Vol.2015-MPS-103 No.64
Vol.2015-BIO-42 No.64
2015/6/25

Rule-based Assembly for Short Read Data Set obtained with Multiple Assemblers and k-mer Sizes

Ayako OSHIRO[1,a)] Hitoshi AFUSO[2] Takeo OKAZAKI[3]
Morikazu NAKAMURA[3]

Abstract: Various de novo assembly methods based on the idea of k-mer have been proposed. In despite of the success in these methods, another approach, called as Hybrid approach, that combine different traditional methods to take advantages of them have been proposed. However, the results obtained from traditional methods that used in hybrid approach depend on not only the algorithm or heuristics, but also the selection of user-specific k-mer size. Consequently, the results by hybrid approaches also depend on them. In this paper, we designed new assembly approach, called as Rule-based assembly. It follows the strategy similar to the hybrid approach. But it uses certain rules learned from some characteristics of draft contigs to remove erroneous ones and merges them. To construct effective rules, a learning method based on decision tree, called Complex decision tree was proposed. Comparative experiments were also conducted. The results showed that proposed method outperformed traditional one in certain case.

1. Introduction

Giga sequencers which use parallel processing technique output massive short reads that contains read errors. Various *de novo* assembly methods, such as Velvet[1], ABySS[2] and SSAKE[3], are based on the idea of k-mer. In despite of the efficacy of giga sequence technology, it had been pointed out that short reads that obtained giga sequencer often contains read-error and they occur mis-assembly. To remove such errorneous reads, various methods had been proposed. For instance of such methods, we can cite Trimmomatic[4], ECHO[5] and Quake[6].

On the other hand, some comparative studies for traditional assembly algorithms had been conducted in view of the ability of assembly itself[7][8][9]. In such situation, it had been understood that results of assembly depend on both algorithm of the assembler and parameter value, such like k value for k-mer. As a example that supports the fact, Oshiro,*et.al*[10] showed that complementality between different results obtained by ABySS in different setting of k-mer. To improve accuracy and contingency of the assembly results, some methods take another approach, called as *Hybrid approach*, have been developed. In the approach, the results from different traditional methods under various parameter settings are combined. As example for the methods following the approach, we can cite various methods such as, IDBA[11], IDBA-UD[12], MAIA[13], GAA[14] and CISA[15]. Such approach was applied to not only DNA sequence, but also mRNA sequence by Oases[16].

In despite of the success of the hybrid approach, there is one bottleneck for that. That is the fact that the accuracy or contingency of the conbined results depend on the accuracy of the individual traditional assembly method. Erroneous contigs results the combined assembly that contains error, inevetably. In other words, even though the complete hybriding method was archieved, if the output contigs contains error, correct assembly never be obtained. To solve the problem, i.e., to identify or detect the misassembly of contigs, several characteristics have been observed. For example, assembly errors often appear with the region with low read coverage, or it tends to contain chimeric or recombined reads, and so on. From the common observations for the misassembly, five traditional measures maximum and minimum length of low read coverage regions, maximum and minimum length of low clone coverage regions, and finally compression or expansion of paired end reads, have been designed. In addition to these measures, Choi,*et.al*[17] have been proposed four more measures to detect misassemblies. They also applied machine learning techniques to improve the accuracy of detection by combination of proposed measures. However, the proposed measure focused only the length or the number of clones. For example, the measure GMB is caculated as the difference between the number of good clones and bad clones, and goodness or badness is determiend by thresholding their deviations from the average length. From this definition, this measure does not consider about the frequency of k-mer that might contains informa-

[1] Graduate School of Engineering and Science, University of the Ryukyus
[2] Graduate School of Medical Research, University of the Ryukyus
[3] Faculty of Information Engineering, University of the Ryukyus
[a)] ayami@ms.ie.u-ryukyu.ac.jp

tion about the error. And also, in the paper[17], the learning results were not discussed about. We thought this reason is that they used the machine learning methods that is difficult to make the discriminat rule explicit, such as random forest[18], even though they have high ability to learn to detect the misassemblies.

To tackle this problem, we proposed new characteristics that represents the feature of output contigs. Next, to remove or detect the erroneous combining contigs, certain disctiminat rules were constructed. In the construction of the rules, we extended the traditional decisiton tree method to the one that can consider multiple objective variables. And also, the experiments to show the validity of proposal method was conducted. Finally, consdering and analysing the obtained discriminant rules, some hypothesis about the erroneous or error-free combination of contig were suggested.

2. Methods

2.1 Outline of Hybrid Approach

In this section, we give a blief introduction of hybrid approach. Hybrid approach is, as shown its name, the approach that combines the results from different assembers with different parameer settings. The main purpose is the improvement of accuracy and contingency of resulted contigs. As a example of the method that apply the hybrid approach, we give the explanation about the DAwH[10]. Firstly, the algorithm considers all possible combinations of contigs that have enough overlap length. And then, By applying the discriminat rules that obtained by machine learning algorithm with certain chracteristcs, erroneous combinations would be estimated and removed. In this approach, there are two important points. First, what characteristics should be used for the learning of discriminant rules?. And second, what algorithm is appropriate for the learning to improve accruracy and easy to interpret the discriminant results? Next, we give the definition of characteristics and learning algorithm.

