Tentative project topics

DS202: Algorithmic Foundations of Big Data Biology

Chirag Jain
Head, Algorithmic Techniques in Computational Genomics (ATCG) Lab
Assistant Professor, CDS

Today’s agenda

1. Projects
    1. Guidelines
    2. Deliverables
    3. Tentative topics
2. Discussion of midterm questions

Timeline

Feb 22 — Guidelines & tentative topics

Feb 22 — Finalise topic & team

Mar 6 — Submit report 1
- Problem description
- Application(s)
- Deliverables
- References

Apr 1 - Apr 10 — Presentations
- One lecture per project
- Teach what you learnt (new algorithms + results)

Apr 10 — Final report, code hosted on your GitHub account, presentation slides

Expectations

1. Self-learn a new research topic (related to big-data biology)
2. Teach the class what you learnt (1 hour lecture)
3. Oral-presentation and submit written-report
    1. Problem description
    2. Significance
    3. Algorithms (design and analysis)
    4. Experiments using real biological data (okay to use open-source software)
    5. Identify scope for further research

Expectations

1. Work as a team (2 members) or solo - your choice
2. All project work must be your own
3. https://sites.google.com/view/ds202/policy

Weightage

1. 10% - Preliminary report
2. 40% - Overall quality of oral presentation (content + communication + Q/A)
3. 25% - Final written report
4. 25% - Efforts towards experimenting with real biological data

Topic-1: Edit distance algorithms with good average-case behaviour

1. We have seen a quadratic-time $O(n^2)$ dynamic programming algorithm to solve edit distance
2. Running $O(n^2)$ algorithm for long sequences can be extremely slow
    1. We may need 95 CPU years to align human genome to mouse genome.
3. Conditional lower bounds suggest that $O(n^{1.99})$ algorithm cannot exist.
4. The runtime of these algorithms is decided by “worst-case” inputs
    1. This makes it necessary to compare every part of one string to every part of another string
5. This serves as a motivation for “average-case analysis”

Topic-1: Edit distance algorithms with good average-case behaviour

Near-Linear Time Edit Distance for Indel Channels

Arun Ganesh
Department of Electrical Engineering and Computer Sciences, UC Berkeley, CA, USA
arunganesh@berkeley.edu

Aaron Sy
Department of Electrical Engineering and Computer Sciences, UC Berkeley, CA, USA
raaronsy@gmail.com

[WABI 2020] https://doi.org/10.4230/LIPIcs.WABI.2020.17

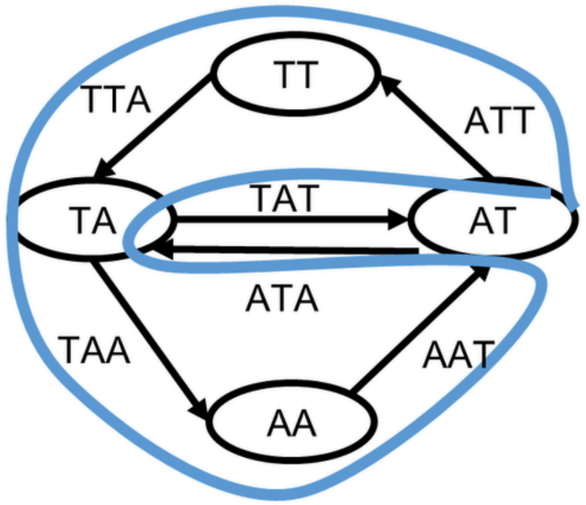
Other references that may be useful
- Paul Medvedev, “Theoretical analysis of edit distance algorithms: an applied perspective”
- Sosic and Sikic, “Edlib: a C/C++ library for fast, exact sequence alignment using edit distance”
- https://github.com/Martinsos/edlib

Topic-2: Solving genome assembly via Eulerian Path

An Eulerian path approach to DNA fragment assembly

Pavel A. Pevzner*, Haixu Tang†, and Michael S. Waterman†‡§

1. Genome reconstruction from short fragments of the genome
2. The above paper inspired several new algorithms and software for genome assembly.
3. Blogs / videos available online



reads *R = {TATTA, TAATA}*

Figure source: [1]

Other references that may be useful
[1] Medvedev and Pop, “What do Eulerian and Hamiltonian cycles have to do with genome assembly?”

