PROTEIN JURASSIC PARK*

Jeffrey Boucher

***Less than 10% Dinosaur content**

Talk Outline

- Talk 1:
  - “How to Raise the Dead: The Nuts & Bolts of Ancestral Sequence Reconstruction”
- Talk 2:
  - Ancestral Sequence Reconstruction Lab
- Talk 3:
  - “Ancestral Sequence Reconstruction: What is it Good for?”

How to Raise the Dead: The Nuts and Bolts of Ancestral Sequence Reconstruction

Jeffrey Boucher

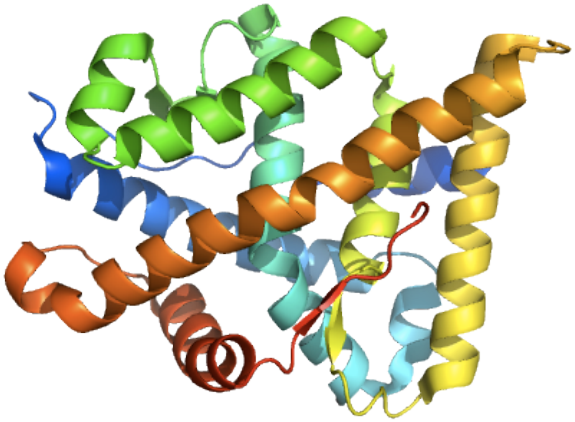
Theobald Laboratory

Orientation for the Talk

- The Central Dogma:

DNA

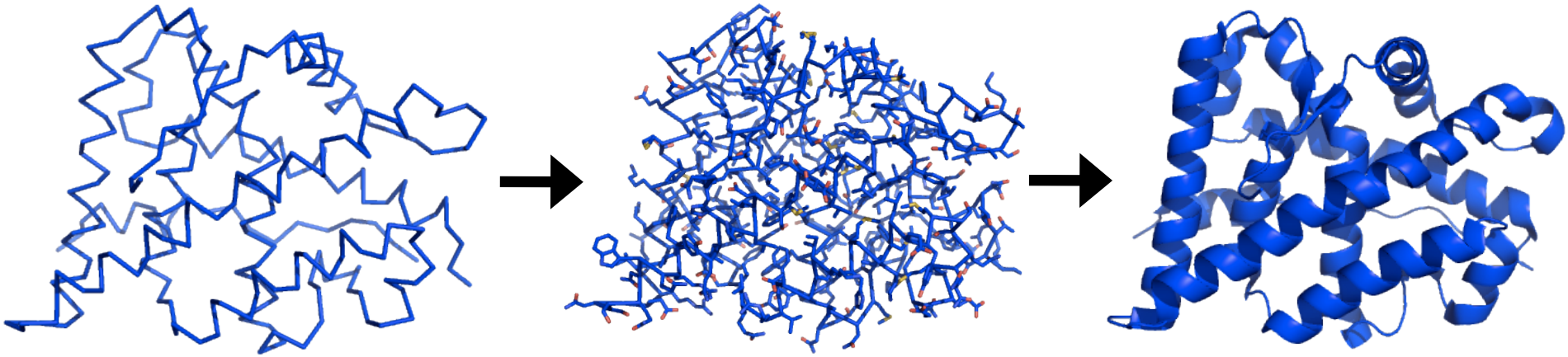
RNA



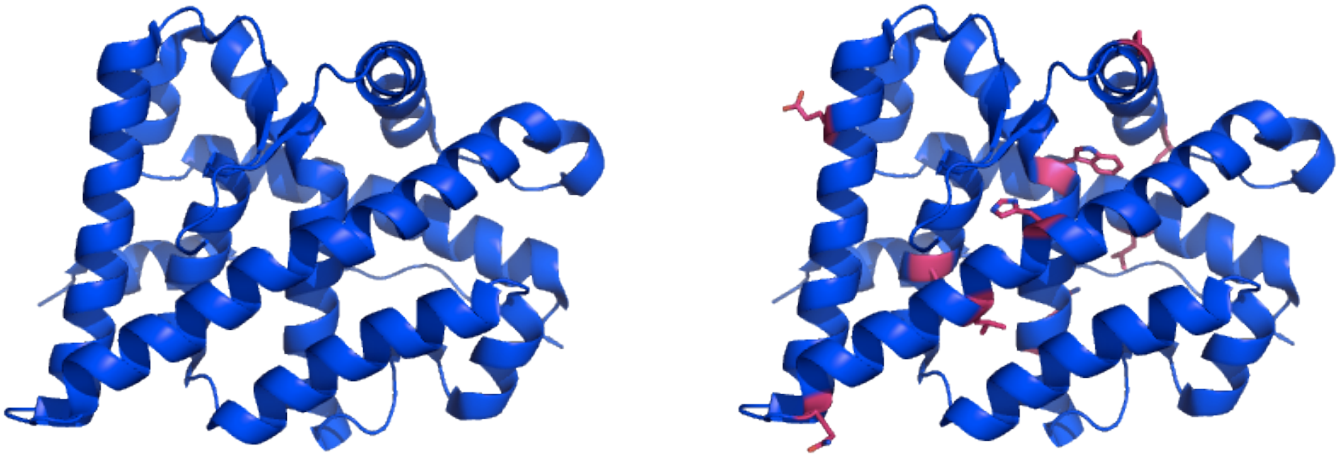
Protein

Orientation for the Talk (cont.)

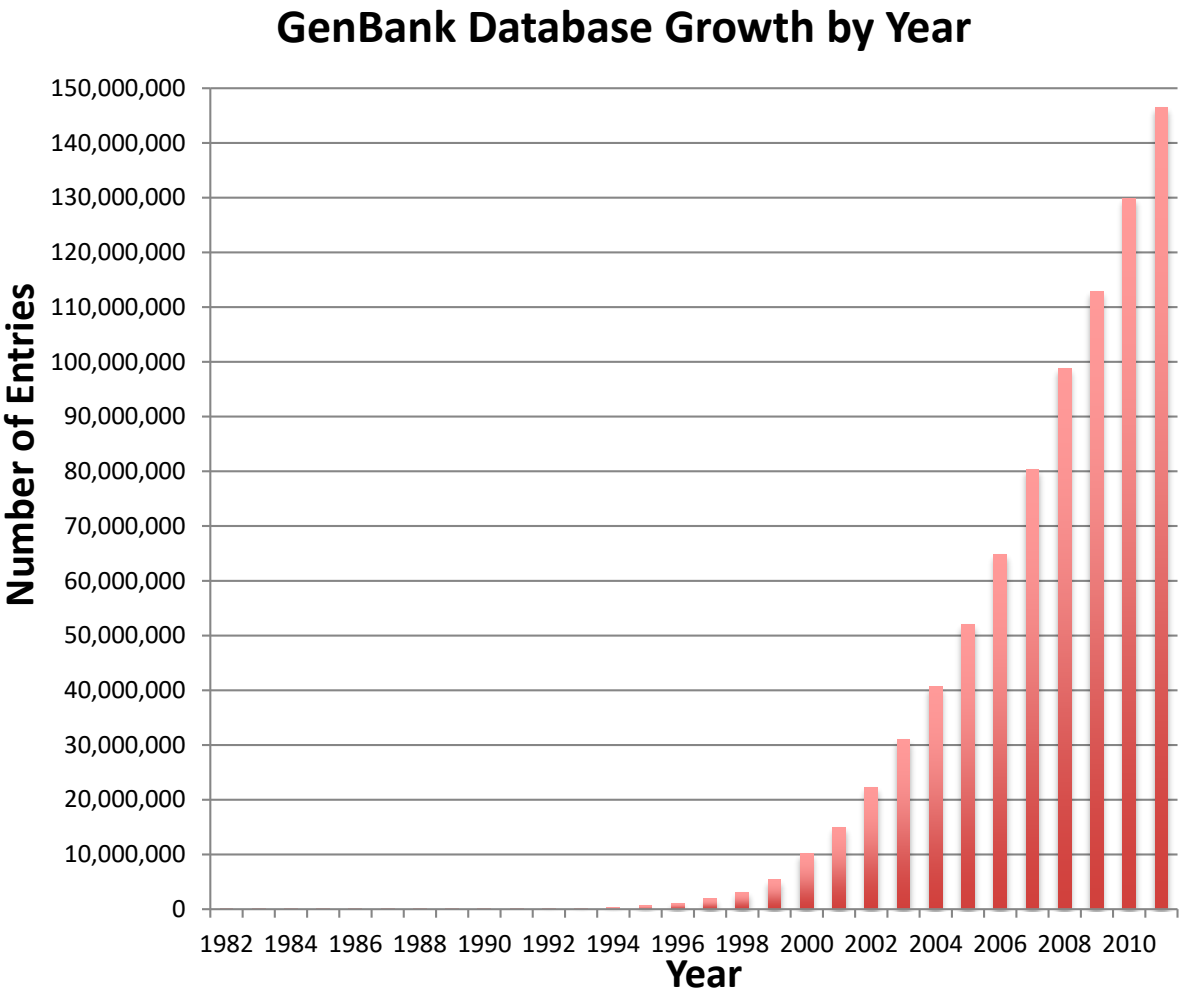
- Chemistry of side chains govern structure/function



- Mutations to sequences occur over time



We Live in The Sequencing Era



Since inception, database size has doubled every 18 months.

http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html

What Can We Learn From This Data?