2.2 Characteristics of *k*-mers

Fist, we give the defitions about the characteristics that used in this research. In traditional machine learning based method such as [17], the characteristics that focus the length of reads or clones were used in the learning step. In the defition of such characteristics, the frequency of k-mers that consist of reads and contigs is not considered. From two facts, that misassembly tends to occur in the regions that have low coverage values and the contigs as the result of assembly are composed by k-mers locally, it is expected that the information about the frequency of the k-mers that contained in reads might have potential to distinguish the correct or incorrect assembly of contigs. Consequently, we decided to use the characteristics that reflect the frequency of k-mers for the learning of disctiminant rules. However, since the frequency of k-mers depend on the coverage of reads, it becomes difficult to use the frequency information from the results obtained from different dataset that differ

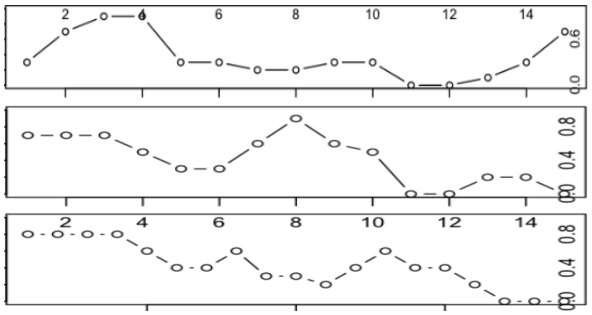

Fig. 1 Example of the plots of p-value corresponding contigs.

the read coverage each other. For example, suppose we have two informations from different datasets. Even though the frequency values are same, like 20, if the read coverages of two datasets are different, one is 100 and the other is 50, we could not simply compare the frequency as same value. To solve this problem, certain method to normalize the difference of read coverage is required. As such method, we used the measure that represents relative order of the frequency. Suppose that c_i denotes the frequency of k-mer k_i and C represents the set of whole k_i. Then the relative order p_{c_i} of k_i is defined as Eq.(1):

$$p_{c_i} = \frac{\|\{c_j \in C | c_i \leq c_j\}\|}{\|C\|} \quad (1)$$

Where $\|\cdot\|$ denotes the number of elements in the set. From now, we call this measure that defined in Eq.(1) as "p-value of k-mer k_i". Using p-value, we can compare the frequencies that obtained from two read sets that differ the read coverage. A example of the plots of p-value certain contigs is shown in Fig.1. In Fig.1, horizontal axis denotes the potions of k-mer in the cotig and vertical axis shows the p-value of corresponding k-mer. As shown in Fig.1, some variation patterns could be seen for each contigs. From these plots, we made hyphothesis that the the waveform of p-value corresponding contig might contains some information about the correctness of combining. In other words, the correctness of combining contigs might be decided by the combination of p-value waveform corresponding ones. To glance the validity of this hypotheis, preliminary experiment was conducted. In the experiment, we used *E.coli* genome data from NCBI[19]. From the dataset, some correct contigs were generated. Under the condition, for four cases, two correct and two incorrect cases, the patterns of p-value waveforms were analysed. Both correct and incorrect cases consisted of two sub-cases, fixing former contig and latter one. The results were shown in Fig.2. From the Fig.2, in both correct and incorrect cases, some similarity between the p-value waveform corresponding contigs without fixing was observed. This results suggest that the existence of appropriate combination of p-value waveform to generate correct combining of contigs, i.e., the combinations of p-value waveforms corresponding contigs might determine whether the result of combined contigs is correct or not. According to this result, we designed some characteristcs about the waveform of the p-value. The designed characteristics

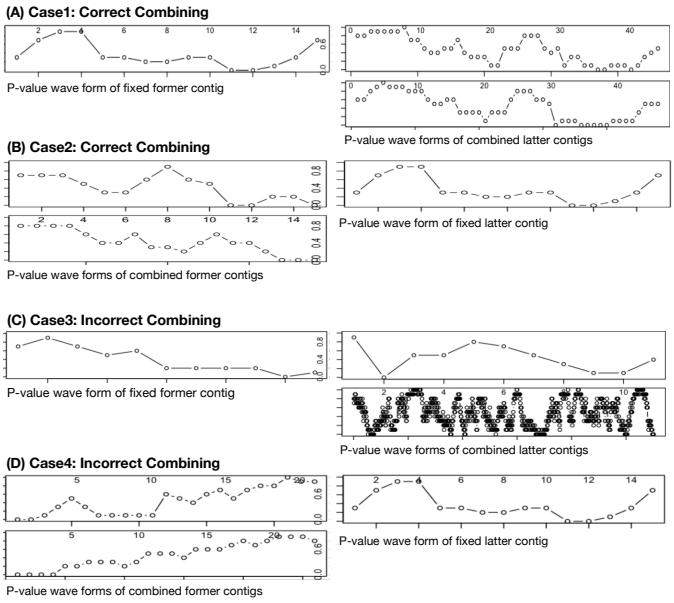

Fig. 2 Results of preliminary experiment: In both cases, correct and incorrect, some similarity between the p-value waveform corresponding contigs without fixing are shown.

Table 1 Designed characteristics represents the fluctuation of p-value waveform.

Fluctuation	$D^{f,l}$	Gradient of waveform
	$I^{f,l}$	Rate of increasing value
	$U^{f,l}_{freq}$	High frequency components
	$W^{f,l}$	Powered value in Fourier transform(F.T)

Table 2 Designed characteristics represents the distribution of p-value waveform.

