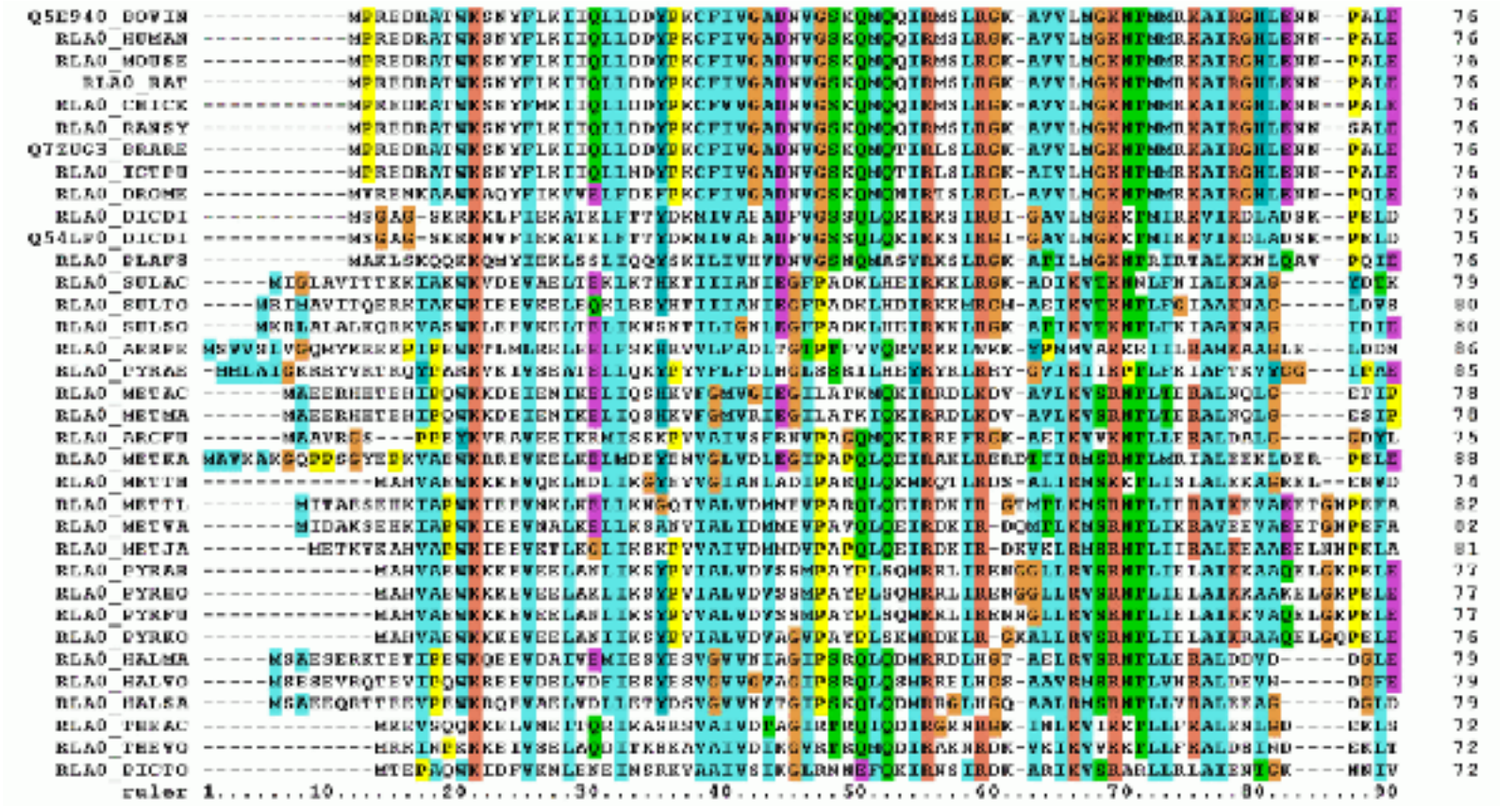
TOPIC 5: Sequence alignment

Outline

- Sequence alignment
- Alignment algorithms
- Whole genome alignment - inferences made from them

Sequence alignment

Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.



A multiple alignment of protein sequences

From Wikipedia

Pairwise alignment

Alignment of two sequences is a relatively straightforward computational problem, but…

- there are many possible alignments
- there can be a very large reference

NOTE: Two sequences can always be aligned and there can be more than one optimal solution

