

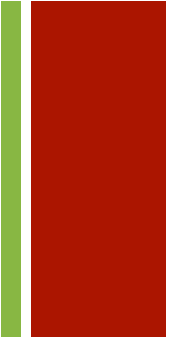


# Computer Science and Biology

Arthi Ramachandran

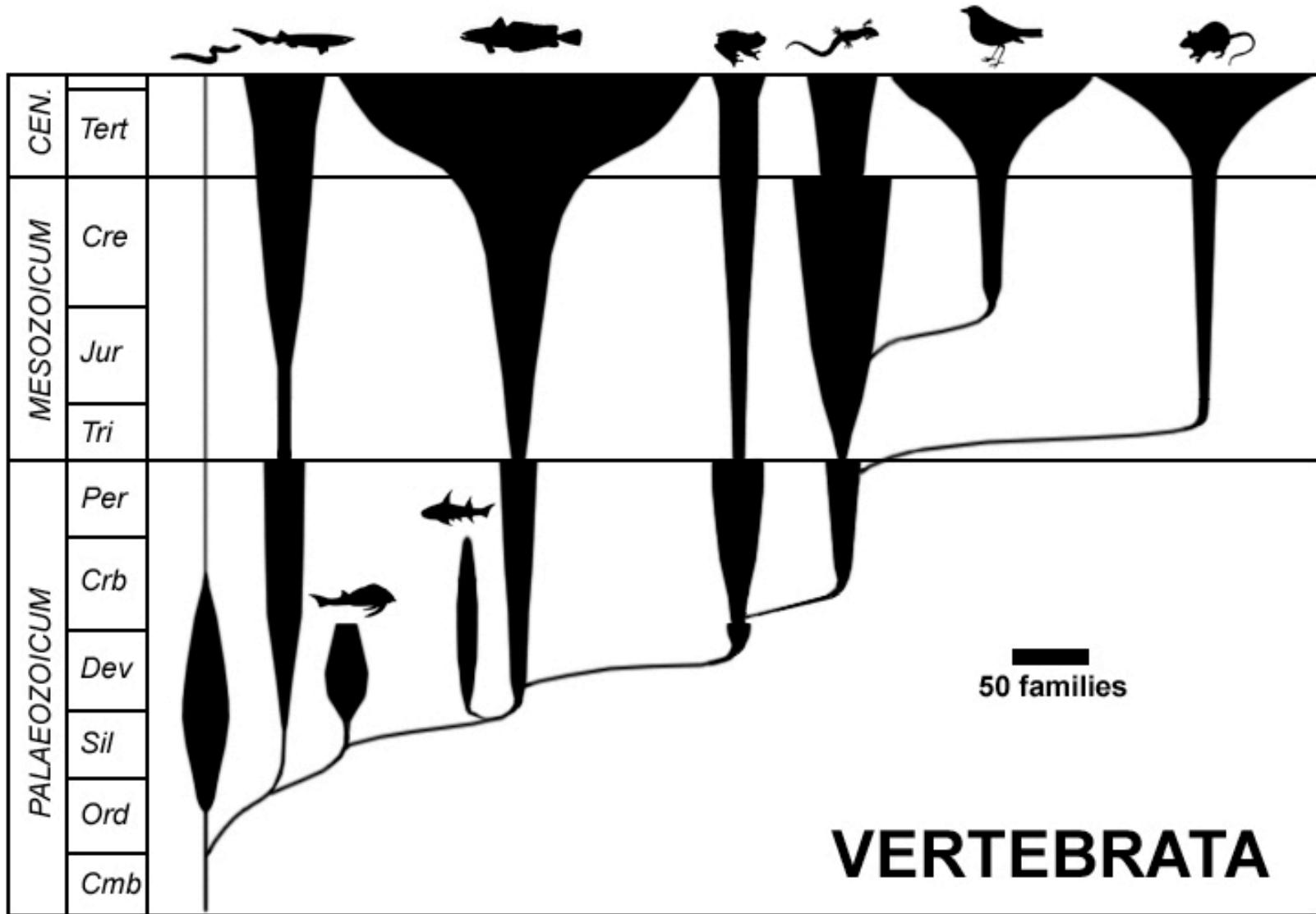
July 7, 2011

# + Outline