Distribution	$L^{f,l}_{freq}$	Low frequency components
	$Q^{f,l}$	P-value with null frequency
	$W^{f,l}_{freq}$	Powered value of frequency distribution in F.T

Table 3 Designed characteristics represents the correlation of p-value waveform.

Correlation	ρ	Correlation
	ρ_{freq}	Correlation between frequency distributions
	$\Phi^{f,l}_{\max}$	Maximum cross-correlation
	H	Hamming distance between frequency distributions
	$R^{f,l}$	Length between end ponint of former and start point of latter contig.

are shown in from Table.1 to Table.3. In Table.1, the gradient of waveform D and the rate of increasing value I are difined as follows:

$$d_i = \begin{cases} -1 & (p_i < p_{i-1}) \\ 0 & (p_i = p_{i-1}) \\ 1 & (p_i > p_{i-1}) \end{cases} \quad (2)$$

$$D = \sum_{i=1}^{n} d_i \quad (3)$$

$$I = \frac{\|\{d_i | d_i = 1\}\|}{\|n\|} \quad (4)$$

Where p_i denotes the p-value of i-th k-mer of a read and n represents the number of k-mers contained in a read.

These characteriscis is unique from the point of view that traditional characteristics focused the length of reads or clones, not frequency. In addition, they are focused not only frequency infromation simply, but also its features as a waveform information. Next, using these characteristcs as explanatory variables, discriminant rules for contig conbining would be constructed.

2.3 Complex Decision Tree

Using previously designed characteriscs, discriminant rules that distinguish whether contig conbining is correct or not. For the construction of discriminant rules, there are various methods such as support vector machine[20] or neural networks[21]. In this study, for the convenience of interpreting the results of discrimination, decision tree making algorithm, C4.5[22] was utilezed. Firsly, to simply assess the validity of application of C4.5 algorithm for discrimination, preliminary comparison experiments were conducted. In this experiments, the data set of *E.coli* from NCBI with reference data same as previous experiment. The steps of the experiments were as follows. First, the read dataset was generated from the already known sequence. Second, the assembly with traditional methods for multiple k values. Third, all correct(consistent with the reference) contigs were collected and whole possible combination among them were constructed. The combination was mede only in the case the considering contigs are overlapped and its length is more than 5. Next, combined contigs were evaluated whether it is consistent with reference or not. Then, the characteristics descrived above were calculated for corrected combined contigs and discriminat rules were constructed using their values as explanatory variables. Finally, the results from traditional methods with and without applying discriminant rules were compared. Since two types of discriminant rules, for correct and incorrect, could be obtained, then we applied for each rules, respectively. Each assembly result was evaluated seven measures, the number of output contigs (#.Output), the number of correct contigs(#.Corr), the ratio of combined correctly (R.Corr), N50 contig length(N50), the ratio of the length of mapped region(R.Mapped), the maximum length of correcly combined contigs(ML.Corr) and the maximum length of incorrectly combined contigs(ML.Incorr). The results were shown in Table.4. As shown in Table.4, maximum length of corretly combined contigs, ML.Corr were elonged by using hybrid approach than traditionals. In the results from the hybrid method with incorrect rules, N50 and the ratio of mapped region, %.Mapped were also improved(shown in bold). On the other hand, the maximum length of incorrectly combined contigs ML.Incorr was elonged drastically in some hybrid methods(shown with underlines). In addition, it was observed the large decreasing of the ratio of correctly combined %.Corr(shown in bold with underlines). Although the effectiveness of hybrid approach and application of the rules were suggested in view of the ra-

Table 4 Results of preliminary comparison.

Method	#.Output	#.Corr	%.Corr	N50	%.Mapped	ML.Incorr	ML.Corr
Velvet(k=15)	20	19	0.95	2963	0.98	34	4815
Velvet(k=17)	12	12	1.00	7889	0.98	-	10850
ABySS(k=16)	54	54	1.00	3048	0.87	-	4817
ABySS(k=18)	40	40	1.00	7891	0.77	-	10852
CISA	38	38	1.00	4044	0.98	-	10852
Hybrid without Rules	671	412	**0.61**	7849	0.99	**15767**	18729
Hybrid with Correct	338	234	**0.69**	7849	0.96	**15767**	18729
Hybrid with Incorrect	444	315	**0.71**	**7906**	**0.99**	**15767**	18729

tio of correct combined contigs %.Corr, both correct and incorrect discriminat rules could not distinguish the corresponding combinations enough. From these results, more accurate correct and incorrect discriminat rules are required to improve the accuracy of resulted assembly. Especially, since the large number of the contigs combined incorrectly was obtained in the experiments, more precise discriminat rules for incorrect combinations are preffered. Such large number of incorrectly combined contigs may be occured by the fact that even though almost combination of contigs are correct, if only one incorrect combination is occured, then whole combination of contigs would be conterminated. And while it was not argued in this experiment, the application of both rules at same time might improve the assembly results.

