[DOI: 10.2197/ipsjtbio.10.9]

Original Paper

Rule-based Assembly for Short-read Datasets Obtained with Multiple Assemblers and *k*-mer Sizes

Ayako Ohshiro[1] Hitoshi Afuso[2] Takeo Okazaki[3] Morikazu Nakamura[3]

Received: October 25, 2016, Accepted: January 27, 2017

Abstract: Various *de novo* assembly methods based on the concept of *k*-mer have been proposed. Despite the success of these methods, an alternative approach, referred to as the hybrid approach, has recently been proposed that combines different traditional methods to effectively exploit each of their properties in an integrated manner. However, the results obtained from the traditional methods used in the hybrid approach depend not only on the specific algorithm or heuristics but also on the selection of a user-specific *k*-mer size. Consequently, the results obtained with the hybrid approach also depend on these factors. Here, we designed a new assembly approach, referred to as the rule-based assembly. This approach follows a similar strategy to the hybrid approach, but employs specific rules learned from certain characteristics of draft contigs to remove any erroneous contigs and then merges them. To construct the most effective rules for this purpose, a learning method based on decision trees, i.e., a complex decision tree, is proposed. Comparative experiments were also conducted to validate the method. The results showed that proposed method could outperformed traditional methods in certain cases.

Keywords: DNA assembly, *k*-mer, hybrid assembly, decision tree, C4.5

1. Introduction

Giga-scale sequencers that use a parallel processing technique provide an output of massive short reads. Various *de novo* assembly methods have been developed based on the idea of *k*-mer, including Velvet [1], ABySS [2], and SSAKE [3]. Despite the efficacy of giga-sequence technology, the short reads obtained often contain reading errors, resulting in misassembly. Therefore, various methods have also been developed to remove such erroneous reads, such as Trimmomatic [4], ECHO [5], and Quake [6].

Some comparative studies of traditional assembly algorithms have been conducted from the perspective of the ability of the assembly itself [7], [8], [9]. These studies have shown that the assembly result depends on both the specific algorithm of the assembler and a parameter value such as the *k* value of the *k*-mer (i.e., the sequence length). Moreover, alternative methods based on a different approach have been developed, which are collectively known as the *hybrid approach*. In this approach, the results from different traditional methods are integrated, such as MAIA [10], GAA [11], and CISA [12]. Furthermore, other methods exist that integrate the results from different *k*-values such as IDBA [13] and IDBA-UD [14]. These approaches have been applied for analyses of both DNA and mRNA sequences using Oases assembly [15].

In spite of the success of the hybrid approach, there remains a major hurdle, which is that traditional hybrid approaches focus on the length improvement of overlapped contigs regardless of their accuracy.

The production of erroneous contigs will inevitably lead to errors in the combined assembly. In other words, even though the complete hybridization method might be achieved, if the output contigs contains an error, correct assembly will never be obtained. To solve this problem, i.e., to identify or detect the misassembly of contigs, several general characteristics of this problem can be considered. For example, assembly errors often appear within a region with low-read coverage, or one that tends to contain chimeric or recombined reads. Based on the common observations in misassembly, five traditional measures have been developed: maximum and minimum lengths of low-read coverage regions, maximum and minimum lengths of low-clone coverage regions, and compression or expansion of paired-end reads. In addition to these measures, Choi et al. [16] have proposed four more measures to detect misassemblies. They also applied machine-learning techniques to improve the accuracy of detection obtained with a combination of the proposed measures. However, the measures proposed thus far have focused only on either the length or number of clones. For example, the measure GMB is calculated as the difference between the number of good clones and bad clones, in which goodness or badness is determined by thresholding their deviations from the average length. Given this definition, this measure does not take into account the frequency of a *k*-mer that might contain information about the error. In addition, in Choi et al. [16], the learning results were not discussed. This is likely because since they used machine-learning methods it is difficult to make a discriminant rule explicit, such as a random forest [17], even though this results in a high ability to learn to detect misassemblies.

