CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Mar 5

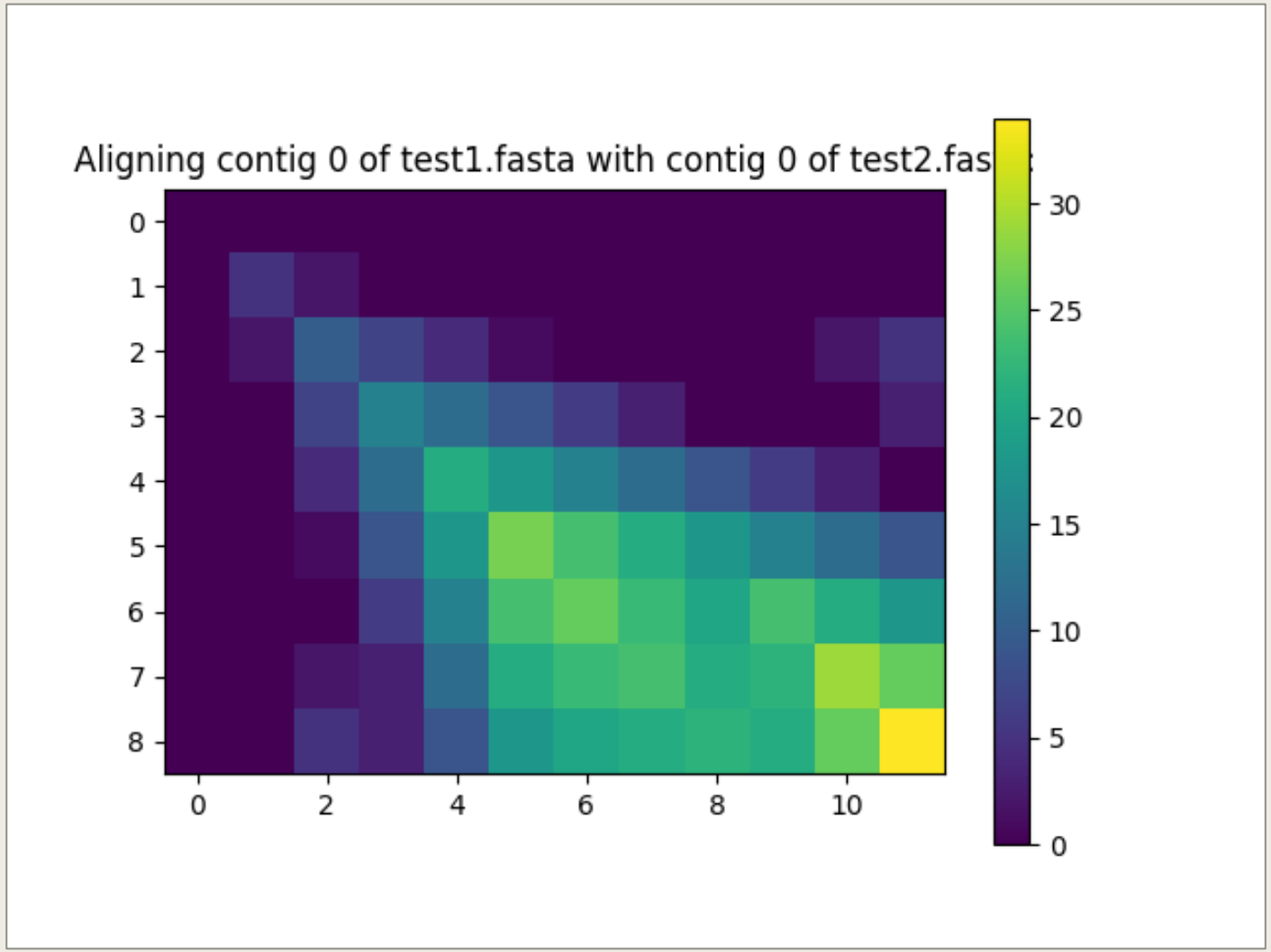
- Lab 3 examples
- Recap Fitch's algorithm (small parsimony)
- Sankoff's algorithm (weighted parsimony)

Notes:

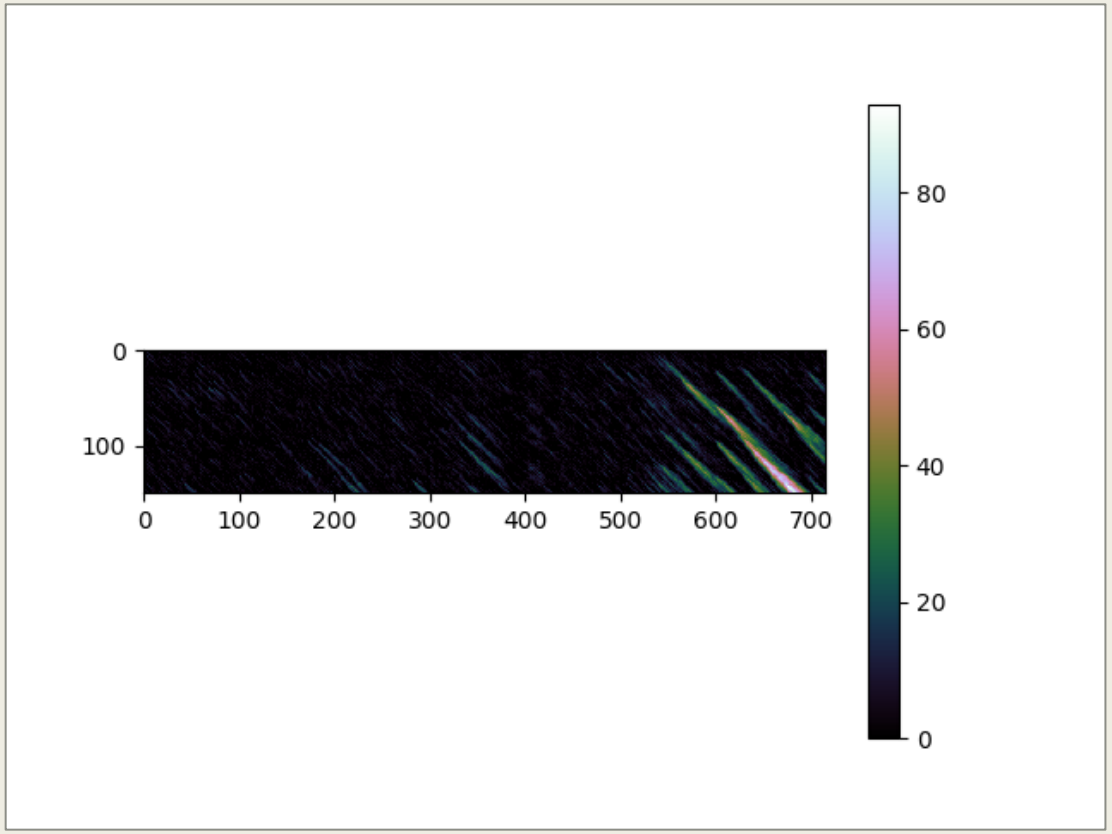
- Office hours TODAY 3-5pm
- I am off campus tomorrow (Tues)

