Sequence alignment

Meaning and significance

Outline

- What is sequence alignment?
- What are the applications of alignment?
- What does it *mean* to *align* sequences?

What is sequence alignment?

Pattern matching

...CATCGATGACTATCCG...
ATGACTGT

suffix trees, Burrows-Wheeler Transform,...

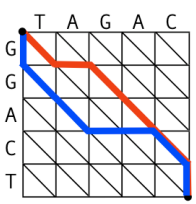
Database searching

CATGCTTGCTGGCGTAAA

CATGCATGCTGCGTAC
CATGGTTGCTCACAAGTAC
CATGCCTGCTGCGTAA
TACGTGCCTGACCTGCGTAC
CATGCCGAATGCTG

BLAST

Optimization problem



Needleman-Wunsch, Smith-Waterman,...

Statistical problem

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int_{\theta'} P(D|\theta')}$$

Pair HMMs, TKF, Karlin-Altschul statistic...

Applications of sequence alignment

- **Sequence assembly** – computing overlaps of reads with sequencing errors
- **Evolutionary functional analysis** – identifying evolutionarily-related sequences and conserved positions
- **Protein structure prediction** – use alignment of query protein to reference protein with known structure

DNA sequence edits

- Substitutions: A**C**GA → A**G**GA
- Insertions: ACGA → AC**CGGA**GA
- Deletions: A**CGGA**GA → AGA
- Transpositions: A**CGG**AGA → AAG**CGG**A
- Inversions: AC**GGA**GA → AC**TCC**GA

Alignment scales

- For proteins and short DNA sequences (gene scale) we will generally only consider
  - Substitutions: cause *mismatches* in alignments
  - Insertions/Deletions: cause *gaps* in alignments
- For long DNA sequences (genome scale) we will consider additional events
  - Transposition
  - Inversion
- In this course we will focus on the case of short sequences

What is a pairwise alignment?

- We will focus on *evolutionary* alignment
- matching of *homologous* positions in two sequences
- positions with no homologous pair are matched with a *space* '–'
- A group of consecutive spaces is a *gap*