[1] Graduate School of Engineering and Science, University of the Ryukyus, Nakagami, Okinawa 903–0213, Japan

[2] Graduate School of Medical Research, University of the Ryukyus, Nakagami, Okinawa 903–0215, Japan

[3] Faculty of Information Engineering, University of the Ryukyus, Nakagami, Okinawa 903–0213, Japan

We previously proposed a double-assembly method that merges the different results obtained by Velvet and ABySS under different settings of *k*-mer, and employed discriminant rules with a characteristic distribution of the *k*-mer coverage value for overlapping contigs, named DAwCC (Double Assembly method with Characteristics of *k*-mer's coverage for contig) [18]. In the process of DAwCC, a decision tree is utilized to construct the discriminant rule. The derived decision tree is composed of correct rules for overlapping contigs correctly, and incorrect rules for overlapping contigs incorrectly. However, evaluation of the performance of DAwCC revealed that the use of correct binding rules could not provide completely correct overlapping contigs and the incorrect binding rules could not remove all of the incorrect overlapped contigs completely. Therefore, improvement of the derivation of discriminant rules remains as an essential task in the development of DAwCC.

In recent years, ensemble machine-learning algorithms have been proposed for integrating traditional classifiers. Breimen proposed Bagging prediction [19], which divides a training data set into sub-data sets, generates classifiers for each of them, and finally generates a classifier based on the majority of multiple classifiers. In addition, Random Forest is a random explanatory variable that is selected for each divided sub-data set, and a final classifier is ultimately generated based on the majority from multiple classifiers, as in Bagging. Boosting [20] was also proposed, which involves updating the explanatory variable by weighting with respect to each misclassified piece of data. Some comparative studies for the decision tree algorithms [21] and ensemble machine-learning algorithms [22] have been performed. Although these methods show superior learning ability, the configuration of a classifier and its effective explanatory power remains unknown. Therefore, it is necessary to refer to the contents of the classifier to achieve improvement of the decision tree.

In this paper, we propose a method for construction of a complex decision tree with the use of multiple objective variables and combinations of positive and negative rules, in order to derive the discriminant rules of overlapped contigs for DNA double assembly. First, we describe our originally (traditional) proposed method, DAwCC. Second, we evaluate two aspects of development of this method: one is the possibility of using multiple objective variables based on their distribution, and the other considers integration of positive and negative rules. Based on these considerations, we propose a double-assembly method with a complex decision tree named DAwCDT. Finally, to confirm the effectiveness of DAwCDT, comparative verification was conducted between the new method and the traditional assembly methods ABySS, Velvet, DAwCC, as well as the traditional hybrid assembly method CISA.

2. Methods

2.1 DAwCC (Double Assembly method with Characteristics of the *k*-mer Contig Coverage)

In this section, we provide a brief introduction of our previous method DAwCC [10], that is necessary for description of new approach DAwCDT described at Section 2.2. Double assembly involves integrating the results obtained from different assembly methods with different *k*-mer settings, and considers all possible combinations of contigs that have a sufficient overlap, meaning exact match length. Then, by applying the discriminant rules obtained by the machine-learning algorithm with certain characteristics, erroneous combinations would be estimated and removed. Since the frequency of *k*-mers depends on the coverage of reads, it becomes difficult to use the frequency information from the results obtained from another dataset that differs with respect to read coverage. To normalize the difference in read coverage, we used a measure that represents the relative order of the frequency. Suppose that c_i denotes the frequency of *k*-mer k_i, and C represents the set of the whole k_i. Then, the relative order p_{c_i} of k_i is defined as Eq. (1):

$$p_{c_i} = \frac{|\{c_j \in C | c_i \leq c_j\}|}{|C|} \tag{1}$$

Where $|\cdot|$ denotes the number of elements in the set. Hereafter, we call this measure defined in Eq. (1) as the "*k*-percentile k_i". Using the *k*-percentile, we can compare the frequencies obtained from some read sets that differ in read coverage. We determined some characteristics related to the waveform of the *k*-percentile, which are shown in **Tables 1**, **2**, **3**.

In Table 1, the gradient of waveform D and the rate of the increasing value I are defined as follows:

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

$$D = \sum_{i=1}^{n} d_i \tag{3}$$

$$I = \frac{|\{d_i | d_i = 1\}|}{|n|} \tag{4}$$

Where p_i denotes the *k*-percentile of the *i*-th *k*-mer of a read, and n represents the number of *k*-mers contained in a read. These

Table 1 Designed characteristics representing the fluctuation of the *k*-percentile 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 *k*-percentile waveform.

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

Table 3 Designed characteristics representing the correlation of *k*-percentile waveforms.

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 the end point of the former and the start point of the latter contig.

characteristics are unique from the point of view that traditional characteristics are focused on the length of reads or clones rather than on the frequency. In addition, these characteristics are focused on not only simple frequency information but also on its features with respect to waveform information.