Methods of alignment

By hand

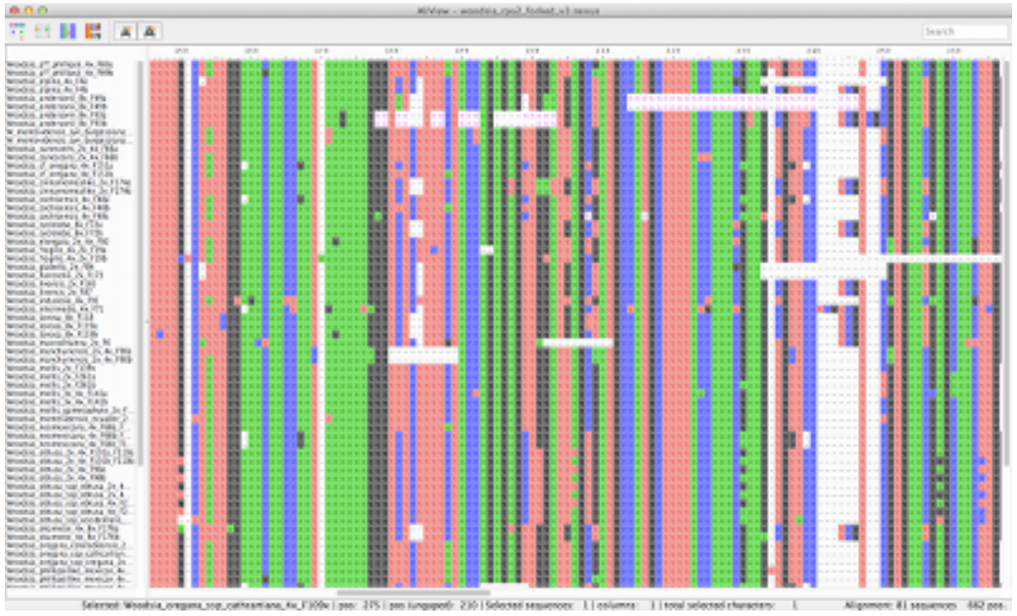
- Can be accurate, but a bit fishy

Mathematical approach

- Dynamic programming (slow, but optimal)

Heuristic methods (fast, but approximate)

- BLAST, short read aligners, CLUSTALW, MAAFT

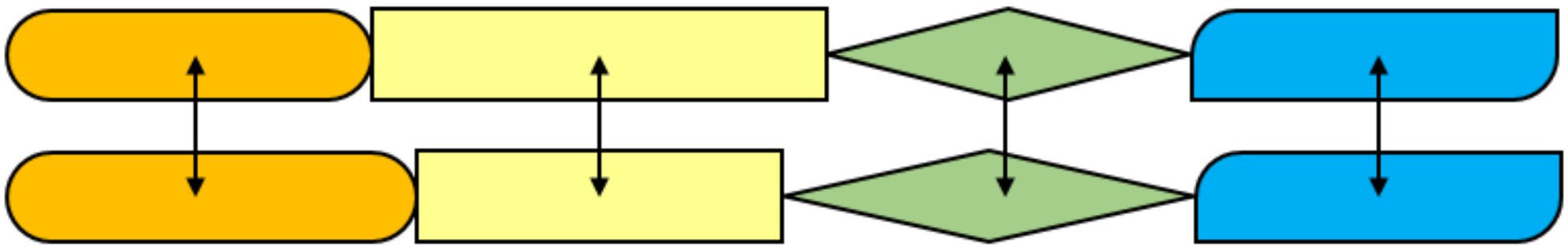


Dynamic programming

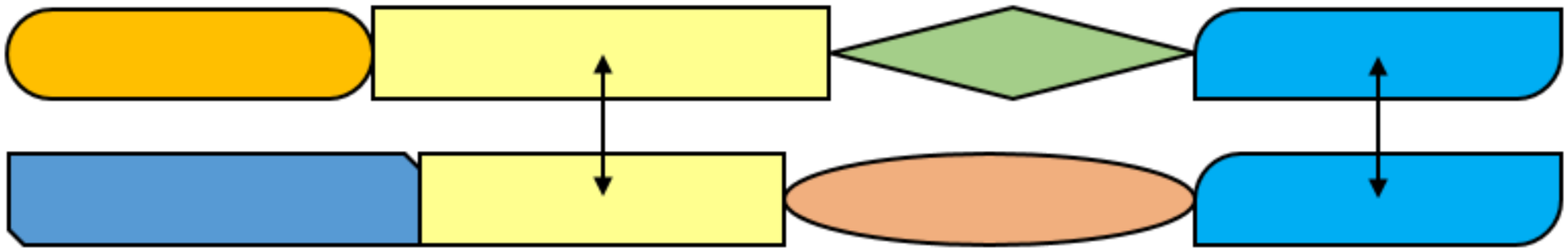
Dynamic programming is a general programming technique

It structures a large search space into a succession of stages

- The initial stage contains trivial solutions to sub-problems
- Each partial solution in a later stage can be calculated by recurring a fixed number of partial solutions in an earlier stage
- The final stage contains the overall solution



Global Alignment



Local Alignment

https://upload.wikimedia.org/wikipedia/commons/0/03/Alignment-Comparison-En.png

Global vs Local alignments

Here's a fun demo of the two main algorithms:

https://gtuckerkellogg.github.io/pairwise/demo/

Global vs Local alignments

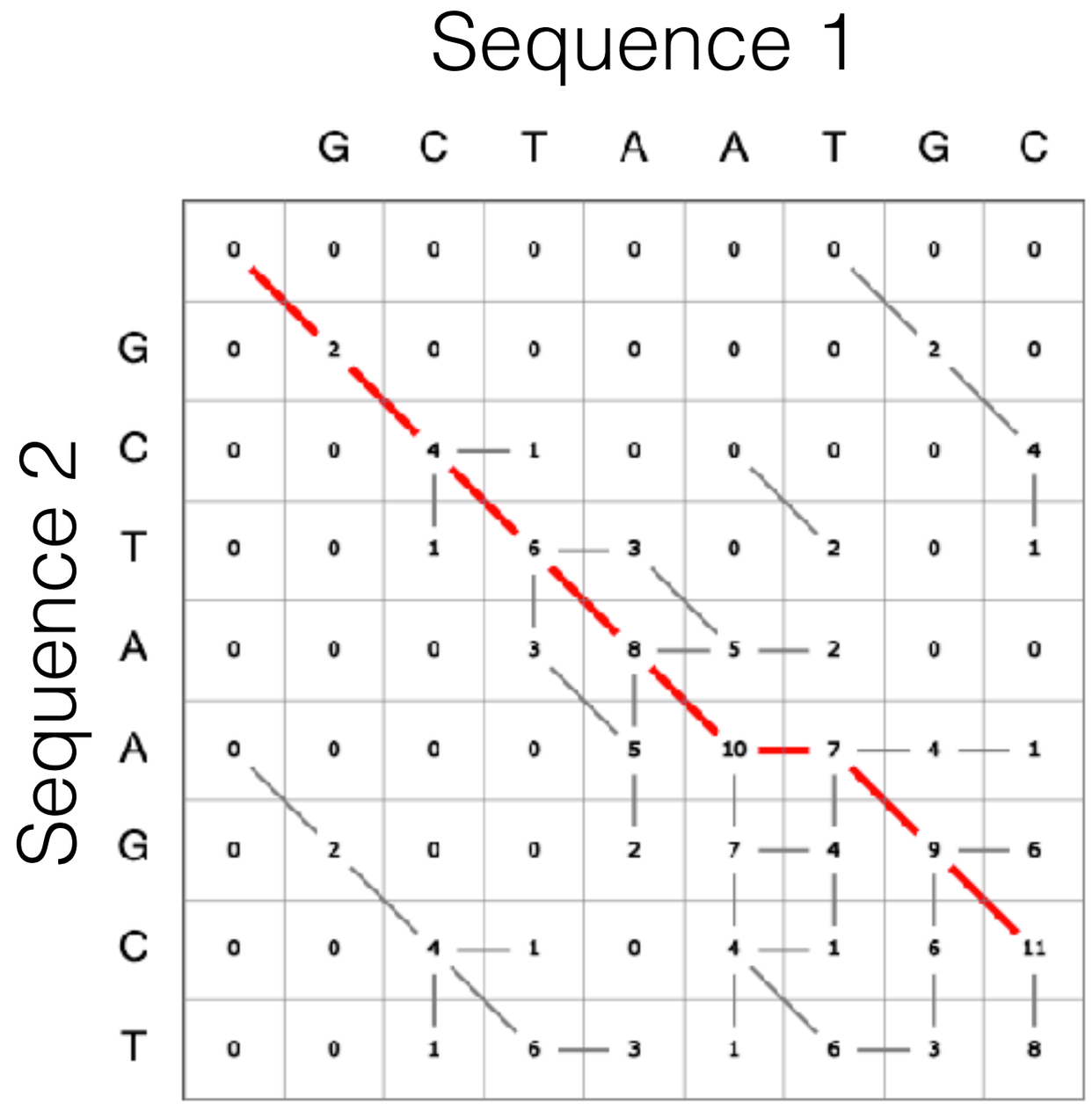
Global alignment algorithms start at the beginning of two sequences and add gaps to each until the end of one is reached (Needleman-Wunsch algorithm).

Local alignment algorithms finds the region (or regions) of highest similarity between two sequences (e.g. the Smith-Waterman algorithm).

Basic principles of dynamic programming

There are too many comparisons to try them all so instead:

- Build alignment path matrix
- Stepwise calculation of score values
- Backtracking (evaluation of optimal path)



Scoring methods

Scoring systems:

- Each symbol pairing is assigned a numerical value, based on a symbol comparison table.
    - nucleotides
    - amino acids (PAM, BLOSUM)

Gap penalties:

- Opening: The cost of introducing a gap.
- Extension: The cost to elongate a gap.

Gap penalties

- Too little gap penalty gives nonsense non-homologous alignments.
- Gaps are common, so too high gap penalty removes real alignments.
- There are multiple gap penalty functions (e.g. constant, linear and “affine”)
- The “affine”is the most commonly used gap penalty function (e.g. BLAST and BWA use it)

