CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Feb 2

- Evaluation of assemblies
- Start: string alignment

Notes:

- Sequence alignment reading posted (from Durbin)

Recap: issues with de Bruijn graph assembly

1) Repeats of length (k-1) or longer
2) Gaps in coverage
3) Differences in coverage
4) Sequencing errors

Evaluating Assemblies

Assembly evaluation

- **N50**: for a set of contigs, N50 is the greatest length such that at least half the bases of the assembly are in a contig with length N50 or longer

1a. Contigs, sorted according to their lengths.

Image: The Molecular Ecologist

Assembly evaluation

- **N50**: for a set of contigs, N50 is the greatest length such that at least half the bases of the assembly are in a contig with length N50 or longer

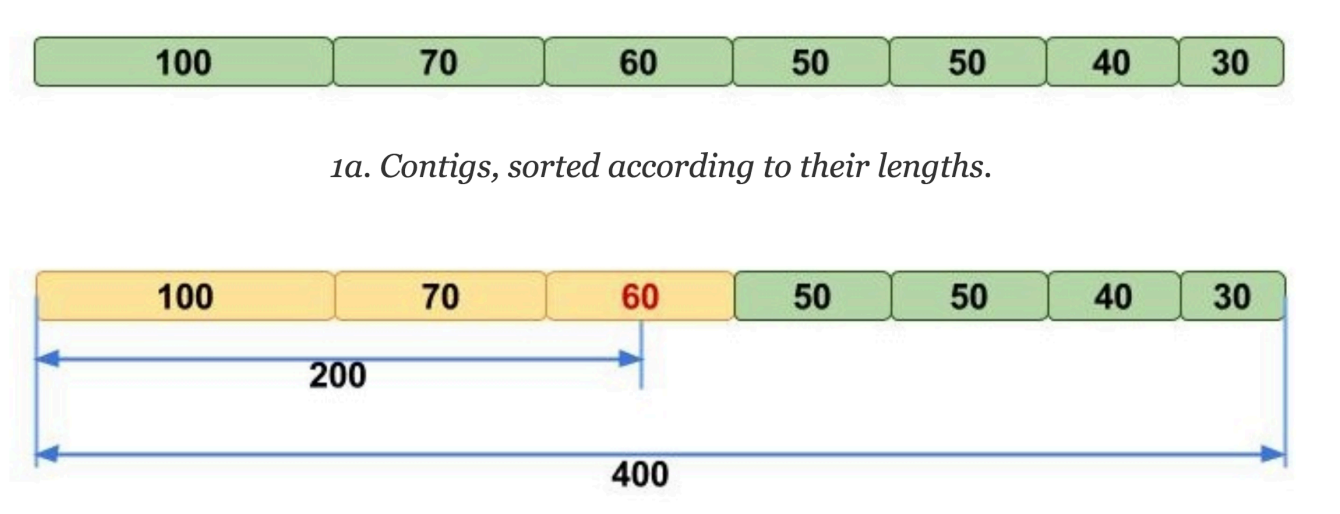
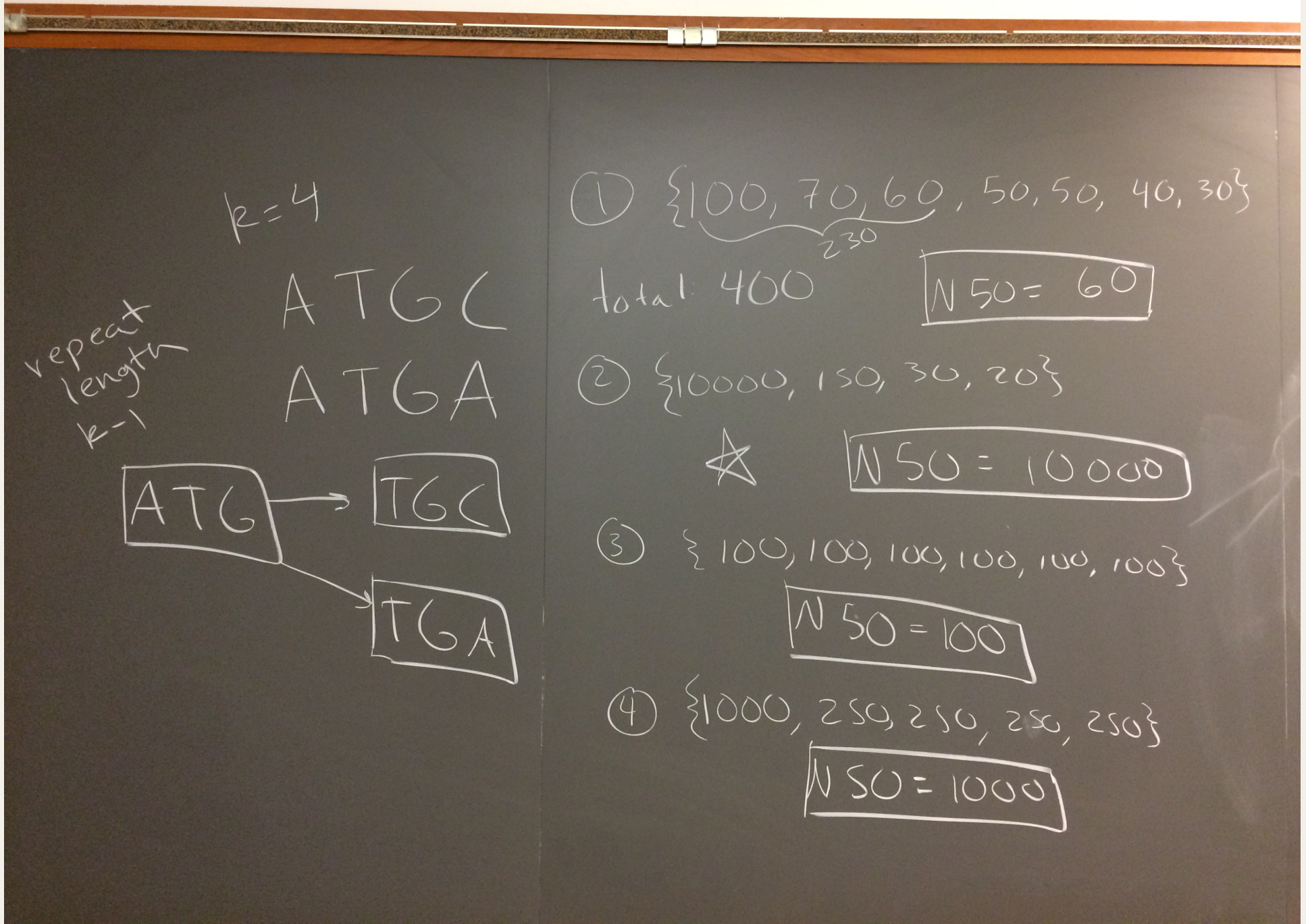