Next, using these characteristics as explanatory variables, discriminant rules for determining whether or not a contig combination is correct were constructed. There are various available methods for the construction of discriminant rules, such as support vector machine [23] or neural networks [24]. In Ref. [10], for the convenience of interpreting the results of discrimination, we selected a decision tree-making algorithm, C4.5 [25]. First, to simply assess the validity of the application of the C4.5 algorithm for discrimination, preliminary comparison experiments were conducted. In this experiment, the dataset of *E. coli* from NCBI reference data was used, as in the experiment described above. The steps of the experiments were as follows. First, the read dataset was generated from the already known sequence. Second, contigs were derived from the assembly using traditional methods for multiple k values. Third, all of the correct (consistent with the reference) contigs were collected, and all possible combinations among them were constructed. The combination was made only for cases in which contigs overlapped with more than 5 bases. Next, combined contigs were evaluated with respect to consistency with the reference. Then, the characteristics described above were calculated for correcting the combined contigs and discriminant rules were constructed using these values as explanatory variables. Finally, the results from traditional methods with and without applying discriminant rules were compared. Since two types of discriminant rules could be obtained for correct and incorrect contig pairs, we applied these separately for each rule.

As a comparative result, the the maximum length of correctly combined contigs, N50 and the ratio of the mapped region were improved by using a DAwCC approach compared to the traditional hybrid approach. On the other hand, the maximum length of incorrectly combined contigs, was drastically increased in some cases with DAwCC. In addition, there was a large decrease in the ratio of correctly combined contigs. Although these results suggest the effectiveness of DAwCC in view of the ratio of correct combined contigs, neither the correct nor incorrect discriminant rules could sufficiently distinguish between the corresponding combinations. Therefore, these results indicate that more accurate rules for discriminating between correct and incorrect combinations are required to improve the accuracy of the resulted assembly. In particular, since a large number of contigs that were combined incorrectly was obtained in the experiments, more precise discriminant rules for incorrect combinations are particularly needed. Such a large number of incorrectly combined contigs may occur due to the fact that even though almost all of the combinations of contigs might be correct, if only one incorrect combination is obtained, then the whole combination of contigs would be contaminated. Therefore, although this argument was not considered in this experiment, the application of both rules simultaneously might improve the assembly results.

2.2 DAwCDT (Double Assembly Method with a Complex Decision Tree

It received the problems described in the previous section, we proposed complex decision tree generating high performance discriminant rules. Whereas traditional decision trees has a single objective variables and a plurality of explanatory variables, complex decision tree has a plurality of objective variables. In this section, we describe objective variables of complex decision tree that is a new approach proposed in this paper.

In the traditional method based on k-mer, common heuristics were applied such that k-mers with smaller frequencies originate from reading errors by the sequencer. Moreover, certain heuristics have been used for the length of overlap, called "overlap-layout census." ith these heuristics, two reads with long overlapping region are considered as a pair that are correctly combined. According to these heuristics, information about the k-mer with a low frequency and long overlapping region among contigs might provide a way to distinguish the correctness or incorrectness of combinations. Based on this idea, two more characteristics, the minimum frequency of the k-mer and the length of the overlapping region between contigs, were added as objective variables for generating decision tree. For now, these two characteristics are represented as $\%_{\min}$.Cover and L.Overlap, respectively. To verify the effect of these additional characteristics on the discriminant task, the distributions of each characteristic with both the correct and incorrect dataset was evaluated. In this evaluation, 2511 overlapped contig pairs derived from ABySS and Velvet with several k values were used. The plots of the distribution for each characteristic are shown in **Figs. 1** and **2**.

As shown in Fig. 1, for both correct and incorrect cases, the two distributions showed large variance. However, the distribution corresponding to the incorrect case was more strongly skewed. Furthermore, as shown in Fig. 2, a clear difference between the two distributions was observed. From these results, we considered that elimination of contaminated incorrect contigs in the correct contigs group may be possible with the use of the feature of incorrect contigs. In other words, the combination of correct rules and incorrect rules may be possible.