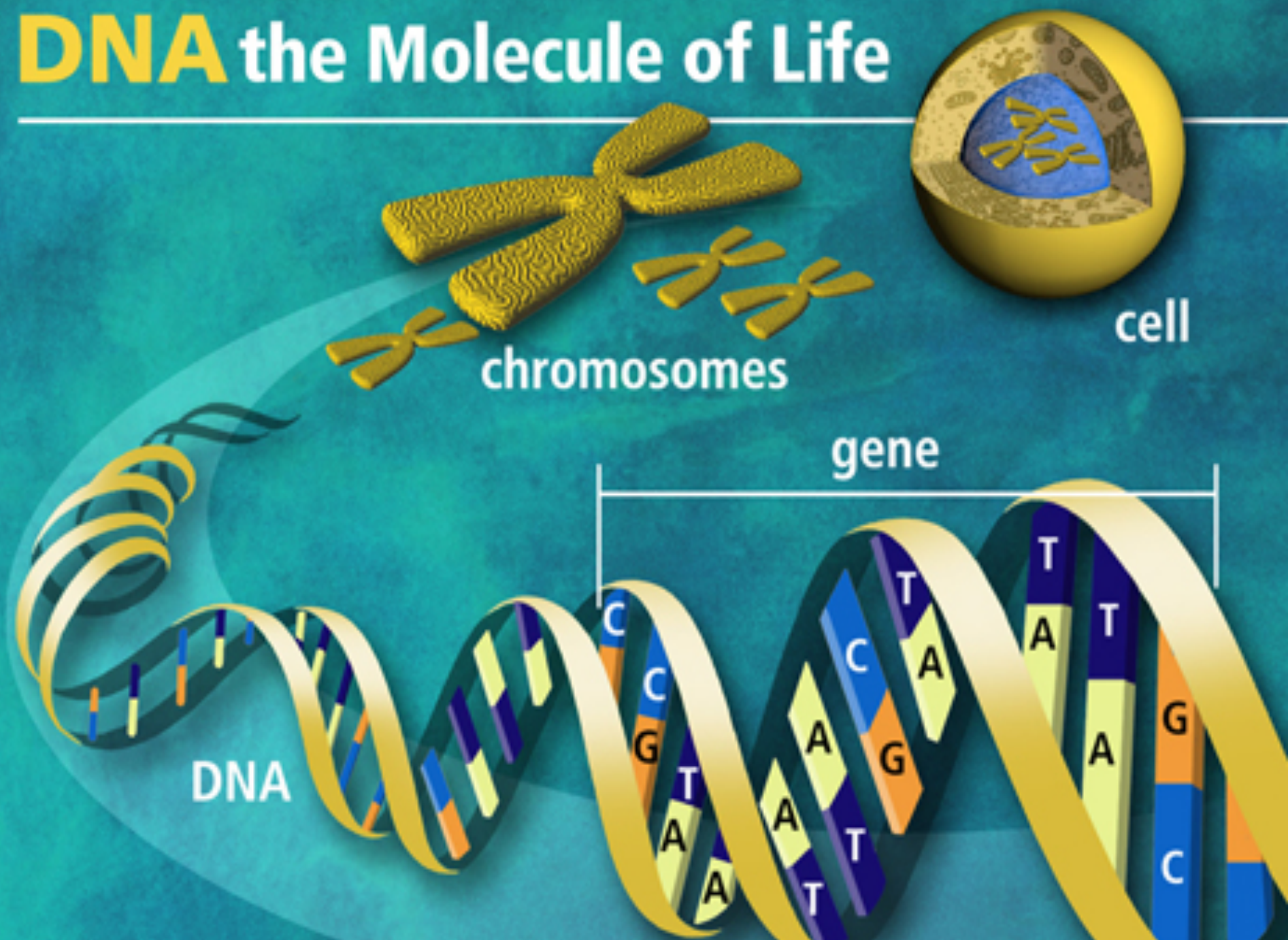


- Phylogeny
- What is DNA?
  - Massive codes, tons of information
- What is Computational Biology?
- CS Unplugged
- Phylogenetics Activity
- Conclusion

