Analysis of Sequence Variation

GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2025-05-07

Sequence Alignment

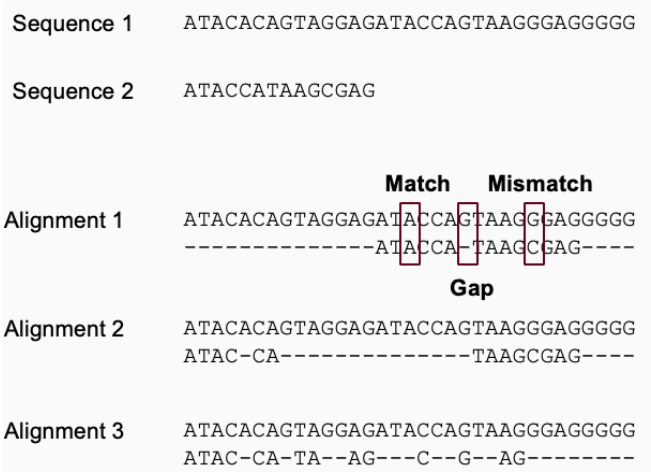
Sequence Alignment

- Provides a measure of relatedness
- Alignment quantified by similarity (% identity)
- Useful for any sequential data type:
  - DNA/RNA
  - Amino acids
  - Protein secondary structure
- High sequence similarity might imply:
  - Common evolutionary history
- Similar biological function

What Alignments Can Tell Us

- Homology - Orthologs, Paralogs
- Genomic identity/origin of a sequence/individual
- Genome/gene structure
- Genic structure (exons, introns, etc)
  - RNA 2D structure
  - Chromosome rearrangements/3D structure

DNA Sequence Alignment Example



Scoring/Substitution Matrices

- Given alignment, how "good" is it?
- Higher score = better alignment
- Implicitly represent evolutionary patterns

	A	C	G	T	-
A	2	-3	-1	-3	-3
C	-3	2	-3	-1	-3
G	-1	-3	2	-3	-3
T	-3	-1	-3	2	-3
-	-3	-3	-3	-3	NA

```
ATACCAGTAAGGGAG
ATACCA-TAAGAGAG
```
Score = 22

```
ATACCAGTAAGG-GAG
ATACCA-TAAG-AGAG
```
Score = 19

```
ATACCA-GTAAGGGAG
A-TACCATAAGAGAG-
```
Score = -20

Sequence Alignment Algorithms

- **Global** alignments - beginning and end of both sequences must align
- **Local** alignments - one sequence may align anywhere within the other
- Multiplicity:
  - Pairwise alignments (2 sequences)
  - Multiple sequence alignment (3+ sequences)

Global Alignment

Both sequences are aligned from end to end

```
AAANTAIYYDPNPDMP A--
   NTAI-YDPN--M-
```

Interior sequences are aligned as well as possible

```
AERAKDNLCRLEHTTLRKVTAAANTAIYYDPNPDMPVVAEDQEWVNVYYEM
A-----N------T-----------AI-YD--P------------N----M
```

However, sequences of vastly different length can produce meaningless alignments

Local Alignment

Alignment may begin and end at any position

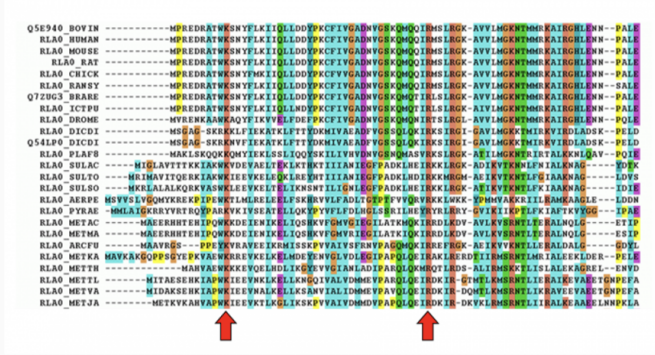
```
AAANTAIYYDPNPDMP -
 AANTAI-YDPN--M-
```

```
AERAKDNLCRLEHTTLRKVTAAANTAIYYDPNPDMPVVAEDQEWVNVYYEM
----------------------AANTAI-YDPN--M----------------
```

Local alignment may produce better alignments when sequence lengths differ greatly

Multiple Sequence Alignment

Like pairwise alignment, but with N sequences



Sequence consensus among many species suggests evolutionary pressure

Methods for Multiple Sequence Aligment (MSA)

1. **Progressive Alignment Algorithms**:
   - *ClustalW*: A widely used progressive alignment tool with a guide tree strategy.
   - *Clustal Omega*: An enhanced version of ClustalW with improved speed and accuracy.
2. **Iterative Alignment Algorithms**:
   - *MAFFT (Multiple Alignment using Fast Fourier Transform)*: Uses iterative refinement with consistency scores.
   - *MUSCLE (Multiple Sequence Comparison by Log-Expectation)*: Utilizes progressive alignment followed by iterative refinement.

Methods for MSA (Continued)

3. **Hidden Markov Models (HMMs)**:
   - *HMMER*: Based on HMMs, used for alignment and homology detection.
   - *SAM (Sequence Alignment and Modeling System)*: Combines HMMs with profiles for database searches.
4. **Probabilistic Alignment Methods**:
   - *ProbCons*: Generates a probabilistic alignment using a Bayesian framework.
   - *PRANK*: Considers sequence and alignment uncertainty in alignment generation.

