Name: Nigel Nicholas

SID: 1155088791

CSCI32

20 Assignment 3

1. (a) number of different 1-gapped-3-mer patterns :

k+gCk 4k = 4C3 43 = 256 different patterns.

(b) List all 1-gapped-3mers supported by s1 = GCGTCCGAC and s2 = CGACGCGAC

Table for s1:

|  |  |
| --- | --- |
| 1-gapped 3-mer | Number of Occurrence |
| \*CGA | 1 |
| \*GAC | 1 |
| C\*AC | 1 |
| CG\*C | 2 |
| CGA\* | 1 |
| GCG\* | 1 |

Table for s2 :

|  |  |
| --- | --- |
| 1-gapped 3-mer | Number of Occurence |
| \*CGA | 1 |
| \*GAC | 2 |
| C\*AC | 2 |
| CG\*C | 2 |
| CGA\* | 2 |
| GCG\* | 1 |

(c) Computing similarity =

1x1(\*CGA) + 1x2(\*GAC) + 1x2 (C\*AC) + 2x2(CG\*C) + 1x2 (CGA\*) + 1x1(GCG\*)

= 1 + 2 + 2 + 4 + 2 + 1

= 12

(d)

Entry on table represents mismatch

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S1 \ S2 | ACGC | CGAC | CGAC | CGCG | GACG | GCGA |
| CCGA | 2 | 3 | 3 | 3 | 4 | 1 |
| CGAC | 3 | 0 | 0 | 2 | 4 | 4 |
| CGTC | 3 | 1 | 1 | 2 | 4 | 4 |
| GCGT | 2 | 4 | 4 | 4 | 3 | 1 |
| GTCC | 3 | 3 | 3 | 3 | 2 | 3 |
| TCCG | 2 | 4 | 4 | 3 | 2 | 2 |

(e) Counting similarity from the mismatch table of k+g-mer.

Since the formula of similarity from each table is: k+g-mCk with m as mismatch. In order to have contribution to similarity, then you can only have .

Therefore looking into table 1 by 1 to find mismatch of at least 1.

CCGA & GCGA -> 1

CGAC & CGAC -> 4

CGAC & CGAC -> 4

CGTC & CGAC -> 1

CGTC & CGAC -> 1

GCGT & GCGA -> 1

---------------------------+

total score = 12.

So the similarity score is the same with **(c)** .

(f) Which value of g and k to use without having to try out multiple combinations to get the best similarity scores.

We can try using the method of (d) which is creating a g+k-mer table. And try to increase 1 by 1 ( i.e , do 2-mer table, 3-mer table, 4-mer table, … n-table \* n = length of of shortest sequence). Then after getting the max similarity score of each k-mer table check if after increasing k the best similarity score decrease we should stop searching and take the value of the previously found maximum of similarity score with its corresponding g-gap k-mer combination.

Secondly in order to compute reduce number of computation after deciding the g+k –mer table, see which number of mismatch is most abundant in the table. The abundant number of mismatch should be the gap, because then all those can be turned into points whilst maintaining a high contribution point for low mismatches.

1. Given binary features for all X1, X2  and Y.
2. –
3. X1 and X2  are independent when conditioned on Y. This means that

The 8 Equations:

Now using the theory of

We can get another 8 equations:

We can try out combinations of any X1 and X2 given the same condition ( Y=0 or Y=1).

Take for example,

Since,

By different result therefore,

And therefore prove that the set of these combination of 8 equations show 2 variables are not conditionally independent on Y.