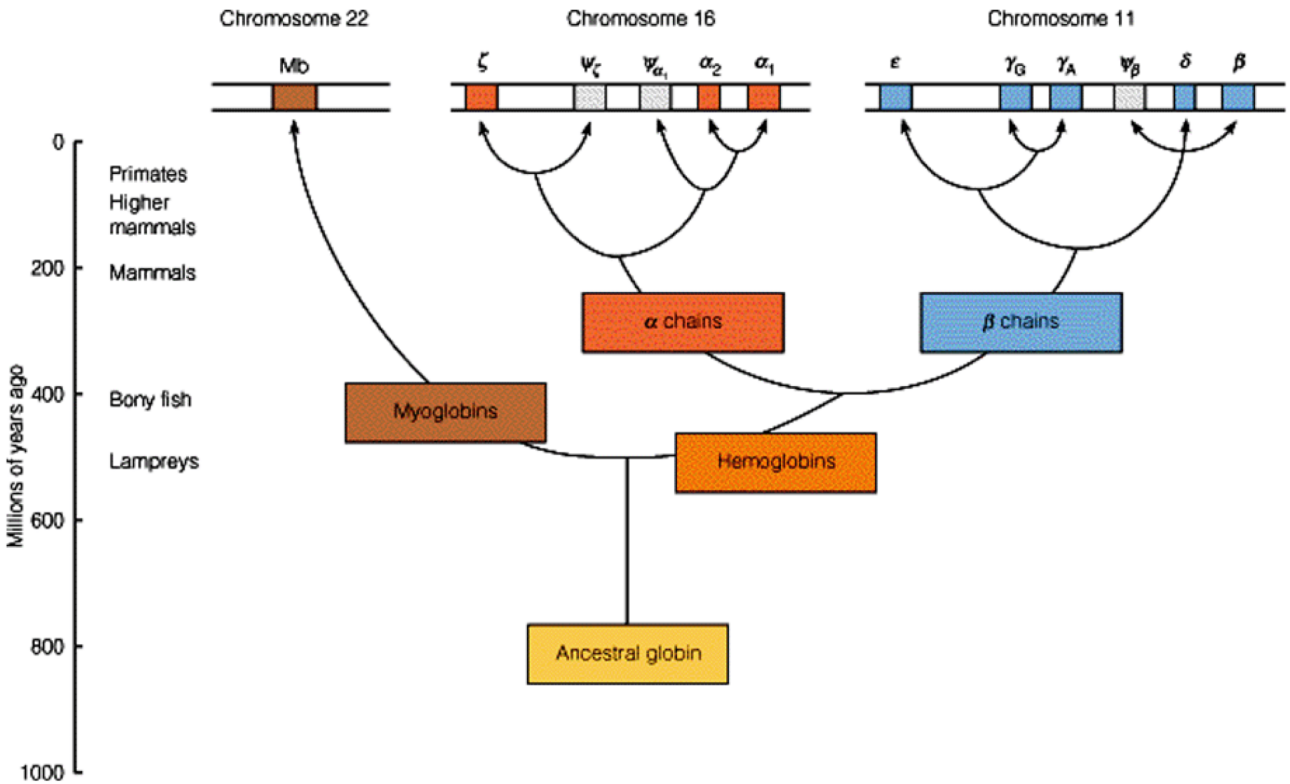
```
CA--GATTCGAAT
CGCCGATT---AT
```

gap

The Role of Homology

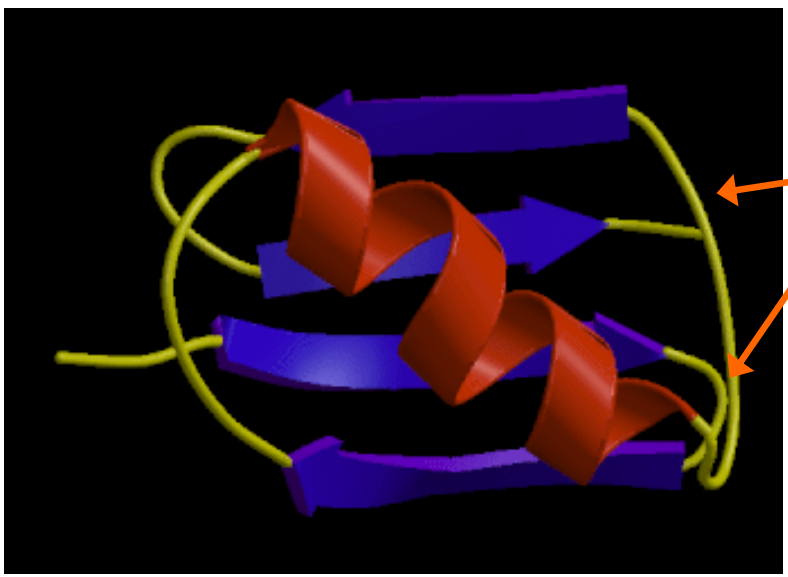
- *character*: some feature of an organism (could be molecular, structural, behavioral, etc.)
- *homology*: the relationship of two characters that have descended from a common ancestor
- homologous characters tend to be similar due to their common ancestry and evolutionary pressures
- thus we often infer homology from similarity
- thus we can sometimes infer structure/function from sequence similarity

Homology Example: Evolution of the Globins



Insertions/Deletions and Protein Structure

- Why is it that two "similar" sequences may have large insertions/deletions?
  - some insertions and deletions may not significantly affect the structure of a protein

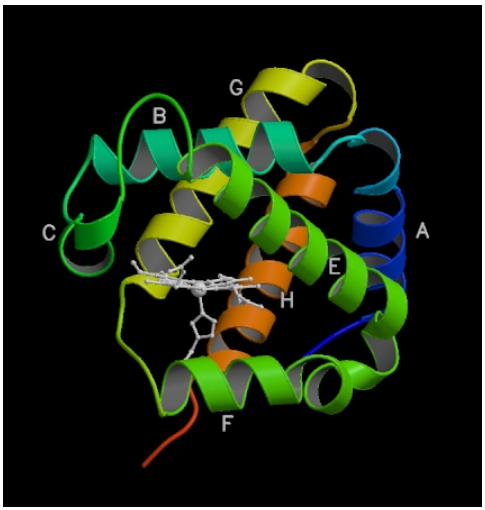


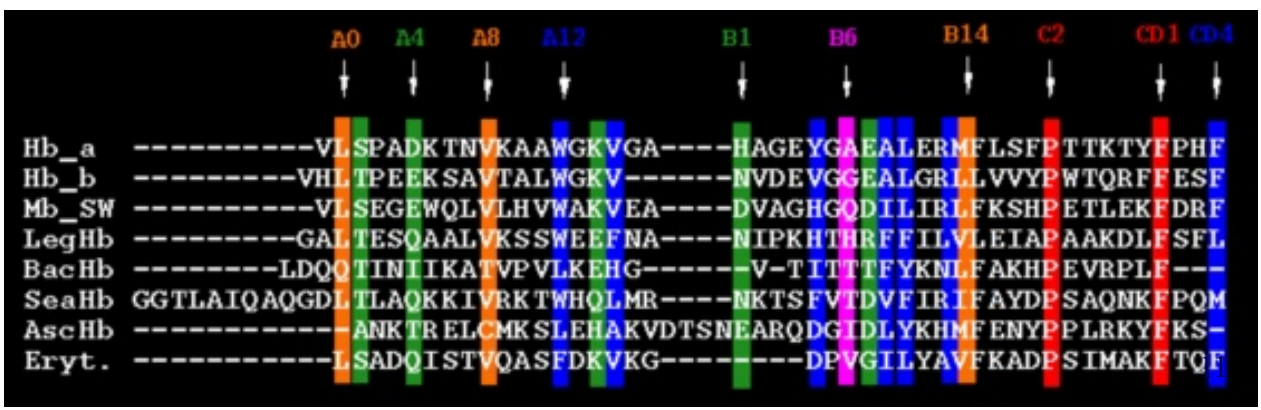
loop structures: insertions/deletions here not so significant

Example Alignment: Globins

- figure at right shows prototypical structure of globins

- figure below shows part of alignment for 8 globins (-'s indicate gaps)





Summary

- Meaning of the term ”sequence alignment”
- The task of evolutionary pairwise alignment
  - Homology
- Representations of alignments
- Sequence edits