- Individually…not much

  >gi|93209601|gb|ABF00156.1| pancreatic ribonuclease precursor subtype Na [Nasalis larvatus]
  MALDKSVILLPLLVVVLLVLGWAQPSLGRESRAEKFQRQHMDSGSSPSSSSTYCNQMMK
  RRNMTQGRCKPVNTFVHEPLVDVQNVCFQEKVTCKNGQTNCFKSNSRMHITDCRLTNG
  SKYPNCAYRTTPKERHIIVACEGSPYVPVHFDASVEDST

- Too many sequences to characterize individually
  - Today:

    1.5 E 8 sequences ÷ 7 E 9 people = 1 sequence/50 people
  - By 2019

    1.2 E 9 sequences ÷ 7.5 E 9 people = 1 sequence/6 people

Bioinformatics!

- Bioinformatic methods developed to deal with this backlog

- Methods covered:
  - Sequence Alignment (& BLAST)
  - Phylogenetics
  - Sequence Reconstruction

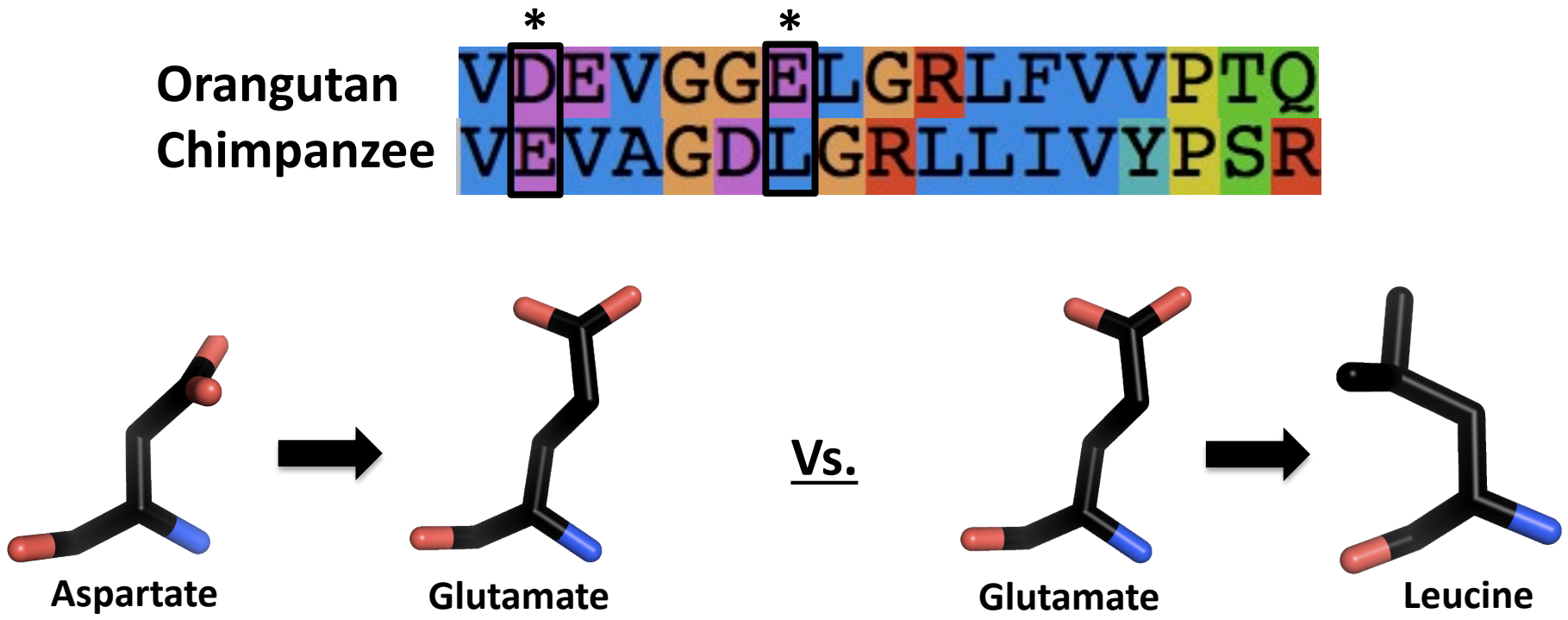
Sequence Alignment

- How can we compare sequences?

Orangutan VDEVGGELGRLFVVPTQ
Chimpanzee VEVAGDLGRLLIVYPSR

1 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0 = 5

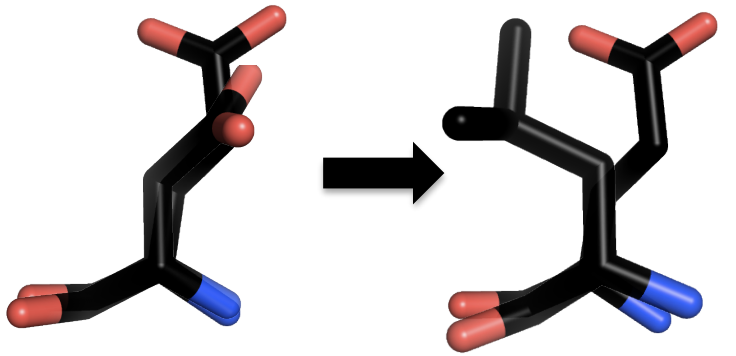
Not All Mismatches Are Created Equal



- How can scoring function account for this?

Substitution Matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4



Calculating A Substitution Matrix

- How are the rewards/penalties determined?

- Determined by log-odds scores:

