Vol.2010-BIO-22 No.1
2010/7/28

A study of DNA assembly algorithm using shortest common superstring problem

Ayako OHSHIRO †1　Takeo OKAZAKI†2
Hitoshi AFUSO†1　Morikazu NAKAMURA †2

1. Introducion

Sequence analysis such as sequence alignment and motif detection are depend on accurate DNA sequence. Generally, DNA sequence is read by DNA sequencer. Recently, because of development of second generation sequencer technology, it became possible to get massive short read sequence quickly. On the other hand, output short sequences from second generation sequencer occur misassembly as shown in Figure 1. Therefore, accuracy of DNA assembly is not high. This problem make de novo assembly difficut. In fact, second generation sequence is used for allocation procrdure to reference sequence mainly. To solve this problem, several studies have been made on short sequence assembling. Rene L. Warren[1] proposed assembling algorithm with prefix tree, and Daniel[2] used de Bruijn graphs. But how to evaluate the restored sequence has not been surveyed enough. My research purpose is to survey about searching algorithm for longest simple path using length of overlap region and restored sequence as indexes, and to propose a discriminant method that improves accuracy of assemble.

†1 Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus
†2 Faculty of Engineering, University of the Ryukyus

2. Sequence assembly

In second generation sequencer procedure, whole sequence is cut into many short sequences randomly as shown in Figure 2. Regarding each read sequences as nodes and existence of overlap between sequences as links, an adjacent graph shall be generated.

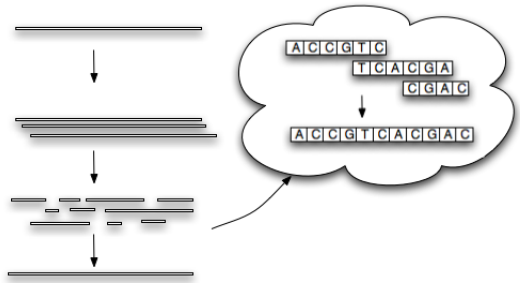

図 1 Flow of DNA sequence assembly

Overlap length play the role of weight of link. On this adjacent graph, one path corresponds to one restored sequence. Appropriate paths may have high value of sum of overlaps. So, by searching appropriate path, we can expect to restore the cut original sequence. In this searching, we regarded a simple path that has high value of sum of overlap length over edges included in that path as appropriate path. Warshall-Floyd algorithm[3] can be applied for this searching, although this algorithm aims to find shortest simple path. We can extend this algorithm to the longest simple path searching, updating the distace and path matrix by each step in this algorithm as follows.

d_{ij}:Distance between vertex i and j
p_{ij}:Path between i and j
h: Number of path step
n: Number of nodes

(1) The distance from each node to adjacency node by 1 step (when h=1) is put to distance matrix. The node from start node is put to path matrix.

(2) If $d_{ij} > d_{kj}$, d_{ij} is overwritten as $d_{ik} + d_{kj}$ and p_{ij} is overwritten as p_{ik} in same way.

(3) $h = h + 1$, go to step2. When $2^h \geq n - 1$, longest searching is finished.

(4) Distance and path matrixes are outputted finally.

By using above algorithm, we can restore the cut original sequence from the adjacent graph of short read sequences.

Here is a numerical example with n=5 in Figure 3 and 4.

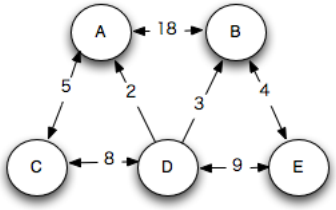

図 2 Adjacency graph

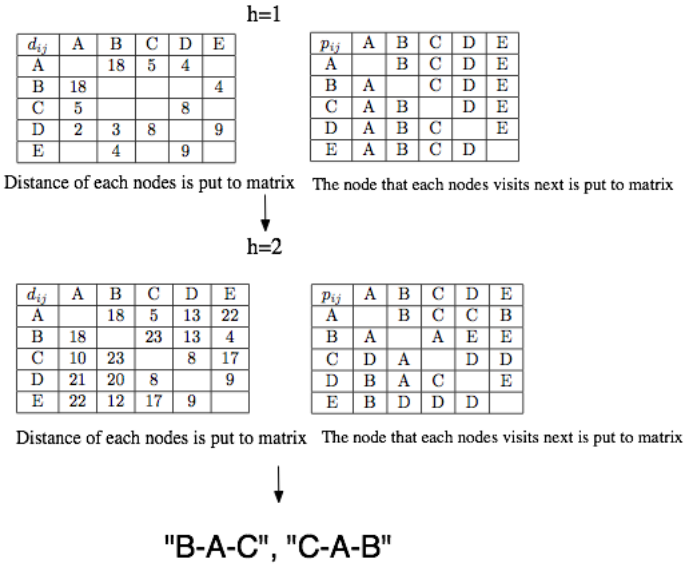

図 3 Distance and path matrix

3. Discriminant procedure with Sapport Vector Machine

It sometimes happens that restored sequences by the proposal assembly algorithm are not included in the original sequence. So it is required to distinguish the output sequence whether consistent or inconsistent for accurate assembly of short read sequences. As the parameters that have relationship with the discriminant, we have overlap length and length of restored sequence. For deciding discriminant rule from these parameters, we focused on machine learning technique, especially on the supervised learning. To get the training data, we generated artificial sequence data from a DNA sequence whose base allocation is already known. Random walk in the adjacent graph from the artificial sequence data will produce the unbiased assembled sequence data. We can distinguish the assembled data "consistent" or "inconsistent" by comparing to the original sequence

and regard them as training data. Finally, we can obtain the discriminant rule by applying training data to machine learning. Actual procedure for getting discriminant rule is as follows.

