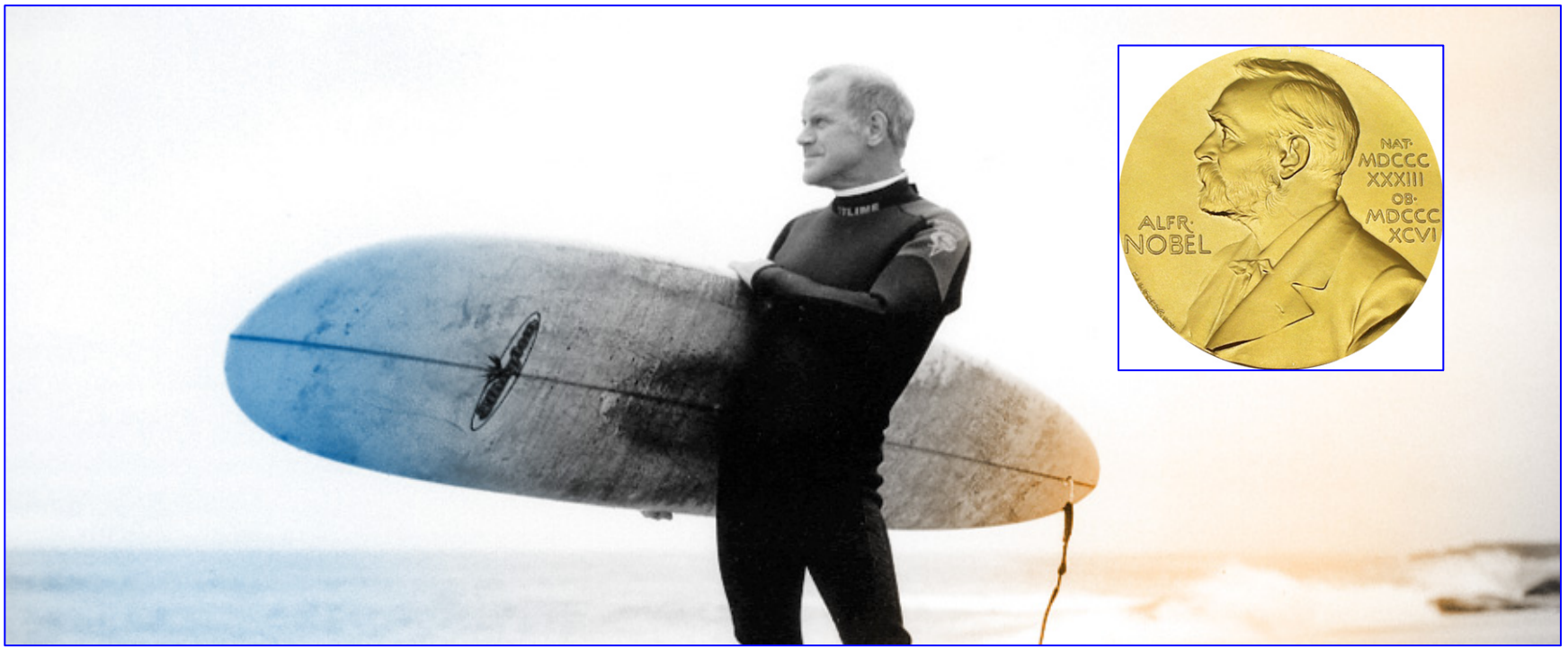
To answer your question …

- Question: You didn’t cover xxxxx today. Does that mean it won’t be on the exam?
- Answer: It might be on the exam.

Technology

- Decks 01 and 06
  - Amplification (PCR, no cloning)
  - Sequencing
  - Paired-end

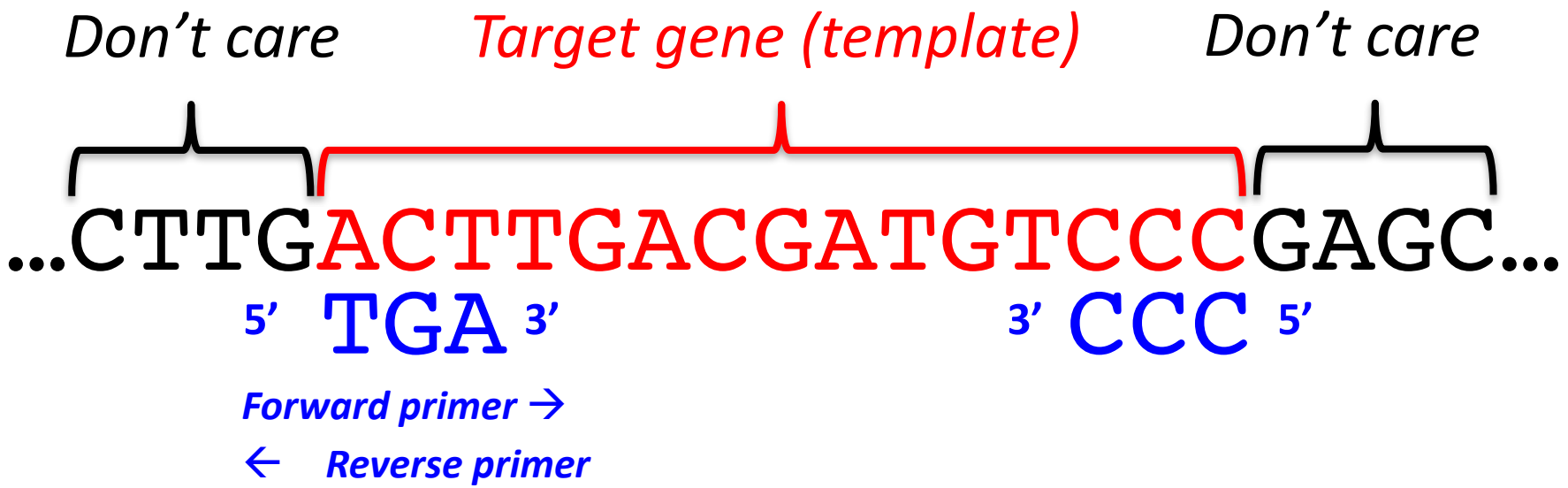
1983: Kary Mullis develops Polymerase Chain Reaction (PCR)