$$S_{i,j} = \log \frac{p_{i,j}}{q_i * q_j}$$

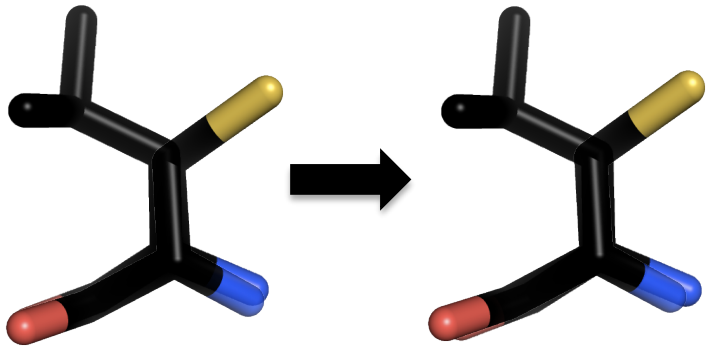
← **Why not just $p_{i,j}$?**

$p_{i,j}$ is probability amino acid i transforms to amino acid j

q_i & q_j represent the frequencies of those amino acids

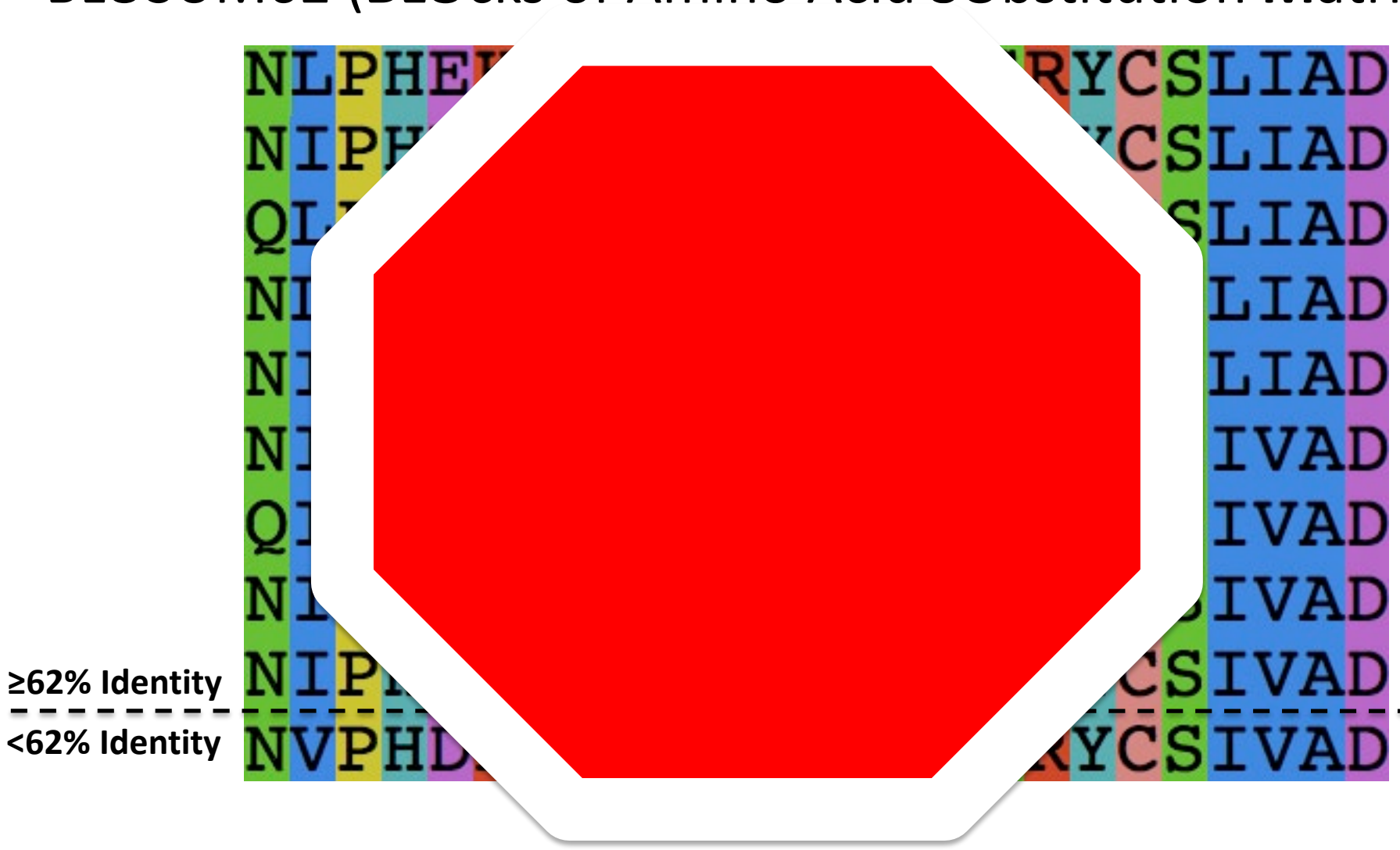
Neither Are All Matches

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4



Cysteine Cysteine

BLOSUM62 (**BLO**cks of Amino Acid **SU**bstitution **M**atrix)

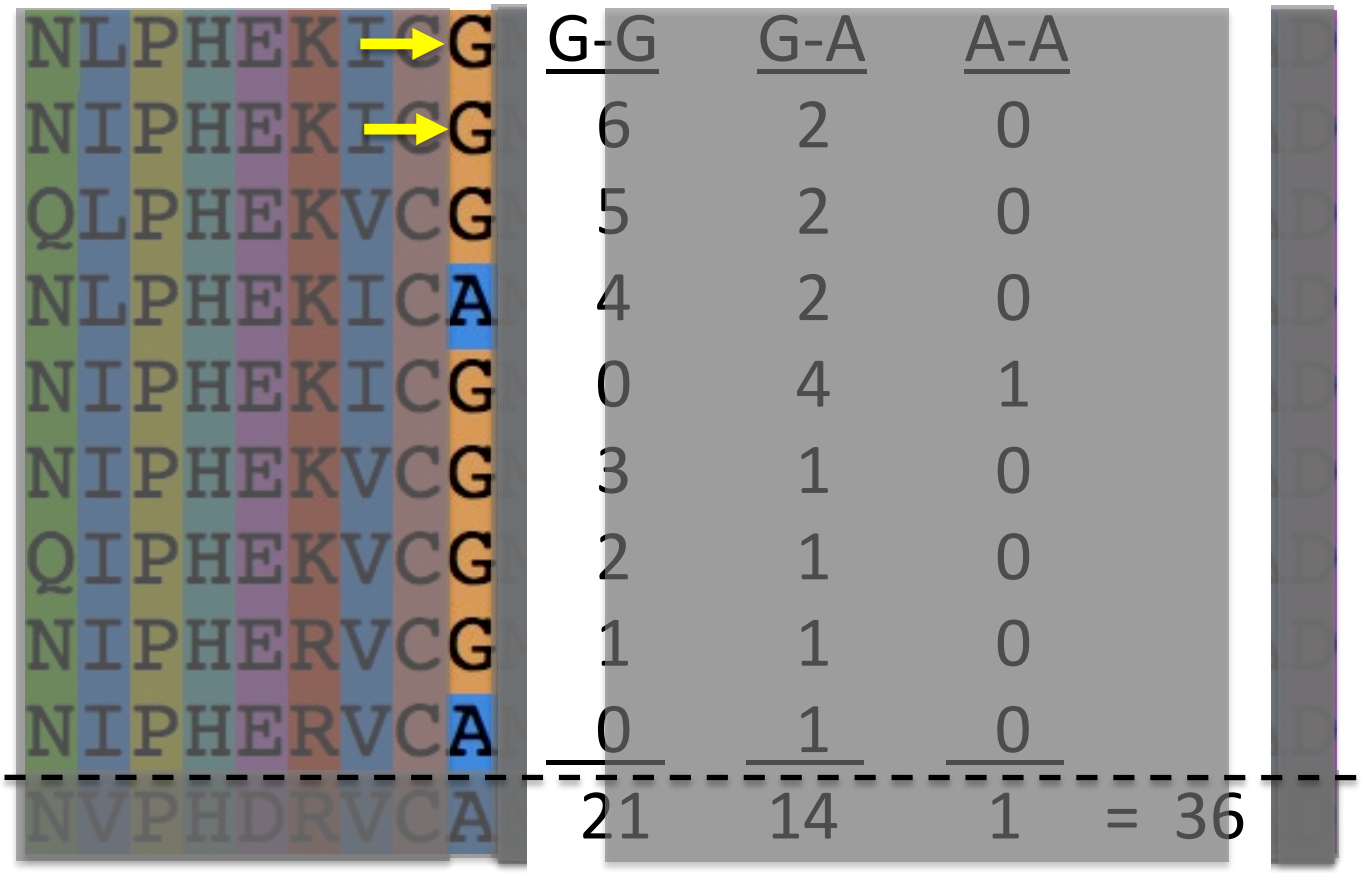


How did you get an alignment?

Blocks used align well with 1/0 scoring function

You're talking about 'How to Make an Alignment'!

BLOSUM62 Matrix Calculation



G-G	G-A	A-A	
6	2	0	
5	2	0	
4	2	0	
0	4	1	
3	1	0	
2	1	0	
1	1	0	
0	1	0	
21	14	1	= 36

≥62% Identity

<62% Identity

$$S_{i,j} = \log \frac{p_{i,j}}{q_i * q_j}$$

$p_{G,A}$ = 14/900 = 0.016

q_G = 7 + 9 = 16/225 = 0.071

q_A = 2 + 9 + 9 = 21/225 = 0.093

Pairwise Alignment Examples

- No Gaps allowed:

- Gap Penalty of -8:

  - Penalty heuristically determined

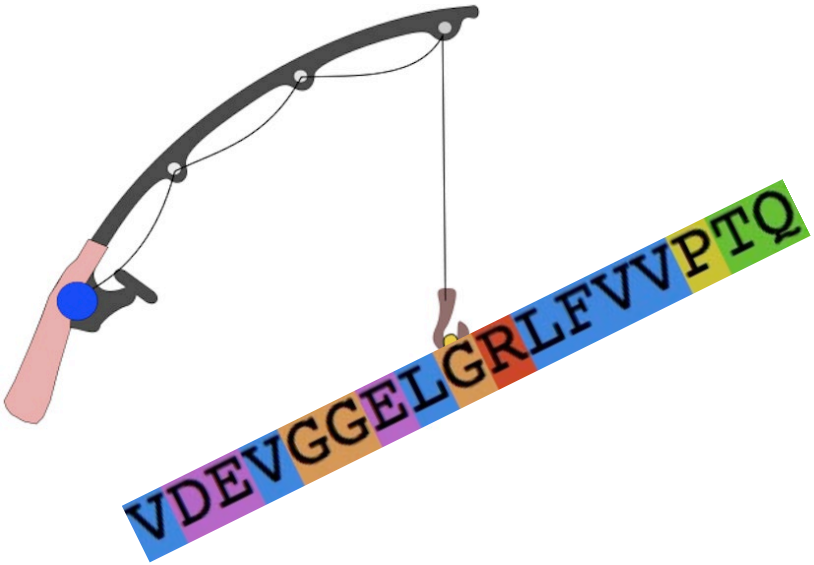
Pairwise Alignment Examples (cont.)