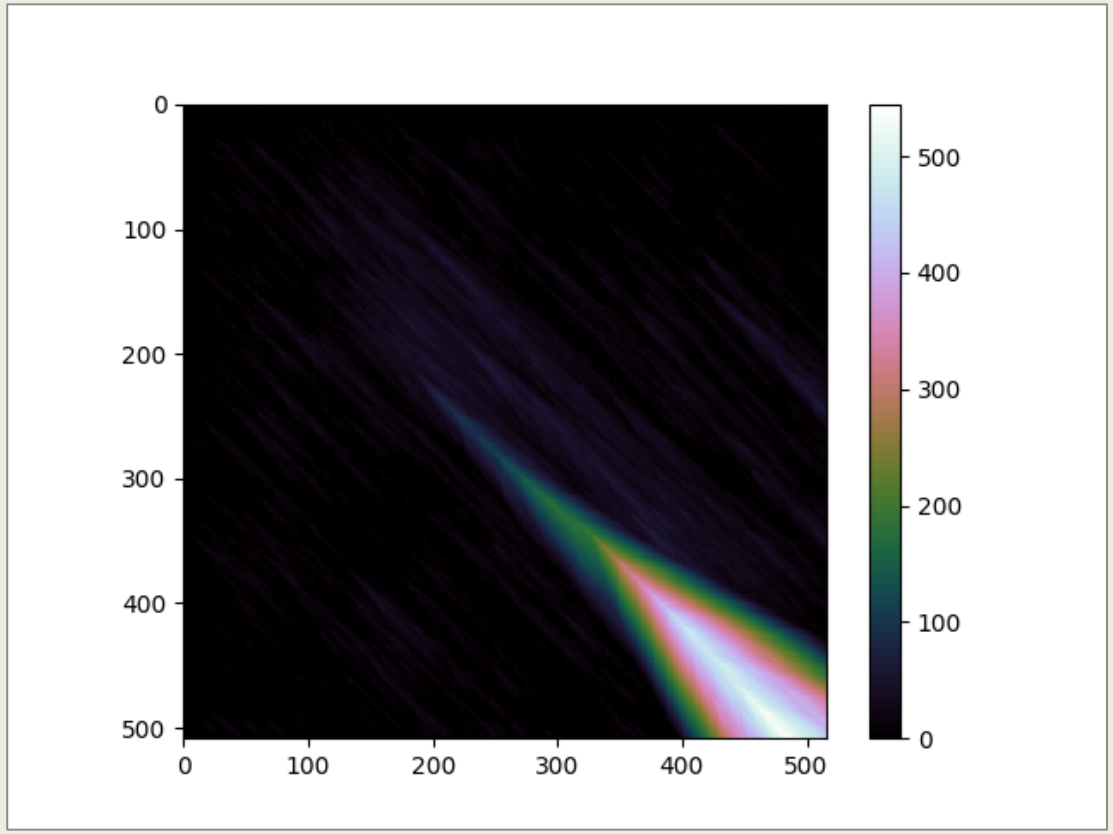
Lab 3 examples

Test 1 and test 2: Quinn and Kelly



Kinase and Calcium: Emily and Charlotte



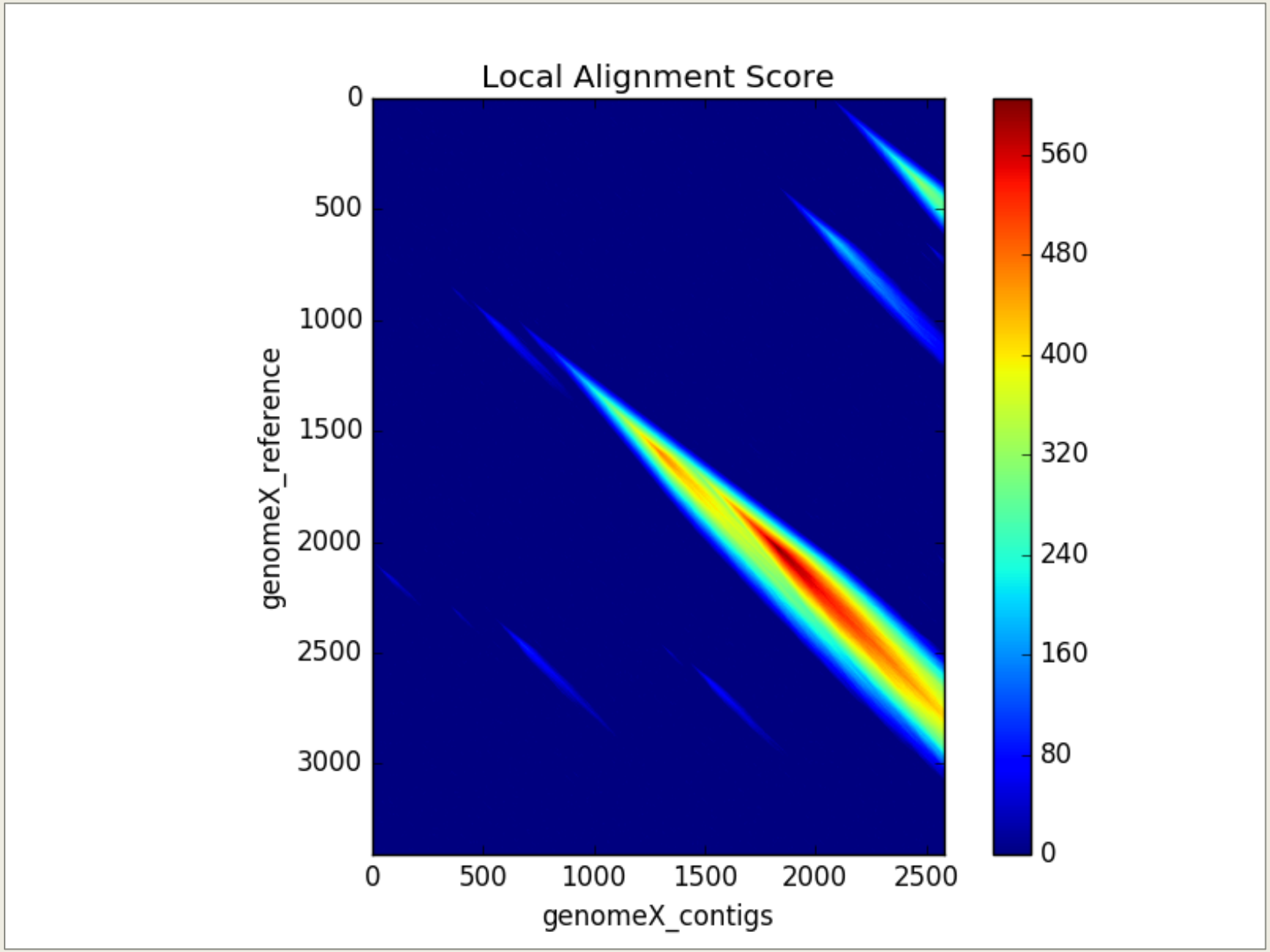


BLAST analysis: Sam and Hunter

- On genome X contigs and genome X reference:
- BLAST time: ~4 seconds
- DNAAligner: ~35 seconds

- BLAST's scores were higher
- Start and end indices about the same

Genome X contigs and reference: Tyler and Nathan



Best FOXP2 analysis: Kate and Dayo

```
A   300    TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCK    350
           ||||||||||||||||||||||||||||||||||||||||||||||||||
B   297    TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCK    347
```