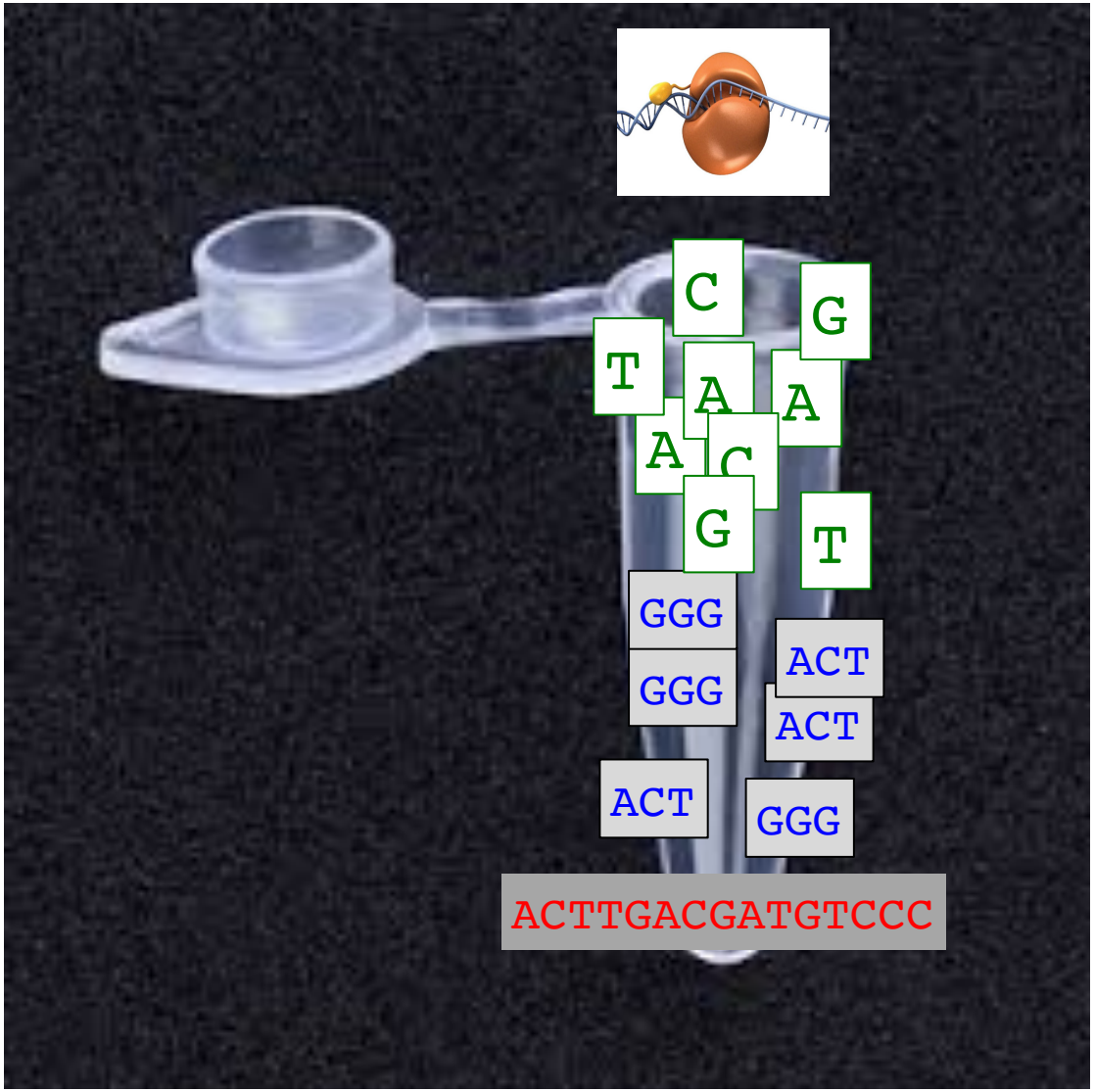
No more plasmids, no more bacteria

PCR Primers (short single strands)

- If you know the sequence of the gene you want to amplify…
- Or actually just the beginning and end of the sequence …
- Make 2 primers
  - Reverse complement of first n bases of the gene
  - Last n bases of the gene
  - n ~= 10 – 30 (but n = 3 in these simple examples)



The PCR Tube: where it all happens



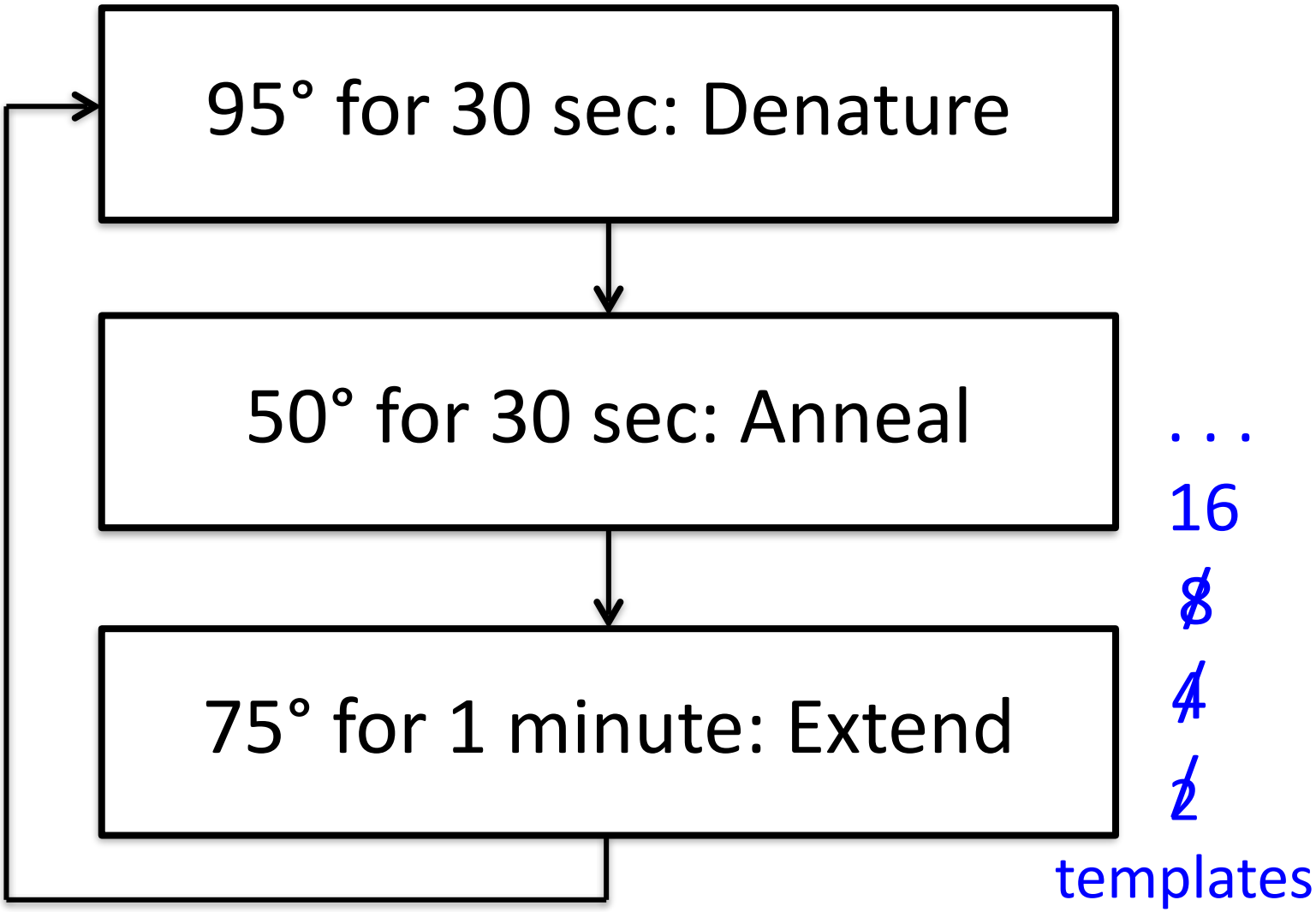
Add DNA template

Add many copies of primers

Add lots of As, Cs, Gs & Ts

Add DNA polymerase and other enzymes

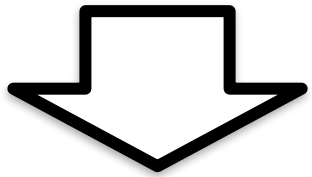
Now just cycle the temperature
Temps & times are rough approximates



PCR Cycle, Step 1: Denature (Separate the 2 DNA strands)

