2019-04-15 Sequence Alignment

Monday, April 15, 2019 3:01 PM

- Similar idea to edit distance algorithm
- Goal of sequence alignment is to align two strings based on three factors
  - Reward for a match
  - Penalty for a mismatch
  - Penalty for a "gap"
- Commonly used in genetics (aligning DNA sequences), activity recognition (AI / ML)

Example:
1. AABABBA
2. ABAABA

A	A	B	A	B	B	A
A	-	B	A	A	B	A

no transformations, just alignment

possible result when mismatch isn't so bad or when gap isn't a problem

A	A	B	A	-	B	B	A
A	-	B	A	A	B	-	A

possible result when mismatch is very bad

Test of Knowledge

Match: 1
Mismatch: -1
Gap: -2

A	B	B	A
B	A	B	-

Score: -3

Match: 1
Mismatch: -1
Gap: -1

A	B	B	A	-
-	B	-	A	B

Score: -1

Bottom-Up Matrix

Gap = -2, mismatch: -1, match: 1

		B	A	B
	0	-2	-4	-6
A	-2	-1	-1	-3
B	-4	-1	-2	0
B	-6	-3	-2	-1
A	-8	-5	-2	-3

Above us: take the gap
Left of us: take the gap
Diagonal: take match or mismatch

B	A	B	-
A	B	B	A

Bottom-Up Matrix

Gap = -1, mismatch: -1, match: 1

		B	A	B
	0	-1	-2	-3
A	-1	-1	0	-1
B	-2	0	-1	1
B	-3	-1	-1	0
A	-4	-2	0	-1

-	B	-	A	B
A	B	B	A	-

- Unlike levenshtein, needleman-wunch restricts back paths.
- Thus, we need to track not only cost of each cell, but how we got there as well
    - To do so, don't store INTs in your mem, store complex classes that have:
        - Direction pointers AND cell value