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：Unyanee Poolsap

Title：Computational methods for predicting single and interacting RNA secondary structures

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

Partner Institute: Department of Biology and Computer Science Program in Bioinformatics Boston University

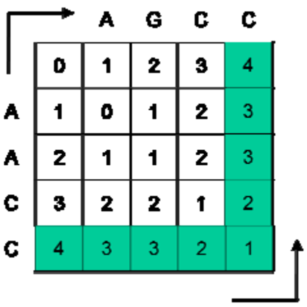
Duration: February 19 - May 18, 2010

Report:

The laboratory I had visited during the training program is the Laboratory for Biocomputing and Informatics led by Associate Professor Gary Benson. This lab is a part of the Department of Biology and Computer Science, Program in Bioinformatics, Boston University. The environment in this lab is very relaxed. There are only three members, Dr. Benson and 2 researchers. Dr. Benson has his private office on the 9th floor of the building, but most of the time he comes to work at the lab room on the 1st floor. He works closely with his collaborators. There is no weekly seminar; however, professor and researchers discuss about their work and progress directly every day. The lab members work as a team in the same projects. The first project is developing computer program (web server) to find tandem repeats and make a tandem repeats database. The second project is discovering structure variance in human genomes. I had participated in both projects.

For the first project, I worked on studying how to speed up the global pair-wise edit-distance, sequence alignment using dynamic programming (DP) algorithm. We encoded the input sequences, and divided the DP table into small subtables, then aligned each subtable. Our assumption is that, if we can find a relationship of an encoded input and its corresponding output, we may be able to reduce the computation time of DP by making an input-to-output table and looking up for the output without computing DP. In addition, since some tables are independent, we can compute them in parallel, which may help to reduce the computation time.

We encoded the input sequences before aligning them to reduce the space of DP inputs. The encoding method is as follows: a sequence of length n is parsed. The first character is encoded to a digit ‘0’ and all occurrences of that character are also encoded to ‘0’. Then, the next character is encoded to ‘1’ and all occurrences of it are also encoded to ‘1’. Repeat until all characters are encoded. For example, ACCGG is encoded to 01122, and GCTAT is encoded to 01231. We align all combinations of pair of encoded input sequences. We counted the number of input and number of output from the alignment result. In this context, an input refers to an encoding of the concatenation of two input sequences (vertical sequence followed by horizontal sequence). An output is a concatenation of score from the last row and the last column of DP table. For example in Fig. 1 input is 0011 1200 and output is 3321233



	A	G	C	C
0	1	2	3	4
A 1	0	1	2	3
A 2	1	1	2	3
C 3	2	2	1	2
C 4	3	3	2	1

Figure 1: Example of DP table

Report (Continued) :

(read direction indicated by an arrow). We found that different sequences which are encoded to the same encoding always produce the same output. However, some different encodings produce the same output. Table below shows the number of all possible input sequences, number of encodings and number of outputs for input sequence of length 2 to 6.

Sequence length	# input	# encoding	#output
2	$4^4 = 256$	15	9
3	$4^6 = 4{,}096$	187	48
4	$4^8 = 65{,}536$	2,795	287
5	$4^{10} = 1{,}048{,}576$	43,947	1,860
6	$4^{12} = 16{,}777{,}216$	700,075	12,325

We also tried another way to encode the input sequences. We determine the relationship of pairs of nucleotide from the two input sequences. If nucleotides in a pair are match, it is encoded to a digit ‘0’, otherwise it is encoded to a digit ‘1’. For example, the encoding for input sequences AC and GA is 0100 (A:G = 0, A:A = 1, C:G = 0, C:A = 0). However, in this encoding method, some different encodings also produce the same output and the length of string is n^2, which is longer than the previous method. Therefore, this method may not be useful.