2.4 Rule-based Assembly with Complex Decision Tree

In traditional method based on k-mer, heuristics such that k-mers have smaller frequencies are originated from read-error by sequencer was applied in common. And also, about the length of overlap certain heuristic have been used, called as "overlap-layout census". In this heuristics, two reads have long overlapped region are considered as a pair correctly combined. According to these heuristics, information about the k-mer that has low frequency and longer overlapped region among contigs might tell us the way to distinguish correctness or incorrectness of combinations. From this idea, added two more characteristics, the minimum ratio of the frequency of k-mer and the length of overlapped region between contigs. From now, these two characteristics are represented as $\%_{\min}$.Cover and L.Overlap, respectively. To check the effect of these additional characteristics to the discriminant task, the distributions of each characteristic with both correct and incorrect dataset. The plots of distribution for each characteristic are shown in Fig.3 and Fig.4. As shown in Fig.3, in both correct and incorrect case two distributions had large variance. However, the distribution corresponding incorrect case was strongly skewed. And also in Fig.4, the difference between two distributions was observed. From these results, it was shown that these chracteristics contain some information to distinguish correct or incorrect combinings.

Since $\%_{\min}$.Cover and L.Overlap are quantitative variables, multiple regression analysis was utilized to construct the disctiminant rules[23]. As explanatory variables, the characteristics shown in Table.1 to Table.3 were used.

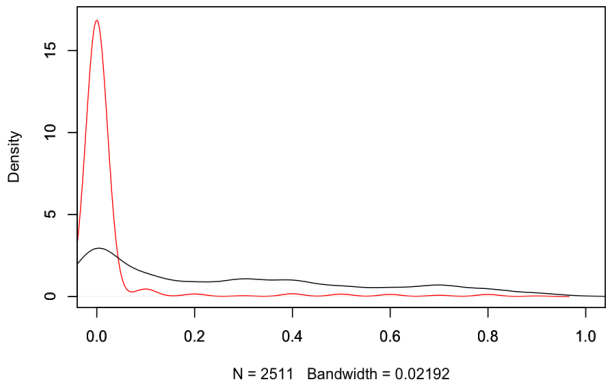

Fig. 3 Distribution of $\%_{\min}$.Cover. The density function was estimated with kernel density estimation method. Horizontal axis denotes the minimum ratio of the frequency. Black and red lines represent the desity function for correct case and the one for incorrect, respectively.

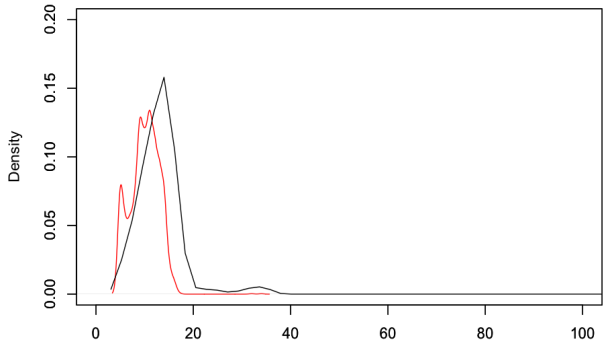

Fig. 4 Distribution of L.Overlap. The density function was estimated with kernel density estimation method. Horizontal axis denotes the minimum ratio of the frequency. Black and red lines represent the desity function for correct case and the one for incorrect, respectively.

$\%_{\min}$.Cover and L.Overlap were analysed individually as objective variable. The analysis was based on model selection using AIC[24]. However, the discriminat ability of resulted rules from the analysis was not enough(detailed values are not shown). This result might be originated from the high variance of objective variable or they are difficult to be discriminate linearly.

To solve this problem, decision tree algorithm was applied. In the application, the characteristics shown in Table.1 to Table.3 were used as explanatory variables and for each variable, $\%_{\min}$.Cover and L.Overlap, decision trees were generated. Eventually, to include the information contained $\%_{\min}$.Cover and L.Overlap, the correspoinding k-mers and results obtained from the decision trees that these two variable had been used as objectives were evaluated whether

correct or incorrect. We named this hierarchical and multi objective variables based decision tree as "Complex Decision Tree(CDT)", as a extension of traditional decision tree for one objective variable.

Using this CDT, we proposed new assembly method, named "Rule-based Assebly". The steps of the algorithms as follows. First the contigs are generated using some assembly algorithm with various parameters. Second, comparing the contigs to the reference, filtering out the incorrect contigs are removed. Next, from the all correct contigs characteristics shown in Table.1 to Table.3 are calculated. Using these characteristics values and additional parameters, minimum ratio of frequency and the length of overlap regions, complex decision tree composed with multiple objective variables is constructed. These steps are for the generation of the disctiminat rules to distinguish the combined contigs are correct or not. And finally, using obtained discriminant rules, cut or remove the some possible combinations of contigs that have overlap longer than 5. After the removement, the schafolds would be generated from remaining contig combinations.

3. Experiments

3.1 Experiments for Rule Construction and Evaluation

To show the validity of the rules, constructions of discriminant rules and evaluation of them were conducted. Especially, to confirm the ability of our proposal part, decision trees for the $\%_{min}$.Cover and L.Overlap were generated. Consequently, the corresponding k-mers were evaluated in view of consistency(correctness) to the reference. After the evaluations, we designed new discriminat rules to distinguish the combined contig is correct or not. Finally we made another evaluation for the designed rules in benchmark dataset.

The generated decision trees were constructed three rules. First, if $\%_{\min}$.Cover is greather than 0, then the corresponding combination of contigs would be correct with the probality 99.6%. Second, if the value of L.Overlap in combined contigs is greater than 14, the combination would be correct with the probability 93.3%. And finally, if $\%_{\min}$.Cover is less than 0 and L.Overlap is smaller than 14, then the combination would be incorrect with the probability 72.8%. Combining these rules, we designed two discriminat rules as follows:

Rule for Correct Combinings(RfC): $\%_{\min}$.Cover is greater than 0 and L.Overlap is longer than 14.