# + Phylogeny

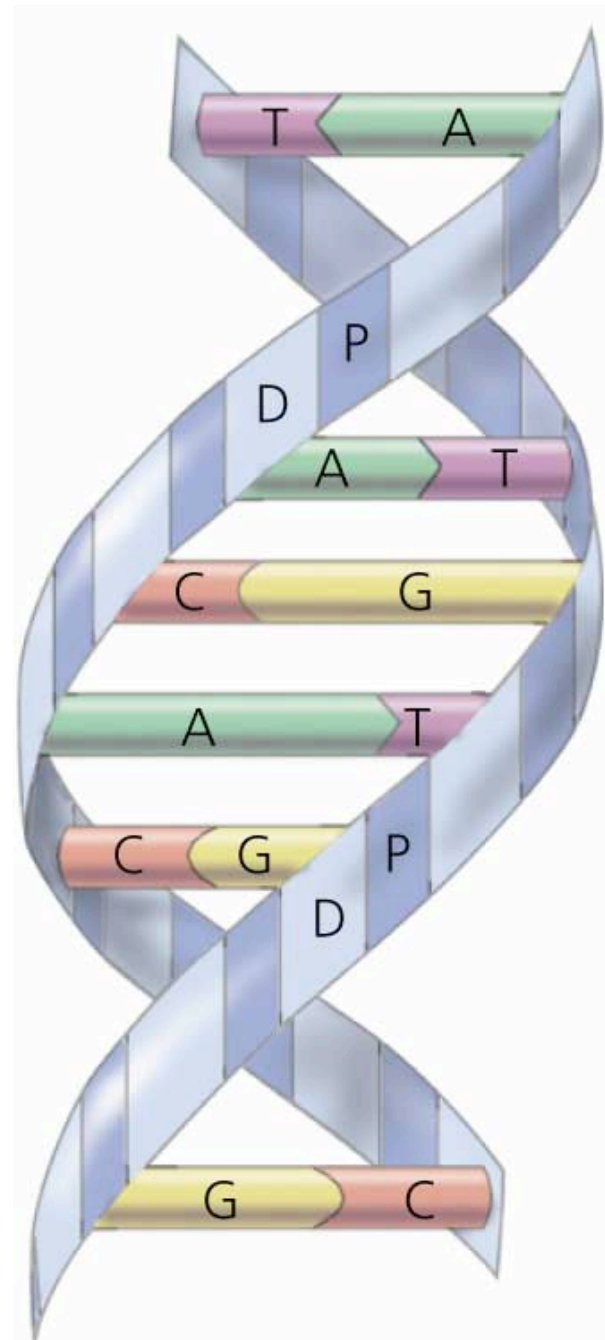


# DNA the Molecule of Life



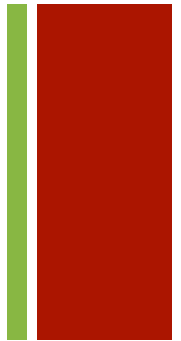
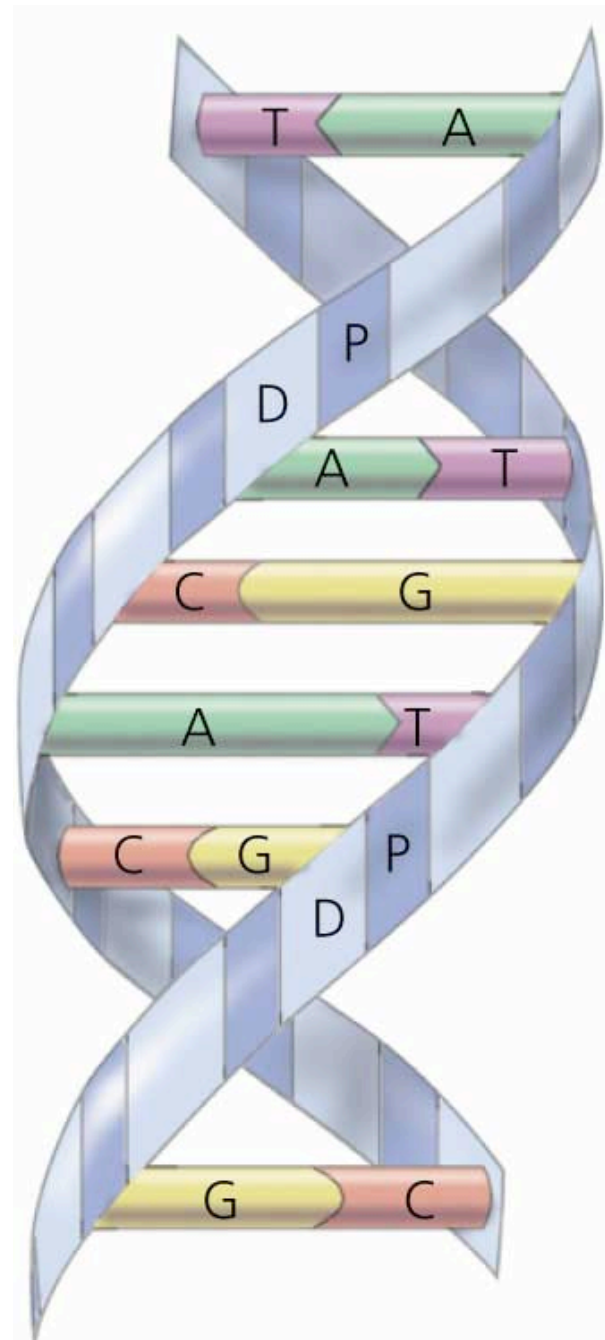
# + What is DNA?