I implemented pair-wise sequence alignment by dividing the DP table into small subtables. As an example, I implemented a DP table for aligning sequences of length 4, called 4x4table. It is divided into 4 subtables of size 2x2. Each subtable is computed according to the following order. First, subtable 1 which is independent of other tables is computed. Then, subtable 2 and subtable 3 are computed using output from subtable 1 as their initial score (See Fig 2, an output is indicated by the colored region). Note that subtable 2 and 3 can be computed in parallel because they are independent to each other. However, I have not yet implemented the parallel computation of these two tables. Next, subtable 4 is computed using output from subtable 2 and subtable 3 as its initial score (See Fig.2). Finally, the output of 4x4table is retrieved from output of subtable 2, 3, and 4.

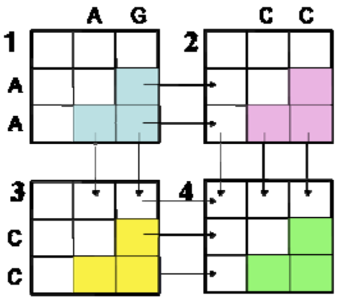


Figure 2: Sequence alignment by subtables

The second project I have been working on is discovering structural variation from next-generation sequencing data. The aim of this project is to discover structural variance in human genome compared with the reference genome. The basic structural variations are insertion, deletion and inversion. For insertion, the distance between the positions that a pair of reads mapped to the reference genome (mapped distance) is less than the average insert size. If the mapped distance of a read pair is greater than the average insert size, it is called deletion. Inversion occurred when there exist two pairs of which the two reads map in the same direction and one read in both pairs span beyond its breakpoint. I developed several python scripts to summarize some statistical information and extract read pairs of interest from the input files in export format and SAM format (Sequence Alignment/Map format which is a generic format for storing large nucleotide sequence alignments) or BAM format (a compressed version of SAM).

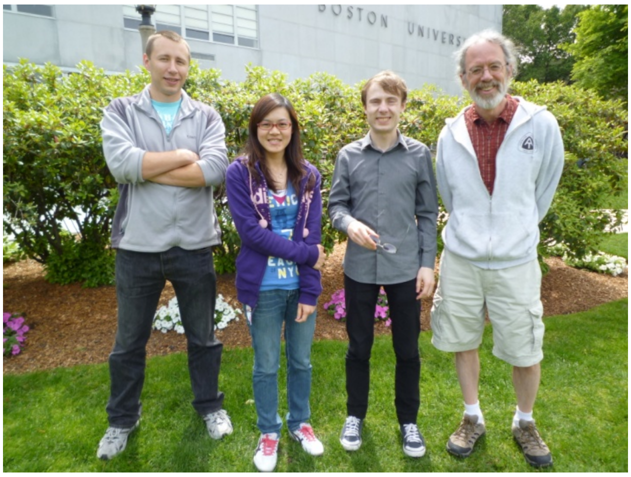
Plan (Continued)

Besides research, during the program, I had a chance to attend two open seminars at Boston University, which I found that they were interesting. The first one was "The third generation sequencing: Single-Molecule Real Time Biological Sequencing" by Dr. Stephen Turner, Founder & Chief Technology Officer of Pacific Biosciences company, and the second one: "Life after Ph.D.: Finding the Right Postdoctoral Position" by BU AGEP and the graduate student organization. Furthermore, I had a chance to join the Friday-party for graduate students in Bioinformatics program. It was a good opportunity to make friend with other students.

Although the work that I had done during the training program does not relate with my own research project directly, I gained experience in dealing with the real human genome data which are very huge, and require high computation time. Small mistakes can slow down the analysis processes. As a part of the research team, I have to have more responsibility on my part and need a good communication with other people. In addition, I had developed many computer programs to help analyzing data. This helps me to improve my algorithm development and programming skill.

Finally, I would like to thank the Program Director: Prof. Minoru Kanehisa for giving me this opportunity, Prof. Hiroshi Mamitsuka and Prof. Tatsuya Akutsu for helping me through the preparation process. Also thanks to Associate Prof. Gary Benson for accepting me to work in his lab and his kind advices throughout the program. Last but not least, I would like to thank the lab members for all their helps regarding to technical problems and daily life.

Enjoy the Friday party at Bioinformatics program student lounge.



With lab members on the last day at Boston University (from left to right: Yevgeniy Gelfand, Eugene Scherba and Assoc. Prof. Gary Benson)