Rule for Incorrect Combinings(RfI): $\%_{\min}$.Cover is less than 0 and L.Overlap is shorter than 14.

As the first step of the evaluation two rules RfC and RfI, the benchmark dataset was generated using *E.coli* genome. We constructed a read set by artificial samplings. Then, the contigs were obtained by traditional assembly method. After the assembly, we filtered out incorrect contigs comparing reference data. As the result, 647 correct contigs were obtained as a benchmark dataset. In the Table.5 and Table.6, Acc denotes the accuracy rate. It was caculated as follows:

Table 5 The discriminat result obtained by RfC.

Responce \ Answer	Correct	Incorrect	Acc
Correct	394	3	0.937
Incorrect	14	234	

Table 6 The discriminat result obtained by RfI.

Responce \ Answer	Correct	Incorrect	Acc
Correct	400	8	0.951
Incorrect	24	215	

$$\mathrm{Acc} = 1 - \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{N}} \tag{5}$$

Where N denotes the total number of contigs. FP represents the number of cases that rule discriminated incorrect contigs as correct ones and FN does the number of cases opposite to FP.

As shown in Table.5 and Table.6, both rules could distinguish correctness of combined contigs with high correct answer rate. The correct answer rates of both rules were greater than 0.9. In addition, the rule for incorrect combinings(RfI) resulted higher correct ratio. From the results, it was expected that by using RfI, more accurate removement of incorrect combining of contigs could be archieved.

3.2 Experiments for Rule Application

As next step, to show the ability of whole proposed algorithm, comparison experiments were conducted. In the experiments, as traditional methods, Velvet, ABySS and CISA were utilized. In addition, as hybrid methods, hybrid method without rules, two methods that with correct and incorrect rules, respectively and proposed method were compared. As the training dataset, we generated 30,000 reads from *E.coli* sequence its length is 30,000. Length of each read was 50, i.e., the depth of reads of this dataset was 50. In the hybrid methods, ABySS and Velvet were used. Settings of k value were 16 and 18 for ABySS, and 15 and 17 for Velvet.

As the results of discriminant rule constructions, 36 rules for correct combinings were generated. They were composed by 10 rules from the decision tree that used the length of overlap region as objective, 17 rules from the one used the minimum ratio of frequency and 9 rules from traditional C4.5 decision tree. Obtained 32 rules for incorrect combinings consisted of 9 rules from the length of overlap, 6 rules from the minimum ratio and 17 from ordinary C4.5 algorithm. We labeled the rules that obtained from the length of overlap as Ovl$X$, where X denotes the index of corresponding rule. As similar to that, the labels MnRX for the rules obtained from the minimum ratio of frequency were assined. The rules generated from the traditional C4.5 algorithm were labeled as Trd$X$.

Applying obtained rules, we found some effective combinations of rules to remove incorrect combined contigs. The rules and the length of incorrectly combined contigs were shown in Table.9. From the effective rules shown in Table.9, we used three compositions, First, Ovl2 and Ovl5. Second,

Table 7 The list of positive rules from complex decision tree.