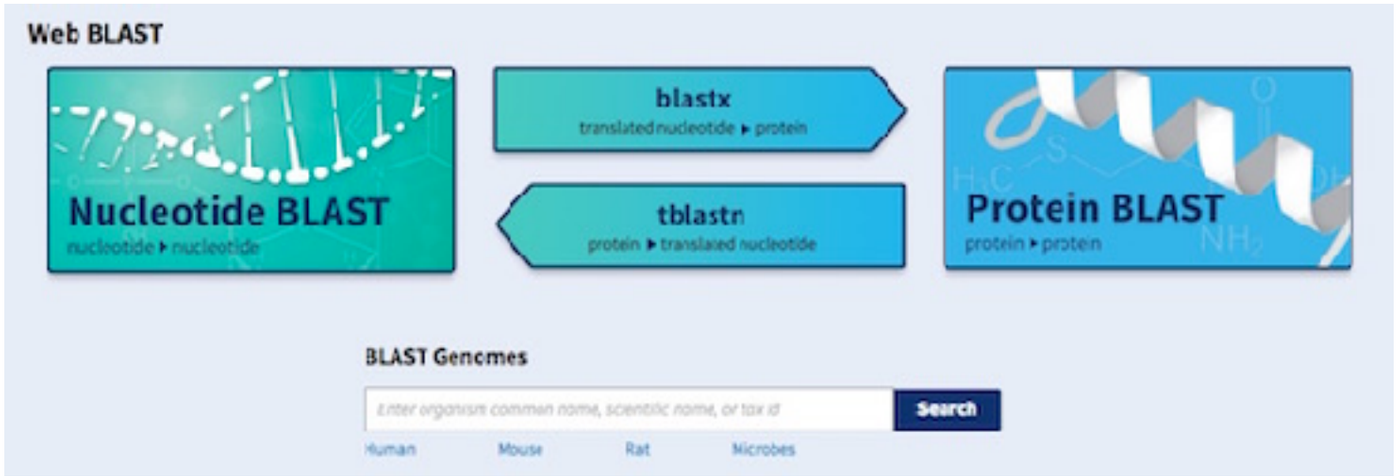
The “affine” gap penalty function

$$GP = A + BL$$

Where A is the penalty for opening a gap, B is the penalty for extending a gap and L is the length of the gap

BLAST - Best Local Alignment Search Tool

A great tool for comparing a small number of sequences against a database (e.g. NCBI BLAST)

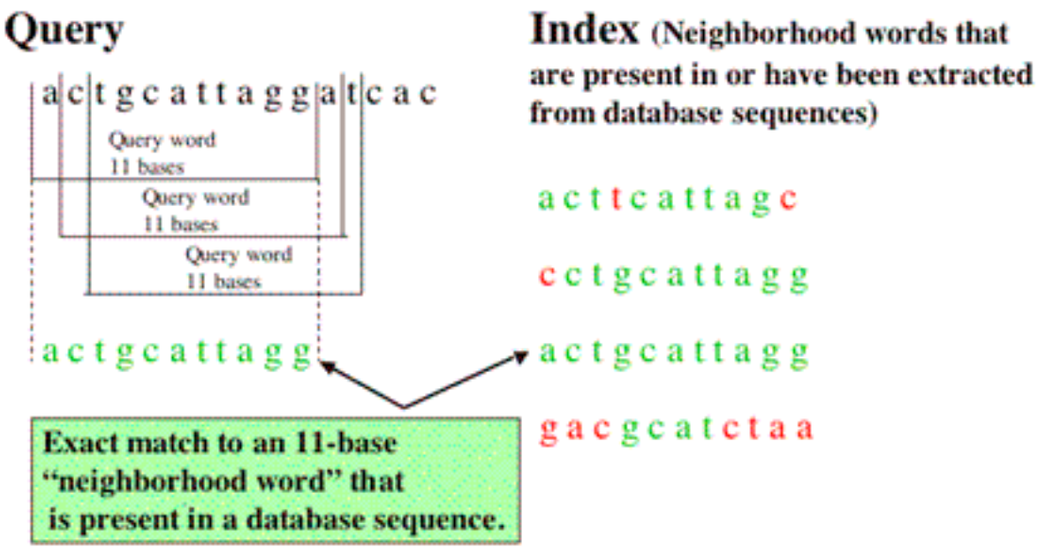


- An example of a “hashed seed-extend algorithm”

BLAST - Best Local Alignment Search Tool

Designed to identify homologous sequences.

First finds highly conserved or identical sequences which are then extended with a local alignment

