Assignment 4 BioInformatics

Filippo Casari

November 2022

1 Point 1

Using the excess sequence similarity, the similarity searching algorithms BLAST and FASTA find homologous DNA sequences and proteins. The excessive ancestry-homology between two DNA or amino acid sequences is what causes this resemblance. Comparing the amino acid sequences of proteins rather than DNA sequences is the most efficient method of similarity searching. In order to compare two sequences and generate extremely precise statistical estimates regarding the similarity between sequences, both BLAST and FASTA employ a scoring strategy. The primary distinction between BLAST and FASTA is that while FASTA looks for similarities across less comparable sequences, BLAST focuses on discovering ungapped, locally optimum sequence alignments.

2 Point 2

When generating PPMs based on a limited dataset, pseudocounts (or Laplace estimators) are frequently used to prevent matrix entries from having a value of 0. This enables the probability to be determined for new sequences by multiplying each column of the PPM by a Dirichlet distribution (that is, sequences which were not part of the original dataset).

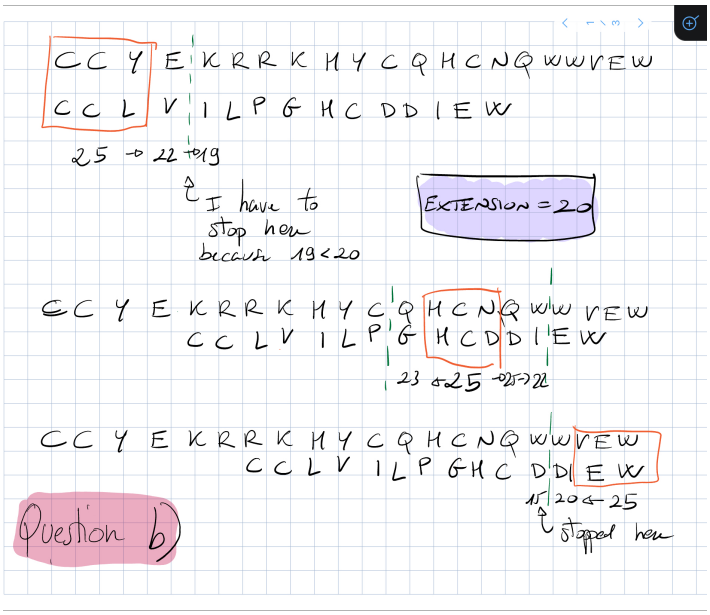
3 Point 3

To create the wmer matrix I wrote a program in python.

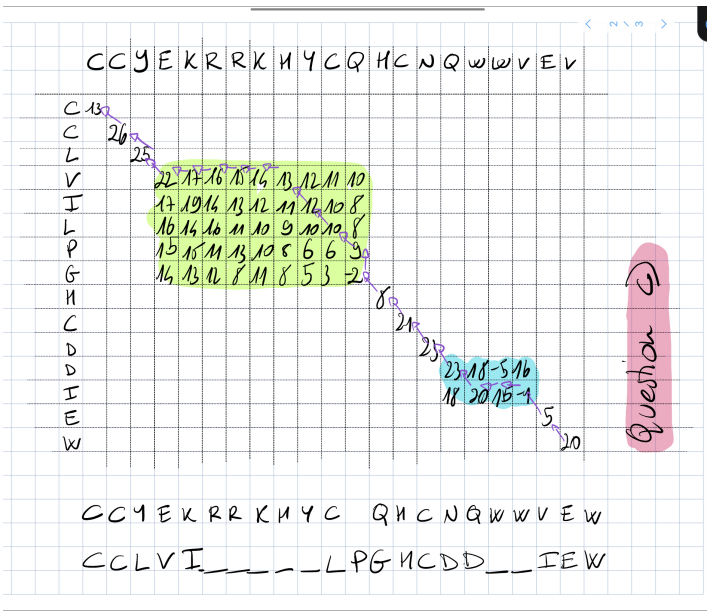
3.1 Building the wmer matrix, question a

	query	CCL	CLV	LVI	VIL	ILP	LPG	PGH	GHC	HCD	CDD	DDI	DIE	IEW
0	CCY	25	10	-4	-4	-7	-9	-5	-9	7	6	-9	-8	-3
1	CYE	7	9	-7	-5	-4	-8	-7	-4	-4	12	-11	1	-7
2	YEK	-9	-9	-7	-8	-5	-4	-6	-6	-2	-2	-4	-6	2
3	EKR	-9	-9	-10	-9	-10	-7	-3	-7	-5	-6	-3	-1	-6
4	KRR	-10	-9	-10	-10	-9	-9	-4	-6	-6	-7	-7	-5	-6
5	RRK	-11	-10	-9	-10	-8	-8	-6	-6	-5	-7	-7	-5	-7
6	RKH	-10	-11	-10	-9	-9	-6	5	-6	-4	-6	-7	-5	-6
7	KHY	-7	-7	-8	-8	-9	-8	-1	5	-6	-7	-3	-7	-1
8	HYC	-8	-5	-6	-7	-9	-9	-8	13	3	-10	-6	-5	-11
9	YCQ	8	-8	-5	-5	-4	-7	-5	-9	15	-7	-10	-3	-5
10	CQH	7	7	-9	-7	-6	-5	4	-5	-7	12	-8	-7	-3
11	QHC	-8	-7	-8	-9	-10	-7	-6	21	-6	-8	-3	-7	-8
12	HCN	6	-8	-7	-10	-8	-7	-4	-7	25	-5	-8	-3	-11
13	CNQ	9	6	-8	-6	-7	-6	-3	-5	-5	15	-5	-5	-3
14	NQW	-7	-7	-10	-8	-9	-8	-7	-4	-7	-7	-1	-4	14
15	QWW	-10	-8	-8	-8	-9	-9	-7	-10	-9	-13	-8	-6	9
16	WWV	-9	-2	-1	-5	-8	-10	-11	-7	-12	-14	-6	-11	-9
17	WVE	-9	-7	-1	-2	-3	-8	-8	-10	-2	-7	-13	5	-9
18	VEW	-6	-7	-5	-1	-3	-3	-9	-9	-12	-4	-5	-11	25

