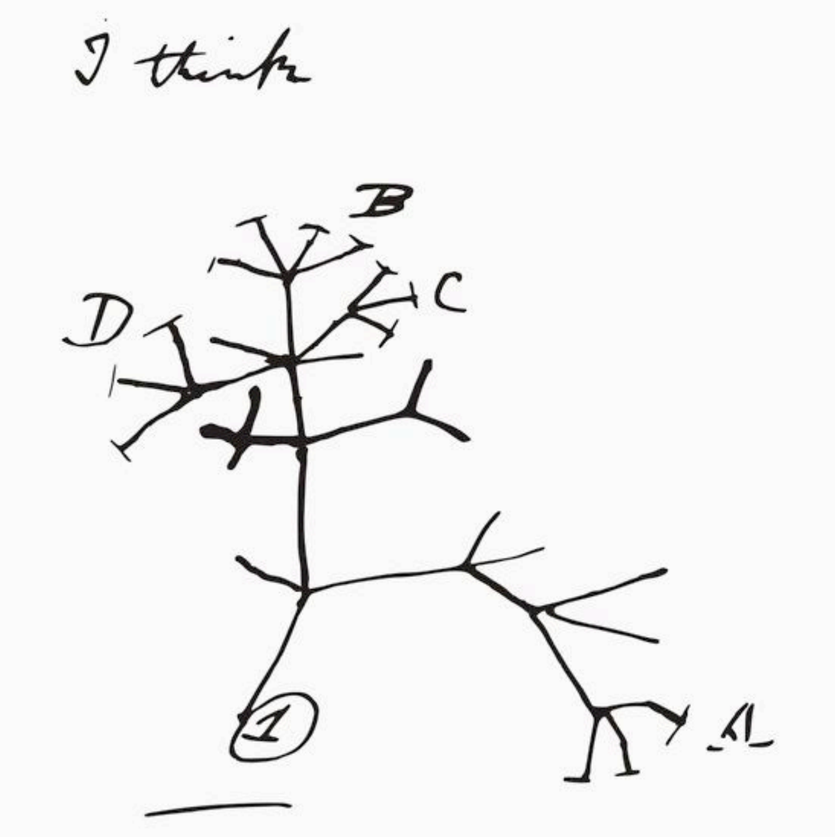
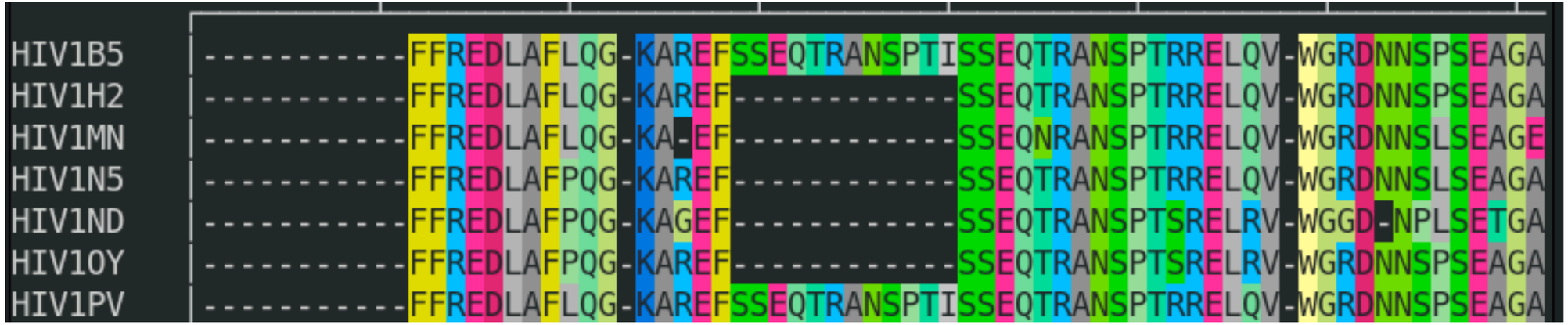
Biological sequence alignment

Jakob Nybo Nissen, 2021-10-21

Sequences are homologous



- Most nucleotide sequences come from copying existing sequences = sequences are related
- Two different sequences can be closely related enough for the similarity to be apparent = homology
- Matching "equivalent" bases in two homologous sequences is called *alignment*
- Fundamental for any comparative analysis



Pairwise alignment

- How should we match sequences? What is a good match?
- Edit distance:

  - cot -> coat
  - cost -> coat
  - coatl -> coat?
- Levenshtein distance algorithms are effective!
  - Used in fuzzy string searches all over the Internet, well-studied problem
  - ... but it's not very biologically plausible :(

Needleman-Wunch algorithm

We can make a dynamic programming algorithm which allows us to add more biologically realistic matching.