- If gap penalty is too low…

- Alignment of multiple sequences similar method

(& BLAST)

- Alignment can identify similar sequences
- BLAST (Basic Local Alignment Search Tool)



- How does alignment compare to alignment of random sequences?
  - E-value of 1_E-3 is a 1:1000 chance of alignment of random sequences

Homology vs. Identity

- Significant BLAST hits inform us about evolutionary relationships

- Homologous - share a common ancestor
  - This is binary, not a percentile
  - Identity is calculated, homology is a hypothesis
  - Homology does not ensure common function

Visual Depiction of Alignment Scores

- Suppose alignment of 3 sequences…

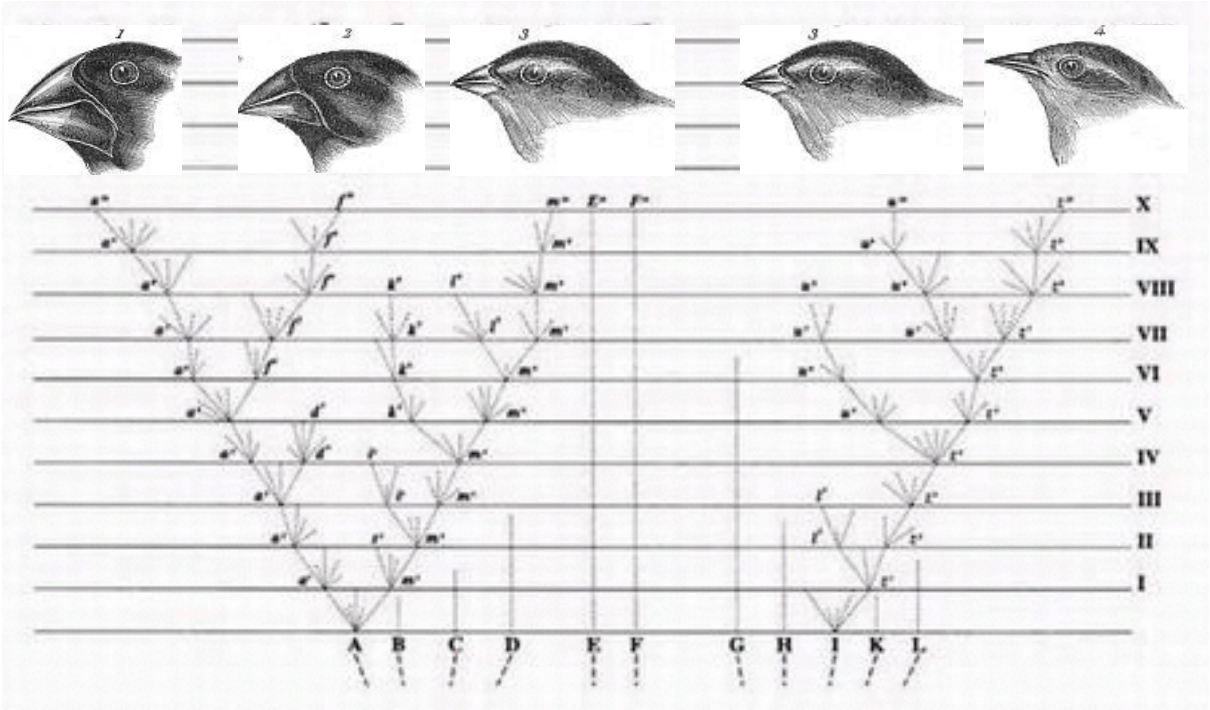
Orangutan
Chimpanzee
Mouse

O	19	40	-
C			
M	18	-	40
	M	C	O
	-	18	10

Phylogenetics

- Relationships between organisms/sequences

- *On the Origin of Species* (1859) had 1 figure:



Phylogenetics

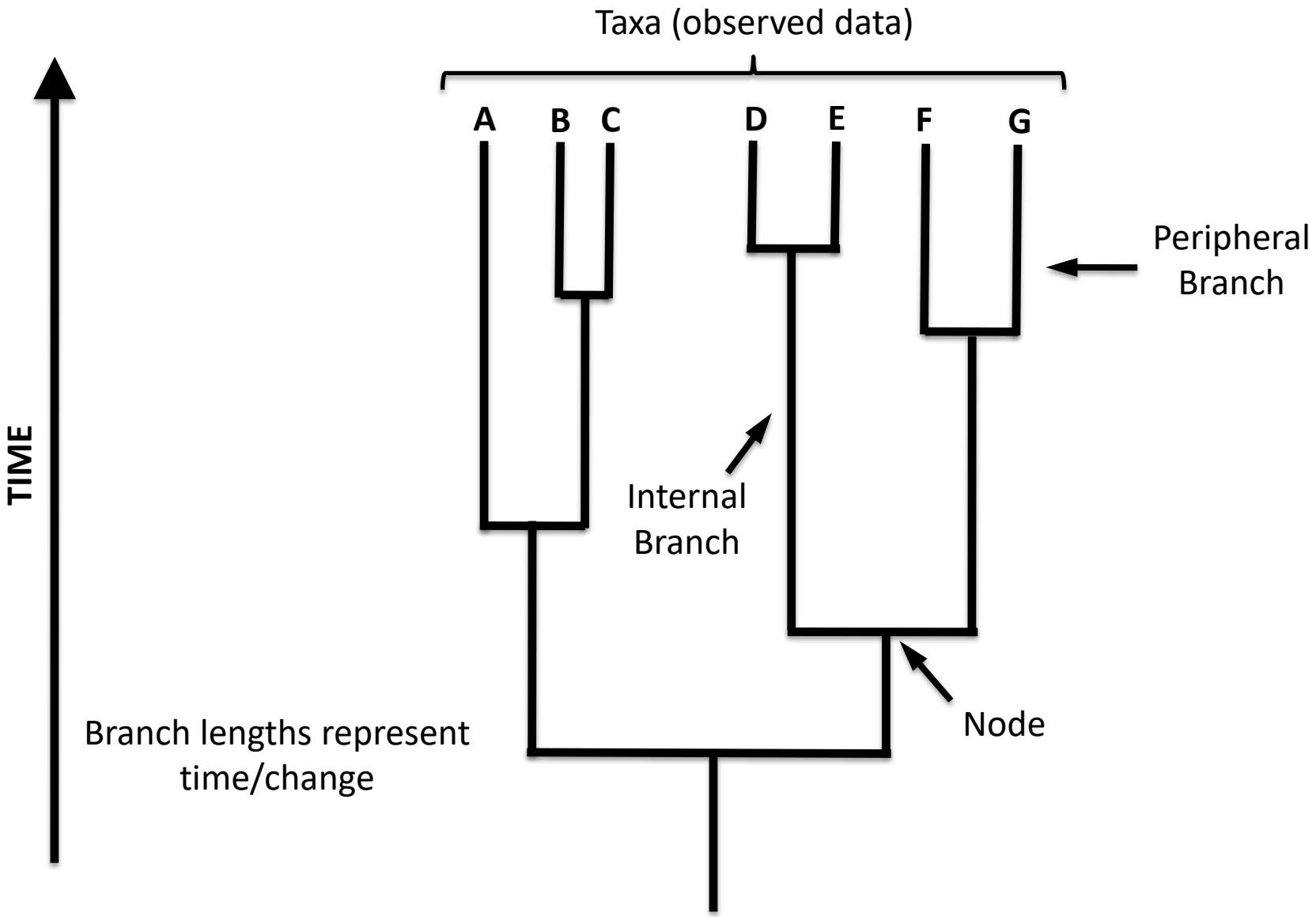
- Prior to 1950s phylogenies based on morphology





- Sequence data/Analytical methods
  - Qualitative → Quantitative

Phylogeny


Taxa (observed data)
A
B
C
D
E
F
G
Peripheral Branch
Internal Branch
Node
TIME
Branch lengths represent time/change