- Double helix polymer
- Backbone
  - *Deoxyribose* (sugar)
  - Phosphate
- Bases
  - Adenine
  - Cytosine
  - Guanine
  - Thiamine



# + What is DNA?

- Instructions for
  - How to build a human
  - Code for who and what we are
- 1 human = 3 *billion* bases
- Humans can't analyze that much information by hand

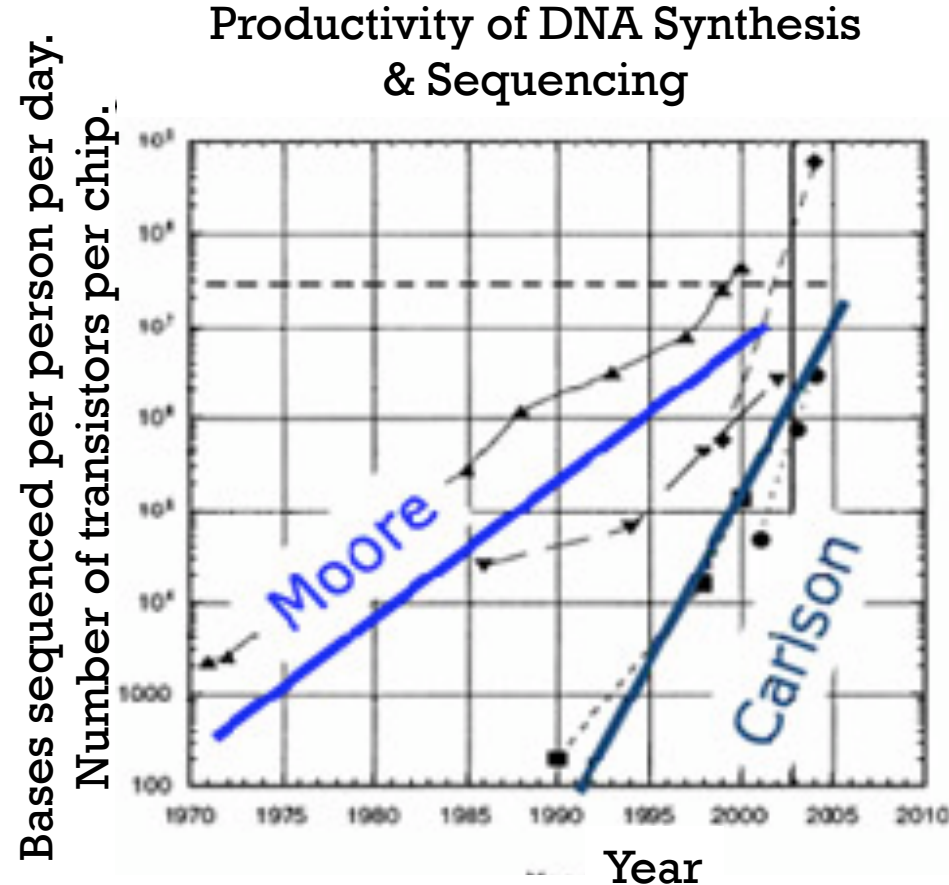




# + How much data is there?



- Moore's law:
  - # transistors on chip doubles in 2 years
- Sequencing
  - 2001: 1<sup>st</sup> human genome
    - Billions of dollars
    - 10 year effort
  - Now:
    - Thousands of dollars
    - Several days
  - Opportunity for research



# + Computational Biology

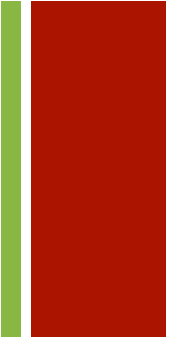


- To analyze DNA, we need to use ideas from
  - Computer science
  - Mathematics
  - Statistics



# + Computational Biology is everywhere!

- Phylogenetics
  - How different species are related
- Genomics
  - DNA sequence analysis
- Medical informatics
  - Data mining medical records
- Computational neuroscience