ACTTGACGATGTCCC

TGAACTGCTACAGGG

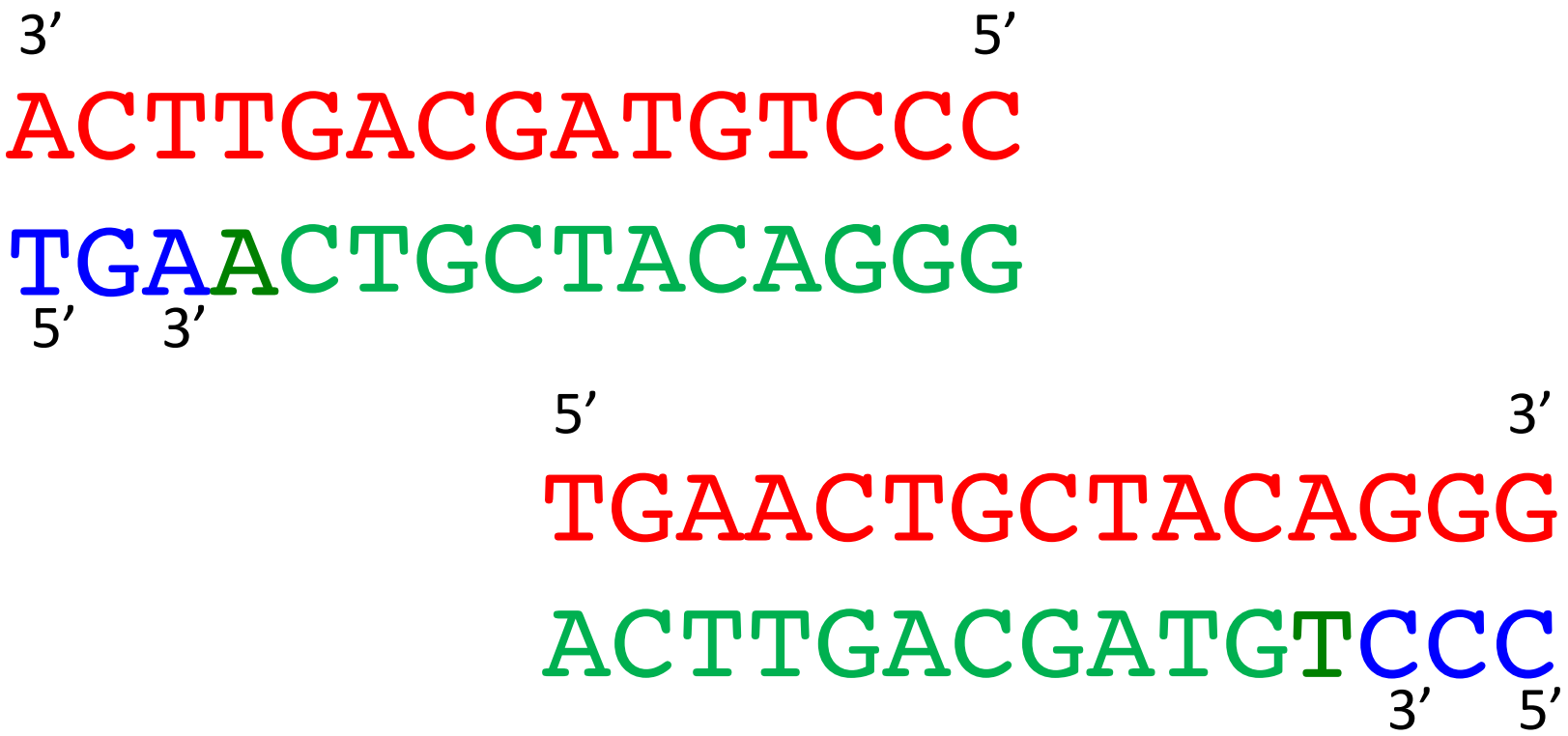


ACTTGACGATGTCCC

TGAACTGCTACAGGG

PCR Cycle, Step 2: Anneal (Attach primers to 5’ ends)

PCR Cycle, Step 3: Extend
(Free-floating bases extend primers at 3')



