WASEDA UNIVERSITY

Bioinformatics Homework Assignments

2023 Homework Assignments

Name: Li Dongchen
Furuzuki Laboratory
Email: lidongchen@fuji.waseda.jp
早稲田大学 情報生産システム研究科
Graduate School of Information, Production and Systems, Waseda University

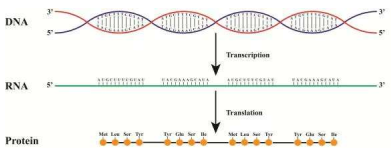
WASEDA UNIVERSITY

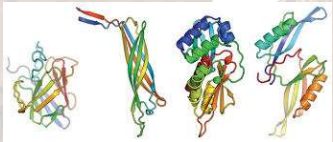
Homework_1 Assignments

Write a report describing *Cell*, *DNA*, *RNA*, Proteins and discussing their relations, from information point of view.

Key points

DNA ⟷ (transcription) RNA → (translation) Proteins → (function) Cell





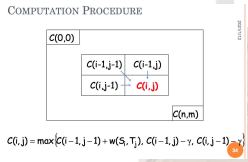
* Images are from the Internet

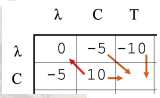
WASEDA UNIVERSITY

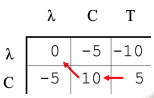
Homework_2 Assignments

Sequence Alignment Using Dynamic Programming(reference PPT3-1)

Scoring scheme : $\begin{cases} +1 & \text{for letter that match} \\ -1 & \text{for mismatches} \\ -1 & \text{for gaps} \end{cases}$







		C	O	E	L	A	C	A	N	T	H
	0.0	-1.0	-2.0	-3.0	-4.0	-5.0	-6.0	-7.0	-8.0	-9.0	-10.0
P	-1.0	-1.0	-2.0	-3.0	-4.0	-5.0	-6.0	-7.0	-8.0	-9.0	-10.0
E	-2.0	-2.0	-2.0	-1.0	-2.0	-3.0	-4.0	-5.0	-6.0	-7.0	-8.0
L	-3.0	-3.0	-3.0	-2.0	0.0	-1.0	-2.0	-3.0	-4.0	-5.0	-6.0
I	-4.0	-4.0	-4.0	-3.0	-1.0	-1.0	-2.0	-3.0	-4.0	-5.0	-6.0
C	-5.0	-3.0	-4.0	-4.0	-2.0	-2.0	0.0	-1.0	-2.0	-3.0	-4.0
A	-6.0	-4.0	-4.0	-5.0	-3.0	-1.0	-1.0	1.0	0.0	-1.0	-2.0
N	-7.0	-5.0	-5.0	-5.0	-4.0	-2.0	-2.0	0.0	2.0	1.0	0.0

```
COELACANTH
P-ELICAN--

COELACANTH
-PELICAN--
```

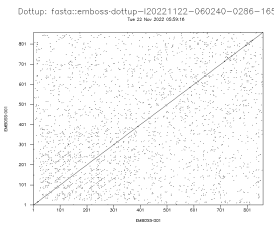
3

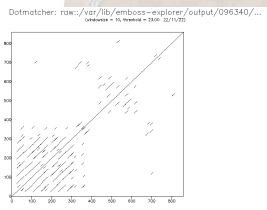
WASEDA UNIVERSITY

Homework_2 Assignments

Two different dot matrix analysis servers to analyze the sequence of the human low density lipoprotein receptor(np_000518)

1. https://www.ncbi.nlm.nih.gov/protein/NP_000518.1 get the origin
2. https://www.ebi.ac.uk/Tools/seqstats/emboss_dottup/ generate the result
3. https://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher generate the result





4

WASEDA UNIVERSITY

Homework_2 Assignments

Examine the two dot matrices and use them to answer the following questions:

a) What does the long diagonal from one corner to the other represent?

The long diagonal indicates that there exists a perfect match from the beginning to the end between the two sequences. (Because the two sequences are the same.)

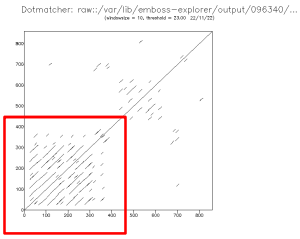
5

WASEDA UNIVERSITY

Homework_2 Assignments

b) What do the shorter diagonals (mostly in the lower left corner) indicate about this protein?

