Add simple logo if possible

SubLiMat

Substitution matrix benchmarking with pairwise alignment

Problem statement

- Substitution matrices are hypothesis of evolutionary change.
- There are many substitution matrices used in the literature.
- Their effects in different types of sequences isn’t easy to determine

T C T T A G

5’ A G A A T C 3’ A

5’ A C A A T C 3’ B

AGAATC_

AAATCC

Goals

- Given two genetic sequences of size n:
- Generate a pairwise alignment: match nucleotides based on a substitution matrix. Get a score
- Evaluate multiple substitution matrices: Rank the quality of alignments produced by different substitution matrices.

Symbols

Symbol	Meaning
A,G	Purine nucleotides
T,C	Pyrimidine nucleotides
↔	Transition
/↔/	Transversion
JC	Jukes Cantor model
K80	Kimura model
TN93	Tamura Nei model
HKY85	Hasegawa-Kishino-Yano model
Mb	Mega bases

Terminology

Term	Definition
SNP	Atomic positional unit that carries DNA genetic information
Codon	Atomic positional unit that carries protein information
Genetic Unit of Information (GenUI)	Atomic positional unit of genetic information (includes SNP and Codon)
Quality	The capacity of an alignment to maximize the pairing of same GenUIs, while avoiding mismatches.

Example of calculation

Add multiple cases. One default, one easy, One hard

Case 1. base case

Input

```
sequence1 = “A"
sequence2 = “A"
```

```
   A   T   G   C
[0,-3,-1,-3], # A
[-3,0,-3,-1], # T
[-1,-3,0,-3], # G
[-3,-1,-3,0]  # C
```

Output

matrix	score
JC	1.0
K80	1.0
HKY85	1.0
TN93	1.0
baseline	0.0

Example of calculation

Add multiple cases. One default, one easy, One hard

Case 1. easy case

Input

```
sequence1 = “G"
sequence2 = “A"
```

```
   A   T   G   C
[0,-3,-1,-3], # A
[-3,0,-3,-1], # T
[-1,-3,0,-3], # G
[-3,-1,-3,0]  # C
```

Output

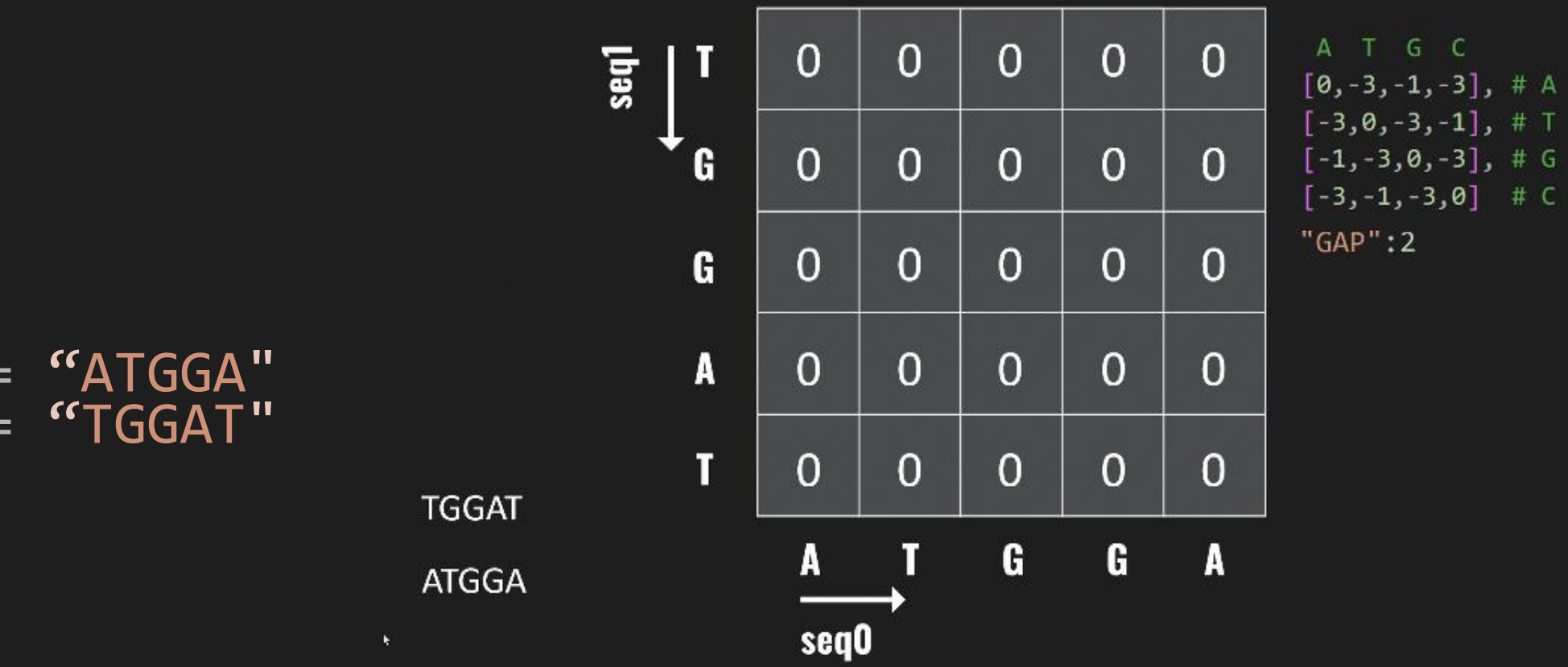
matrix	score
JC	-0.33
baseline	-1.00
K80	-1.00
HKY85	-1.00
TN93	-1.00

Example of calculation

Case 1. algorithm

Input

```
sequence1 = “ATGGA"
sequence2 = “TGGAT"
```



Assumptions

Each genUI exists only once for a given alignment.