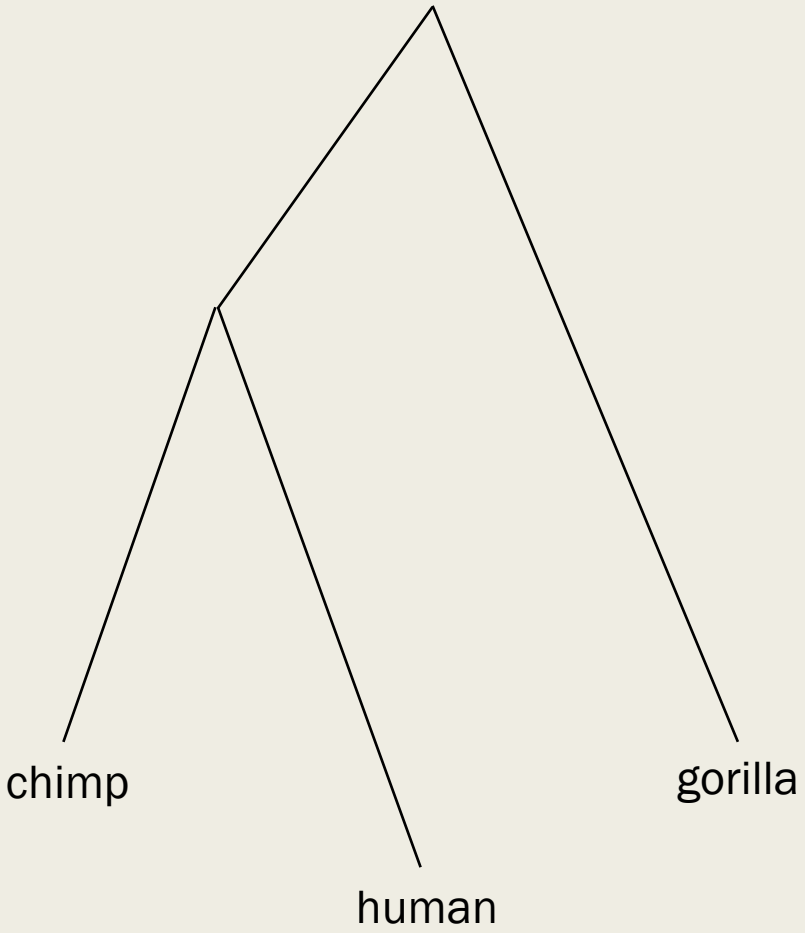
```
A   300    TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCK    350
           ||.||||||||||||||||||||.||||||||||||||||||||||||||
C   299    TSSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHTLYGHGVCK    349
```

```
B   298    SSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKW    348
           ||.||||||||||||||||||||.|||||||||||||||||||||||||
C   300    SSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHTLYGHGVCKW    350
```

C has two amino acid substitutions with respect to A and B => C is human

FOXP2 analysis

- Species A&B: 3620
- Species A&C: 3634
- Species B&C: 3617



A&C are most closely related =>
A&C are human & chimp and B is gorilla

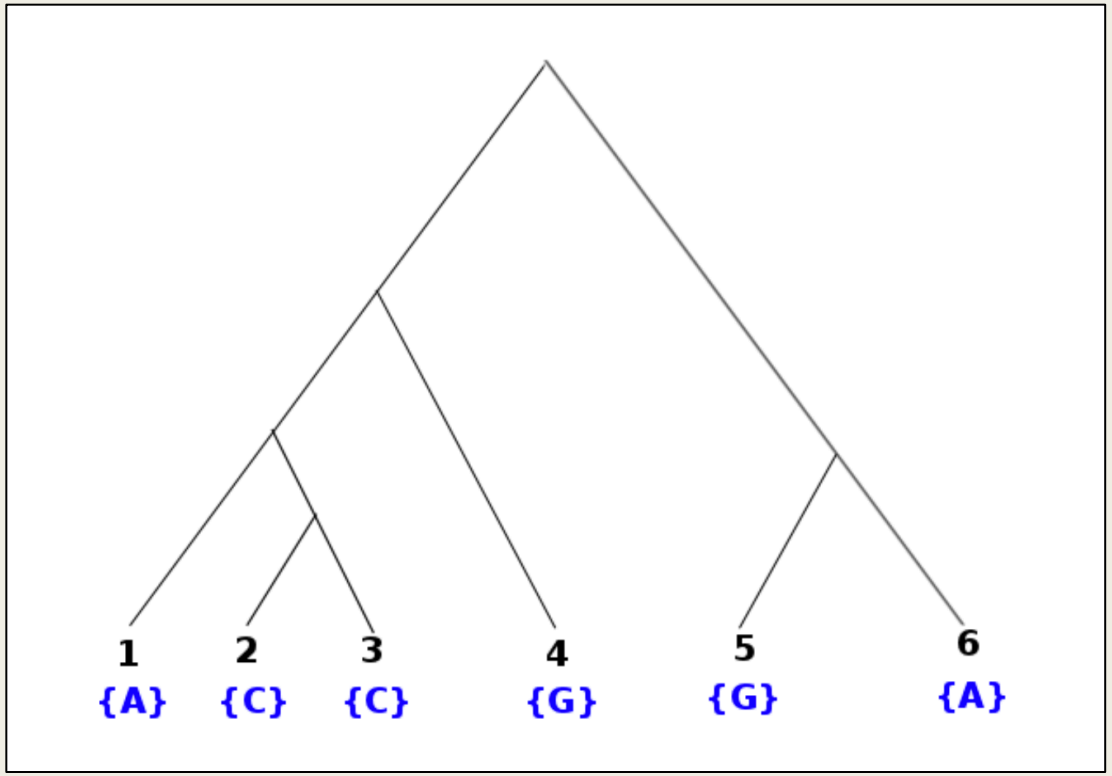
Out of A&C, C is further from B (gorilla) =>
C is human