Let's begin simple:

- DNA match => 1 point
- DNA mismatch => -1 point
- DNA indel => -1 point

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	G
G								
A								
T								
T								
A								
C								
A								

Needleman-Wunch algorithm

- How many points should a match/mismatch give?

- We can review sequences with a known evolutionary history and *empirically* assign scores proportional to their log-likelihood

```
julia> EDNAFULL
SubstitutionMatrix{DNA, Int64}:
     A  C  M  G  R  S  V  T  W  Y  H  K  D  B  N
  A  5 -4  1 -4  1 -4 -1 -4  1 -4 -1 -4 -1 -4 -2
  C -4  5  1 -4 -4  1 -1 -4 -4  1 -1 -4 -4 -1 -2
  M  1  1 -1 -4 -2 -2 -1 -4 -2 -2 -1 -4 -3 -3 -1
  G -4 -4 -4  5  1  1 -1 -4 -4 -4 -4  1 -1 -1 -2
  R  1 -4 -2  1 -1 -2 -1 -4 -2 -4 -3 -2 -1 -3 -1
  S -4  1 -2  1 -2 -1 -1 -4 -4 -2 -3 -2 -3 -1 -1
  V -1 -1 -1 -1 -1 -1 -1 -4 -3 -3 -2 -3 -2 -2 -1
  T -4 -4 -4 -4 -4 -4 -4  5  1  1 -1  1 -1 -1 -2
  W  1 -4 -2 -4 -2 -4 -3  1 -1 -2 -1 -2 -1 -3 -1
  Y -4  1 -2 -4 -4 -2 -3  1 -2 -1 -1 -2 -3 -1 -1
  H -1 -1 -1 -4 -3 -3 -2 -1 -1 -1 -1 -3 -2 -2 -1
  K -4 -4 -4  1 -2 -2 -3  1 -2 -2 -3 -1 -1 -1 -1
  D -1 -4 -3 -1 -1 -3 -2 -1 -1 -3 -2 -1 -1 -2 -1
  B -4 -1 -3 -1 -3 -1 -2 -1 -3 -1 -2 -1 -2 -1 -1
  N -2 -2 -1 -2 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1
```

Needleman-Wunch algorithm

- What about gaps (deletions / insertions)?
- Biologically, deletions are rare, larger deletions are only somewhat rarer.
- First gap symbol is expensive (say, -12), subsequent gaps are cheaper.
- We call this *affine gap score model*

- By disallowing negative scores and beginning at argmax, we get Smith-Waterman algorithm
- Both S/W and N/W are provably optimal! But not 100% biologically accurate.
- There have been developed better algorithms since...
    - ... But all the ones I know about are just fast approximations of S/W!

BioAlignments.jl

Abstracts over pairwise alignment with 4 parameters:

Algorithm (N/W)

Query/subject seq

Model (affine gap score model)

```
julia> pairalign(
           GlobalAlignment(),
           dna"TAGCTAG", dna"TACCAG",
           AffineGapScoreModel(EDNAFULL, gap_open=-12, gap_extend=-2)
       )
PairwiseAlignmentResult{Int64, LongDNASeq, LongDNASeq}:
  score: 7
  seq: 1 TAGCTAG 7
         || | ||
  ref: 1 TACC-AG 6
```

Internal alignment layout

```
julia> dump(x, maxdepth=4)
PairwiseAlignmentResult{Int64, LongDNASeq, LongDNASeq}
  value: Int64 7
  isscore: Bool true
  aln: PairwiseAlignment{LongDNASeq, LongDNASeq}
    a: AlignedSequence{LongDNASeq}
      seq: LongDNASeq
      aln: Alignment
        anchors: Array{AlignmentAnchor}((6,))
        firstref: Int64 1
        lastref: Int64 6
    b: LongDNASeq
```

Output is a little simplified

Wait, why store it internally as an `Alignment` instead of just two `LongDNASeq` with gap symbols?