A Tale of Two Proteins

- Significant sequence similarity & the same structure

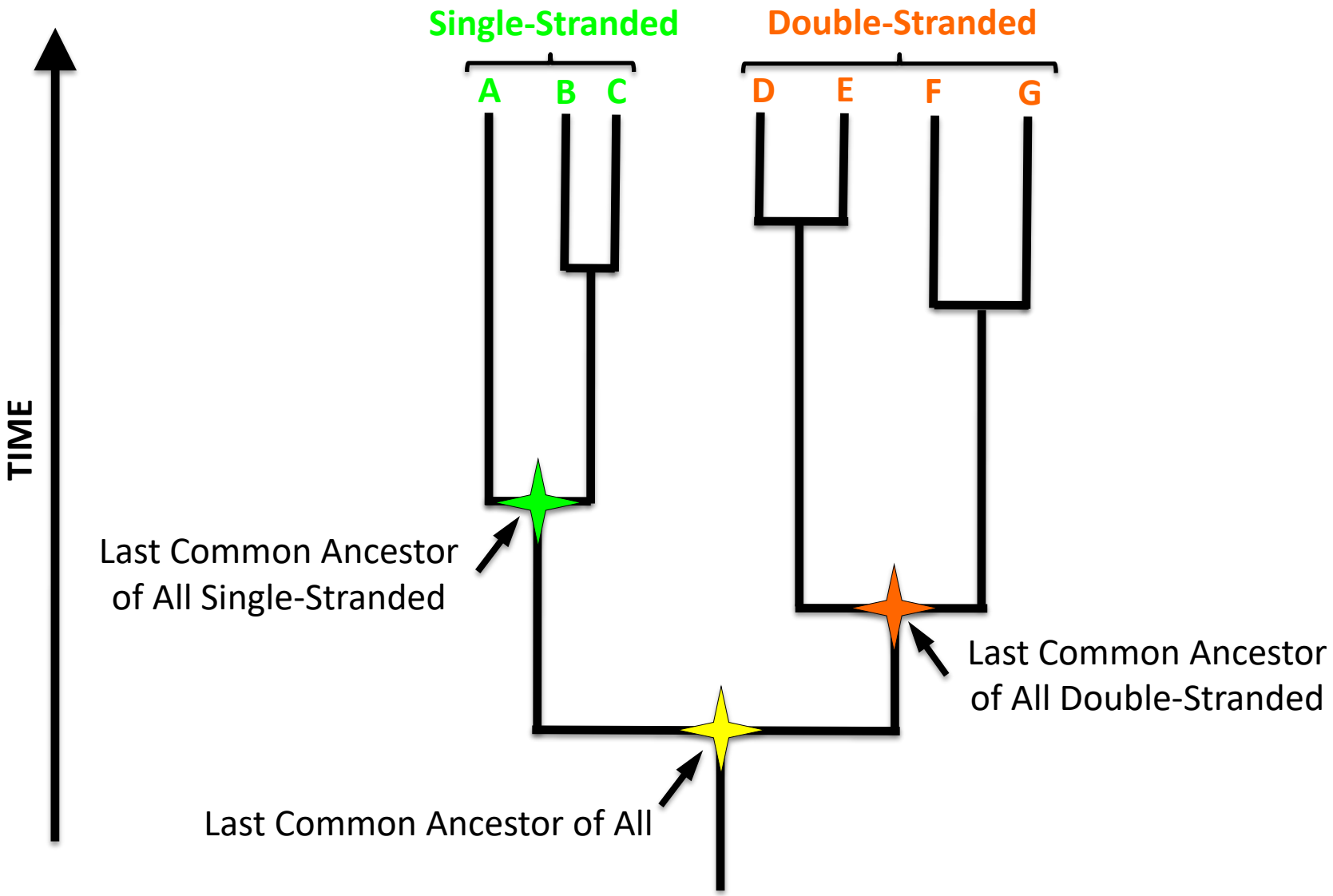
Protein X

-Binds Single Stranded RNA

Protein Y

-Binds Double Stranded RNA

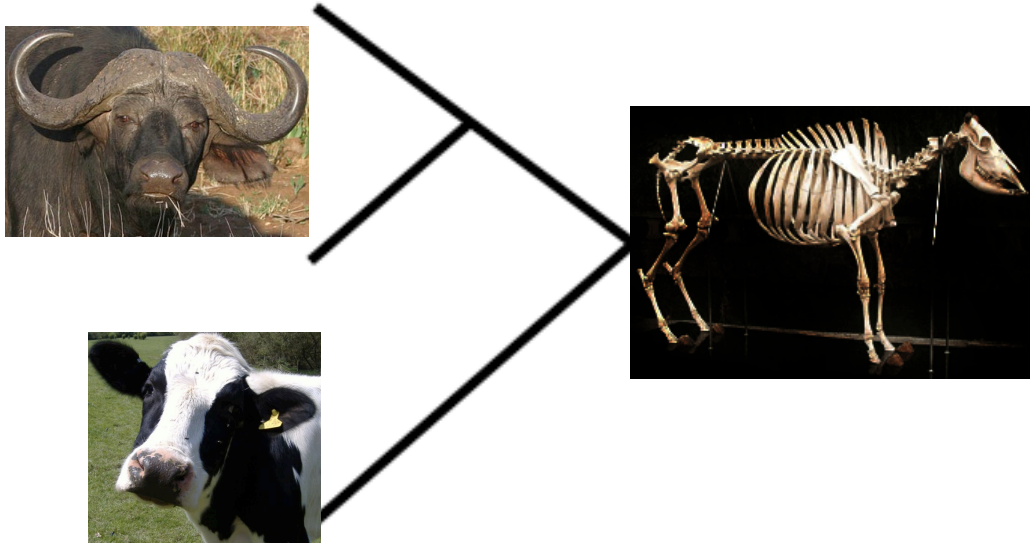
“Gene”alogy


Single-Stranded
Double-Stranded
A
B
C
D
E
F
G
TIME
Last Common Ancestor
of All Single-Stranded
Last Common Ancestor
of All Double-Stranded
Last Common Ancestor of All

Back to the Future

- Resurrecting extinct proteins 1st proposed Pauling & Zuckerkandl in 1963

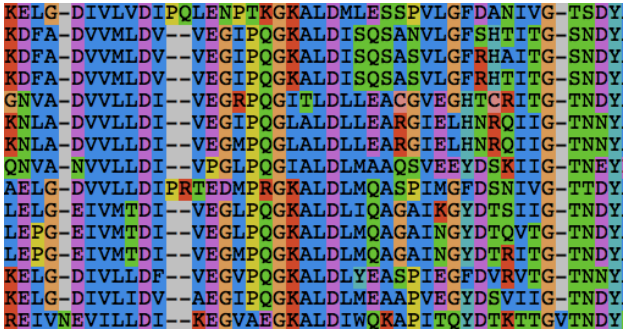
- In 1990, 1st Ancestral protein reconstructed, expressed & assayed by S.A. Benner Group
  - RNaseA from ~5Myr old extinct ruminant



What Took So Long ?

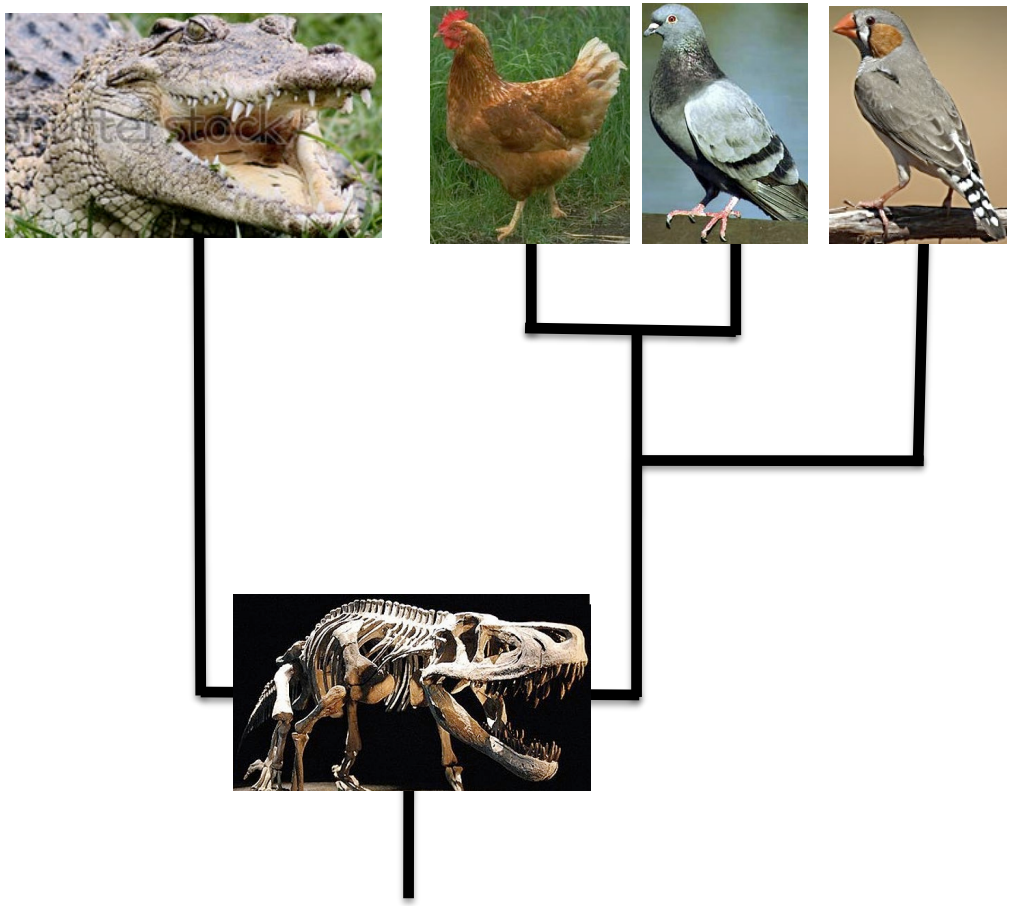
How to Resurrect a Protein

1) Acquire/Align Sequences



2) Construct Phylogeny

(from Chang et al. 2002)