(1) Prepare the whole sequence whose base allocation is already known.
(2) Replicate the given sequence like PCR processing.
(3) Generate short read sequences from whole sequence by cutting according to the feature of second generation sequencer.
(4) Generate adjacency matrix based on information of overlap region for each sequence.
(5) Generate restored sequence in random walk algorithm by appropriate length considering the length of whole sequence.
(6) Obtain parameters such as length of restored sequence and overlap region.
(7) Distinguish restored sequences "consistent" or "inconsistent" by comparing to whole sequence.
(8) Increase the the number of parameters by calculating simple formulation such as logarithm and maltiplication of length of restored sequence and overlap region.
(9) Obtain the target discriminant rule from Support Vector Machine with the training data that consists of above parameters value and consistency judgement.

In the discriminant method above, all parameters' value can be derived from the each assembled sequences even if we don't know the original sequence allocation. So we can apply the consequent discriminant rule to the actual assembling case. The restored sequence from the extended Warshall-Floyed algorithm in section 2 has a possibility that there is some inconsistent sequence. Appling the discriminant rule to this restored sequence, we can remove a part of inconsistent sequence. Then we will get higher accuracy for short read sequence assembly.

4. Experiments and Results

To confirm the effect of our proposal, we carried out numerical experiments.

We prepared corera virus sequence[4] which has 719 bases length. After the replication of original sequence, we produced 400 short read sequences by cutting randomly. Each short read sequence has 30-40 bases length, because of analogy for SOLiD environment. An adjacent graph was constructed based on the overlap region, and 5000 restored sequences were generated by 20 steps random walk procedure. Comparing the restored sequences and the original sequence, 52 restored sequences were consistent and other sequences were inconsistent. Length of overlap region, length of restored sequence, logarithm of them and multiplication of them were calculated for each restored sequence. With the consequent discriminant rule from SVM process for this training data, we applied to the training data itself. (Table. 2) To check the generality of this discriminant rule, we generated another 5286 restored sequences with the same whole sequence. Table3 shows that the learning effect was almost same between two datasets. For proposal extended Warshall-Floyed algorithm, we prepared 4 data sets as follows.

case1 100 short read sequences from the same corera virus sequence.
case2 400 short read sequences from the same corera virus sequence.
case3 100 short read sequences from the different corera virus sequence.
case4 400 short read sequences from the different corera virus sequence.

The different corera virus sequence has 705 bases length. We got 122, 164, 108 and 170 restored sequence for Case1, Case2, Case3 and Case4 respectively. Table 4, 5, 6 and 7 show the discriminant results.

	correct	incorrect
consistent	# seq judged as correct for "consistent"	# seq judged as correct for "inconsistent"
inconsistent	# seq judged as wrong for "consistent"	# seq judged as wrong for "inconsistent"

表 1 Discriminant result form

	correct	incorrect
consistent	25	27
inconsistent	11	4937

表 2 Accuracy of test data for training data

	correct	incorrect
consistent	25	37
inconsistent	25	5199

表 3 Accuracy of train data for open data

	cocorrect	incorrect
consistent	8	1
inconsistent	33	80

表 4 100 sequences from training data

	correct	incorrect
consistent	13	2
inconsistent	24	125

表 5 400 sequences of from training data

	correct	incorrect
consistent	23	0
inconsistent	56	29

表 6 100 sequences of open data

	correct	incorrect
consistent	1	8
inconsistent	0	161

表 7 400 sequences of open data

From these results, proposed procedure could remove many of inconsistent restored sequences, although some consistent sequences were removed too. Comparing 100 and 400 sequences case, more number of short read sequences gave better results. And even different original sequence case got the effect of the discriminant procedure.

Vol.2010-BIO-22 No.1
2010/7/28

5. Coclusion

For short read DNA sequences, we proposed an assembly procedure that included extended Warshall-Floyed algorithm and discriminant method with SVM. Experimental results showed some effectiveness in the sense of accuracy improvement. But we could not remove inconsistency absolutely. For future tasks, we will improve the searching algorithm so that the discriminant function shall be included.

参 考 文 献

1) Rene L. Warren , Granger G. Sutton , Steve J. M. Jones and Robert A. Holt : Assembling millions of short DNA sequences using SSAKE, Bioinformatics, vol.23 no.4, pp.500-501, (2007)
2) Daniel R. Zerbino and Ewan Birney : Algorithms for de novo short read assembly using de Bruijn graphs, Genome Research, vol.18, pp.821- 829, (2008)
3) Cormen, Thomas H.; Leiserson, Charles E., Rivest, Ronald L. : The Floyd-Warshall algorithm, pp.558-565 MIT Press and McGraw-Hill. (1990)
4) DNA data base (DDBJ), http://www.ddbj.nig.ac.jp/
5) EXPERIMENTAL MEDICINE,vol.26 no.7, Youdosya, (2008)