Since $\%_{\min}$.Cover and L.Overlap are quantitative *variables*, *multiple regression analysis* was utilized to construct the discriminant rules [26]. As explanatory *variables*, the characteristics shown in Tables 1–3 were used. $\%_{\min}$.Cover and L.Overlap were

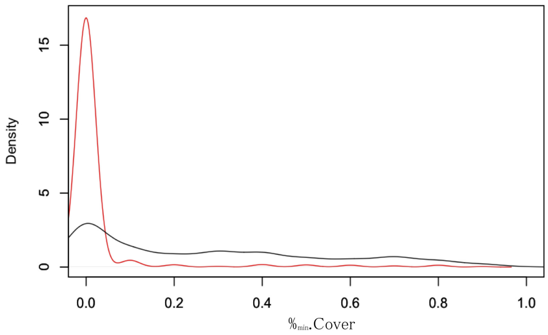


Fig. 1 Distribution of $\%_{\min}$.Cover. The density function was estimated with the kernel density estimation method. The horizontal axis denotes the minimum ratio of the frequency. Black and red lines represent the density function for correct and incorrect cases, respectively.

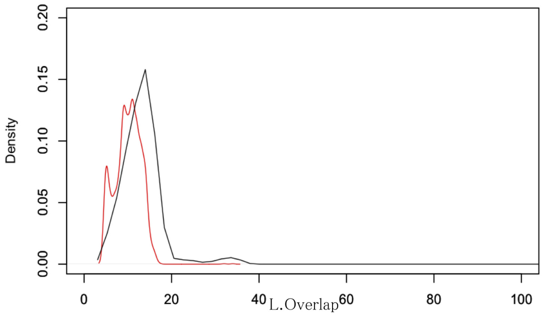


Fig. 2 Distribution of L.Overlap. The density function was estimated with the kernel density estimation method. The horizontal axis denotes the minimum ratio of the frequency. Black and red lines represent the density function for correct and incorrect cases, respectively.

Table 4 Discriminant function of *variables*, $Coef_{reg}$, $Multiple^{cor}_{coef}$, and $Coef_{det}$ for $\%_{\min}$ and L.Overlap.

	variables	$Coef_{reg}$	$Multiple^{cor}_{coef}$	$Coef_{det}$
$\%_{\min}$	$\Phi^{f,l}_{\max}$	-6.230×10^{-6}	0.012	0.009
	ρ	5.141×10^{-6}		
	Intercept	1.191×10^{1}		
L.Overlap	$\Phi^{f,l}_{\max}$	5.558×10^{-8}	0.214	0.203
	D^l	-4.299×10^{3}		
	I^f	-3.893×10^{-1}		
	I^l	-3.242×10^{-1}		
	W^f	6.061×10^{-5}		
	W^l	7.211×10^{-5}		
	U^f_{freq}	-1.409×10^{-4}		
	U^l_{freq}	-1.700×10^{-4}		
	Intercept	4.461×10^{-1}		

analyzed individually as objective variables. *Multiple regression analysis* outputs the discriminant function by parameter selection with the AIC [27], *multiple correlation coefficient*, and *determination coefficient*, which represents the fit to the discriminant functions. **Table 4** shows the selected variables according to the *determination coefficient* for a regression function named *variables*, partial regression coefficient named $Coef_{reg}$, *multiple correlation coefficient* named $Multiple^{cor}_{coef}$, and *determination coefficient* named $Coef_{det}$.

If the $Multiple^{cor}_{coef}$ and $Coef_{det}$ were closer to 1.0, the fit of the discriminant function was determined to be higher. Table 4 shows that the $Multiple^{cor}_{coef}$ values of $\%_{\min}$ and L.Overlap were 0.012 and 0.214, and the $Coef_{det}$ values of $\%_{\min}$ and L.Overlap were 0.009 and 0.203, respectively. Therefore, the adequacy of the discriminant functions for $\%_{\min}$.Cover and L.Overlap were not high. The reason may be the high variance of the objective variable, as shown in Figs. 1 and 2, and it is difficult to represent the discriminator by a linear classifier. To solve this problem, the decision tree algorithm was applied. In the application, the characteristics shown in Tables 1–3 were used as explanatory variables, and decision trees were generated for each variable, $\%_{\min}$.Cover and L.Overlap. Eventually, to include the information contained in $\%_{\min}$.Cover and L.Overlap, the corresponding k-mers and results obtained from the decision trees (for which these two variables were used as objectives) were evaluated with respect to whether or not they were correct. We named this hierarchical and multi-objective variable-based decision tree as a "complex decision tree (CDT)," which represents an extension of a traditional decision tree for a single-objective variable. Furthermore, using discriminant rules that the combination of correct incorrect rules from CDT, we proposed a new assembly method, named "rule-based assembly." The steps of the algorithms are as follows. First, the contigs are generated using some assembly algorithm with various parameters. Second, comparing the contigs to the reference, the incorrect contigs are filtered out and removed. Next, the all-correct contigs characteristics shown in Tables 1–3 are calculated. Using these characteristic values and additional parameters, based on the minimum ratio of the frequency ($\%_{\min}$) and length of overlapping regions (L.Overlap), a CDT composed of multiple-objective variables is constructed. These steps are used for the generation of the discriminant rules to distinguish whether or not the combined contigs are correct. In this step, correct and incorrect rules are obtained from each of the three decision trees (about $\%_{\min}$, L.Overlap, and traditional). A group of correct rules and incorrect rules from CDT was labeled as Positive rules and Negative rules. Finally, using the obtained discriminant rules that positive and effective negative rules, possible combinations of contigs with an overlap longer than 5 bases are provided or removed. After the removal, scaffolds would be generated from the remaining contig combinations.