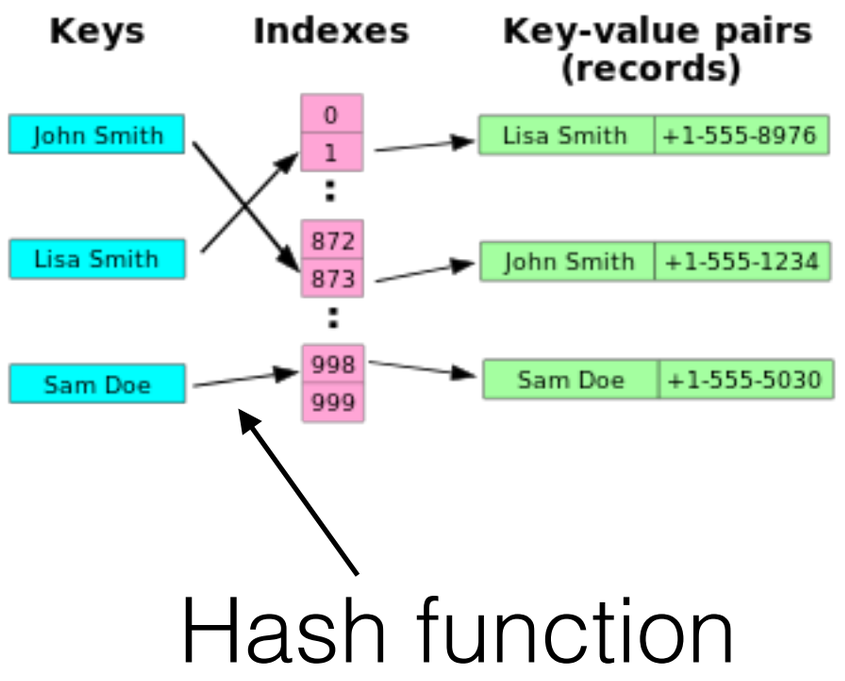


Hashed seed-extend algorithm

- A “hash” is a structure used in computer programming

- It is a way of storing information in a look-up table

- Allows efficient searching



Seed-extend algorithm

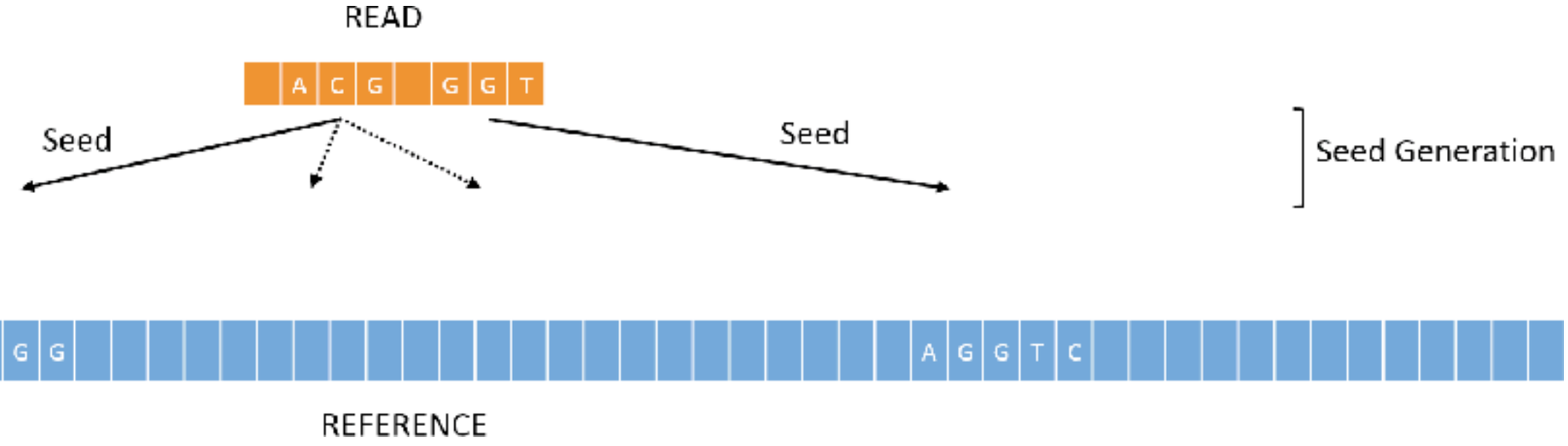
READ

A C G G T

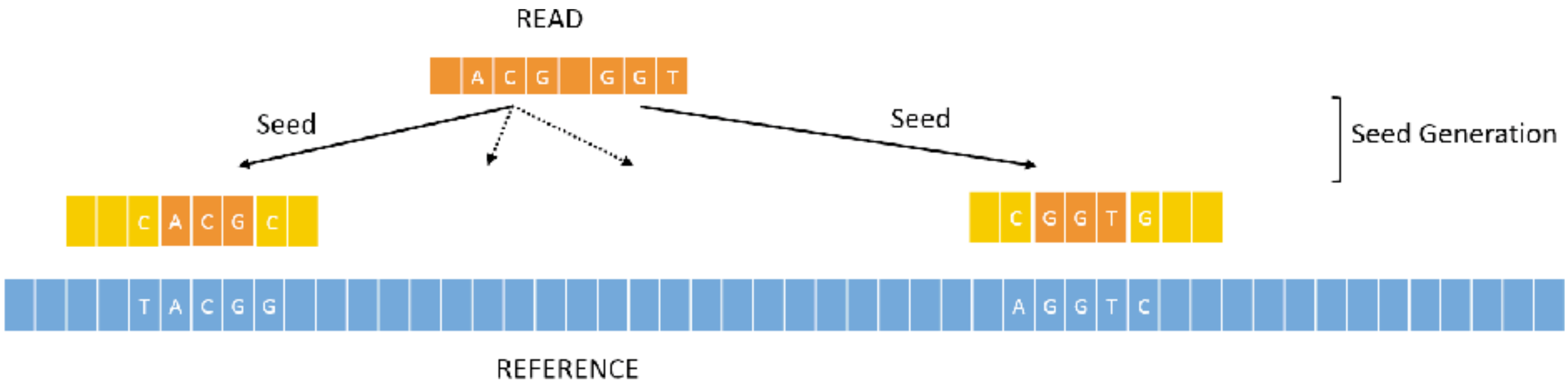
T A C G G A G G T C