The shorter diagonals represent the local approximate matches between the two sequences. The reason that those shorter diagonals mostly occur in the lower left corner is that there are many similar patterns before the position of 400.



6

WASEDA UNIVERSITY

Homework_3 Assignments

1. Answer the following questions about scoring an alignment.

A. Calculate the score for the DNA sequence alignment shown below, using the scoring matrix below. Use an affine gap penalty to score the gaps, with -11 for opening the gap and -1 for each additional position in the gap. ("Affine gap penalty" refers to a situation when the gap opening and gap extension penalties are not the same. The gap opening penalty should be greater than the gap extension penalty.)

B. How would the score change if you were to use a nonaffine gap penalty? To answer the question, try a nonaffine penalty of -2, and then -6.

```
GACTACGATCCGTATACGCACA---GGTTCAGAC
||||||    ||||||||||||   |||||||||
GACTACAGCTCGTATACGCACACATGGTTCAGAC
```

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

A: 2x27 – 5x4 – 11 – 2 = 21

B: 2x27 – 5x4 – 3x2 = 28(for nonaffine penalty -2)

2x27 – 5x4 – 3x6 = 16(for nonaffine penalty -6)

7

WASEDA UNIVERSITY

Homework_3 Assignments

2. Below is part of the BLOSUM62 matrix. Answer the following questions using this matrix.

A. Using this matrix, two aligned cysteines (C) would receive a score of 9 while two aligned threonines (T) would only receive a score of 5. What can you conclude about cysteine relative to threonine?

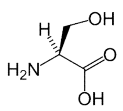
B. A serine (S) aligned with a cysteine (C) would receive a negative score (-1) while a serine aligned with a threonine would receive a positive score (1). Offer a possible explanation for this in terms of physicochemical properties of the amino acid side chains.

Part of BLOSUM62 matrix:

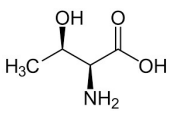
	C	S	T
C	9		
S	-1	4	
T	-1	1	5

A: cysteines (C) is rarer than threonines (T) , The higher the match score the rarer the match

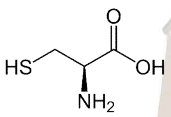
B: The Mercapto is difficult to generate hydrogen bond than Hydroxyl



serine



threonines



cysteines

8

WASEDA UNIVERSITY

Homework_3 Assignments

C. The bestng an alignment score. Use the BLOSUM62 matrix alignment of the two amino acid sequences "LDS" and "LNS" is obvious (it's shown below). Given a scoring system, you could easily calculate an alignment score. Set up a matrix and use the dynamic programming algorithm to "prove" that this is the best alignment by calculating x (see Powerpoint notes) to score aligned residues, and use a gap penalty of -1. (You may hand write the matrix in your homework rather than typing it if you like.)

```
seq1    LDS
         | |
seq2    LNS
```

L	4			
S	-2	4		
N	-3	1	6	
D	-4	0	1	6
	L	S	N	D

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

BLOSUM62 matrix

9

WASEDA UNIVERSITY

Homework_3 Assignments

C. The bestng an alignment score. Use the BLOSUM62 matrix alignment of the two amino acid sequences "LDS" and "LNS" is obvious (it's shown below). Given a scoring system, you could easily calculate an alignment score. Set up a matrix and use the dynamic programming algorithm to "prove" that this is the best alignment by calculating x (see Powerpoint notes) to score aligned residues, and use a gap penalty of -1. (You may hand write the matrix in your homework rather than typing it if you like.)

```
seq1    LDS
         | |
seq2    LNS
```

	λ	L	D	S
λ	0	-1	-2	-3
L	-1	4	3	2
N	-2	3	5	4
S	-3	2	4	9

$$x(LDS, LNS) = 9$$

10

WASEDA UNIVERSITY

Homework_3 Assignments

3. Find the optimal global alignment of the two sequences *Seq1*: THISLINE and *Seq2*: ISALIGNED based on the BLOSUM62 matrix with linear gap penalty of -4.