3. Experiments

3.1 Experiments for Rule Construction and Evaluation

To confirm the validity of the determined rules, the discriminant rules were constructed and evaluated. In particular, to confirm the performance of our proposed approach, decision trees for the $\%_{\min}$.Cover and L.Overlap were generated. Consequently, the corresponding k-mers were evaluated in view of consistency (i.e., correctness) to the reference. After the evaluations, we designed new discriminant rules to distinguish whether or not the combined contig is correct. Finally, we conducted another evaluation of the designed rules using a benchmark dataset.

The generated decision trees were constructed and generated three rules. First, if $\%_{\min}$.Cover is greater than 0, then the corresponding combination of contigs would be correct with the probability 99.6%. Second, if the value of L.Overlap in the combined contigs is greater than 14, the combination would be correct with the probability 93.3%. 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 discriminant 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, were generated for the benchmark dataset using the *E. coli* genome. We constructed a read set by artificial samplings. Then, the contigs were obtained by the traditional assembly method. After the assembly, we filtered out incorrect contigs and compared the result to the reference dataset. As a result, 647 correct contigs were obtained as the benchmark dataset.

In **Tables 5** and **6**, Acc denotes the accuracy rate, which was

Table 5 The discriminant result obtained by the RfC rule.

Answer Response	Correct	Incorrect	Acc
Correct	394	3	0.937
Incorrect	14	234	

Table 6 The discriminant result obtained by the RfI rule.

Answer Response	Correct	Incorrect	Acc
Correct	400	8	0.951
Incorrect	24	215	

caculated as follows:

$$\text{Acc} = 1 - \frac{\text{FP} + \text{FN}}{\text{N}} \qquad (5)$$

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

As shown in Tables 5 and 6, both rules could effectively distinguish the correctness of combined contigs with a high correct answer rate; i.e., greater than 0.9. In addition, applying the rule for an incorrect combination (RfI) resulted in a higher correct ratio. From these results, it is expected that employing RfI would yield more accurate removal of the incorrect combination of contigs.

3.2 Experiments for Rule Application

As the next step, comparison experiments were conducted to verify the ability of the whole proposed algorithm. As traditional methods, Velvet, ABySS, and CISA were utilized. In addition, as hybrid methods, a hybrid method without rules, two methods with correct and incorrect rules, respectively, and the newly proposed method were compared. As the experimental data, we used *E. coli* K-12 MG1655 verified CISA. Because it is difficult to handle the whole genome for computation time cost, the length of 30,000 base sequence was cut. The length of each read was 50, i.e., the depth of reads of this dataset was 50. In the double assembly methods, ABySS and Velvet were used. Since the purpose of new approach is obtainment more discriminant rules than traditional decision tree, the same training dataset is required in the process of generating complex decision tree. As the training dataset, the *k* value was set to 16 and 18 for ABySS, and to 15 and 17 for Velvet. As the test dataset, read data was resampled at the same condition.

The results of discriminant rule constructions generated 36 rules for correct combinations. These were composed of 10 rules from the decision tree that used the length of the overlap region as the objective, 17 rules from the tree that used the minimum ratio of the frequency, and 9 rules from the traditional C4.5 decision tree. Similarly, 32 rules were obtained for incorrect combinations, consisting of 9 rules from the length of overlap, 6 rules from the minimum ratio, and 17 rules from the ordinary C4.5 algorithm. We labeled the rules obtained from the length of overlap as Ovl*n*, where *n* denotes the index of the corresponding rule. Similarly, the label MnR*n* was assigned for the rules obtained from the minimum ratio of frequency, and the rules generated from the traditional C4.5 algorithm were labeled as Trd*n*. The list of positive and negative rules are shown in **Tables 7** and **8**.