REFERENCE

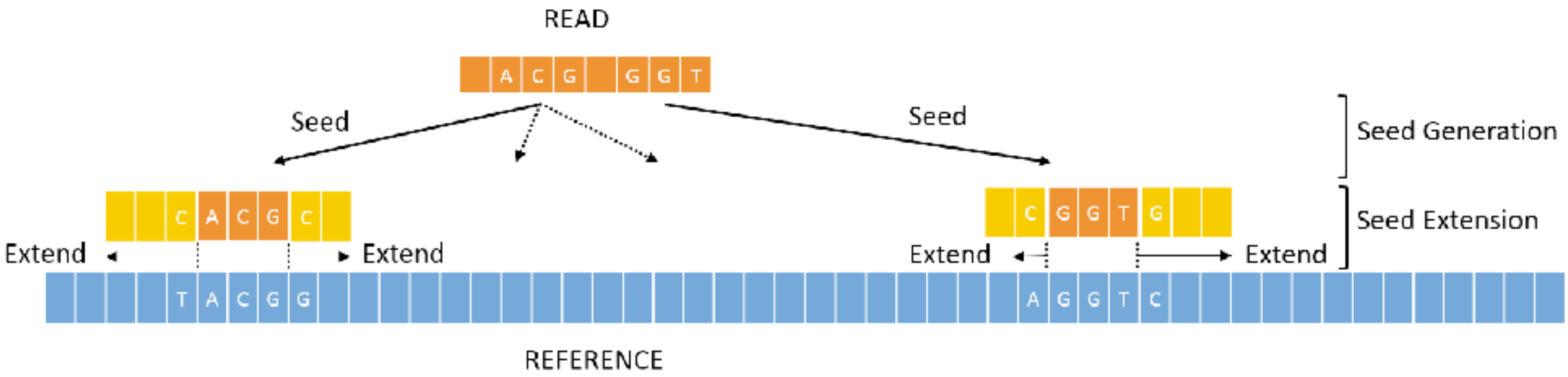
Seed-extend algorithm



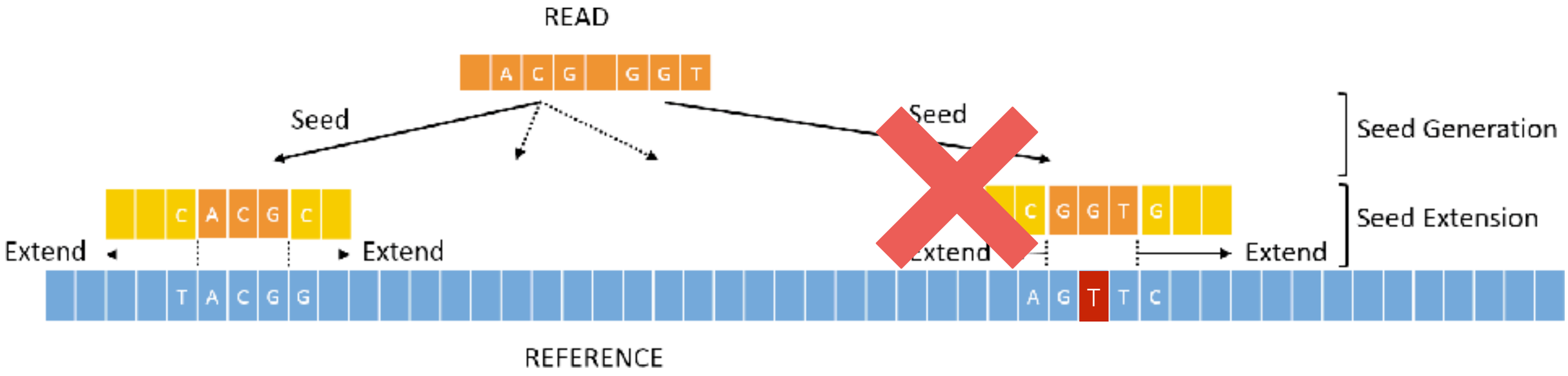
Seed-extend algorithm



Seed-extend algorithm



Seed-extend algorithm



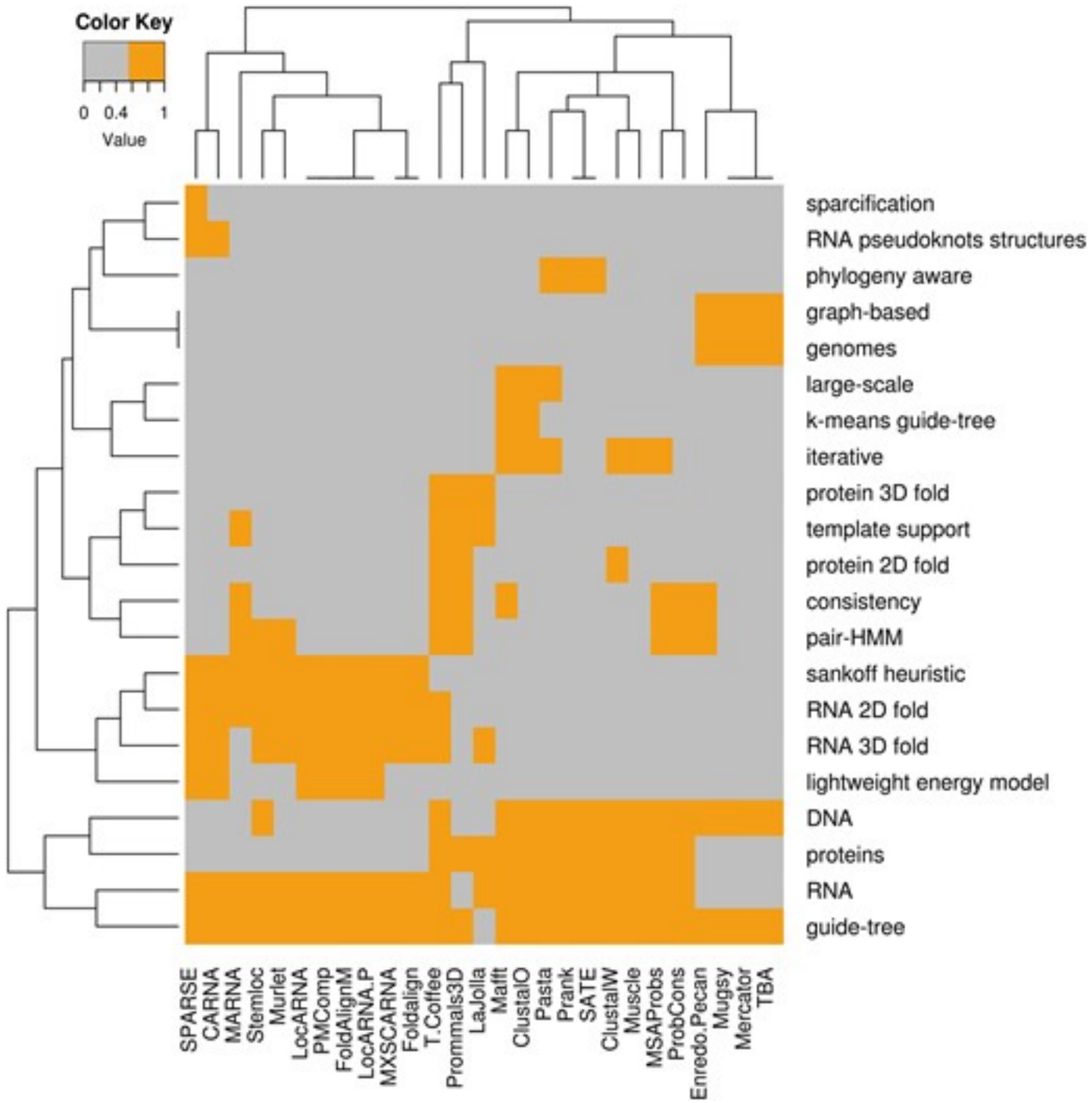
Multiple Alignment

In high-throughput genomics, we are usually using pair-wise alignments (albeit often many millions of pairs)