Ovl1	$D^l \leq -10, \rho \leq 2.91$
Ovl2	$33 < \Phi_{\text{freq}}, D^f \leq 2, 0.6 < U^l_{\text{freq}}, U^l_{\text{freq}} \leq 5.1$ $L^l_{\text{freq}} \leq -0.10$
Ovl3	$\rho \leq 2.91, 1488.1 < U^l_{\text{freq}}, -0.10 < L^l_{\text{freq}}$
Ovl4	$D^f \leq 2, Q^f \leq 0.2, 0.63 < \rho, \rho \leq 0.96$ $-0.10 < L^l_{\text{freq}}$
Ovl5	$8 < H, 0.96 < \rho, \rho \leq 2.91, L^l_{\text{freq}} \leq 0.503$
Ovl6	$\rho \leq 0.28, -0.10 < L^l_{\text{freq}}$
Ovl7	$D^f \leq 2, Q^f \leq 0.2, \rho \leq 0.96, I^l \leq 0.17, L^l_{\text{freq}} \leq -0.10$
Ovl8	$Q^l < 0.5, U^f_{\text{freq}} \leq 1.3, 5.1 < U^l_{\text{freq}}, L^l_{\text{freq}} \leq -0.10$
Ovl9	$Q^l \leq 0.5, \rho \leq 2.91, L^l_{\text{freq}} \leq -0.10$
Ovl10	$D^f \leq 2, \rho \leq 0.96, -0.10 < L^l_{\text{freq}}$
MnR1	$0.1 < Q^l, \rho \leq 5.52, I^l \leq 0.38, 8.003 < W^f_F, L^f_{\text{freq}} \leq 9.5$
MnR2	$R \leq 0.1, 643.86 < W^f_F >, 708.6 < U^f_{\text{freq}}$
MnR3	$-2 < D^f, 0.1 < Q^l, \rho \leq 5.52, I^l \leq 0.384, 8.00 < W_F$
MnR4	$\rho_{\text{freq}} \leq 380631, 2.83 < W^f_F, 566.37 < W^l_F, L^l_{\text{freq}} \leq 1490.2$
MnR5	$R \leq 0.3, D^f \leq 1, 0.1 < Q^l, Q^l \leq 0.3, 1.12 < \rho, 5.52 < \rho, I^l \leq 0.38$
MnR6	$4.20 < \rho_{\text{freq}}, D^f \leq 1, \rho \leq 1.12, I^f \leq 0.5, I^l \leq 0.5, U^f_{\text{freq}} \leq 401.4$
MnR7	$1.22 < \rho, I^l \leq 0.38, 9.6 < U^f_{\text{freq}}, U^l_{\text{freq}} \leq 4.2$
MnR8	$D^f \leq 1, H \leq 5, \rho \leq 1.12, I^f \leq 0.5, I^l \leq 0.5$
MnR9	$0.3 < R, R \leq 0.6, 0.01 < Q^l, I^l \leq 0.38, 2.83 < W^f, W^f \leq 8.00$
MnR10	$0.1 < Q^l, \rho \leq 5.528.003 < W^f, 6.6 < U^l_{\text{freq}}$
MnR11	$\Phi_{\text{freq}} < 898574, 566.37 < W^l, U^f_{\text{freq}} \leq 708.6, U^l_{\text{freq}}\ leq 1490.2$
MnR12	$7 < H, 0.1 < Q^l, \rho \leq 5.52, I^l \leq 0.38, 2.83 < W^f$
MnR13	$0.8 < Q^l, 5.52 < \rho, W^f \leq 643.86, U^l_{\text{freq}} \leq 130$
MnR14	$D^l \leq 1, \rho \leq 6.55, 0.38 < I^l, I^l \leq 0.5, 2.83 < W^f$
MnR15	$I^f \leq 0.125, I^l \leq 0.5, W^f \leq 2.83, W^l \leq 25.3$
MnR16	$H \leq 2, 0.38 < I^l$
MnR17	$0.2 < Q^l, I^l \leq 0.38, 643.9 < W^f$
Trd1	$R \leq 0.1, 708.6 < U^f_{\text{freq}}$
Trd2	$\leq 0.1, 898574 < \Phi_{\text{freq}}, U^l_{\text{freq}} \leq 1905.7$
Trd3	$D^f \leq 1, D^l \leq 1, Q^l \leq 0.3, \rho \leq 2.91, U^f_{\text{freq}} \leq 401.4$
Trd4	$\rho \leq 2.91, L^l_{\text{freq}} \leq -0.10$
Trd5	$R \leq 0.1, 182.32 < \rho, I^f < 0.23$
Trd6	$5 < H, 182.32 < \rho$
Trd7	$0.1 < R, 2908.5 < \Phi_{\text{freq}}, \Phi_{\text{freq}} \leq 6787.5, 2.91 < \rho$
Trd8	$0.6 < Q^f, 0.24 < I^l, W^l \leq 5.33$
Trd9	$W^f \leq 331.4$

Table 8 The list of negative rules from complex decision tree

Ovl1	$H \leq 8, 0.96 < \rho, \rho \leq 2.91, U^l_{\text{freq}} \leq 1488.1, -0.10 < L^l_{\text{freq}} > -0.10$
Ovl2	$2 < D^f, L^l_{\text{freq}} \leq -0.7$
Ovl3	$1.3 < U^f_{\text{freq}}, 5.1 < U^l_{\text{freq}}, U^l_{\text{freq}} \leq 14.9, L^l_{\text{freq}} \leq -0.10$
Ovl4	$2 < D^f, -0.10 < L^l_{\text{freq}}$
Ovl5	$2.91 < \rho$
Ovl6	$2 < D^f, 1 < D^l$
Ovl7	$7 < D^f$
Ovl8	$-10 < D^l, 0.5 < Q^l, L^l_{\text{freq}} \leq 0.10$
Ovl9	$R \leq 0.2$
MnR1	$0.5 < I^l$
MnR2	$116387.2 < \Phi_{\text{freq}}, \Phi_{\text{freq}} \leq 898574$
MnR3	$0.1 < R, H \leq 5, 0.6 < Q^l, I^f \leq 0.227$
MnR4	$R \leq 0.6, D^f \leq -2, Q^f \leq 0.8, 9.5 < U^f_{\text{freq}}, 4.2 < U^l_{\text{freq}}$
MnR5	$\Phi_{\text{freq}} \leq 4.2, 5 < H$
MnR6	$R^l_{inc} \leq 0.5$ [0.601]
Trad1	$0.1 < R, 6787.5 < \Phi_{\text{freq}}, 0.291 < \rho, \rho \leq 0.321, I^l \leq 0.24$
Trad2	$0.1 < R, 6787.5 < \Phi_{\text{freq}}, 0.6 < Q^f, \rho_{\text{freq}} \leq 0.58, \rho \leq 0.182, 5.33 < W^l$
Trad3	$0.1 < R, 6787.5 < \Phi_{\text{freq}}, 0.6 < Q^f, -0.031 < \rho_{\text{freq}}, \rho_{\text{freq}} \leq 0.56, 0.291\rho, \rho \leq 0.182$
Trad4	$R \leq 0.1, 489 < \Phi_{\text{freq}}, -1 < D^f, 0.3 < I^f, W^l \leq 5464.053$
Trad5	$56.78 < \Phi_{\text{freq}}, D^l \leq 1, Q^l \leq 0.1, 0.291 < \rho$
Trad6	$R \leq 0.1, W^l \leq 5464.053, U^f_{\text{freq}} \leq 708.6, 1905.7 < U^l_{\text{freq}}$
Trad7	$R \leq 0.1, H \leq 5, 0.182 < \rho, I^f < 0.23$
Trad8	$R \leq 0.1, \Phi_{\text{freq}} \leq 898574, 331.414 < W^{f,l}, U^f_{\text{freq}} \leq 708.6$
Trad9	$-5 < D^l, 3 < H, 0.291 < \rho, 10.1 < U^f_{\text{freq}}, U^f_{\text{freq}} \leq 18.2$
Trad10	$1 < D^f, \rho \leq 0.291, 4.21 < W^f, I^l_{\text{freq}} \leq 7.6, -0.10 < L^l_{\text{freq}}$
Trad11	$1 < D^f, 3 < H, 0.1 < Q^l, 0.291 < \rho, U^f_{\text{freq}} \leq 18.2$
Trad12	$0.1 < R, \Phi_{\text{freq}} \leq 2908.5, 0.291 < \rho, 19.155 < W^l$
Trad13	$0.1 < R, \Phi_{\text{freq}} \leq 2908.5, 1 < D^l, 0.291 < \rho$
Trad14	$D^l \leq 1, 5 < H, Q^l \leq 0.3, 401.4 < U^f_{\text{freq}}, -0.10 < L^l_{\text{freq}}$
Trad15	$1 < D^l, Q^f \leq 0.1, \rho \leq 0.291, 4.21 < W^f$
Trad16	$0.3 < Q^l, \rho \leq 0.291, 4.214 < W^f$
Trad17	$1 < D^l, \rho_{\text{freq}} \leq -0.17, 4.21 < W^f, -0.101 < L^l_{\text{freq}}$