	λ	T	H	I	S	L	I	N	E
λ	0	-4	-8	-12	-16	-20	-24	-28	-32
I	-4	-1	-5	-4	-8	-12	-16	-20	-24
S	-8	-3	-2	-6	0	-4	-8	-12	-16
A	-12	-7	-5	-3	-4	-1	-5	-9	-13
L	-16	-11	-9	-3	-5	0	1	-3	-7
I	-20	-15	-13	-5	-5	-3	4	0	-4
G	-24	-11	-17	-9	-5	-7	0	4	0
N	-28	-19	-18	-13	-8	-8	-4	6	4
E	-32	-23	-22	-17	-12	-11	-8	2	11
D	-36	-27	-26	-21	-16	-15	-12	-2	7

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

T H I S—L I —NE—
——I SA L I G NE D

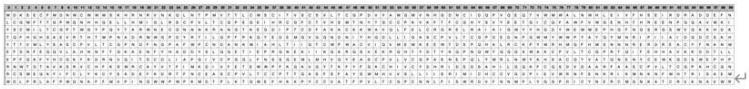
11

WASEDA UNIVERSITY

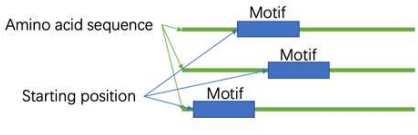
Homework_4 Assignments

1. **Identifying protein motifs using Expectation Maximization (EM) algorithm.**

Suppose we have 200 protein sequences of length 100, in the dataset. Each character represents an amino acid. The length of the motif is 10. The figure below shows the first 10 samples in the dataset.



Find the starting positions and the PSSM of the motif



12

WASEDA UNIVERSITY

Homework_4 Assignments

Expectation Maximization(EM algorithm)

EM is a **converge algorithm** but cant converge to a **global maximum**

Likelihood:

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} p(x_i|\theta), \quad \theta \in \Theta$$

Maximum likelihood estimation:

$$H(\theta) = \ln L(\theta) = \ln \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta)$$

$$\hat{\theta} = \operatorname{argmax} L(\theta)$$

13

WASEDA UNIVERSITY

Homework_4 Assignments

EM algorithm for motif identification

-------ABCABCABC----------

-ABCABCABC----------

----------ABCABCABC-

---ABCABCABC----------

--------ABCAECABC---------

Challenge 1: positions

Challenge 2: not necessarily the same

PSSM:

Position	1 2 3 4 5 6
Sequence 1	ATGTCG
Sequence 2	AAGACT
Sequence 3	TACTCA
Sequence 4	CGGAGG
Sequence 5	AACCTG

Convert multiple alignment to a raw frequency table

Pos.	1	2	3	4	5	6	Overall freq.
A	0.6	0.6	—	0.4	—	0.2	0.30
T	0.2	0.2	—	0.4	0.2	0.2	0.20
G	—	0.2	0.6	—	0.2	0.6	0.27
C	0.2	—	0.4	0.2	0.6	—	0.23

Normalize the values by dividing them by overall freq.

Pos.	1	2	3	4	5	6	Overall freq.
A	2.0	2.0	—	1.33	—	0.67	0.30
T	1.0	1.0	—	2.0	1.0	1.0	0.20
G	—	0.74	2.22	—	0.74	2.22	0.27
C	0.87	—	1.74	0.87	2.61	—	0.23

Convert the values to log to base of 2

Pos.	1	2	3	4	5	6
A	1.0	1.0	—	0.41	—	-0.58
T	0.0	0.0	—	1.0	0.0	0.0
G	—	-0.43	1.15	—	-0.43	1.15
C	-0.2	—	0.8	-0.2	1.38	—

14

WASEDA UNIVERSITY

Homework_4 Assignments

EM algorithm for motif identification

Parameter: the motif (PSSM)

Hidden variable: the probability of each position to be the starting point

E-step:

Using the values in the PSSM, the probability of finding the pattern at every possible position in each sequence is calculated.

M-step:

Providing new information about the likely location of the pattern in each sequence.

15

WASEDA UNIVERSITY

Tips

There is no problem with using ChatGPT, but writing code by yourself is very important for programmers.

Please check your homework (total 4 times).

You can submit it before 2024.1.19 if you forget.

There will be punishment after 2024.1.19

16

WASEDA UNIVERSITY

Thanks for your listening

17