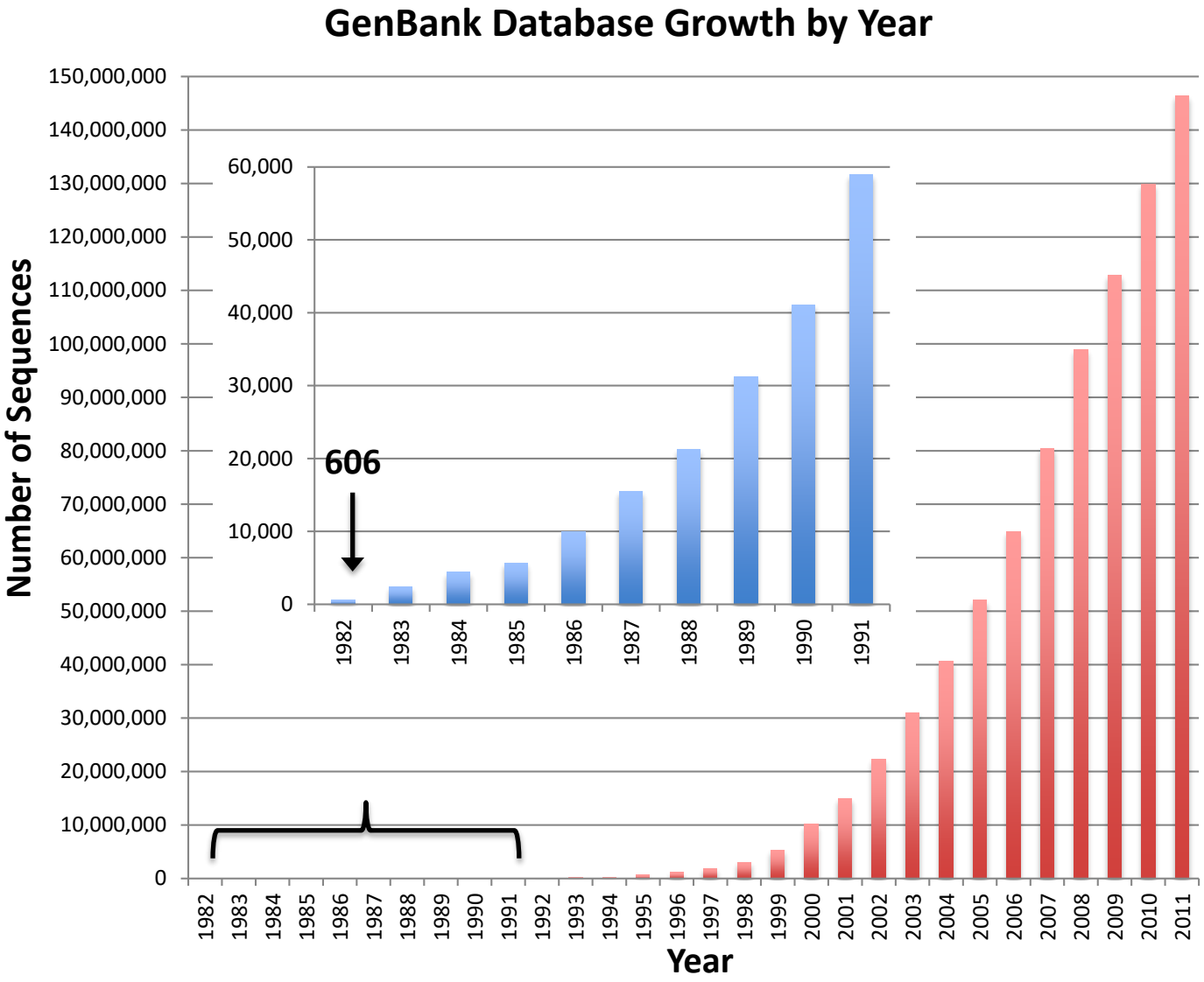
3) Infer Ancestral Nodes

4) Synthesize Inferred Sequence

So Really…What Took So Long?

- Advances in 3 areas were required:
  - Sequence availability
  - Phylogenetic reconstruction methods
  - Improvements in DNA synthesis

Sequence Availability



http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html

- Advances in 3 areas were required:

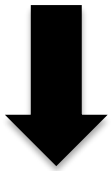
  ✓ Sequence availability

  – Phylogenetic reconstruction methods

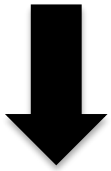
  – Improvements in DNA synthesis

Advances in Reconstruction Methods

Consensus



Parsimony



Maximum Likelihood

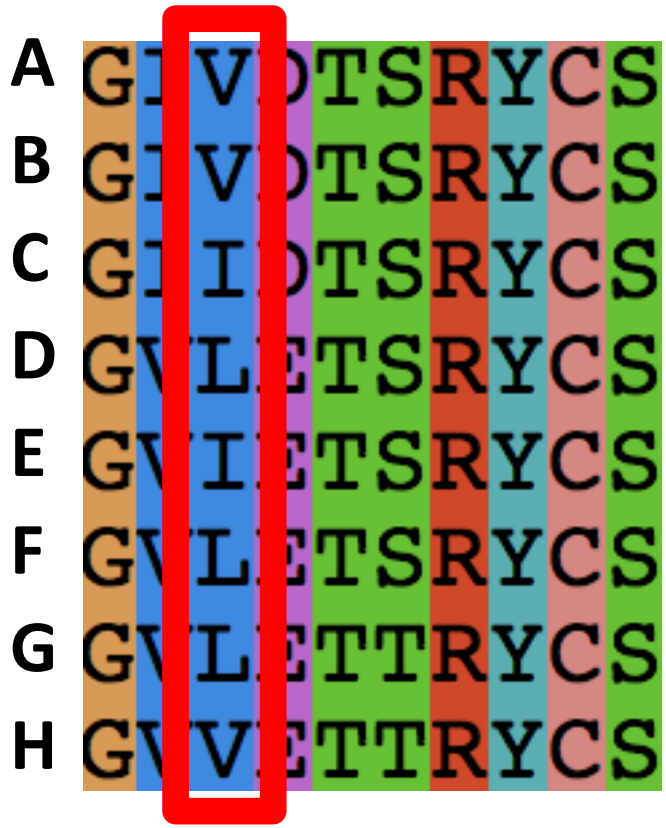
Consensus

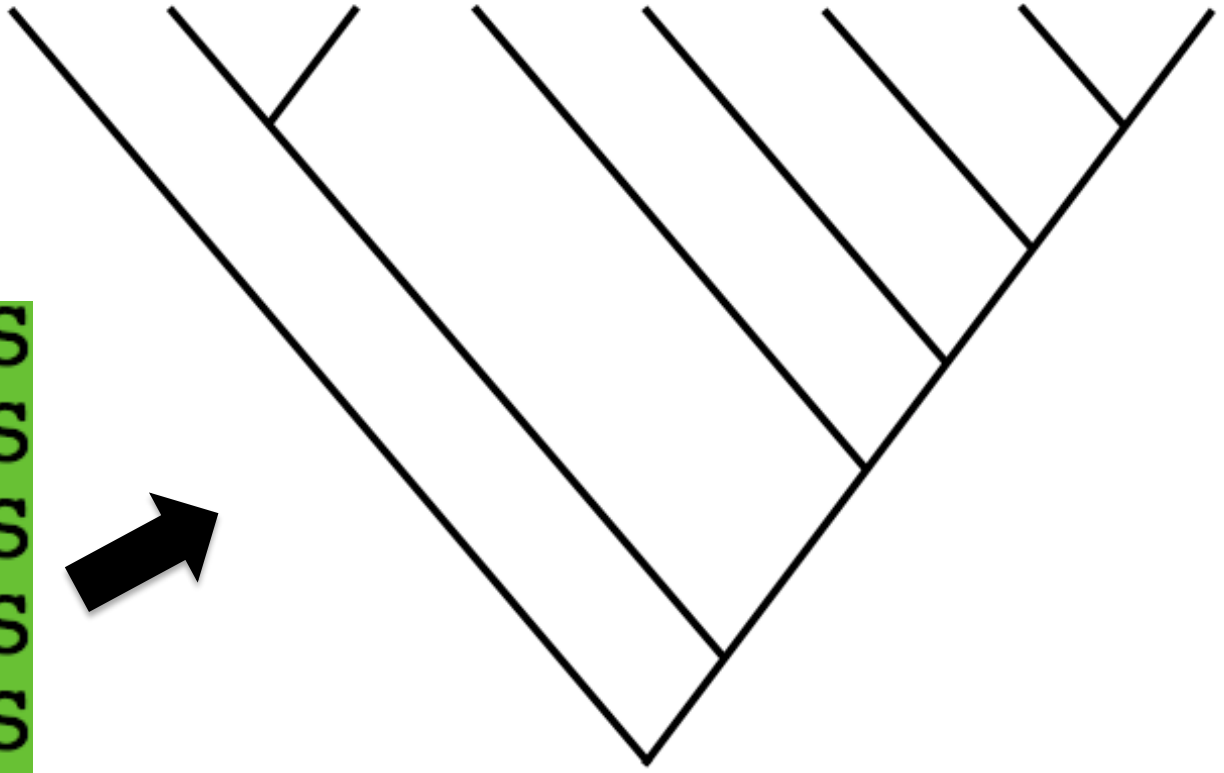
- Advantage: Easy & fast
- Disadvantages: Ignores phylogenetic

Parsimony

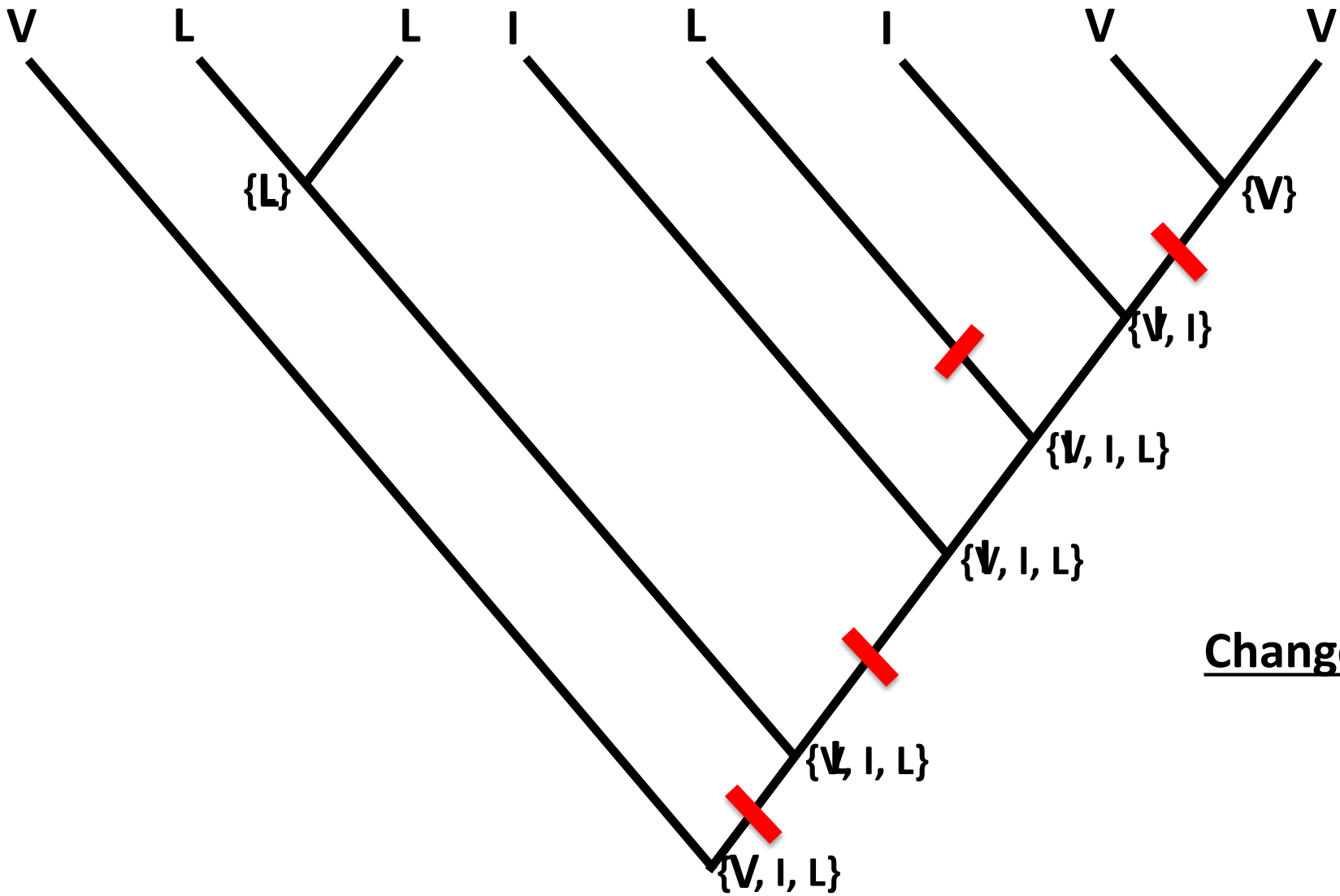
- Parsimony Principle
  - Best-supported evolutionary inference requires fewest changes
  - Assumes conservation as model
- Advantage:
  - Takes phylogenetic relationships into account
- Disadvantage:
  - Ignores evolutionary process & branch lengths

Parsimony





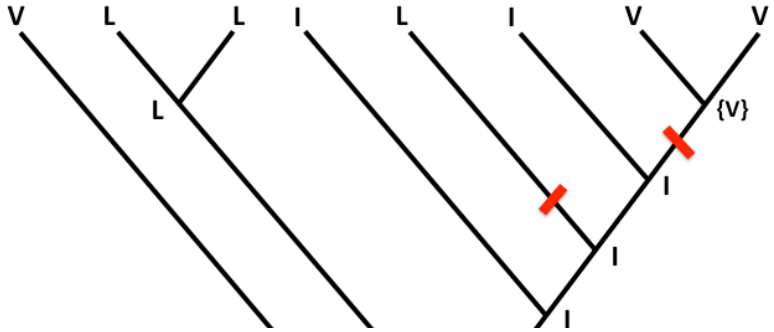
Parsimony

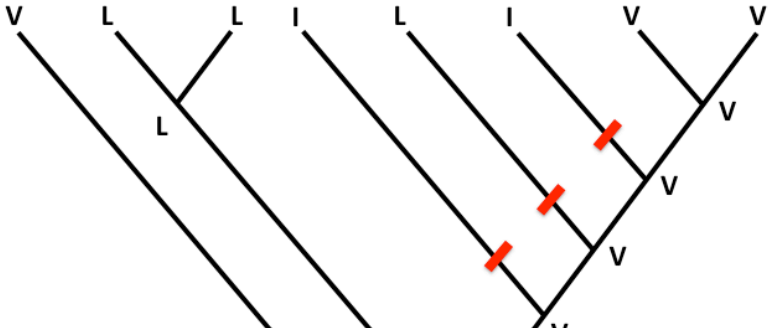


Changes = 4

Example adapted from David Hillis

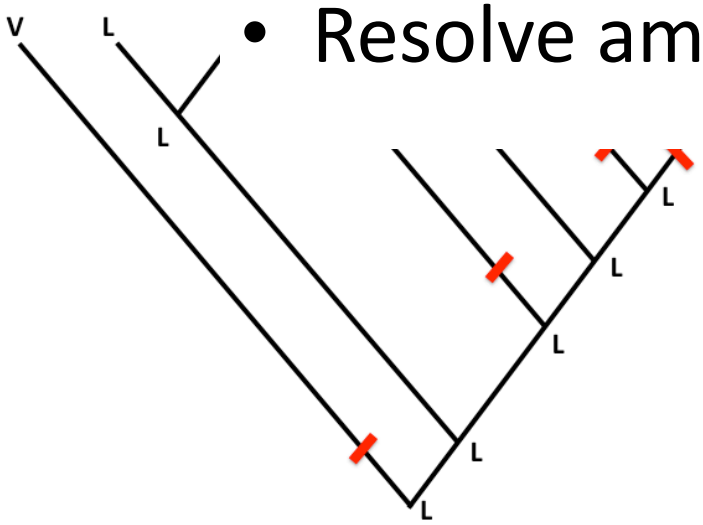
Parsimony - Alternate Reconstructions

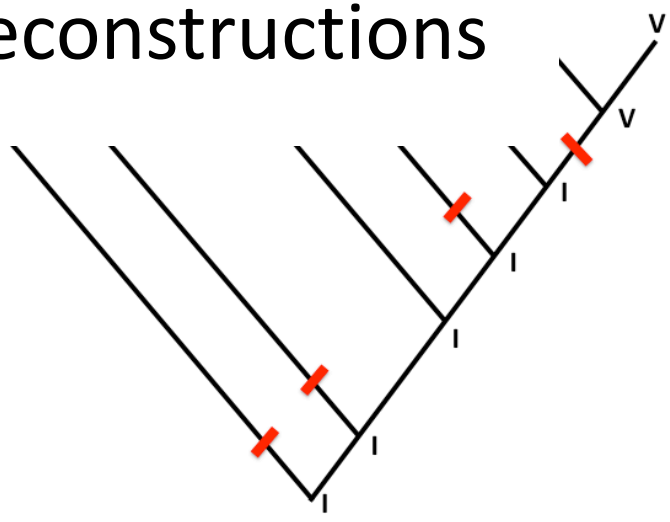




- Is conservation the best model?