- No need to copy both sequences
- What if the sequences use an alphabet without gap symbols?

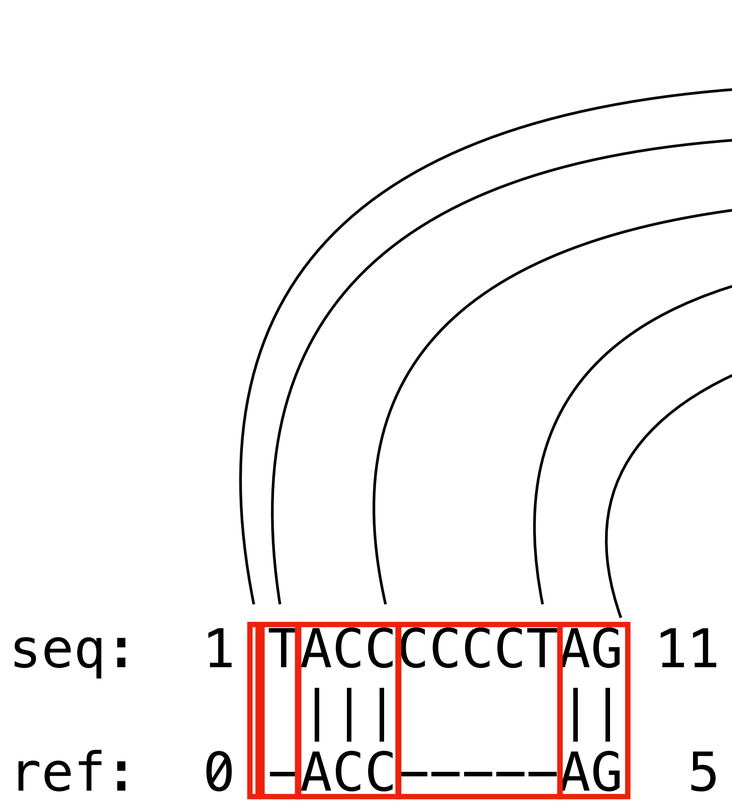
What's in an Alignment anyway?

Internal alignment layout

```
aln: Alignment
  anchors: Array{AlignmentAnchor}((5,))
  firstref: Int64 1
  lastref: Int64 6
```

?????

```
5-element Vector{AlignmentAnchor}:
AlignmentAnchor(0, 0, '0')  OP_START
AlignmentAnchor(1, 0, 'I')  OP_INSERT
AlignmentAnchor(4, 3, '=')  OP_SEQ_MATCH
AlignmentAnchor(9, 3, 'I')  OP_INSERT
AlignmentAnchor(11, 5, '=') OP_SEQ_MATCH
```



Internal alignment layout

```
5-element Vector{AlignmentAnchor}:
 AlignmentAnchor(0, 0, '0')  OP_START
 AlignmentAnchor(1, 0, 'I')  OP_INSERT
 AlignmentAnchor(4, 3, '=')  OP_SEQ_MATCH
 AlignmentAnchor(9, 3, 'I')  OP_INSERT
 AlignmentAnchor(11, 5, '=') OP_SEQ_MATCH
```

- A type of run-length encoding
- This can be written "1I3=5I2=".
- Known as a CIGAR string, "Concise Idiosyncratic Gapped Alignment Report"
- We will see CIGARs in next exercise!

Alignment file formats

- Multiple poorly-specified formats (of course) with bizarre and arbitrary restrictions
- Just stick with FASTA: Insert gap symbol "-" as padding to make the sequences aligned

```
>Seq1
-AGCCTAGGAGAA
>Seq2
TAGC--AGGAAGA
```

- In BioSequences.jl, alphabet types have a gap symbol (`DNA_Gap`, `RNA_Gap`, `AA_Gap`) for this purpose

Questions?

Exercise 4