Documentation

Author - Simran Bansal

Part 1 - Finding the frequency of repeats

The database for this part was taken from UniProt for the proteome sequence of bacterias.

For explanation purposes, let's take into consideration a bacteria with the following protein sequences -

Protein name	Protein sequence
Prot 1	ABA
Prot 2	ABABA
Prot 3	ABB
Prot 4	BBB

This part included finding the number of times a part of the protein sequence of a particular length was being repeated -

Consider Prot 1 with sequence ABA

Protein part	Frequency of repetition
A	2
B	1
AB	1
BA	1
ABA	1

Consider Prot 2 with sequence ABABA

Protein part	Frequency of repetition
A	3
B	2

AB(**AB**ABA, AB**AB**A)	2
BA(A**BA**BA, ABA**BA**)	2
ABA(**ABA**BA, AB**ABA**)	2
BAB(A**BAB**A)	1
ABAB(**ABAB**A)	1
BABA(A**BABA**)	1
ABABA	1

Consider Prot 3 with sequence ABB

Protein part	Frequency of repetition
A	1
B	2
AB	1
BB	1
ABB	1

Consider Prot 4 with sequence BBB

Protein part	Frequency of repetition
B	3
BB(**BB**B, B**BB**)	2
BBB	1

If I add all the results, I get the following -

Protein part	Frequency of repetition
A	6
B	8
AB	4
BA	3
BB	3
ABA	3
BAB	1
BBB	1
ABAB	1
BABA	1
ABABA	1

The last table is obtained for each bacteria in the frequencies.csv files.
Next was finding the uniques which are as follows -

- Length 4
    - BAB
    - BBB
- Length 5
    - ABAB
    - BABA
- Length 6
    - ABABA

This result was classified into separate files containing sequences of different lengths.

Part 2 - Finding uniques

The processing of data in this part was started from length 4 as the number of uniques upto length 3 were very few or non-existent.

This can be divided into two steps -

- Back checking
- Cleaning of the sequences data

For the purpose of this example, let's take the following sequences into consideration -

- Length 4 - ABCC, BCCA, BBBB, AAAB, CCAB, AABC, ABCC
- Length 5 - ABCCA, BCCAB, AABCC, AABCB, CCCCA
- Length 6 - ABCCAB, BCBCBC

Step 1 - Back checking

- We start with the second smallest length i.e. length 5
- Split the sequence into two parts -

Length 5	Length 4(Constituent I)	Length 4(Constituent II)
ABCCA	ABCC	BCCA
BCCAB	BCCA	CCAB
AABCC	AABC	ABCC
AABCB	AABC	ABCB
CCCCA	CCCC	CCCA

- Now, we check if the smaller constituents, in this case of length 4 are present in the unique's list of the smaller constituent(Length 4).

Length 5	Length 4(Constituent I)		Length 4(Constituent II)	
	Sequence	Present	Sequence	Present
ABCCA	ABCC	Yes	BCCA	Yes
BCCAB	BCCA	Yes	CCAB	Yes
AABCC	AABC	Yes	ABCC	Yes
AABCB	AABC	Yes	ABCB	No
CCCCA	CCCC	No	CCCA	No

- Three cases are formed -

    - Case I - Both subsequent are present as in ABCCA, BCCAB, AABCC
    - Case II - One of the constituents is present as in AABCB
    - Case III - None of the constituents is present CCCCA
- The only sequences that are accepted are, are the ones in case I.
- So, the new list of length 5 repeats biomes -
  ABCCA, BCCA, AABCC
- Similarly, we repeat it for length 6.

The final list of repeats after this step become -

- Length 4 - ABCC, BCCA, BBBB, AAAB, CCAB, AABC, ABCC
- Length 5 - ABCCA, BCCAB, AABCC
- Length 6 - ABCCAB

Step 2 - Cleaning the repeats

- We again start with the second smallest length i.e. length 5
- Split the sequence into two parts -

Length 5	Length 4(Constituent I)	Length 4(Constituent II)
ABCCA	ABCC	BCCA
BCCAB	BCCA	CCAB
AABCC	AABC	ABCC

- Now, we remove the length 4 constituents from the lit of length 4 sequences. The resulting length 4 list becomes
  BBBB, AAAB
- Since, some repeats like BCCA can be present in more than two sequences, they cannot be removed in the previous step.
- Similarly, for length 6

Length 6	Length 5(Constituent I)	Length 5(Constituent II)
ABCCAB	ABCCA	BCCAB

- The resulting length 5 list becomes -
  AABCC

The final result after the completion thus becomes,

- **Length 4 - BBBB, AAAB**
- **Length 5 - AABCC**
- **Length 6 - ABCCAB**

The above is final result obtained.

All the values taken in the above examples are only for explanation purposes only. In the real examples, the number of unique sequences increase as we increase the length of the sequence.