Lecture 4: Introduction Sep 13

Lecturer: Yu Li

What are the sequence data?

-DNA: Composed of A,T,C,G, complementary double strand, approximately 3 billion of these base pairs

-RNA of AUCG

-Protein sequence: Usually composed of 20 amino acids ØMultiple sequence alignment

How can we find the best alignment?

-Straightforward way: enumeration >> list down all the possibilities >> count the score >> find the highest

-Problem: too many alignments > need dynamic programming

How do we do and why do we need Dynamic Programming?

-Definition of sequence similarity:

Match: A <-> A
Mismatch (Substitution): G <-> T
Gap (Insertion or deletion): C <-> _

-Sequence alignment score:

Scoring matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10

-Break main problems into sub-problems

-Solve the sub-problems optimally and recursively

- Utilize these optimal solutions to build the best overall solution for the initial problem

-Take a flight problem example:

Cheapest (KAUST, CUHK) = Cheapest { Cheapest (KAUST, Qatar) + Cheapest (Qatar, CUHK) ; Cheapest (KAUST, Dubai) + Cheapest (Dubai, CUHK) ; Cheapest (KAUST, GZ) + Cheapest (GZ, CUHK) }

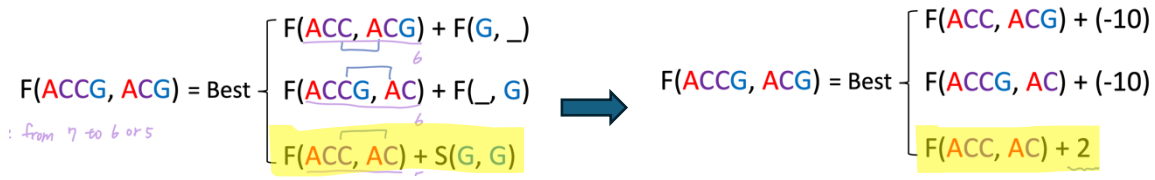
-Direct flight is expensive

-Divide it into several connecting flights

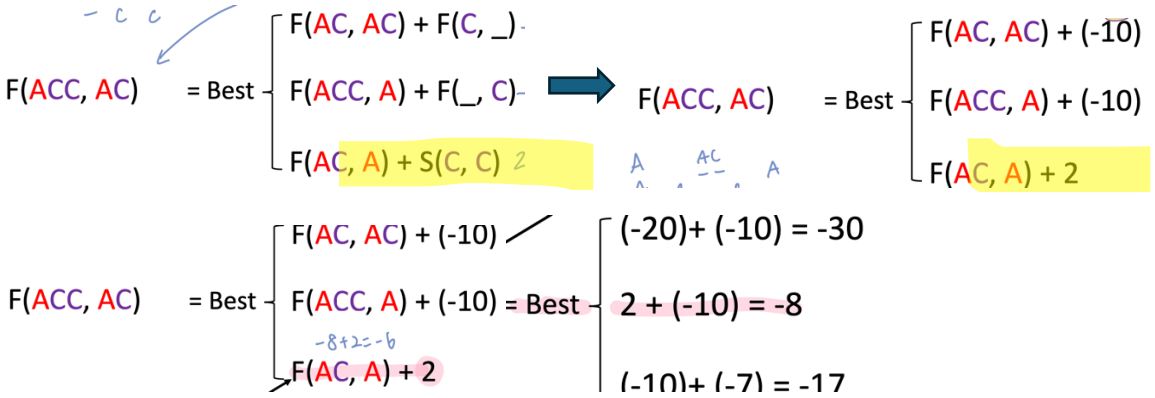
-Compare each trip/way and choose the cheapest one

ACCG, ACG for example, there are 7 bases, so a maximum 6 bases

1. Consider the last pair
   According to the scoring matrix:



Highest score:F(ACC, AC)+2, Devide F(ACC, AC) to



The optimal alignment is ACCG AC_G or ACCG A_CG

The table represents the method:

		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6	-16
G	-30	-18	-6	-3	-4

F(A, A)

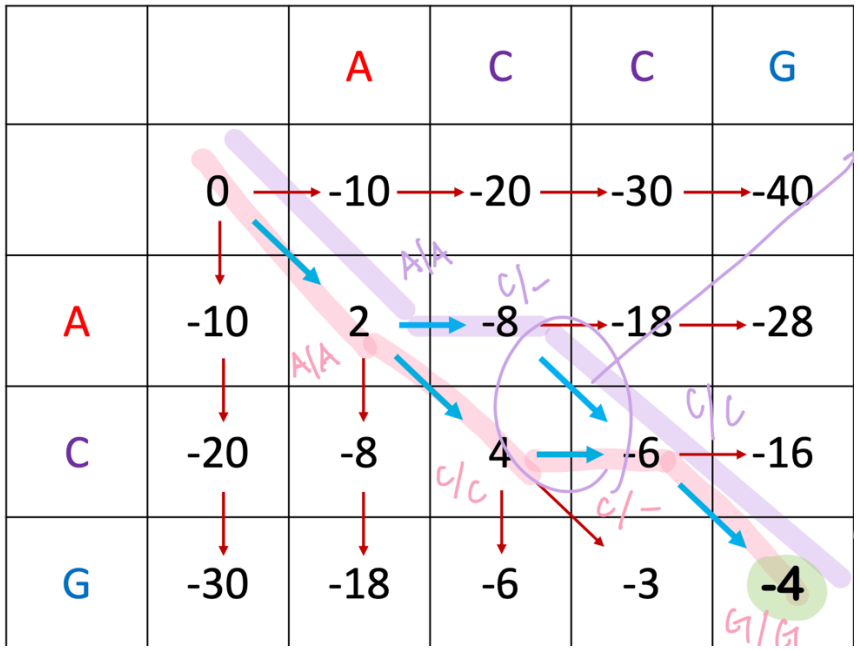
F(A, A)+F(C, C)

F(AC,AC)+F(C_) or

F(AC,A)+F(C,C)

F(ACC,AC)+F(G,G)

Follow the two paths,



Two optimal alignments are ACCG A_CG and ACCG AC_G