Image: The Molecular Ecologist

Why is N50 a bad evaluation metric?

- We could just loop through cycles in our a graph over and over, generating large (incorrect) contigs
- We need a better way to evaluate the quality of assemblies
- Take away: simulated data is every valuable. Take an existing genome, simulate random reads, then try to reconstruct.


k=4
ATGC
ATGA
repeat length k-1
ATG
TGC
TGA
① {100, 70, 60, 50, 50, 40, 30}
230
total 400
N50 = 60
② {10000, 150, 30, 20}
N50 = 10000
③ {100, 100, 100, 100, 100, 100}
N50 = 100
④ {1000, 250, 250, 250, 250}
N50 = 1000

⑤ {1000, 250, 250, 250, 250, 5}

NSO = 250

⑥ - easy to compute
- no ground truth required

⑦

⑧ - assembling sequence that doesn't come from the reference

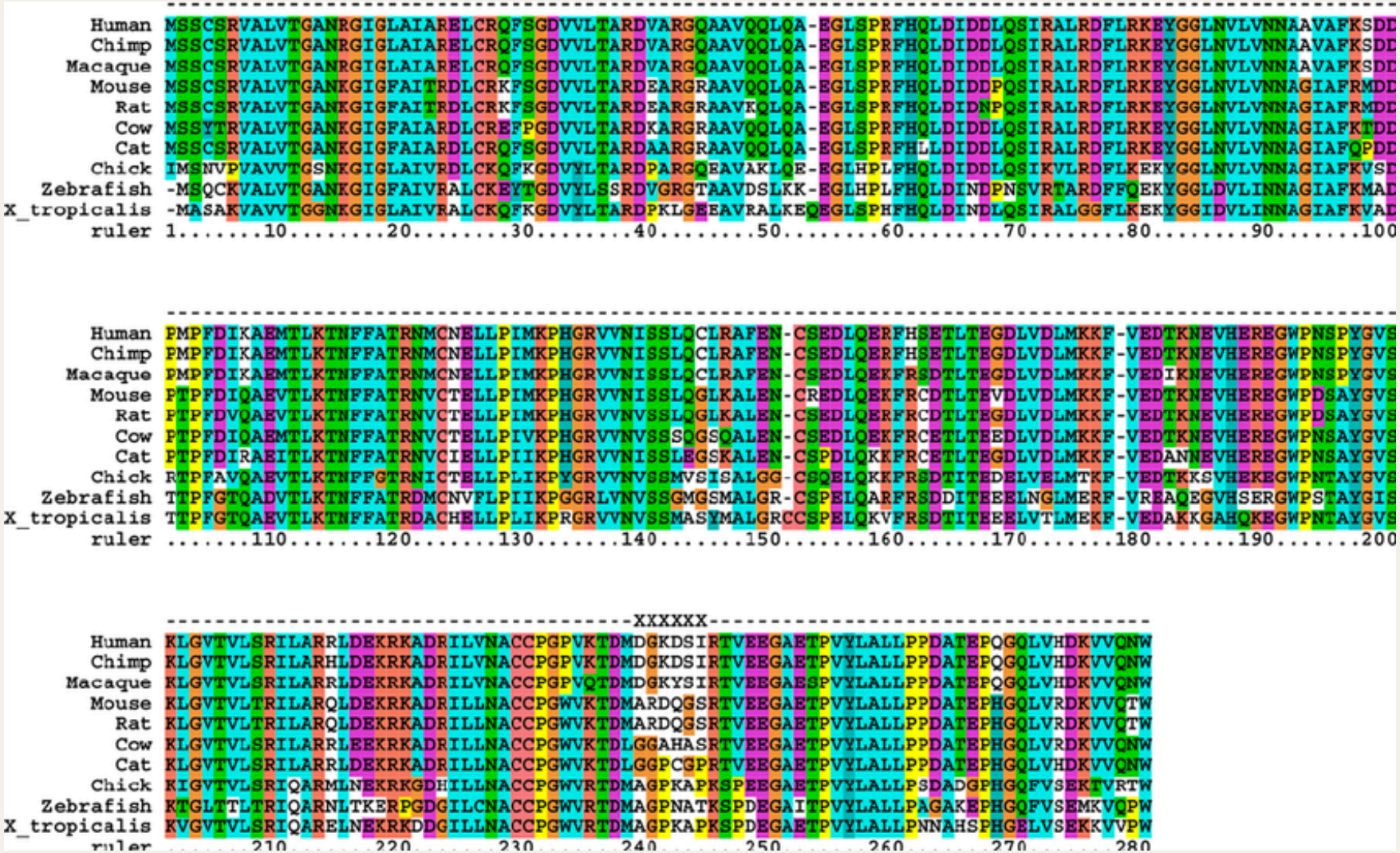
↘ sequence alignment.

Sequence Alginment

Next Topic: sequence alignment

- Goal: given two sequences, what is the best match or “alignment” between them?