- Resolve ambiguous reconstructions





Maximum Likelihood

- Likelihood:

Likelihood = Probability(Data|Model)

  - How surprised we should be by the data
  - Maximizing the likelihood, minimize your surprise

- Example:
  - Roll 20-sided die 9 times:

Maximum Likelihood

Likelihood = Probablity(Data|Model)

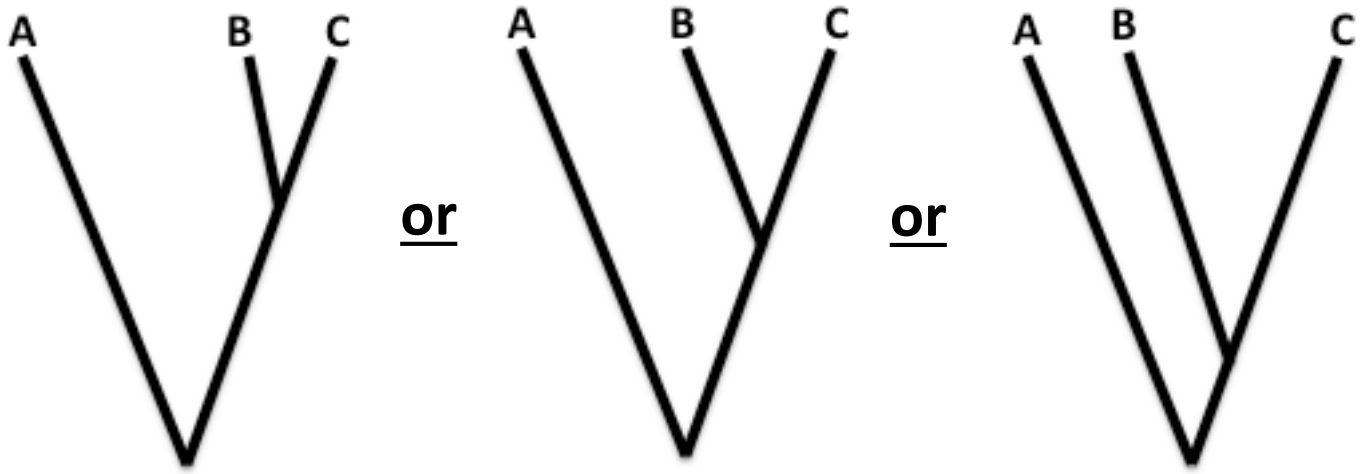
- Fair Die Model:
  - 5% chance of rolling a 20

Likelihood = $(0.05)^9$ = $2E$-11

Assuming trick model maximizes the likelihood

From Dice to Trees

- Likelihood=
  - Data - Sequences/Alignment
  - Model - Tree topology, Branch lengths & Model of evolution



- Choose model that maximizes the likelihood

Improvements Over Parsimony

- Includes of evolutionary process & branch lengths
  - Reduction in ambiguous sites
- Fit of model included in calculation
  - Removes *a priori* choices
  - Use more complex models (when applicable)
- Confidence in reconstruction
  - Posterior probabilities

- Advances in 3 areas were required:

  ✓ Sequence availability

  ✓ Phylogenetic reconstruction methods

  – Improvements in DNA synthesis

Advances in DNA Synthesis

DNA synthesis work starts
1950s

late 1970s
Automated

1983
PCR

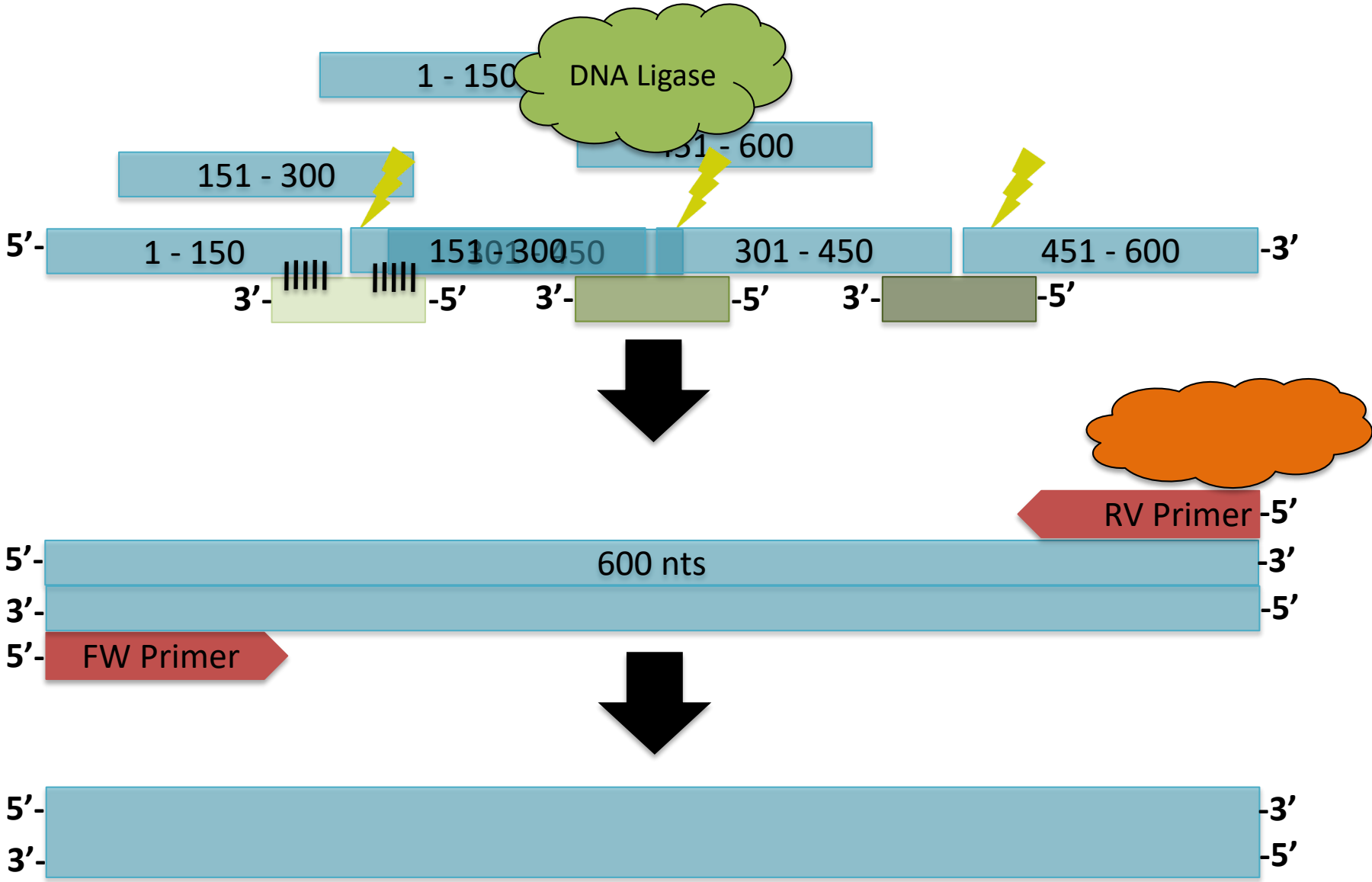
1990
20 nts Fragments

2002
~200 nts Fragments

Advances in Molecular Biology
increased speed & fidelity

PAST → **PRESENT**

How to Synthesize a Gene



Schematic adapted from Fuhrmann et al 2002

On to the Easy Part…