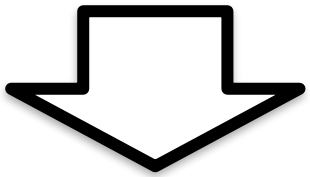
PCR Cycle, Back to Step 1: Denature

ACTTGACGATGTCCC

TGAACTGCTACAGGG

TGAACTGCTACAGGG

ACTTGACGATGTCCC



ACTTGACGATGTCCC

ACTTGACGATGTCCC

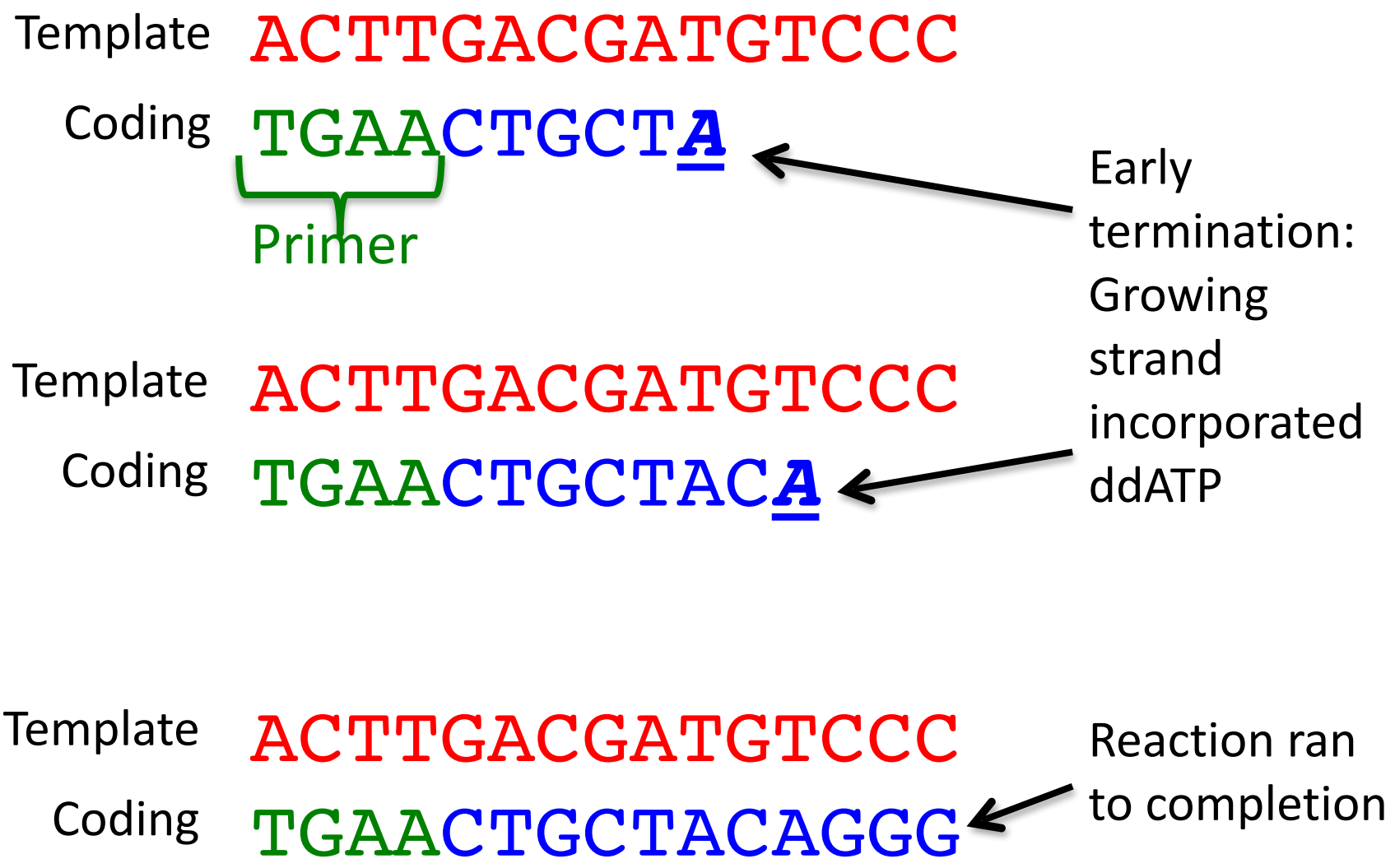
TGAACTGCTACAGGG

TGAACTGCTACAGGG

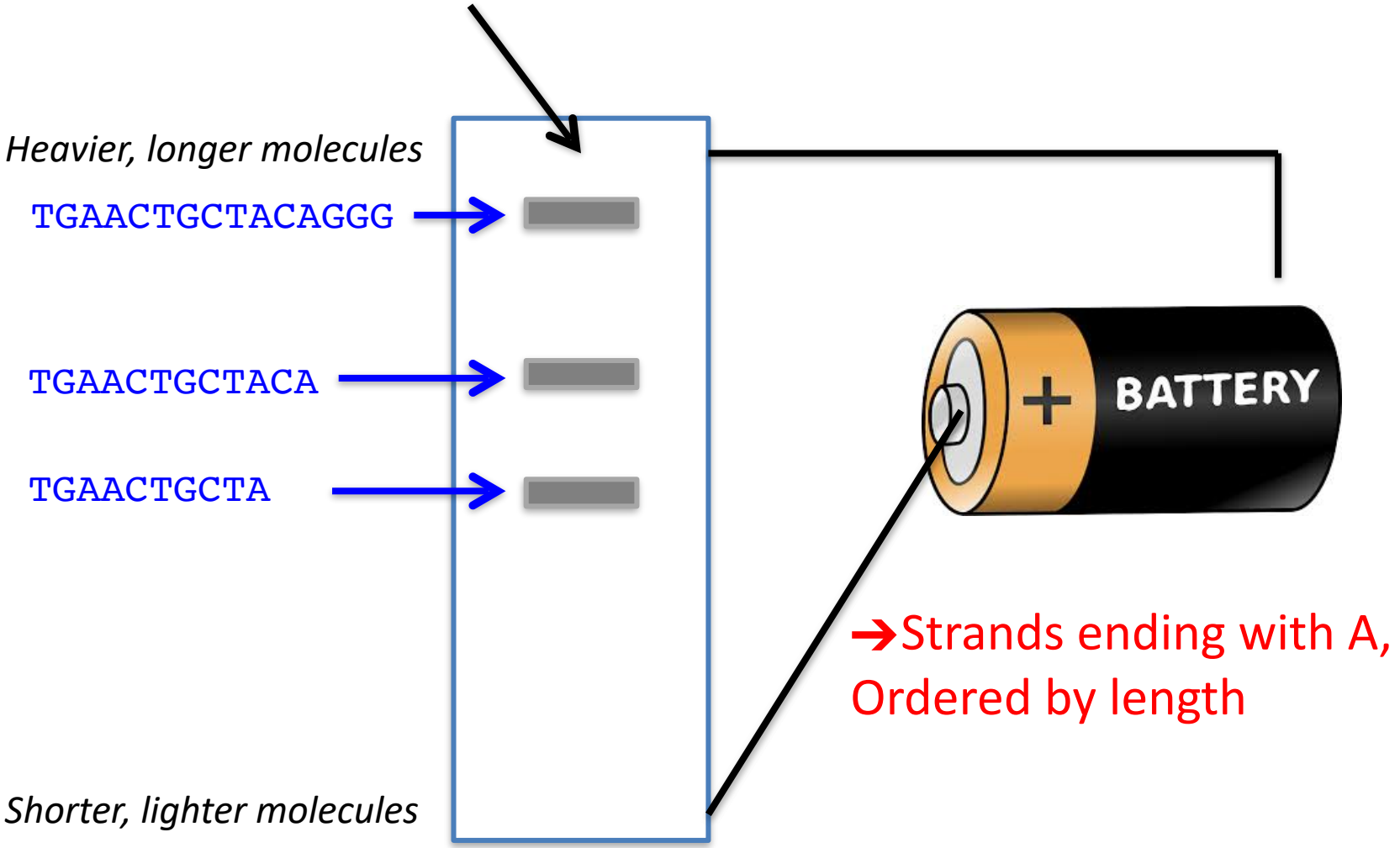
Sanger Sequencing

- ddNTP is “di-deoxy” because compared to RNA pentose, 2 carbons (2’ & 3’) have lost an Oxygen molecule
- There are 4 kinds of ddNTP:
  - Chain-terminating Adenine = ddATP
  - Chain-terminating Cytosine = ddCTP
  - Chain-terminating Guanine = ddGTP
  - Chain-terminating Thymine = ddTTP

After the reaction:

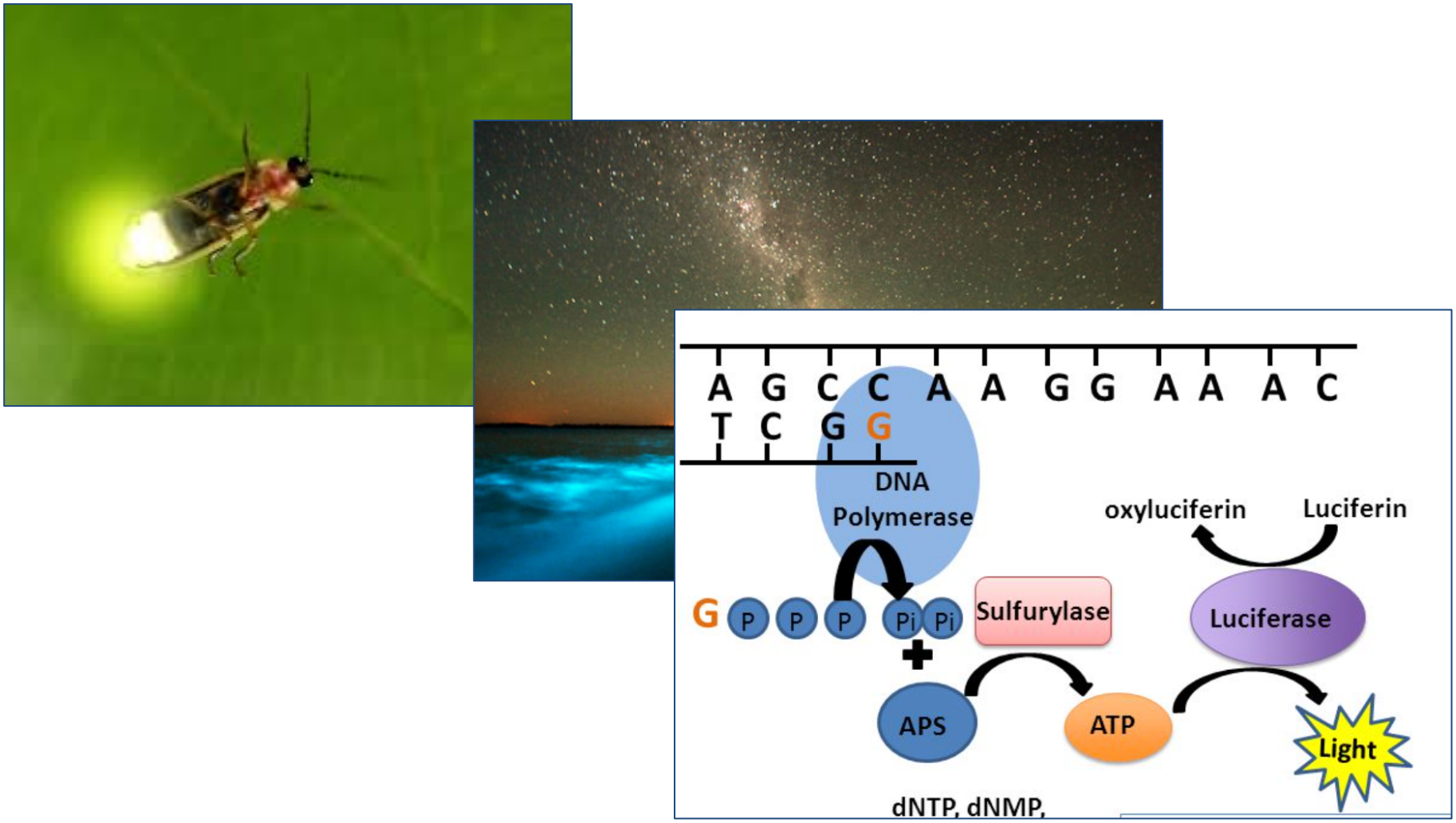


TGAACTGCTA +
TGAACTGCTACA +
TGAACTGCTACAGGG

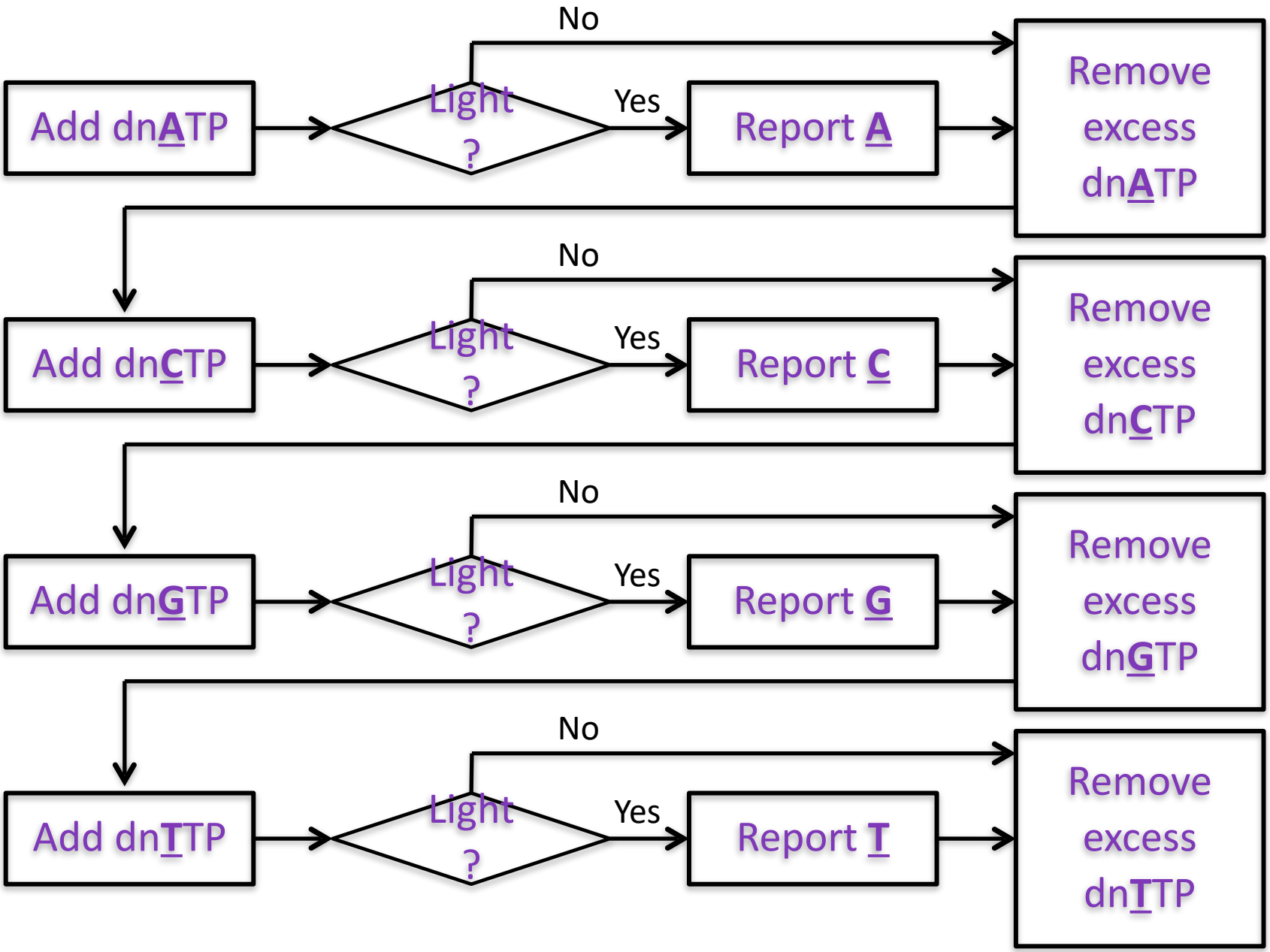


→Strands ending with A,
Ordered by length

1990: Pyrosequencing



Pyrosequencing Cycle



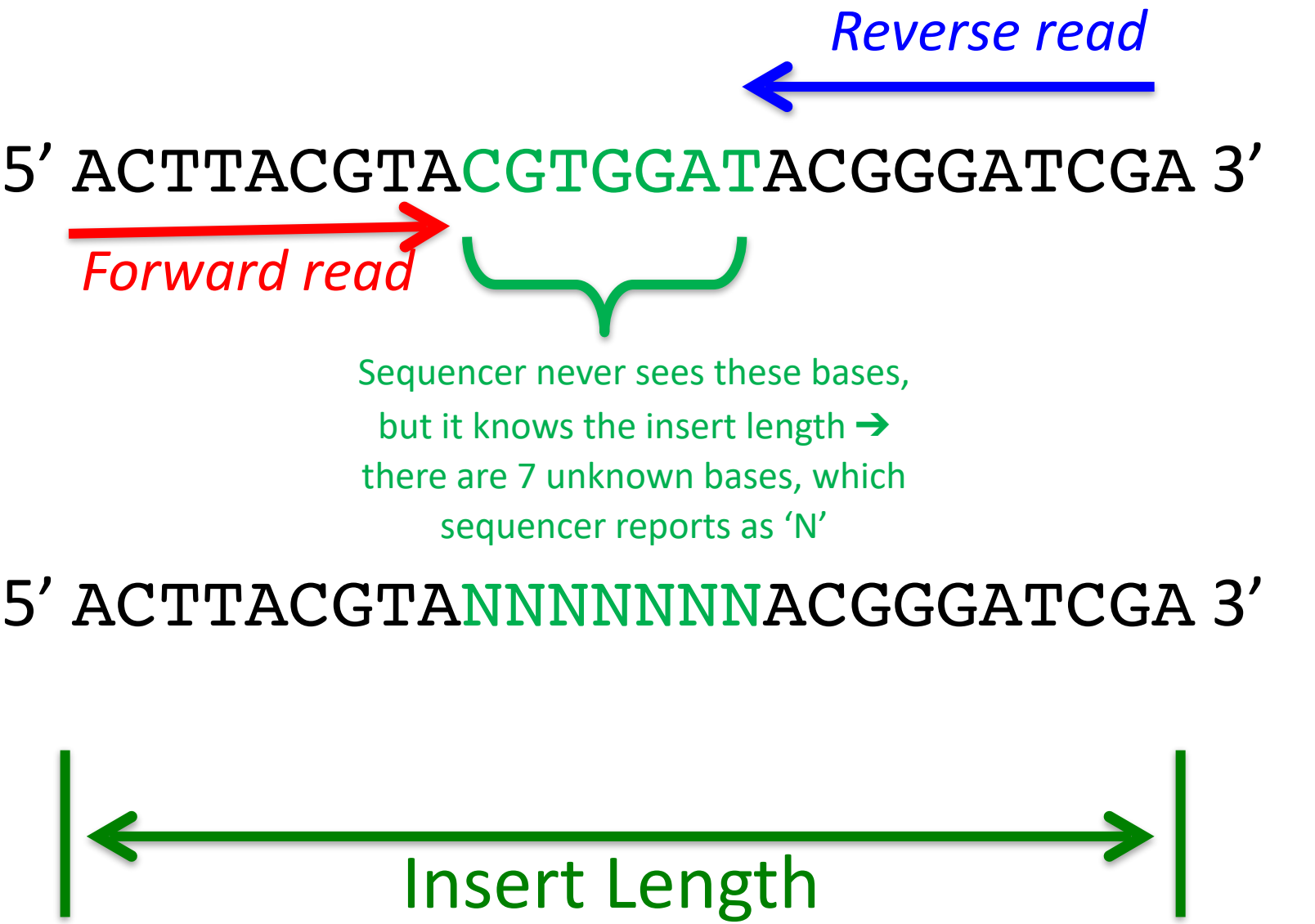
Pyrosequencing is error prone

- Flashes are brief
- Photodetectors are imperfect
- Hard to distinguish XXXXX from XXXXXX
- It's cheap and it's fast, but its quality is lower than Sanger sequencing
- → Use PCR to quickly/cheaply make lots of copies, which are quickly/cheaply sequenced
- → Trust the majority