# + CS Unplugged



- Activities to teach Computer Science at a middle-high school level
  - Without computers
  - No programming
- Examples:
  - Binary arithmetic
  - Artificial Intelligence
  - Cryptography
  - Graph Theory

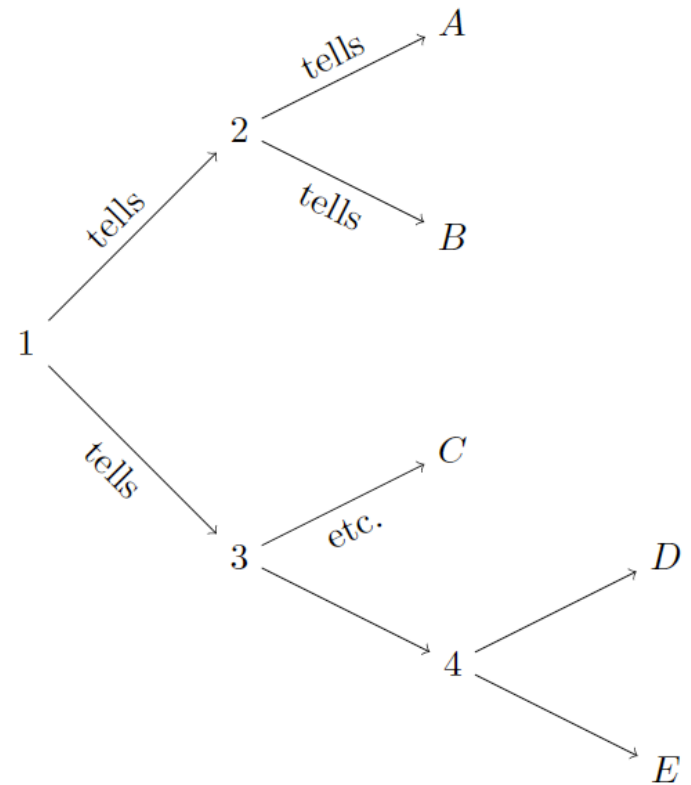
# + CS Unplugged: Computational Biology

- Biology: evolution
- Biology: phylogenetics
- Mathematics: matrices
  - Only High School level math needed



# + Activity

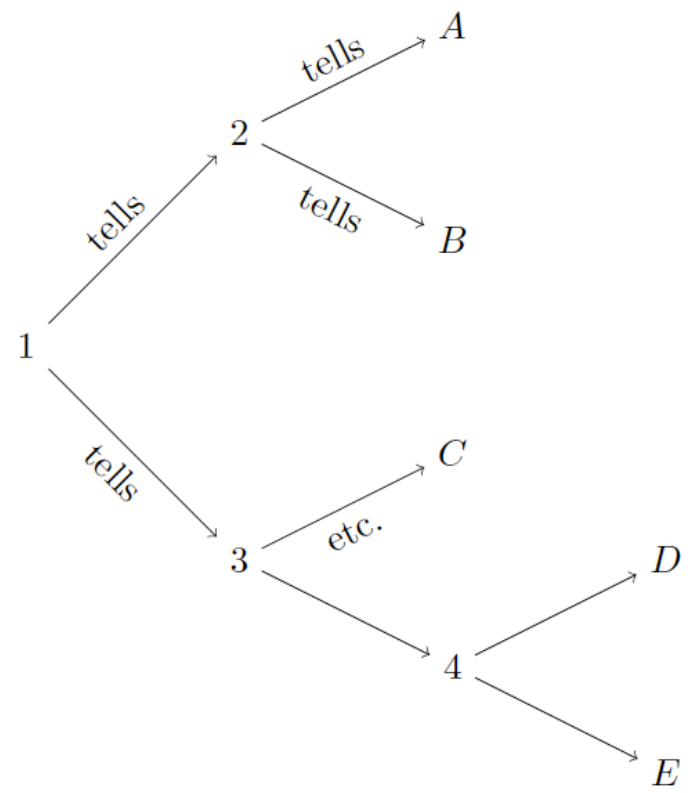
- Play modified version of the children's game "telephone"
- From final set of messages {A, B, C, D, E}, can we reconstruct who told whom and when?



