Database Searching (FASTA)

Lecture – 6.2

Department of CSE, DIU

CONTENTS

1. TP, TN, FP, FN

True Positive, True Negative, False Positive, False Negative

A patient fears that he has Cancer
&
Goes to the doctor for Diagnosis

Possible Scenarios

True Positive

Patient really had cancer
&
Diagnosis came Positive

True Negative

Patient didn't have cancer
&
Diagnosis came Negative

False Positive

Patient didn't have cancer
&
Diagnosis came Positive

False Negative

Patient really had cancer
&
Diagnosis came negative

2. Selectivity and Sensitivity

We will learn about calculating selectivity and sensitivity

Selectivity & Sensitivity

$$\text{Sensitivity} = \frac{True\ Positive}{All\ Positive}$$

$$\text{Selectivity} = \frac{True\ Negative}{All\ Negative}$$

Worked Out Example (Sensitivity)

Dataset	
A	G
C	T
G	T
G	C
A	G
C	G

Search Character = C
Expected = CCC
Outcome = ACC

▷ $\text{Sensitivity} = \frac{True\ Positive}{All\ Positive}$

- Suppose we are searching the character C in entire database
- Each time we encounter a C we should print C
- So the final output of search should be = CCC as there are 3 Cs in the entire dataset. But the outcome is ACC
- So, **All Positive = 3** (as there are 3 Cs in the whole dataset and we are looking for C only)
- **True Positive =** number of Cs in the outcome ACC = 2
- Sensitivity = $\frac{2}{3}$

Worked Out Example (Selectivity)

Dataset	
A	G
C	T
G	T
G	C
A	G
C	G

Search Character = C
Expected = CCC
Outcome = ACC

▷ $\text{Selectivity} = \frac{True\ Negative}{All\ Negative}$

- Suppose we are searching the character C in entire database
- Each time we encounter a C, we should print C
- So the final output of search should be = CCC as there are 3 Cs in the entire dataset. But the outcome is ACC
- So, **All Negative = 9** (Number of entries in the dataset that is not C)
- **True Negative =** number of entries in the outcome ACC that is not C = 1
- Sensitivity = $\frac{1}{9}$

3. Hash Table Used in FASTA

Hash Table Algorithm

Given Data

Query Sequence: JUSTICELEAGUE
Target Sequence: LEAGUEOFASSASINS
Value of K : 1

Step 1 : Build Query Table

1	2	3	4	5	6	7	8	9	10	11	12	13
J	U	S	T	I	C	E	L	E	A	G	U	E

Step 2: Hash Table for Query Sequence

Write all the distinct characters appeared in the Query Sequence Lexicographically and then, beneath that, write the number of the position in which that letter appeared. There can be multiple occurrences.

A	C	E	G	I	J	L	S	T	U
10	6	7 9 13	11	5	1	8	3	4	2 12

Step 3 : Build Target Table

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S

Step 4 : Import the Hash Table for Query Sequence

A	C	E	G	I	J	L	S	T	U
10	6	7 9 13	11	5	1	8	3	4	2 12

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S

Step 5 : Build the Extended Target Table based on Hash Table

A	C	E	G	I	J	L	S	T	U
10	6	7 9 13	11	5	1	8	3	4	2 12

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S
7	5 7 11	7	7	-3 7	1 3 7			1	-7	-8	-2	-10	-9		-13

Entry in Extended Row = Position of the Letter in Hash Table – Position of the Letter in Extended Target Table
Example:

- For L, in Extended Target Table, Entry is 7 (8-1) .
- Similarly For E, the entries are 5 (7-2), 7 (9-2) and 11 (13-2).

Step 6 : Build Offset Table

Draw a table from the minimum to the maximum entry of the extended target table. Then beneath each entry number, write down number of times that entry occurred in extended target table. For example, the entry 7 Occurred 6 times and the entry 1 occurred 2 times.

-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
1			1	1	1	1				1	1			2		1		**1**		6				**1**

Step 7: Build Pre-Final Table

Start both Query and Target sequence from 0 position.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
J	U	S	T	I	C	E	L	E	A	G	U	E			
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S

Step 8 : Build Final Table

- Find out the entry number from the offset table, that occurred maximum number of times (Here 7, which occurred 6 times).
- After that, add that entry number with the previous starting position of target sequence to get the new starting Position of Target Sequence (Previous starting position = 0, Then new starting position of target seq becomes 0 + 7 = 7).

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
J	U	S	T	I	C	E	L	E	A	G	U	E										
							L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S


Thanks!