ゲノム配列を用いた短いメタゲノム配列フラグメントのアノテーション評価
Evaluation of the annotation of short metagenomic sequence fragments using genomic sequences

京都大学化学研究所 バイオインフォマティクスセンター
化学生命科学 金昭

Introduction

BLAST finds regions of local similarity between sequences and it can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. However, BLAST is basically designed for long sequences comparing, how well it works on short metagenomic sequences is still not clear. In this study, we tried to use simulating data to verify the effect of BLAST on short fragments.

Materials and methods

We collected total 728,225 protein sequences belong to 7,784 clusters in the PRK category of CLUSTERS from NCBI FTP as our original database. Next, 61 sequences from different clusters of various sizes were chosen as samples. Every sequence was cut into fragments of 7 kinds of lengths, which were 20mer, 30mer, 40mer, 50mer, 60mer, 70mer, 80mer. We did all the cutting on every sequence from beginning to the end, in order to ensure that all the possible fragments with certain length on this sequence would be collected in the study. Therefore, there're a great many fragments for each sequences of each length.

Secondly, we blasted all the fragments against our database, basically using default e-value 10. Then several methods were developed to filter the blast results.

1) The blast hits with "Identities" > 90% were picked out as group1 data.
2) The blast hits with e-value < 10-5 were picked out as group2 data.
3) The blast hits with alignment_length > 0.8 were picked out as group3 data.

alignment_length = abs(query.end -query.start) / length_of_query_sequence.

Thirdly, the three groups data were used for further analysis called "Hit-back-value" which was to find out how many blast hits belong to the original cluster, the calculation should be:

Hit-back-value = (Number of blast hits belong to original cluster) / (Total blast hits number)

For every sample sequence, it had lots of fragments in different lengths with different results; we calculated the average percentage for every certain length of

sample and put the results on plots, then every single line on the chart stands for one sequence.

Results and future work

As the results plots showed(Figure 1), we could see, for both group1 and group3 data, the "Hit-back" plots are quite similar, which suggested that fragments longer than 50mer would give a perfect result in BLAST search, but the perfect percentage would only appear when fragments were longer than 80mer in group2. During normal BLAST search, we usually filter the results by reducing e-values; however, our study indicated that the same method may not work well when the data are fragments that are shorter than 80mer. In this case, it's better to develop other methods to extract blast hits such as identities or alignment_length used above.

Because of the time limit, we didn't finish testing all the clusters, and only chose one sample sequence to stand for one cluster, which were quite unilateral. We are going to test more sequences and clusters in the future in order to complete the results. What's more, the plots shown left gave the average values for whole sequences, however, fragments picked from different location would show quite different results in BLAST, it's also necessary and interesting to study typical fragments.

Figure1: Parts of Hit-back-values of group 1-3 data. X axis: length of fragments; Y axis: percentage (no more than 100)