A genUI is only relevant by its relative position to other such units.

A strand is always read from the 5’ to 3’ direction.

The quality of an alignment is a strictly comparative metric.

The cost of inserting a gap is constant among substitution matrices.

Inputs and outputs (Data constraints)

Var	Physical Constraints	Software Constraints	Typical Value	Uncert.
seq_A	$seq_A \geq 1$	$\lvert seq_A \rvert \approx \lvert seq_B \rvert$	$\lvert seq_A \rvert = 10^3 \text{genUI}$	10%
seq_B	$seq_B \geq 1$	$\lvert seq_B \rvert \approx \lvert seq_A \rvert$	$\lvert seq_B \rvert = 10^3 \text{genUI}$	10%
S	$S \in \mathbb{R}^{n \times n}, n \geq 4$	$S \in \mathbb{R}^{n \times n}, n \geq 0$	$S \in \mathbb{R}^{4 \times 4}$	10%
F	$F \in \Sigma^{\lvert seq_i \rvert \times \lvert seq_j \rvert}$	$\lvert seq_i \rvert, \lvert seq_j \rvert \geq 1$	$F = 10^3 \times 10^3$	0%
d	$d \in \mathbb{R}$	$d > 0$	$d = -2$	
O_{AB}				

Inputs and outputs (Data constraints)

Var	Definition	Units
seq_A	String of genUIs of arbitrary size to be compared with seqB	SNP/Codon
seq_B	String of genUIs of arbitrary size to be compared with seqA	SNP/Codon
S	Substitution Matrix encoding penalties and rewards for genUI transformations.	--
F	Comparative matrix of seqA and seqB	Index
d	Penalty associated with gap insertions	--
O_{AB}	1D vector encoding the performance of multiple substitution matrices	--

Theoretical models

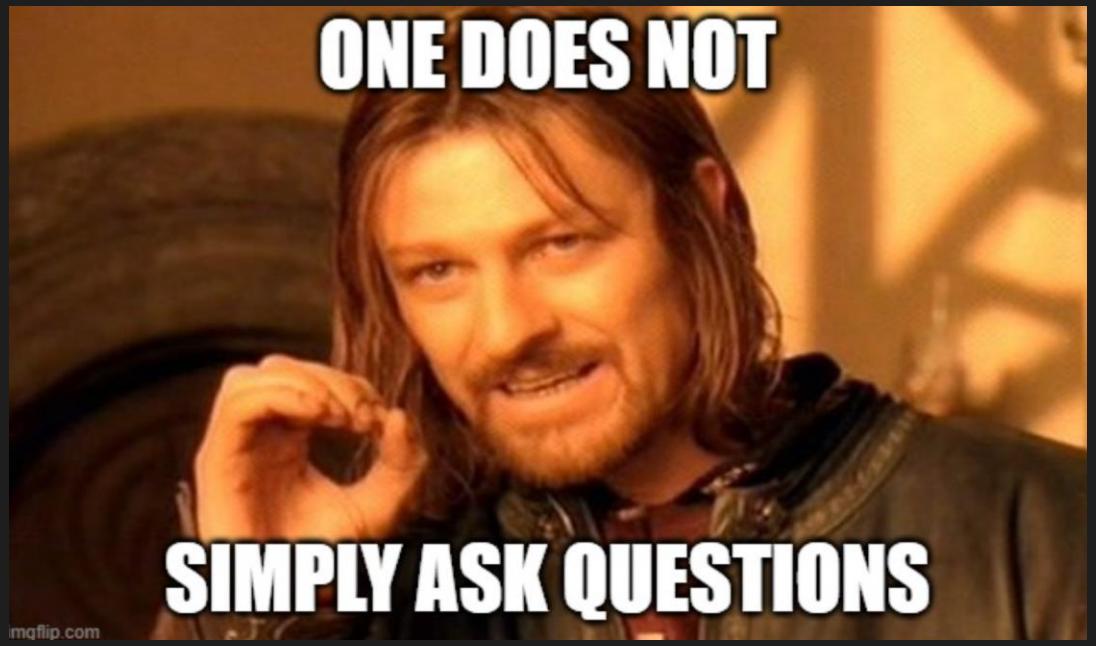
Label	**Needleman-Wunsch recursive model**
Equation	$F_{ij} = \max(F_{i-1,j-1} + S(x_i, y_j), F_{i,j-1} + g, F_{i-1,j} + g)$
Description	g is gap penalty x is a genetic sequence y is a genetic sequence F is 2D comparative matrix of sequences A and B S is a substitution matrix *i* is the index mapping to position in matrix A *j* is the index mapping to position in matrix B
Notes	Recursive function for the pairwise alignment of two genetic sequences to return the best global solution.
Source	wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

Instanced models

Label	Traversing substitution matrices with pairwise alignment
Input	seq_A, seq_B, d
Output	O_{AB}
Input constraints	$\lvert A\rvert, \lvert B\rvert > 0$
Output constraints	
Equation	

$$\forall S \in \mathcal{S} : F_{ij} = \max(F_{i-1,j-1} + S(A_i, B_j), F_{i,j-1} + d, F_{i-1,j} + d)$$

Questions



Questions

Areas where my projects struggles

- Too simple?
- Hard to find more inputs/outputs
- Where should I include the iterative process that involves testing multiple substitution matrices?
- If the substitution matrix S is NOT part of the input, does it still have to be added to the Instanced models?