# + The Biology behind it



- Messages ~ DNA sequences
- Passing of a message ~ passing DNA from one generation to the next
- Changes in message ~ mutations

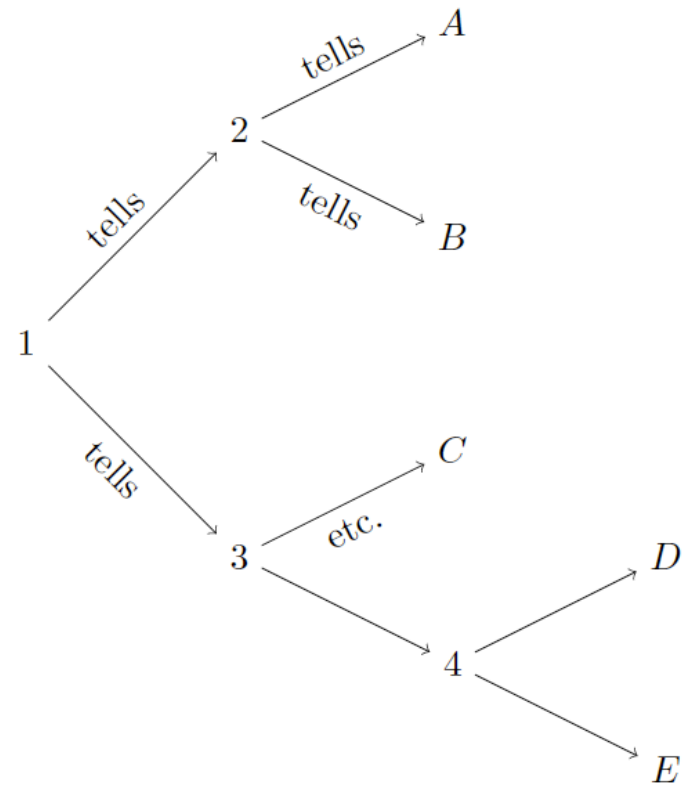


# + The Biology behind it



## ■ Phylogenetics

- From the end DNA sequences (current living organisms) we want to construct the evolutionary tree



# + Activity

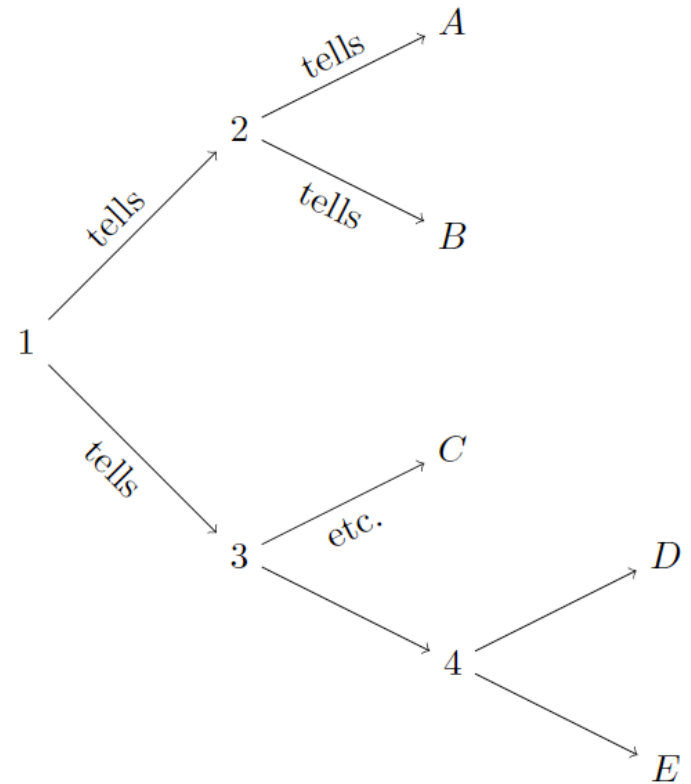


- Set of similar sounding words:
  - {chair, prepare, affair, stair}
  - {school, rule, cool, fool}
  - {dream, stream, theme, beam}



# + Activity

- Construct a message with 5-7 of those words
- Student 1 passes it on to Student 2, 3 and so on