Methods for MSA (Continued)

5. **Structure-Based Alignment**:
   - *MUSTANG (Multiple Structural Alignment by Secondary Structures)*: Aligns based on protein structures considering sequence and structure.
   - *DALI (Distance Alignment Matrix Method)*: Aligns sequences based on structural similarity.

These methods vary in their approaches and are chosen based on factors such as alignment accuracy, computational efficiency, and the characteristics of the input sequences.

ClustalW: A Common MSA Tool

- ClustalW is one of the most widely used tools for multiple sequence alignment.
- It uses a progressive alignment approach.
- Available as standalone software or through a web server.

Example: Aligning TB genomes

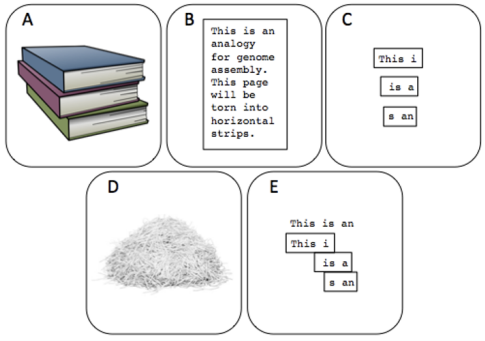
Download the following TB genomes:

- H37Rv
- Mycobacterium tuberculosis str. Erdman
- Combine into single FASTA, first 100 lines:

```
{ head -101 sequence.fasta; head -101 sequence-2.fasta; } \
> combined.fasta
```

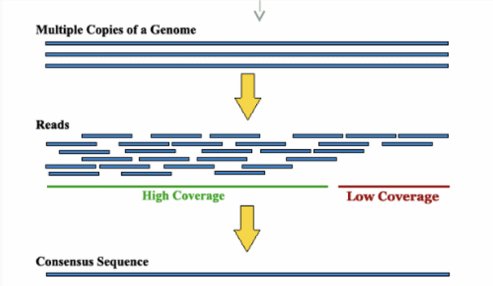
- Analyze using ClustalW

Example: Genome Assembly

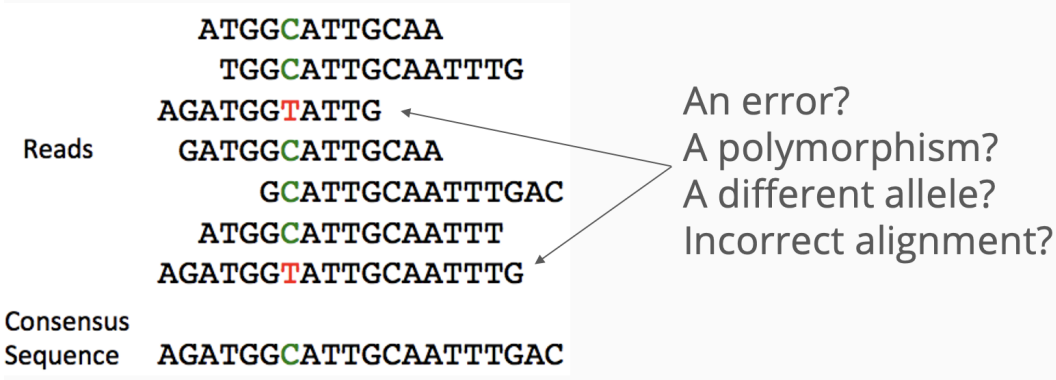


If your genome was a book that had its sentences chopped into fragments, assembly is analogous to reconstructing all the sentences.

We need multiple copies of each book (genome) to arrive at a *consensus* text (DNA sequence) of the original



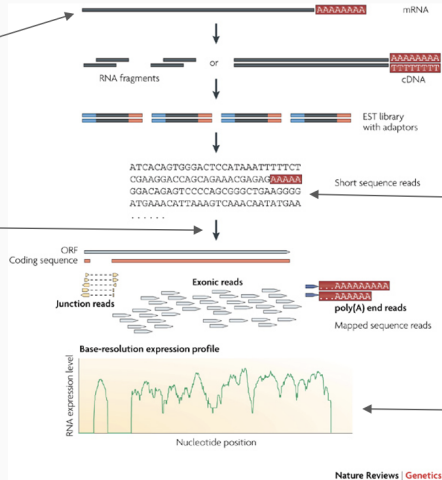
Example: Genome Assembly



Example: mRNA-Seq Analysis

Start with a pool of mRNA molecules

Find all locations where sequences map in genome

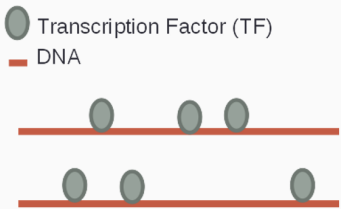


Millions of DNA sequences 30-150 nucleotides long

Count the number of sequences that map to individual regions (e.g. genes)

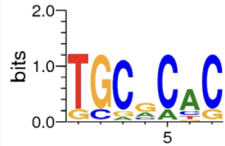
Example: DNA Binding Site Discovery

Identify genomic regions where a particular TF is bound across the entire genome



By extracting and aligning the DNA sequence corresponding to these binding events, we can identify which DNA sequences this TF tends to bind

...TGCTCAC...
...TGCACAC...
...GGCGCAC...
...TGAGCAC...
...TCGCCTC...
...TGCAACG...
...TCCCACG...
...GGTACTC...



Session info

##