- **Global alignment:** align the entire sequences start to finish
- **Local alignment:** find portions of the two sequences with high similarity
- **Homologous**: sequences that are similar due to descent from a common ancestor
- Usually we are aligning homologous sequences (not sequences from completely different regions of the genome)

Example alignments: human, chimp, macaque + other species



New Insights into the Enigma of Immunoglobulin D

Why sequence alignment?

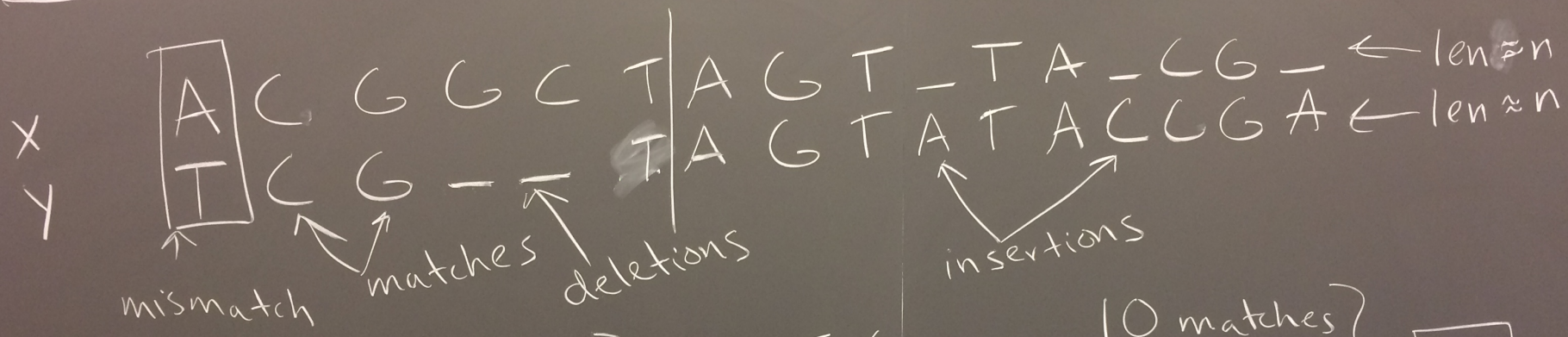
- Understand evolutionary relationships between different species
- In particular: understanding fast-evolving bacterial and viral strains is important for health
- Understand protein function
- Understand diversity at the species level (important for diseases with a genetic component)

Example

- ACGGCTAGTTACG
- TCGTAGTATACCGA

- How should we “line them up” to get the best overlap?