# + Activity Rules

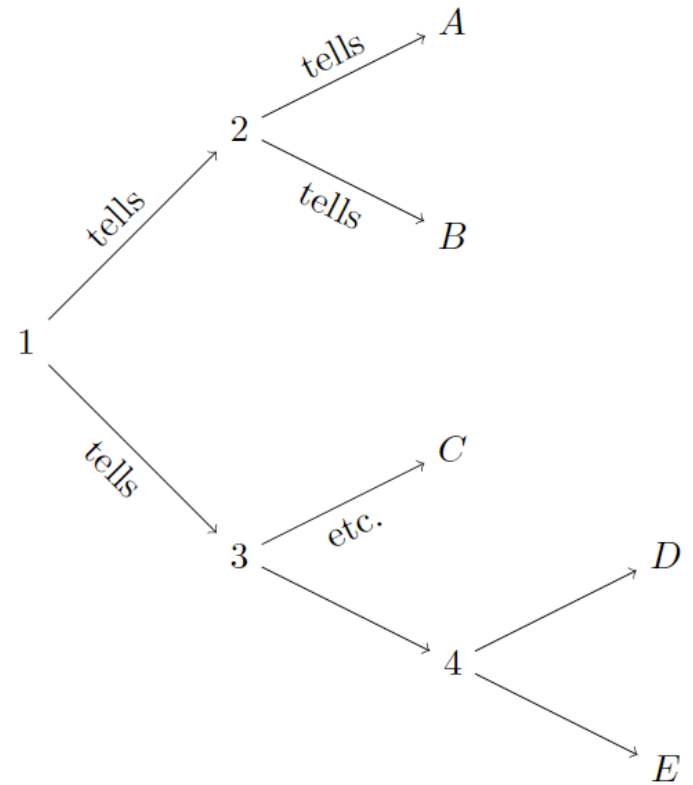


- *Whisper* the message you hear to the next person
- No repeats allowed if you didn't hear the message properly
- Write what you heard down on the given paper

# + Activity



- A, B, C, D, E are the final messages
- Relabel them V, W, X, Y, Z



# + Activity



- V:
- W:
- X:
- Y:
- Z:
- How would you construct the tree structure?

# + Activity



## ■ Edit distance

- How similar are two of the messages?

	<b>V</b>	<b>W</b>	<b>X</b>	<b>Y</b>
<b>W</b>		-	-	-
<b>X</b>			-	-
<b>Y</b>				-
<b>Z</b>				

# + The Biology behind it



- Maximal Parsimony

- Occam's Razor

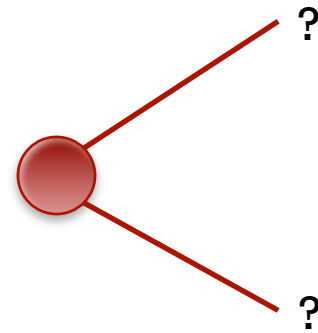
- *Tend towards simpler theories until we can trade some simplicity for increased explanatory power*
    - *the simplest explanation is most likely the correct one*

- Fewest mutations possible in the tree

# + Activity



- Reconstructing the tree
  - Maximal parsimony
- Smallest entry in matrix
  - Group those messages together





# + Activity

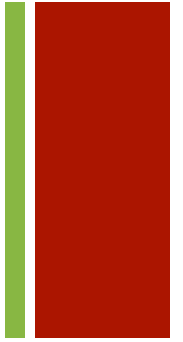


- Revise distance matrix
- Repeat process

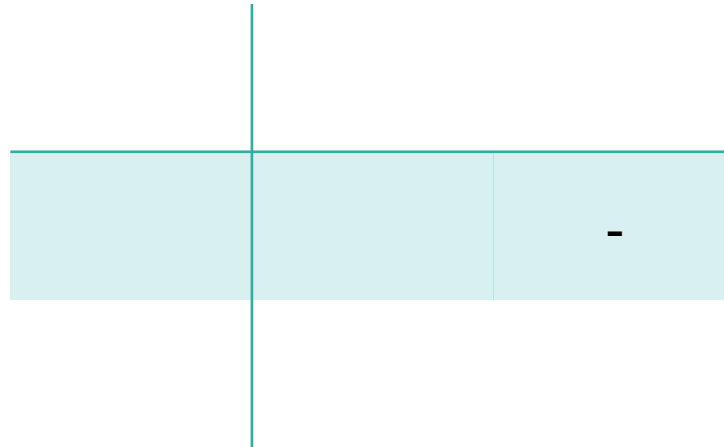
A diagram illustrating a distance matrix. It consists of a light blue rectangular area divided into four quadrants by a vertical and a horizontal line. The top-right and bottom-right quadrants contain a small black dash (-), representing the upper triangle of the matrix. The bottom-left quadrant is empty, representing the lower triangle.