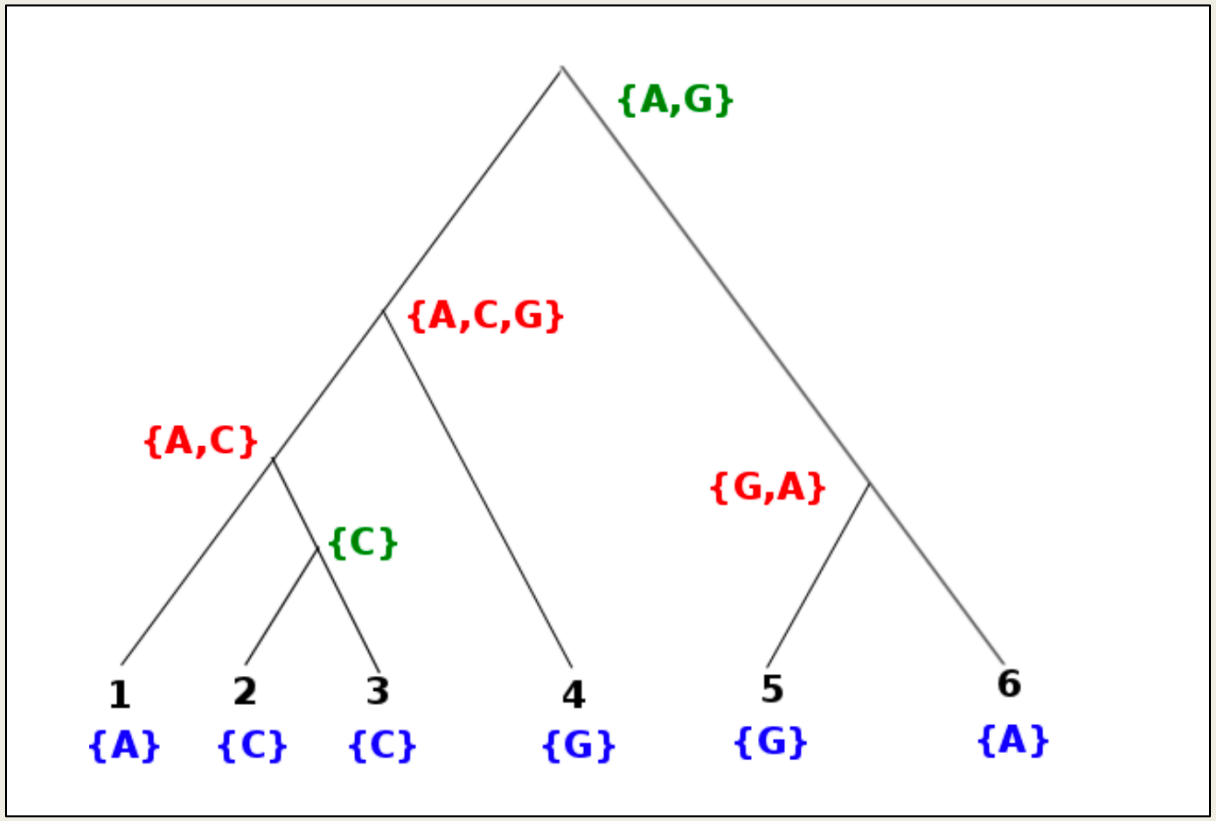
Recap Fitch’s algorithm

Ancestral state reconstruction

- Parsimony: idea that we want the smallest amount of evolutionary change to explain our observed data
- For Fitch's algorithm (often called "small parsimony") we are trying to minimize the number of mutations, where all mutations are equally likely
- We often call the assigned base or amino acid the "character", "state", or "allele" of the vertex (for both leaves and internal vertices)