Sequence Alignment



How do we score?

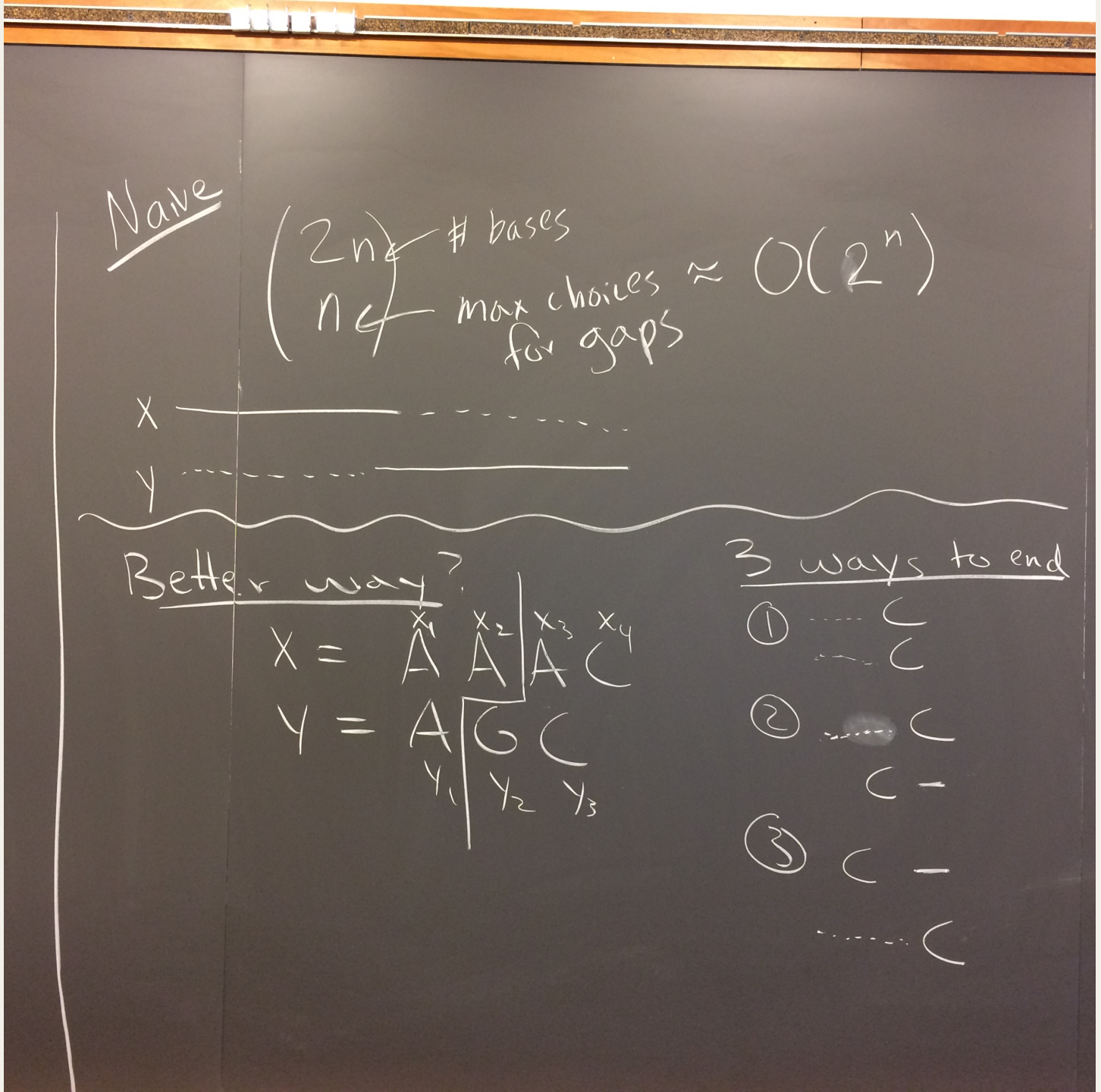
match : +1

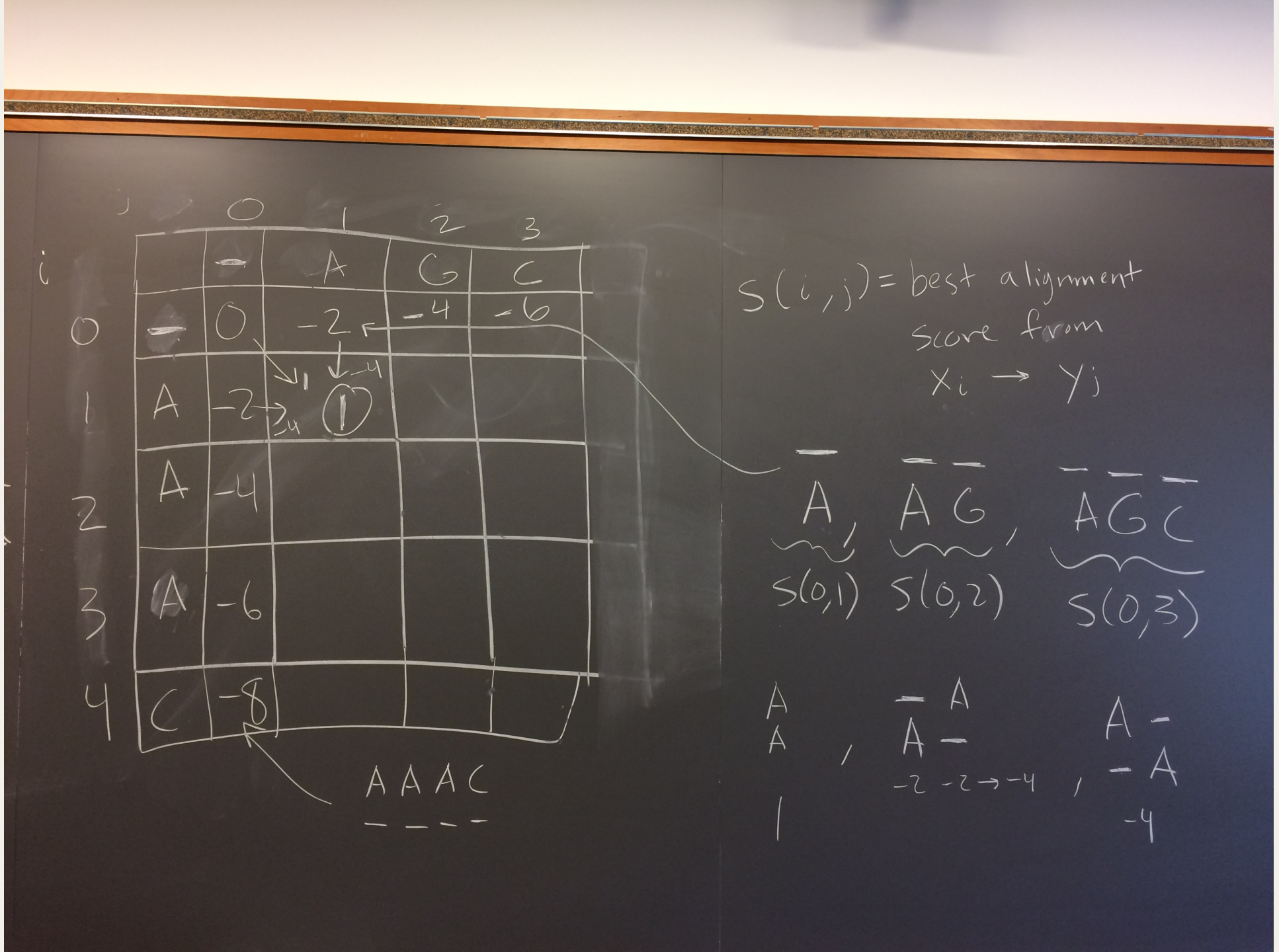
mismatch : -1

gap : -2

$S(x, y) =$ 10 matches, -1 mismatch, -10 gaps → -1

Want the best score (highest)


Naive
(2n # bases
n max choices for gaps ≈ O(2^n)
X
Y
Better way?
X = A A A C
Y = A G C
3 ways to end
1
2
3


j
0
1
2
3
i
A
G
C
0
-2
-4
-6
A
-2
A
-4
A
-6
C
-8
-1
-4
-4
AAAC
S(i,j) = best alignment score from $x_i \rightarrow y_j$
A
AG
AGC
S(0,1)
S(0,2)
S(0,3)
A
A
A
A
A
-2 -2 → -4
A
-4
1