3.2 Question B



3.3 Question C



4 Point 4

To solve the exercise I wrote a program in python. See files attached.

4.1 Frequency / Count of Occurrences

First, I computed the Frequency of the four types of bases: adenine (A), cytosine (C), guanine (G), and thymine (T).

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	A	5	4	3	5	1	2	2	2	3	1	1	3	3	2	3	2	4	1	2	3	2	1	1	1	1
1	T	1	2	1	2	4	2	2	1	1	5	1	1	1	3	2	1	1	2	2	1	1	2	3	4	1
2	C	2	1	4	1	2	4	3	4	3	2	3	2	2	2	1	1	1	1	2	4	2	3	3	1	1
3	G	1	2	1	1	2	1	2	2	2	1	4	3	3	2	3	5	3	5	3	1	4	3	2	3	6
4	SUM	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

4.2 Relative Frequency / Fraction of Occurrences

Then, I divided all the values by 9 (normalized).

	A	T	C	G	SUM
0	0.556	0.111	0.222	0.111	1.0
1	0.444	0.222	0.111	0.222	1.0
2	0.333	0.111	0.444	0.111	1.0
3	0.556	0.222	0.111	0.111	1.0
4	0.111	0.444	0.222	0.222	1.0
5	0.222	0.222	0.444	0.111	1.0
6	0.222	0.222	0.333	0.222	1.0
7	0.222	0.111	0.444	0.222	1.0
8	0.333	0.111	0.333	0.222	1.0
9	0.111	0.556	0.222	0.111	1.0
10	0.111	0.111	0.333	0.444	1.0
11	0.333	0.111	0.222	0.333	1.0
12	0.333	0.111	0.222	0.333	1.0
13	0.222	0.333	0.222	0.222	1.0
14	0.333	0.222	0.111	0.333	1.0
15	0.222	0.111	0.111	0.556	1.0
16	0.444	0.111	0.111	0.333	1.0
17	0.111	0.222	0.111	0.556	1.0
18	0.222	0.222	0.222	0.333	1.0
19	0.333	0.111	0.444	0.111	1.0
20	0.222	0.111	0.222	0.444	1.0
21	0.111	0.222	0.333	0.333	1.0
22	0.111	0.333	0.333	0.222	1.0
23	0.111	0.444	0.111	0.333	1.0
24	0.111	0.111	0.111	0.667	1.0

4.3 Position Specific Scoring Matrix (PSSM)

Then by applying the formula:

$$log_2(\frac{P(pos, nuc)}{P(nuc)})$$

I computed the PSSM as shown below.

AminoAcids	A	T	C	G
0	1.153	-1.171	-0.171	-1.171
1	0.829	-0.171	-1.171	-0.171
2	0.414	-1.171	0.829	-1.171
3	1.153	-0.171	-1.171	-1.171
4	-1.171	0.829	-0.171	-0.171
5	-0.171	-0.171	0.829	-1.171
6	-0.171	-0.171	0.414	-0.171
7	-0.171	-1.171	0.829	-0.171
8	0.414	-1.171	0.414	-0.171
9	-1.171	1.153	-0.171	-1.171
10	-1.171	-1.171	0.414	0.829
11	0.414	-1.171	-0.171	0.414
12	0.414	-1.171	-0.171	0.414
13	-0.171	0.414	-0.171	-0.171
14	0.414	-0.171	-1.171	0.414
15	-0.171	-1.171	-1.171	1.153
16	0.829	-1.171	-1.171	0.414
17	-1.171	-0.171	-1.171	1.153
18	-0.171	-0.171	-0.171	0.414
19	0.414	-1.171	0.829	-1.171
20	-0.171	-1.171	-0.171	0.829
21	-1.171	-0.171	0.414	0.414
22	-1.171	0.414	0.414	-0.171
23	-1.171	0.829	-1.171	0.414
24	-1.171	-1.171	-1.171	1.416

4.4 Question B

Score of the Test Sequence (Human): 4.3. The test sequence is

$$2^{4.3}$$

times more likely than random probability to be part of the group.
Score of the Test Sequence (CHIMPANZEE): 4.71. The test sequence is

$$2^{4.71}$$

times more likely than random probability to be part of the group.