Idea of Fitch's algorithm: intersection if intersection is non-empty, union otherwise





Intersection: no mutation
Union: mutation no matter what we choose

Example of Fitch's algorithm in the literature

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 2987–2994
doi:10.1093/bioinformatics/btt527

Genome analysis

Advance Access publication September 24, 2013

FPSAC: fast phylogenetic scaffolding of ancient contigs

Ashok Rajaraman[1,2], Eric Tannier[3,4] and Cedric Chauve[1,5,*]

[1]Department of Mathematics, Simon Fraser University, Burnaby (BC) V5A1S6, Canada, [2]International Graduate Training Center in Mathematical Biology, Pacific Institute for the Mathematical Sciences, Vancouver (BC), Canada, [3]INRIA Grenoble Rhône-Alpes, Montbonnot 38334, France, [4]Université de Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR5558 F-69622 Villeurbanne, France and [5]LaBRI, Université Bordeaux I, 33405 Talence, France

Associate Editor: Michael Brudno

Black death pandemic in Europe: 1347-1351
Killed 30-60% of Europe's population

In this paper, they used Fitch's algorithm to help reconstruct the genome of the ancient Black death pathogen

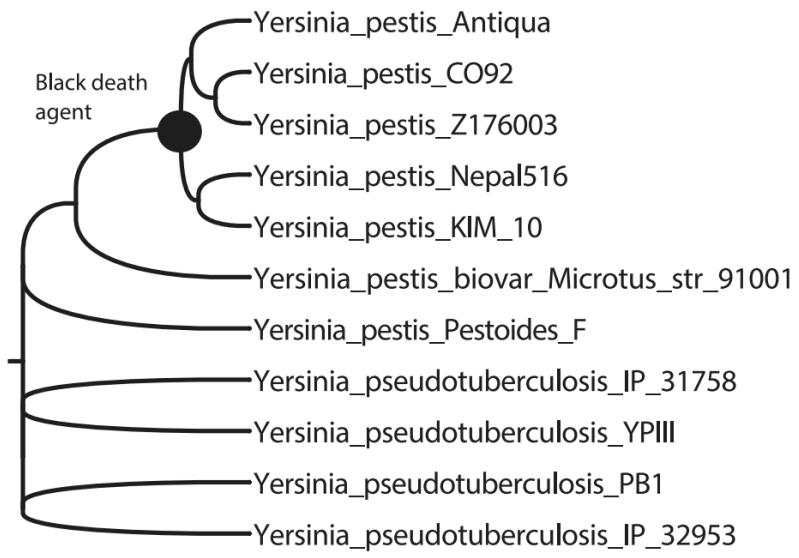
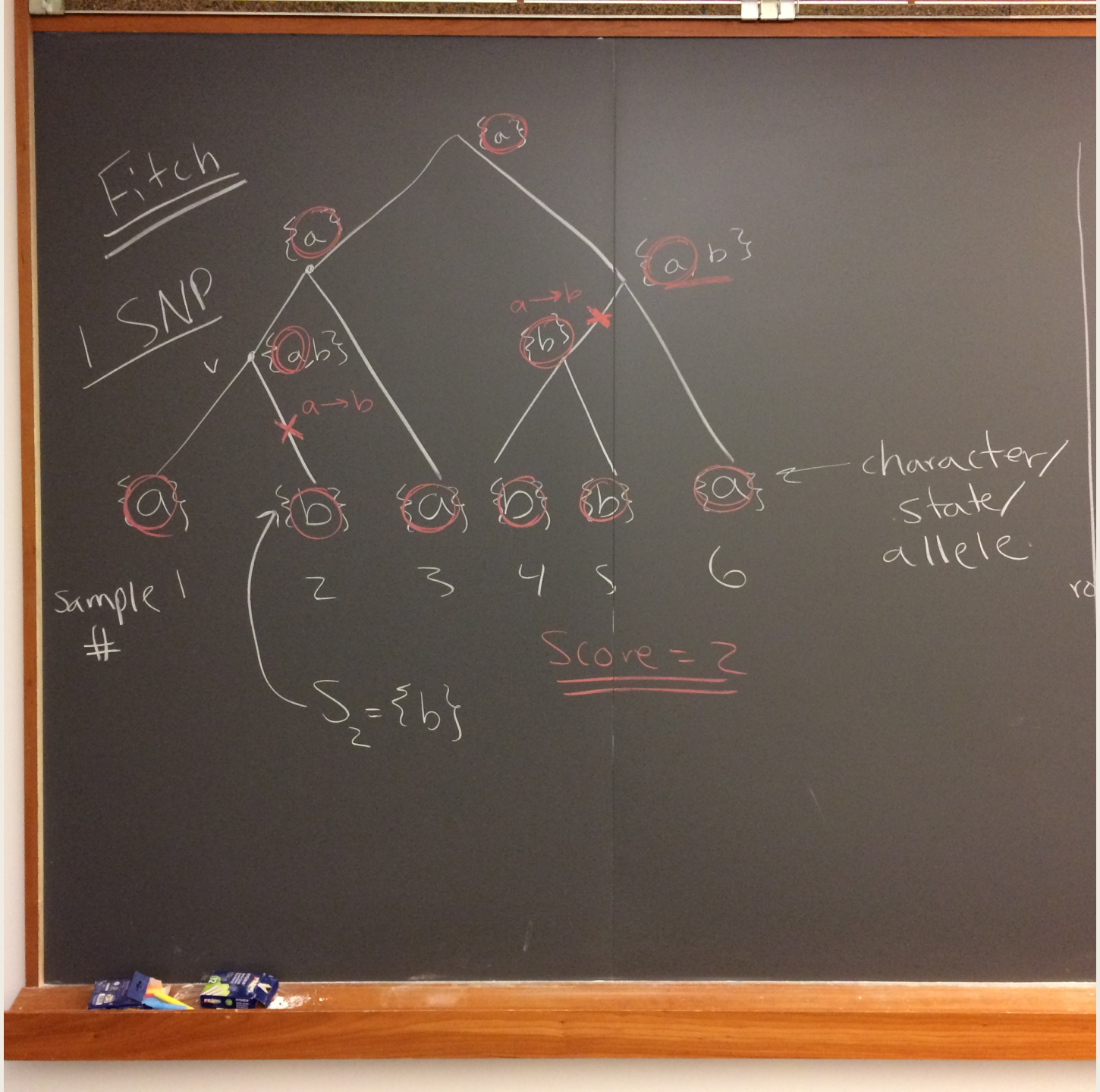