There are many contexts, we may want to compare multiple sequences simultaneously

This is particularly important in phylogenetics, where multi-sequence alignments are the data from which trees are built

Multiple Alignment



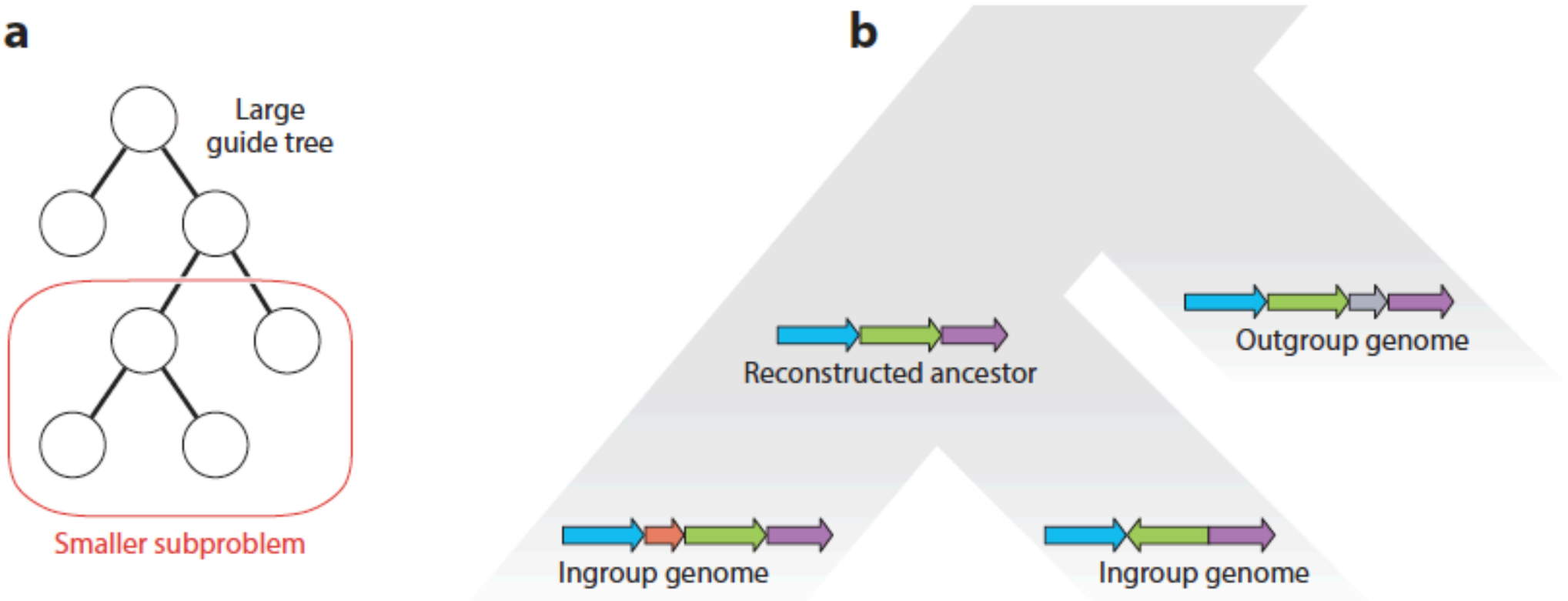
The main algorithmic components of the most widely used software for multiple alignment