Paired-end Sequencing

- 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
  - Result is 2 fastq files

Paired-end non-overlapping reads



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

The Big Ideas

- Technology
  - Amplification
  - Sequencing

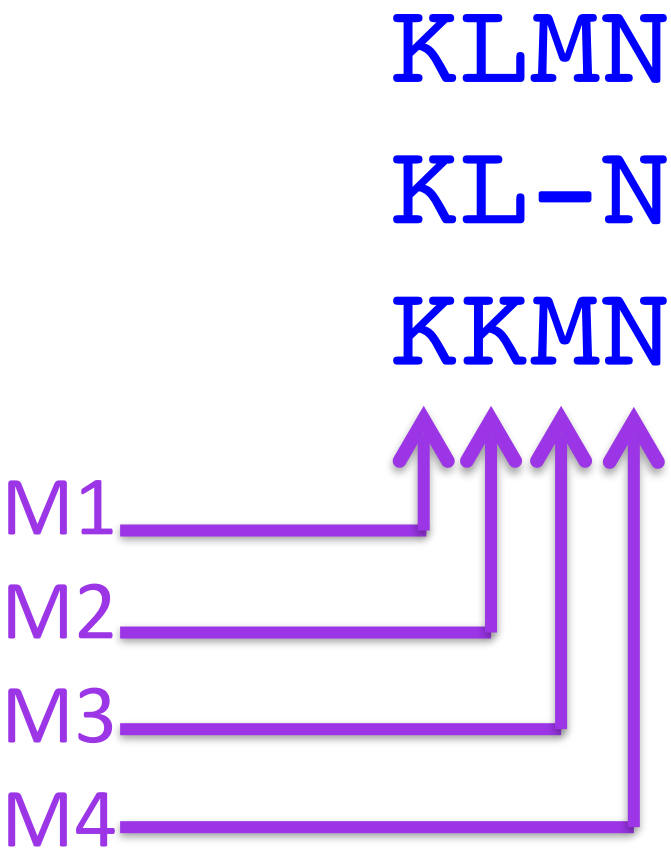
- Classifiers
  - Hidden Markov Models
  - ARBitrator
  - Big-Oh
  - Terminology
- CRISPR
- Statistics
- Metagenomics

Classifiers: HMMs

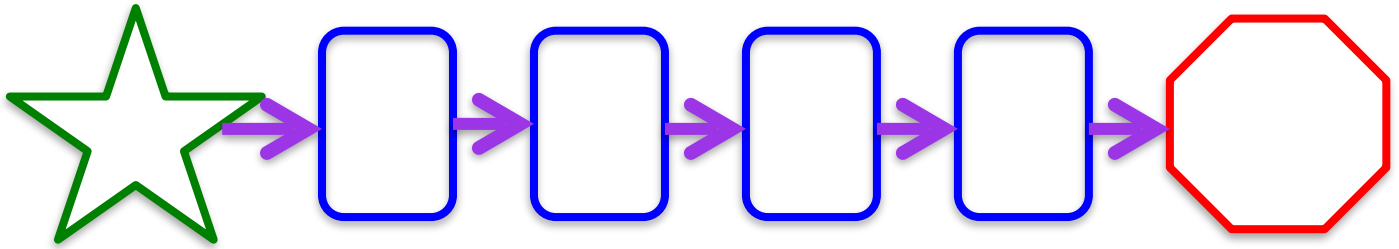
Jack Kirby

- Thor/weather: No
- Algorithms: Forward, Viterbi
- Application: CpG Islands
- Application: pHMMs

