若手研究者インターナショナル・トレーニング・プログラム(ITP)

バイオインフォマティクスとシステムズバイオロジーの国際連携教育研究プログラム 報告書

Name : Zhao Jin

Title : Using simulating data to evaluate the annotation of short metagenomics sequences fragments

Institute: Bioinformatics Center, Institute for Chemical Research, Kyoto University

Partner institute: Bioinformatics Program, Boston University

Duration: January 16, 2012 ~ April 8, 2012

Report:

1. General Report:

With the support of ITP program, I received a chance to visit Boston University as an exchange student, studying in Dr. Daniel Segre's lab for about 3 months.

Segre lab belongs to the Bioinformatics Program, which locates in the 9th floor of Life Science and Engineering Building of BU. Led by Associate Professor Daniel Segre, there're two postdoctoral fellows and over 10 PhD students working in the lab, some of whom are double advised with other professors. Therefore, besides the basic weekly group meeting on every Friday, we can also enjoy the lively discussions in those joint meetings with different labs. I was really surprised by the active communications and open-minded ideas of all the participants every time. They would not only focus on the researches, but also other related studies, interesting references, even something that seems to have nothing to do with the current topics. It confused me at first when they were having a heated discussion on something that totally unconcerned which may last for an hour or more. However, as time went by, I discovered that it's just the discussion that provide great opportunities to encourage everyone to express opinions no matter right or wrong, which make the lab always have enough communications and stay in a relaxed mood. This experience gave me an impressive lesson that science should be strict but also could be flexible and enjoyable.

Besides Boston University, there're many other top universities in Boston, such as Harvard University and Massachusetts Institute of Technology (MIT). These universities communicate with each other frequently, so it's very convenient for the college students to visit different campus in order to attend open seminars or listen to speeches given by famous scholars. For example, there's an informal scientific seminar called Bauer Forum talks held in Harvard University every Wednesday, which was designed to foster communication and collaboration among people with an active interest in genomics and systems biology. In a word, various speeches on advanced researches in such open seminars expanded my horizon greatly, which was also precious experience I gained during ITP program.

2. Research Project:

There are multiple research interests in Segre lab which mainly focus on the dynamics and evolution of metabolism in individual microbes and in microbial ecosystems. At first, I was suggested to work on the functional comparisons between metagenomes across bacteria, using 16S rRNA from the metagenomic datasets. However, we discovered that this project would take several months to process due to the large size of data and complicated methods, so I have to change my project after 2 weeks' study.

The second project was still concerned with metagenomics whose annotation became a very important part in current research. 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.

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 example, we have 100 blast hits, and 50 of them could be found in the original cluster, then the "Hit-back" percent should be 50/100=50%.

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

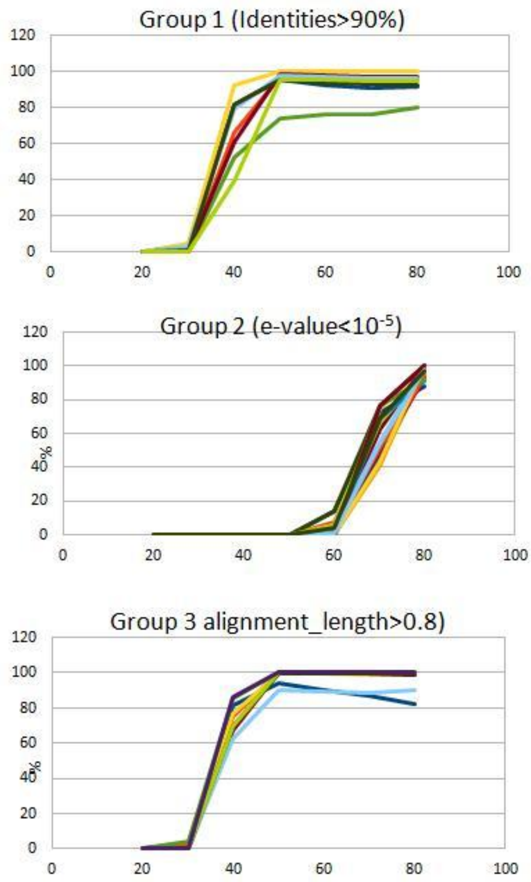


Figure 1:
Parts of Hit-back-values of group1-3 data.
X axis: length of fragments
Y axis: percentage (no more than 100)

plots, then every single line on the chart stands for one sequence.

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.

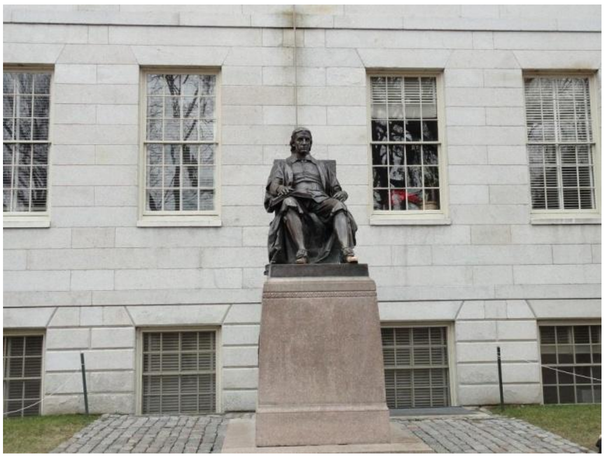
The result has been submitted to ISME14 as co-authorship.

Acknowledgement:

I would like to thank Professor Minoru Kanehisa, Professor Hiroshi Mamitsuka, and Associate Professor Susumu Goto for providing me such a valuable study opportunity. I would also like to express my gratitude to Dr. Daniel Segre, administrator Caroline Lyman, and all the Segre lab members for their kindness and supports during my stay in BU.

Left photo: Dr. Daniel Segre(left), Zhao Jin.

Right photo: left side of the table from front, Dr. Daniel Segre, Christopher Jacobs, Zhao Jin, Dr. Zhuoyun Zhuang, Rama Krishna Simhadri, Dr.Varun Mazumdar; Right side of the table from front, Amrita Kar, Brian Granger, Sara Baldwin, Lina Faller, Arion Stettner.



Left photo: Statue of the Three Lies in Harvard University.

Right photo: MIT Building 10 and the Great Dome.