Topic-3: Algorithm to construct phylogenetic trees using maximum likelihood

Sequence alignment

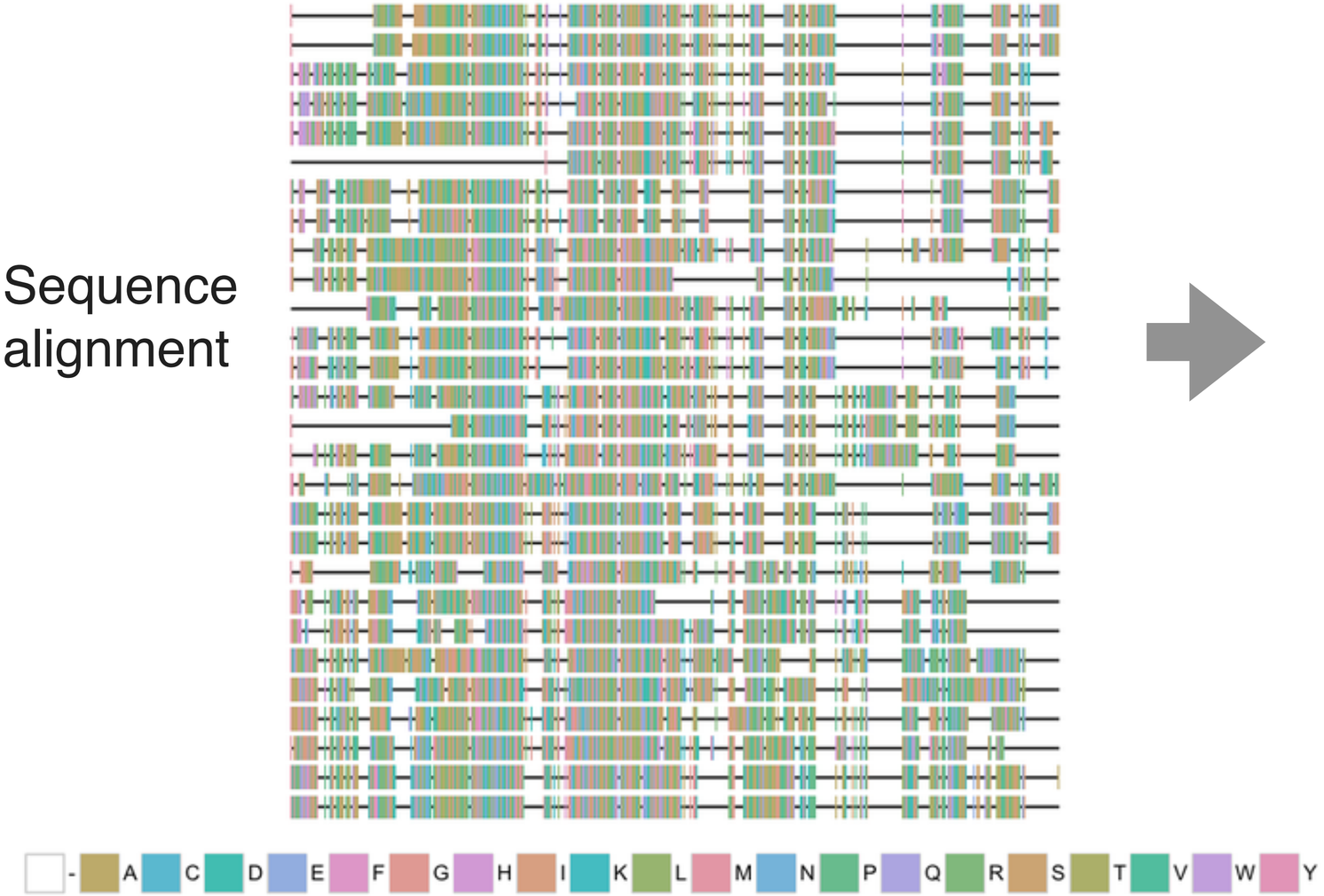


Figure source: Twitter

Evolutionary tree

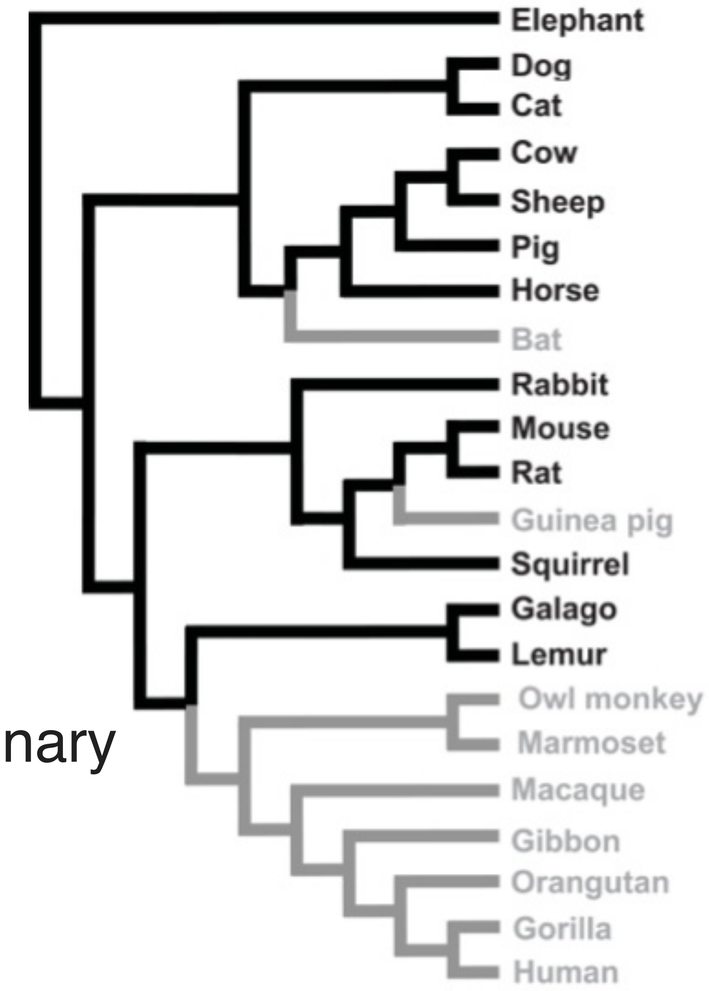


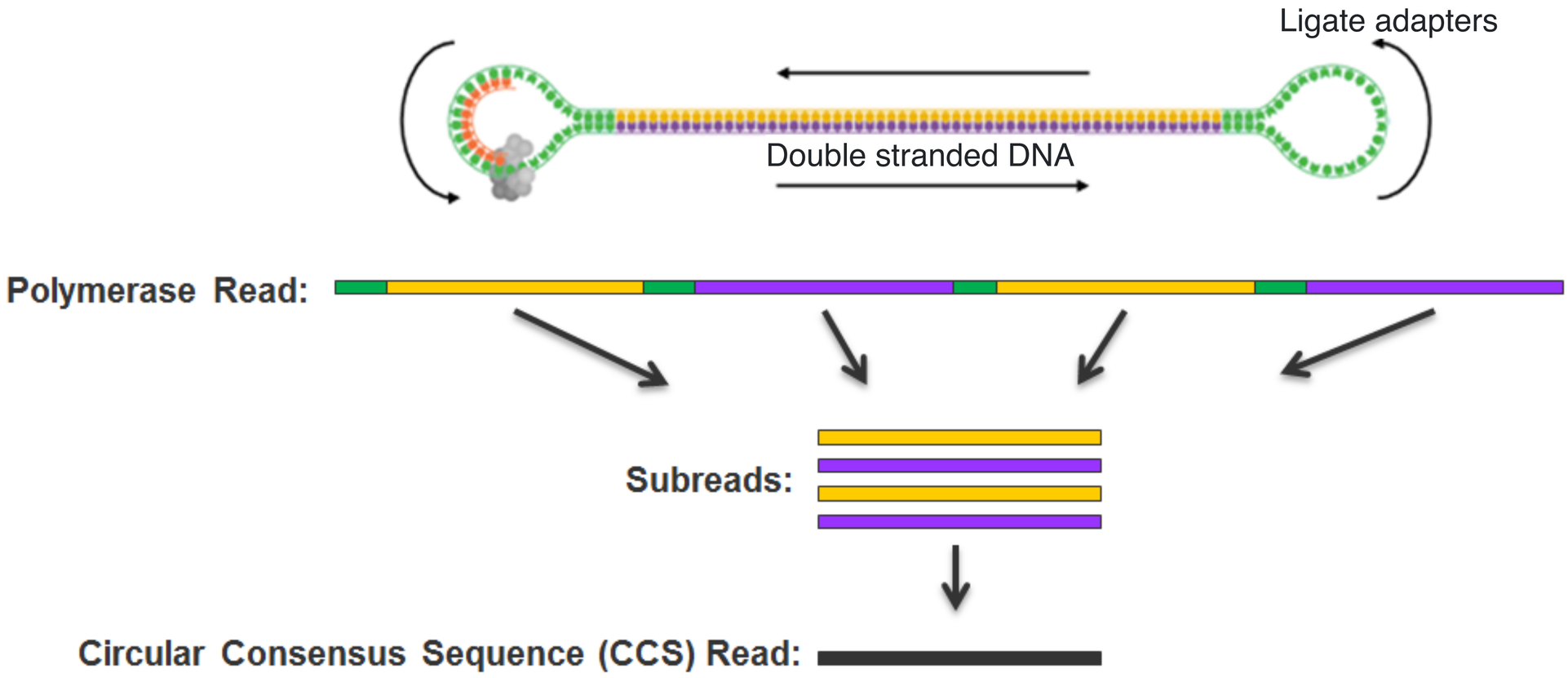
Figure source: realclearscience.com

Topic-3: Algorithm to construct phylogenetic trees using maximum likelihood

1. There are multiple methods to construct phylogenetic trees; maximum likelihood-based methods are most popular and robust
2. Key references:
    1. Durbin textbook (link on course page)
        1. Introduction in section 7.1 and 7.2
        2. Algorithm described in 8.1 - 8.4
3. Pre-requisite: Probability

Topic-4: Transformer-based deep learning model to predict sequence consensus

Developed by Google for PacBio CCS sequencing technology



Topic-4: Transformer-based deep learning model to predict sequence consensus

Developed by Google for PacBio CCS sequencing technology

ARTICLES
https://doi.org/10.1038/s41587-022-01435-7
nature biotechnology
Check for updates

DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer

https://github.com/google/deepconsensus

Topic-5: Compressibility of Burrows-Wheeler-Transform

JUNE 2022 | VOL. 65 | NO. 6 | **COMMUNICATIONS OF THE ACM**

Resolution of the Burrows-Wheeler Transform Conjecture

By Dominik Kempa and Tomasz Kociumaka

https://doi.org/10.1145/3531445