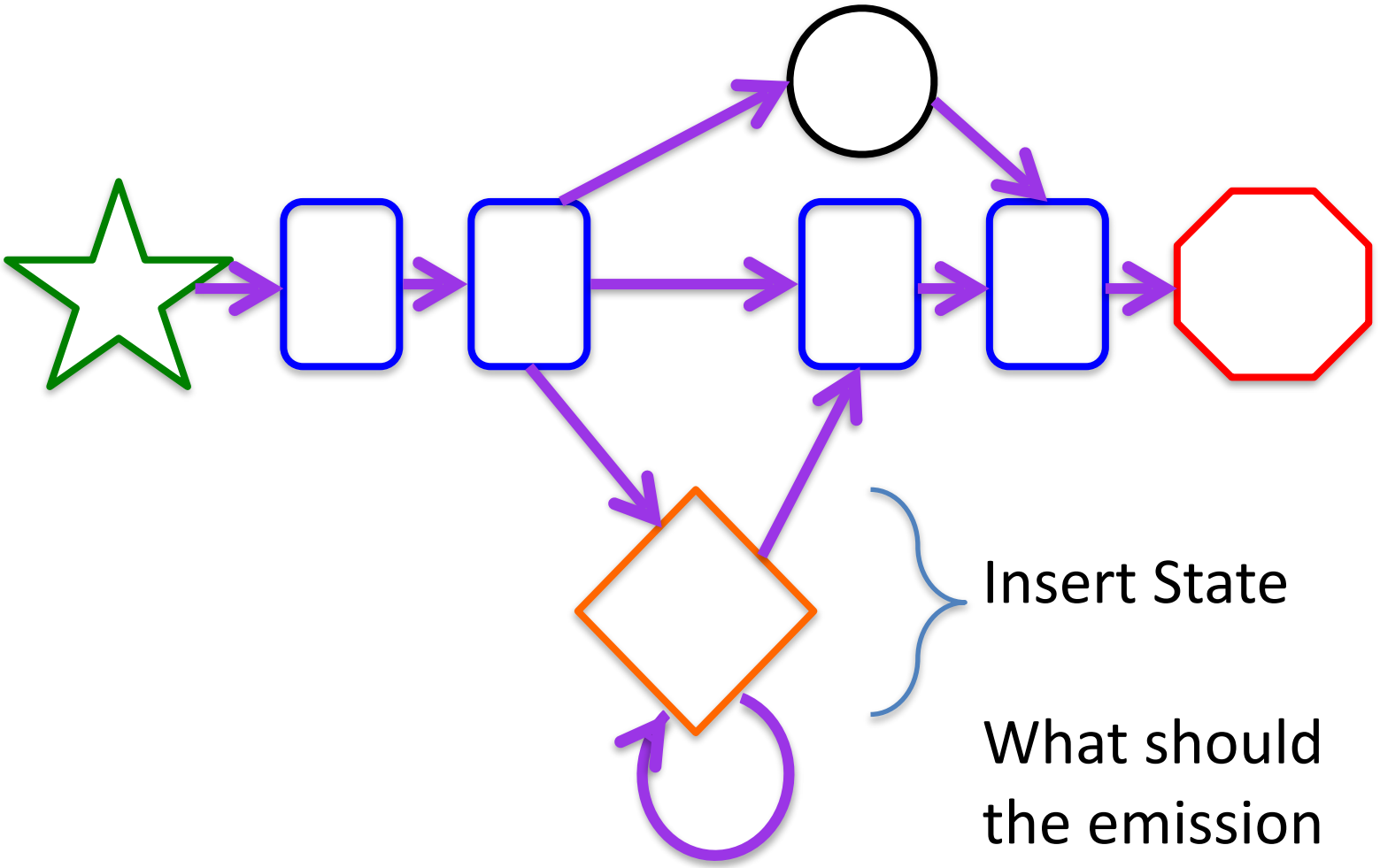
pHMMs: Each alignment column becomes a state



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


Insert State
What should the emission probabilities be?

Classifiers: ARBitrator

- For some genes (?paralogs?), blasting against GenBank isn't accurate enough
- Blast against Conserved Domain database
- Compute superiority
- Compare superiority to a threshold

Classifiers: 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

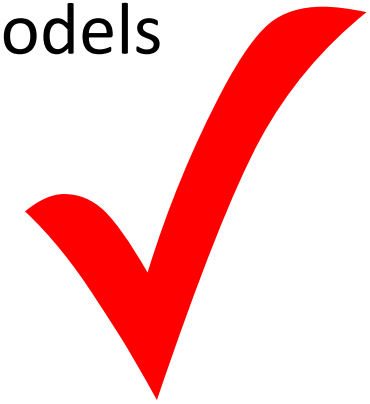
- Your data might be too big to analyze in reasonable time.
- Do an experiment:
  - Use a very small data set (n << actual n).
  - Run the algorithm and measure t (execution time, hopefully reasonable).
  - Now you know t_{little} and n_{little} in the Big-O formula. Solve for k.
- Compute t for the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

Classifiers: Terminology

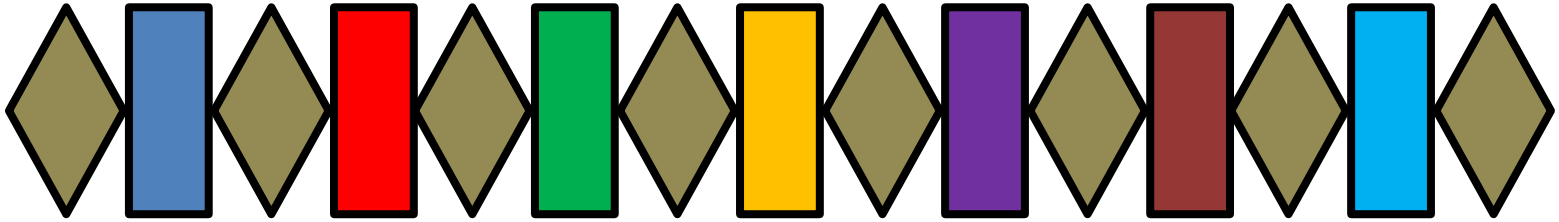
- True positive
- True negative
- False positive
- False negative
- Sensitive
- Specific

The Big Ideas

- Technology
  - Amplification
  - Sequencing
- Classifiers
  - Hidden Markov Models
  - ARBitrator
  - Big-Oh
  - Terminology
- CRISPR
- Statistics
- Metagenomics



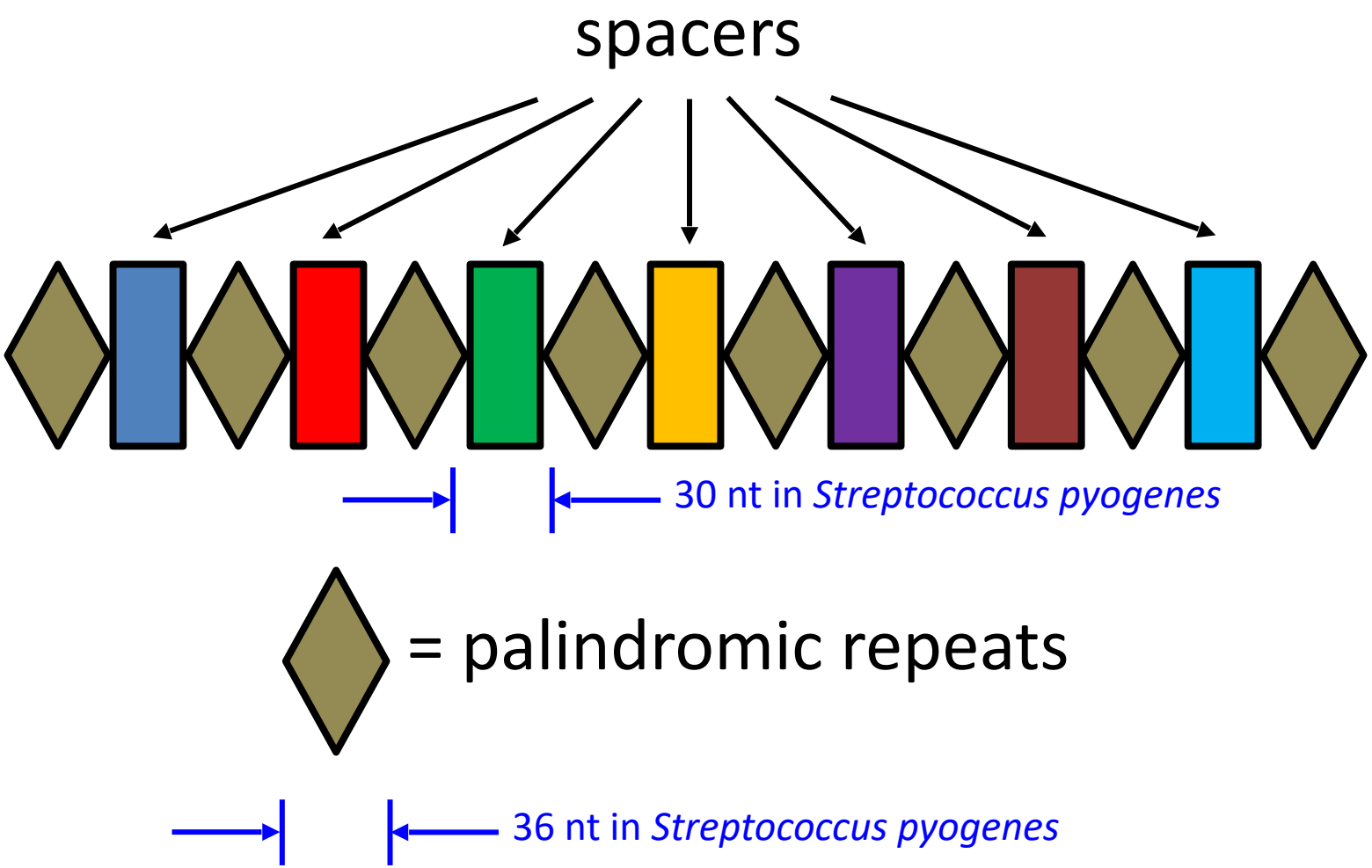
CRISPR: An ancient immune system drives new biotech