Table 9 The list of rules that could remove large incorrect combined contigs

Length of Removed Contigs	Rule ID
15767	Ovl5, Ovl9, MnR6
15767	Ovl4, Ovl5, Ovl6, Ovl7, Ovl9, MnR6
12695	Ovl4, Ovl5, Ovl6, Ovl7, Ovl8, Ovl9, MnR6
10872	Ovl2, Ovl4, Ovl5, Ovl7, Ovl9, MnR6
10860	Ovl2, Ovl7, Ovl9, MnR1
10859	Ovl9, MnR1

Ovl5 and Ovl9. And finally, Ovl5 and MnR1.

Since the most of combined contigs by hybrid methods were longer than 1,500, we added two more evaluation indices, %.Corr$_{1500}$ and %.Mapped$_{1500}$. They represented the ratio of the correct combining contigs longer than 1,500 and the ratio of the mapped correcly with length loger than 1,500, respectively.

The comparison results were shown in Table.10. As shown in Table.10, #.Correct was increased in the case of all rules for correct combinings were used than hybrid with correct only, from 234 to 350. However, %.Correct was decreased from 0.69 to 0.68. It means that applying all positive rules could derive more correct combined contigs than the method used correct rules only. In other words, although the ability of rules for correct contigs were improved, its application also led to the increasing the incorrect combined contigs. By additional applying of effective negative rules with all rules for correct, %.Correct was improved. Especially in the case of All + Ovl5 + Ovl2, its value increased from 0.680 to 0.827. It means that complex decision tree could remove incorrect contig than C4.5 more accurately. This also means that the ability of classification was improved compared to traditional decision tree that uses single objective variable. Furthermore, ML.Incorr of the method with ALL + Ovl5 + Ovl9 and the one with ALL + Ovl5 + MnR1 were decreased than traditional hybrid assembly. This results suggest that the complex decision tree could remove large incorrect contigs. The ML.Corr resulted hybrid methods were increased to 18729. From these results, we can say that hybrid methods could derive longer correctly combined contigs than traditional assembly. In addition, the method ALL + Ovl5 + Ovl9 and All + Ovl5 + MnR1 resulted high

Table 10 Results of performance comparisons between traditionals and hybrid methods

Method	#.Output	#.Correct	%.Correct	N50	%.Mapped	ML.Incorr	ML.Corr	$\%.\text{Correct}_{1500}$	$\%.\text{Mapped}_{1500}$
Velvet(k=15)	20	19	0.950	2963	0.980	34	4815	1.000	0.810
Velvet(k=17)	12	12	1.000	7889	0.980	-	10850	1.000	0.900
ABySS(k=16)	54	54	1.000	3048	0.870	-	4817	1.000	0.720
ABySS(k=18)	40	40	1.000	7891	0.770	-	10852	1.000	0.720
CISA	38	38	1.000	4044	0.980	-	10852	0.465	0.939
Hybrid without Rule	671	412	0.614	7849	0.990	15767	18729	0.465	0.939
Hybrid with Correct	338	234	0.690	7849	0.960	15767	18729	0.570	0.938
Hybrid with Incorrect	444	315	0.710	7906	0.990	15767	18729	0.610	0.940
All Correct Rules(ALL)	512	315	0.680	7356	0.960	15767	18729	0.570	0.938
ALL + Ovl5 + Ovl2	231	191	0.827	5631	0.960	10859	10854	0.890	0.908
ALL + Ovl5 + Ovl9	43	33	0.767	10854	0.628	63	10855	1.000	0.618
ALL + Ovl5 + MnR1	180	147	0.817	3148	0.603	1122	7892	1.000	0.550

value of $\%.\text{Correct}_{1500}$, 1.0. These results also suggest that complex decision tree could generate longer combinations of contigs.