	-	-
		-

# + Activity



- Revise distance matrix



# + Activity



- Is the final tree the same as the original tree?
- If not, why?
  - Random selection to break ties
  - Neutral changes

# + The Math behind it



- Matrix theory
  - Species distance matrix
    - Pairwise distances

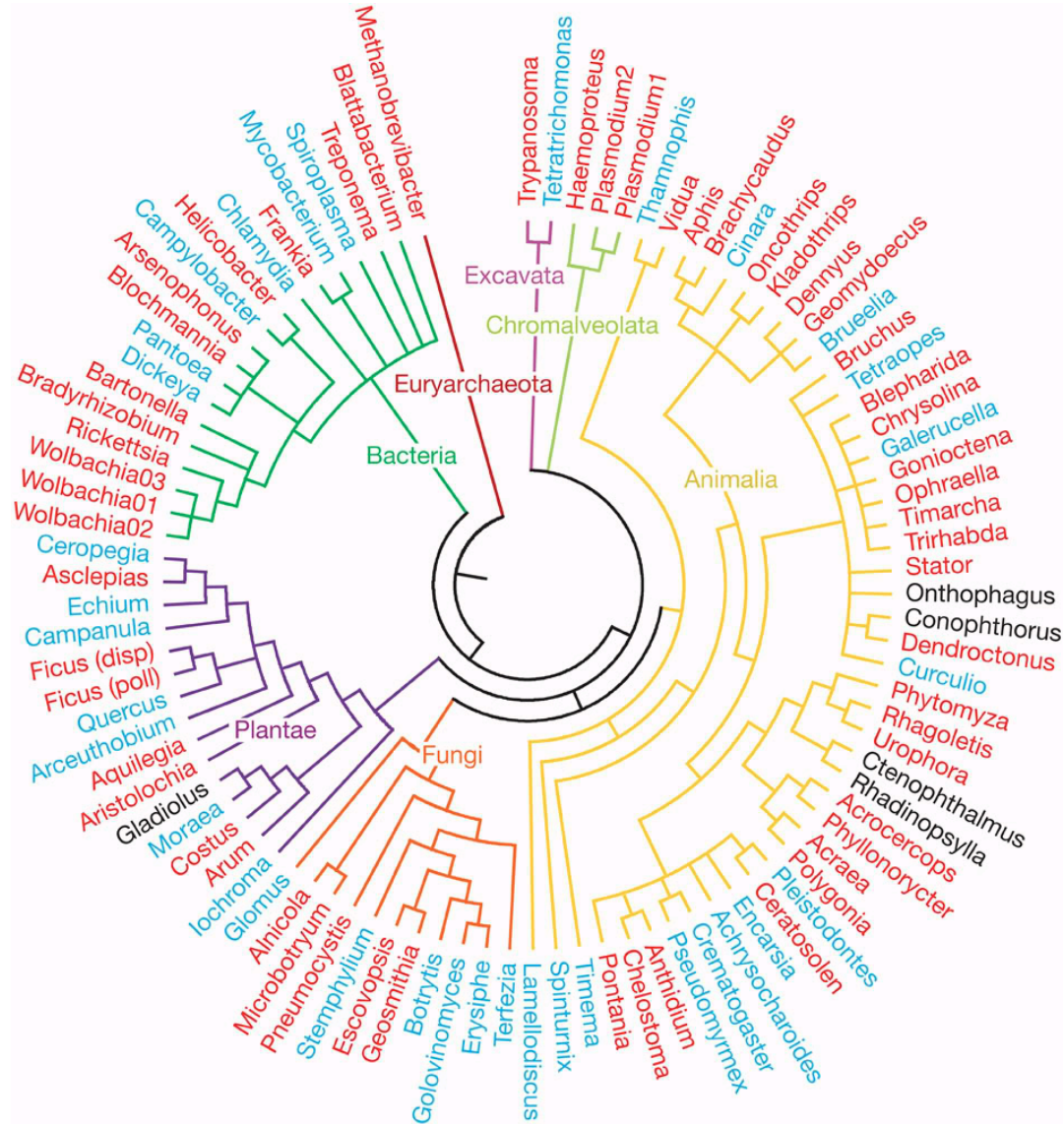
# + The Computer Science behind it



- Easy to solve with 5 messages of short length
  - What if there are hundreds of messages?
  - What if each message is hundreds of words?  
Thousands? Millions?



# Evolutionary tree example



# + Summary



- Just one example of what can be done using computer science in the biological sphere
- Computer Science is used everywhere!



# + Resources



- CS Unplugged
  - <http://csunplugged.org/activities>

Any Questions?