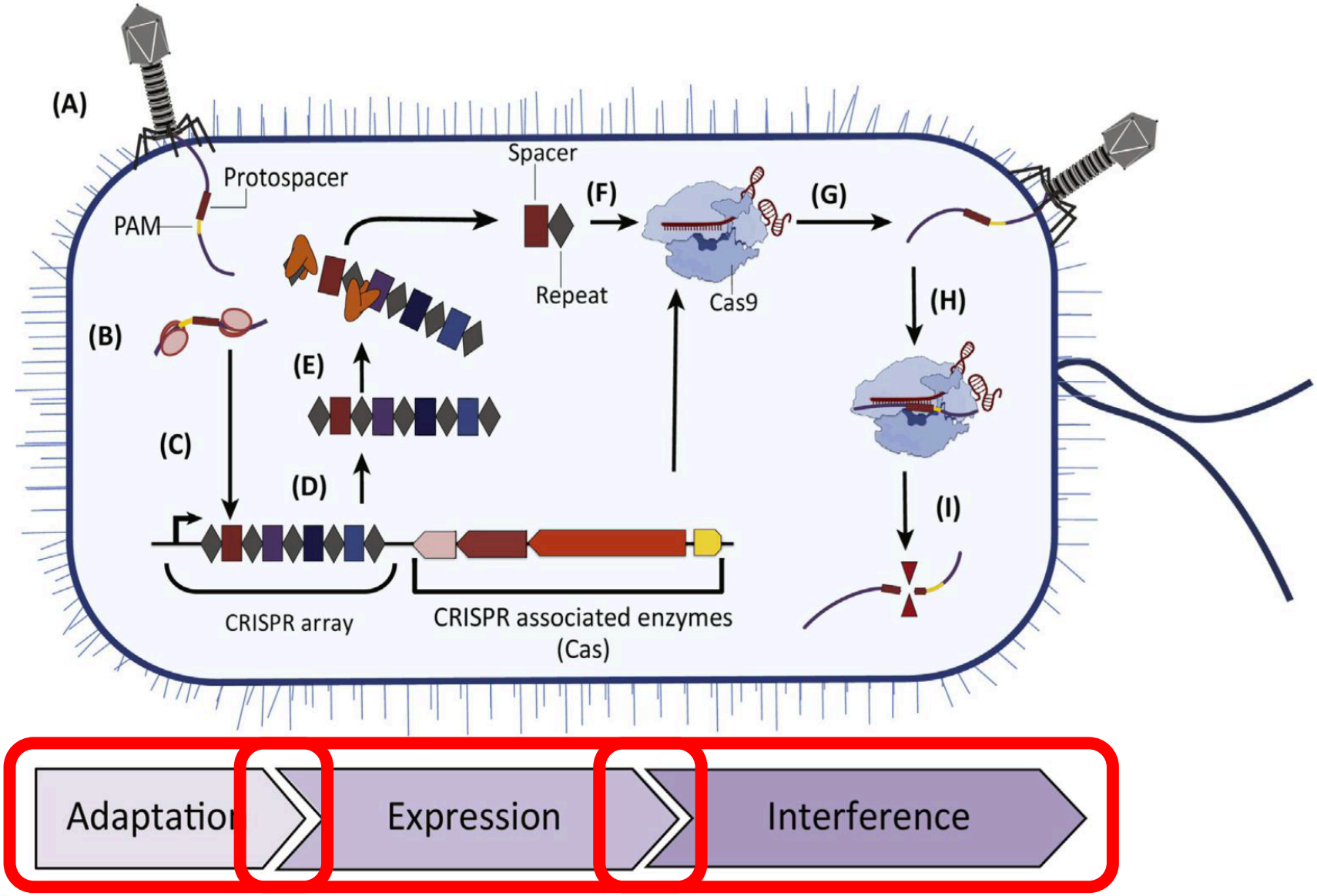
Much material adapted from

"Advanced Bioinformatics for Biotechnology"

Structure of a prokaryotic CRISPR

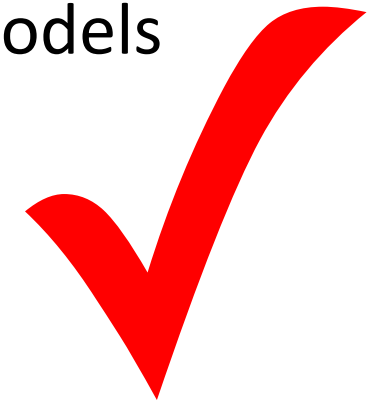


The CRISPR-Cas System: 3 Stages



The Big Ideas

- Technology
  - Amplification
  - Sequencing
- Classifiers
  - Hidden Markov Models
  - ARBitrator
  - Big-Oh
  - Terminology
- CRISPR
- Statistics
- Metagenomics



Statistics

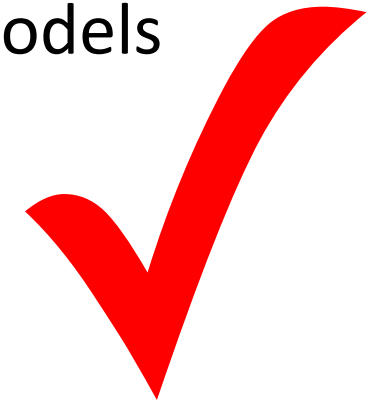
- "Ordinary" probability:
  - Given a hypothesis
  - P(an event | the hypothesis)
- Likelihood:
  - Given observed events
  - P(a hypothesis | the observations)
  - Lets us think about probability of life on Mars
- Null hypothesis H_0
  - "It is known, Khaleesi"

Hypothesis testing

- Null hypothesis H_0
  - "It is known, Khaleesi"
- Analyze observed data, try to reject H_0
- P-value: probability of mistakenly rejecting H_0
- E-value
  - Like p-value for interpreting blast hits
  - probability that the similarity between the query and the subject is due to coincidence, rather than evolutionary relationship.
    - This is a description, not a definition

The Big Ideas

- Technology
  - Amplification
  - Sequencing
- Classifiers
  - Hidden Markov Models
  - ARBitrator
  - Big-Oh
  - Terminology
- CRISPR
- Statistics
- Metagenomics



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")

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

Why we need metagenomics to study invasion

Conservation Biology

Explore this journal >

Decline of a Native Mussel Masked by Sibling Species Invasion

Jonathan B. Geller

First published: June 1999 Full publication history

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 ?

The Big Ideas

- Technology
  - Amplification
  - Sequencing
- Classifiers
  - Hidden Markov Models
  - ARBitrator
  - Big-Oh
  - Terminology
- CRISPR
- Statistics
- Metagenomics