Chatzou et al 2016
Briefings in Bioinformatic
PDF on Website

Aligning whole genomes is a different challenge

Increasingly good reference genomes (see Topic 5) gives us a greater ability to compare genomes

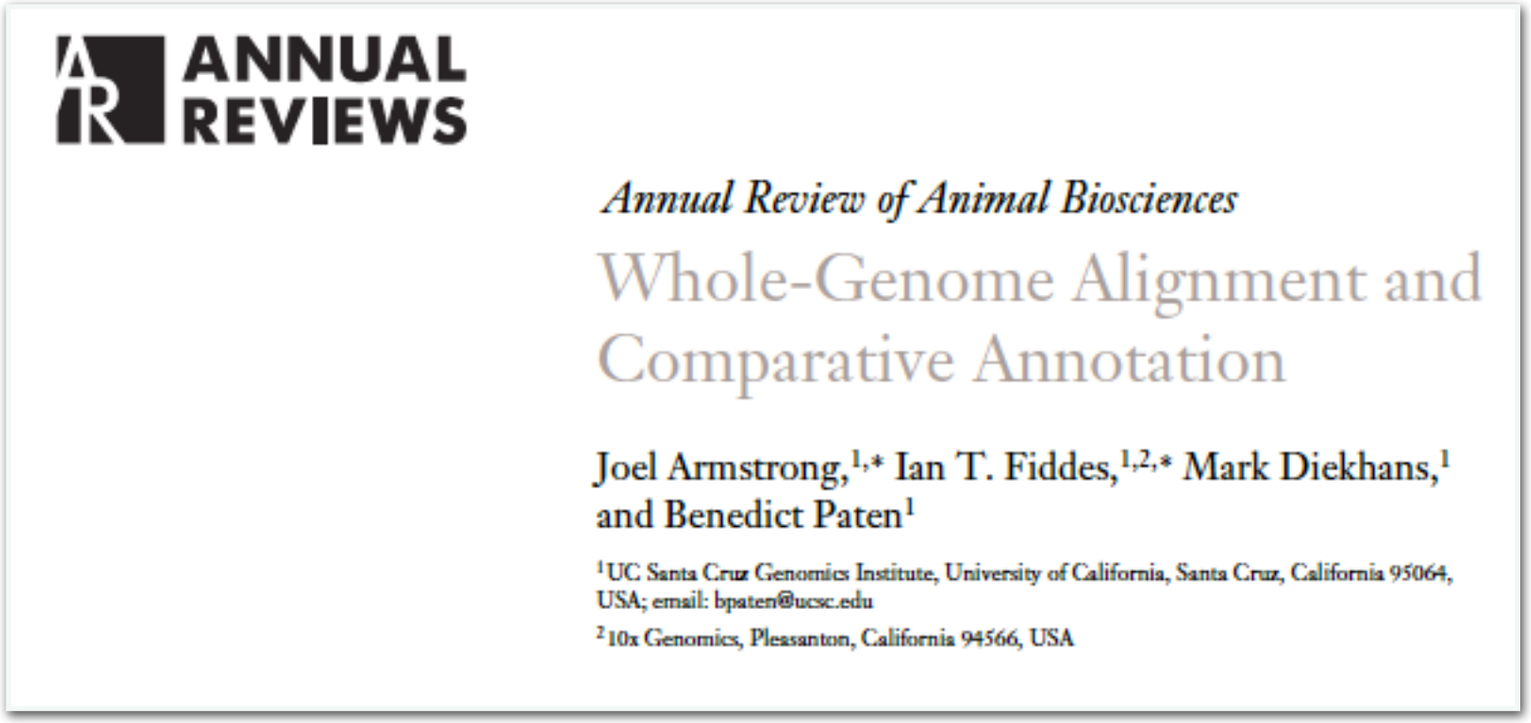
One approach to whole genome alignment is “progressive alignment”:



Alignments done within clades are used to reconstruct ancestral sequences to compare to other clades

Aligning whole genomes is a different challenge

We don't have time to go into it this week, but if you are interested here's a good review to use as a starting point:



ANNUAL REVIEWS

Annual Review of Animal Biosciences

Whole-Genome Alignment and Comparative Annotation

Joel Armstrong,[1,*] Ian T. Fiddes,[1,2,*] Mark Diekhans,[1] and Benedict Paten[1]

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California 95064, USA; email: bpaten@ucsc.edu

[2]10x Genomics, Pleasanton, California 94566, USA

PDF available on the website

Tutorial:
Build k-mers from a sequence
Align sequences using NCBI Blast