The comparative results showed that the ability of hybrid assembly method can generate longer correct overlapped contigs than traditional assembly methods. And also it suggested that the application of discriminant rules generated by complex decision tree to the task for combining obtained contigs, has potential ability to generate more number of correct contigs and improve the accuracy of the resulted combinations of contigs than traditional decision tree.

4. Conclusion

In order to improve the accuracy of the combination of contigs for hybrid assembly method, we proposed complex decision tree with multiple objective variables. The combination of discriminant rules for both correct and incorrect was utilized and it led to the more accurate combinations of contigs. In addition from the result of comparisons with traditional assembly methods, the improvements of the quality indices, the length of correct overlapping contigs, correct ratio, and coverage ratio of large correct contig group, were observed. Furthermore, the ability of discriminant rules was improved compared to the rules simply generated with traditional method. From these result, for error-free read dataset generated artificially, our proposal, complex decision tree could generate more adequate discriminant rules than traditional decision tree algorithm. Consequently, the application of discriminant rules obtained with proposed method to the hybrid methods could archieved more accurate combinations of contigs.

As future tasks, it remains that the application to the dataset contains read errors and performance evaluation in the case. In addition, the more concrete analysis about the generated discriminat rules is also required.

References

[1] Zerbino, D.R. and Birney. E, (2008) Velvet: Algorithms for De novo Short Read Assembly using de Bruijn Graphs, *Genome Res* Vol.18, pp.821-829

[2] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. (2009) ABySS: A parallel assembler for short read sequence data *Genome Research*, Vol.19, Issue.4, pp.1117-23.

[3] Ren L. Warren*, Granger G. Sutton1, Steven J. M. Jones and Robert A. Holt (2006) Assembling millions of short DNA sequences using SSAKE, *Bioinformatics*, Vol.23, Issue.4, pp.500-501.

[4] Anthony M. Bolger, Marc Lohse and Bjoern Usadel, (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data, *Bioinformatics*

[5] Wei-Chun Kao, Andrew H. Chan and Yun S. Song, (2011) ECHO: A reference-free short-read error correction algorithm *Genome Res.* Vol.21, pp.1181-1192

[6] Kelley DR, Schatz MC, Salzberg SL, et al, (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, Vol.11

[7] Jason R. Miller, Sergey Koren, and Granger Sutton, (2010) Assembly Algorithms for Next-Generation Sequencing Data, *Genomics*, Vol.95 pp.315-327

[8] Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng HW (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies, *Bioinformatics*, Vol.15, pp.2031-2037

[9] Steven L. Salzberg, Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, Michael C. Schatz, Arthur L. Delcher, Michael Roberts, Guillaume Marais, Mihai Pop and James A. Yorke (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res.*, Vol.22, pp.557-567

[10] Ayako OHSHIRO, Takeo OKAZAKI, Morikazu NAKAMURA (2014) Double assembly method with characteristics of k-mer's coverage for contig, *IJCSNS International Journal of Computer Science and Network Security*, Vol.14 No.2

[11] Yu Peng, Henry Leung, S.M. Yiu, Francis Y.L. Chin (2010) IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler, *Research in Computational Molecular Biology, 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28*, Vol.6044, pp.426-440

[12] Yu Peng, Henry C. M. Leung, S. M. Yiu and Francis Y. L. Chin (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, *BIOINFORMATICS*, Vol.28 No.11, pp.1420-1428

[13] Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek,Jean-Marc Daran, Marcel Reinders, and Dick de Ridder (2010) Integrating genome assemblies with MAIA, *ECCB*, Vol.26, pp.433-439

[14] Guohui Yao. Liang Ye, Hongyu Gao, Patrick Minx, Wesley C. Warren, George M. Weinstock (2012) Graph accordance of next-generation sequence assemblies, Vol.28, No.1, pp.13-16

[15] Shin-Hung Lin, Yu-Chieh Liao (2013) CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes, Vol.8, Issue.3, e60843

[16] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron and

Ewan Birney (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, *BIOINFORMATICS*, Vol.28, No.8

[17] Jeong-Hyeon Choi, Sun Kim, Haixu Tang, Justen Andrews, Don G. Gilbert and John K. Colbourne (2008) A machine-learning approach to combined evidence validation of genome assemblies, Vol.24, No.6 , pp.744-750

[18] Leo Breiman (2001) Random Forests, *Machine Learning*, Vol.45, pp.5-32

[19] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/

[20] Lance E Palmer, Mathaeus Dejori, Randall Bolanos1 and Daniel Fasulo, (2011) Improving de novo sequence assembly using machine learning and comparative genomics for overlap correction, *BMC Bioinformatics*, Vol.11, Issue.33, doi:10.1186/1471-2105-11-33

[21] Angeleri E, Apolloni B, de Falco D, Grandi L., (1999) DNA fragment assembly using neural prediction techniques, *Int. J. Neural. Syst*, Vol9, Issue.6, pp.523-44.

[22] Quinlan, J. R., (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers

[23] Yong Ma, Shihong Lao, Erina Takikawa and Masato Kawade, Discriminant Analysis in Correlation Similarity Measure Space, Sensing & Control Lab., Omron Corporation, Kyoto, 619-0283, Japan,

[24] Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, Budapest, pp.267-281.