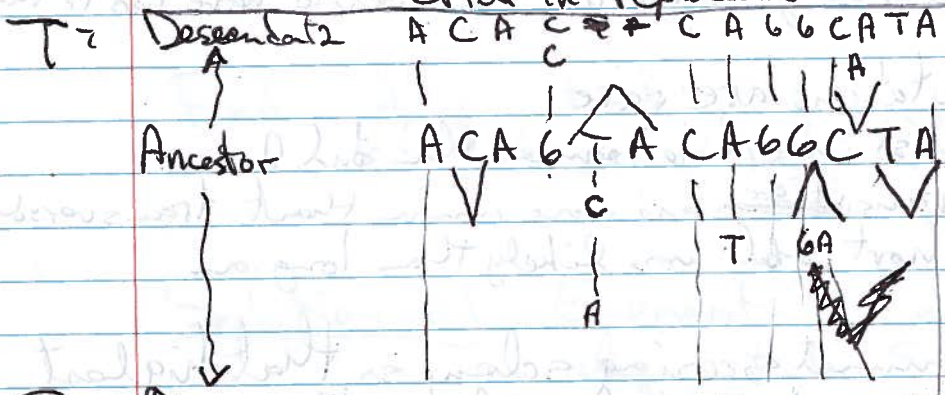
Sept 12 2016

Sequence alignment

Mutation: Change in sequence due to error in replication or unrepaired DNA damage



Substitution: Change one nuc for another

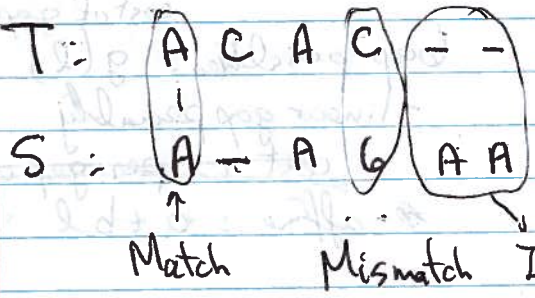
A ↔ 6 } transition
C ↔ T } common

A ↔ C, 6, T transversion (~3 times less common)

Insertion: Insert 1 or more nuc.

Deletion: Delete one or more

Alignment of S, T is obtained by inserting gaps in S and T to obtain S' and T', so that in S' and T', nucleotides derived from same ancestor occur at same position. Gaps reflect nucleotides missing due to del. or insertion



Problem: Ancestor and mutation scenarios are generally not known.

Idealized Pairwise Global Alignment Problem

Given: S, T
Find: Alignment of S and T that is most likely to reflect the evolutionary scenario that would have led to them

Key idea:
1. Mutations are rare
2. Subst. are more common than indel
3. Transitions are more common than transversions
4. Short indel more likely than long ones

⇒ Define alignment scoring scheme so that highest scoring alignment is the one that is the most likely to be correct.

Scoring a given alignment

```
S: A C A G T C - - T A
T: A T A - T A A A T A
   1 -1 1 -2 1 -1 -2 -2 1 1 : -3
```

Score: Subst. score + indel score

Subst. fct. or cost matrix

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} +1 & -2 & -1 & -2 \\ -2 & +1 & -2 & -1 \\ & & \ddots & \\ & & & \ddots \end{pmatrix} \end{array}$$

Gap penalties: $g(\ell)$ = cost of gap of length ℓ

- linear gap penalty: $g\ell = -2\ell$ ~~cost c per gap character~~
- affine: $a + b\ell = -5 - 1\ell$

Pairwise seq. Aln. Problem (linear gap penalty)

Given: $S = S_1 \ldots S_m$, $T = T_1 \ldots T_n$ sequences
M subst cost matrix
c

Find: Alignment of S, T
s.t that score(Aln(S,T)) is maximized

Solution #1: Enumerate and evaluate all possible alignments for S and T, report best

Problem: Way too slow $\Omega(2^{m+n})$

Solution #2: Needleman-Wunsch algo. (1970)

S = AGCT
T = ACA