Fig. 3. Phylogeny of the considered genomes from Bos *et al.* (2011)


Fitch
1 SNP
{a}
{a}
{a b}
{a b}
{b}
a→b
a→b
{a}
{b}
{a}
{b}
{b}
{a}
character/
state/
allele
Sample #
1
2
3
4
5
6
Score = 2
$S_2 = \{b\}$

Sankoff’s algorithm (weighted parsimony)

Sankoff's Algorithm

Input: rooted, binary, phylogenetic tree + leaf labels

Output: internal vertex labels that minimize the weighted parsimony score.

mututional Score

σ (row)	a (col)	b (col)
a	0	2
b	1	0

$\sigma(a,b) = 2$

$\sigma(b,a) = 1$

$A_1 = \begin{matrix} a \\ c \\ g \\ t \end{matrix}$

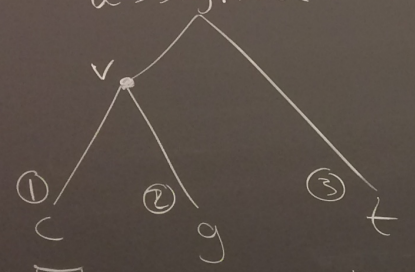
tree

els

ted

Initialization

leaf label characters assigned score 0, other characters assigned score ∞



$A_{leaf}(x) = 0$ if leaf char is x,

$A_{leaf}(y) = \infty$ o.w.

$$A_1 = \begin{array}{c|c} a & \infty \\ c & 0 \\ g & \infty \\ t & \infty \end{array} \quad A_2 = \begin{array}{c|c} a & \infty \\ c & \infty \\ g & 0 \\ t & \infty \end{array} \quad A_3 = \begin{array}{c|c} a & \infty \\ c & \infty \\ g & \infty \\ t & 0 \end{array}$$

Iteration/Recursion

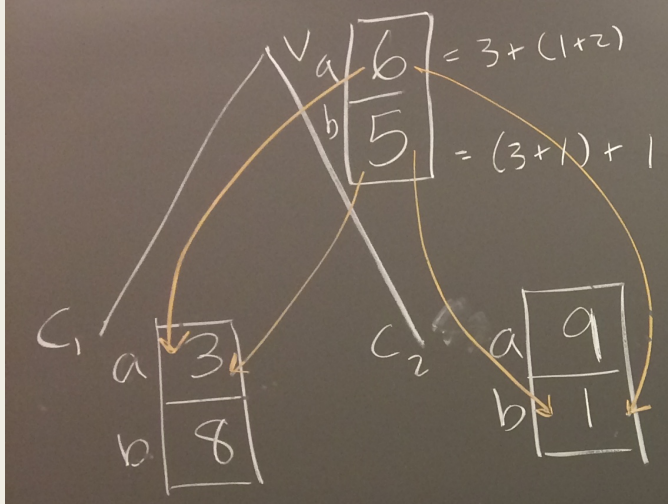
$$A_v(x) = \min_y \{ A_{c_1}(y) + \sigma(x,y) \} + \min_z \{ A_{c_2}(z) + \sigma(x,z) \}$$

$$c_1 \quad \begin{array}{c|c} a & 3 \\ b & 8 \end{array}$$

Termination:

→ choose x s.t. $A_{root}(x)$ is minimized, & assign $C_{root} = x$

→ trace back to find state at all internal vertices



$$A_v(a) = \underbrace{\min\{3+0,\ 8+2\}}_{c_1} + \underbrace{\min\{9+0,\ 1+2\}}_{c_2}$$

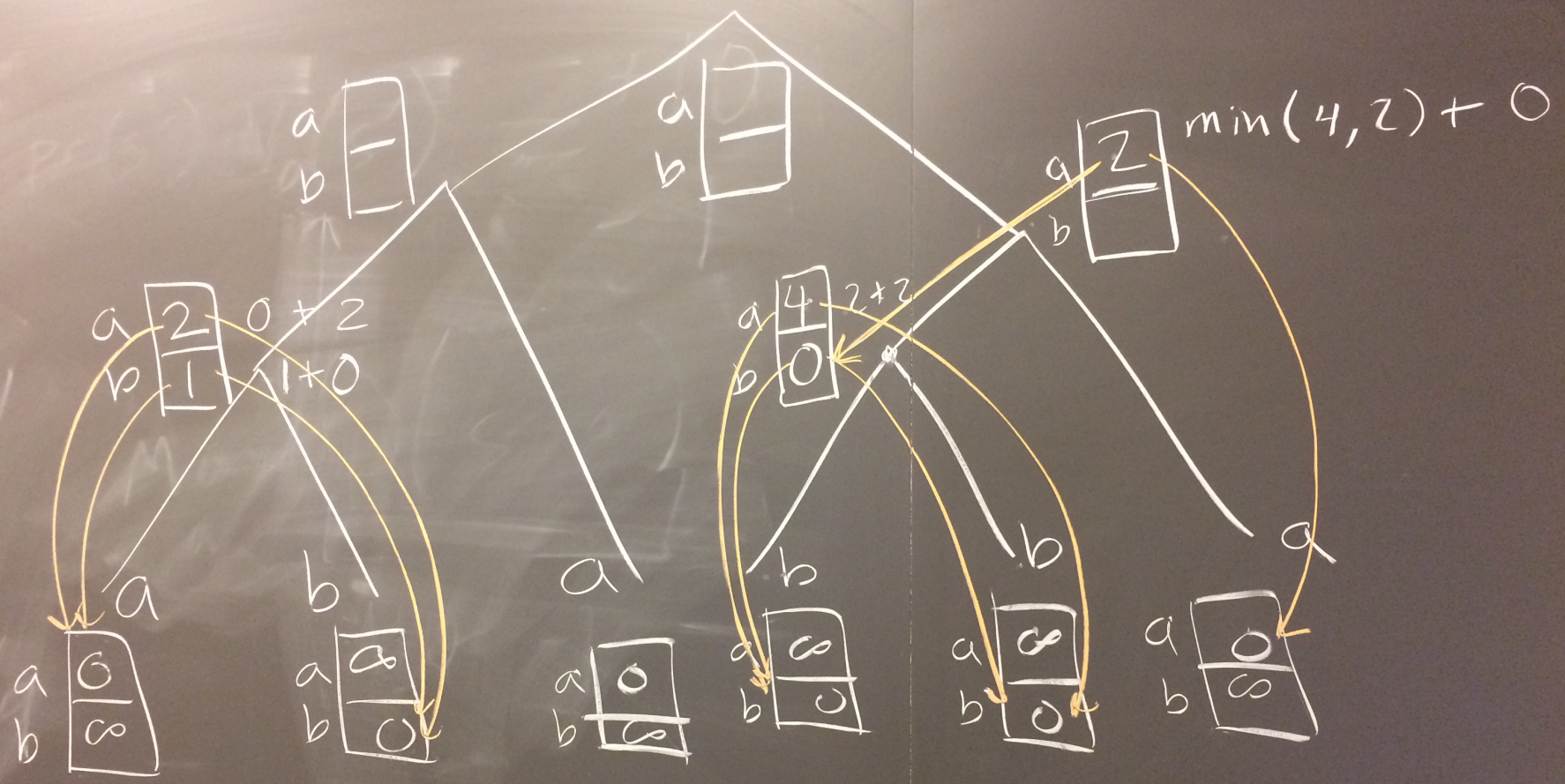
$$= 3 + 3 = 6$$

$$A_v(b) = \min\{\overset{b\to a}{3+1},\ \overset{b\to b}{8+0}\} + \min\{\overset{b\to a}{9+1},\ \overset{b\to b}{1+0}\}$$

$$= 4 + 1 = 5$$

ized,

internal vertices


min(4,2) + 0
a 2
b
a
b
a
b
a 2
b 1
0 + 2
1 + 0
a 4
b 0
2 + 2
a
b
a
b
b
a
b
a
a 0
b ∞
a ∞
b 0
a 0
b ∞
a ∞
b 0
a ∞
b 0
a 0
b ∞