Table 7 The list of positive rules from the 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	$W^f \leq 331.4$

Applying the obtained rules, we found some negative rules that were effective to remove large incorrect combined contigs. The rules and the length of incorrectly combined contigs are shown in **Table 9**. From the effective rules shown in Table 9, we used three compositions: (1) Ovl2 and Ovl5, (2) Ovl5 and Ovl9.,and then (3), Ovl5 and MnR1.

Each assembly result was evaluated with respect to seven measures: the number of output contigs (#.Output), number of correct contigs (#.Corr), ratio of correct combinations (R.Corr), N50 contig length (N50), ratio of the length of the mapped region (R.Mapped), maximum length of correctly combined contigs (ML.Corr), and maximum length of incorrectly combined contigs (ML.Incorr). Since most of combined contigs obtained by the hybrid methods were longer than 1,500, we added two more evaluation indices, $\%.\text{Corr}_{1500}$ and $\%.\text{Mapped}_{1500}$, representing the ratio of the correct combined contigs longer than 1,500 and the ratio of those mapped correctly with a length longer than 1,500, respectively.

The comparison results are shown in **Table 10**. As shown in Table 10, #.Correct was increased in the case when all of the rules for correct combinations were used (315) compared to the hybrid method using only the correct rules (234); however, %.Correct decreased from 0.690 to 0.680. This means that applying all positive rules could result in more correct combined contigs compared to applying a method that uses correct rules only. In other words, although the ability of the rules established for detecting correct contigs was improved, its application also led to increasing the number of incorrect combined contigs. By additionally applying effective negative rules with all of the rules for correct combinations, %.Correct was improved. In particular, in the case of All + Ovl5 + Ovl2, the %.Correct value increased from 0.680 to 0.827. This means that a CDT could remove incorrect contigs more accurately than the C4.5 algorithm. This also means that the ability of classification was improved compared to use of a traditional decision tree that relies on single-objective variables. Furthermore, ML.Incorr of the method with ALL + Ovl5 + Ovl9 and that with ALL + Ovl5 + MnR1 were decreased compared to those obtained with traditional hybrid assembly. This result suggests that the CDT could remove large incorrect contigs. The ML.Corr value of the hybrid methods increased to 18729. From these results, we can conclude that hybrid methods could derive longer correctly combined contigs compared to traditional assembly. In addition, the methods ALL + Ovl5 + Ovl9 and All + Ovl5 + MnR1 resulted in a high value of %.Correct$_{1500}$, 1.000. These results also suggest that the CDT could generate correct combinations of contigs perfectly.

Table 8 The list of negative rules from the 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]
Trd1	$0.1 < R, 6787.5 < \Phi_{\text{freq}}, 0.291 < \rho, \rho \leq 0.321, I^l \leq 0.24$
Trd2	$0.1 < R, 6787.5 < \Phi_{\text{freq}}, 0.6 < Q^f, \rho_{\text{freq}} \leq 0.58, \rho \leq 0.182, 5.33 < W^l$
Trd3	$0.1 < R, 6787.5 < \Phi_{\text{freq}}, 0.6 < Q^f, -0.031 < \rho_{\text{freq}}, \rho_{\text{freq}} \leq 0.56, 0.291\rho, \rho \leq 0.182$
Trd4	$R \leq 0.1, 489 < \Phi_{\text{freq}}, -1 < D^f, 0.3 < I^f, W^l \leq 5464.053$
Trd5	$56.78 < \Phi_{\text{freq}}, D^l \leq 1, Q^l \leq 0.1, 0.291 < \rho$
Trd6	$R \leq 0.1, W^l \leq 5464.053, U^f_{\text{freq}} \leq 708.6, 1905.7 < U^l_{\text{freq}}$
Trd7	$R \leq 0.1, H \leq 5, 0.182 < \rho, I^f < 0.23$
Trd8	$R \leq 0.1, \Phi_{\text{freq}} \leq 898574, 331.414 < W^{f,l}, U^f_{\text{freq}} \leq 708.6$
Trd9	$-5 < D^l, 3 < H, 0.291 < \rho, 10.1 < U^f_{\text{freq}}, U^f_{\text{freq}} \leq 18.2$
Trd10	$1 < D^f, \rho \leq 0.291, 4.21 < W^f, I^l_{\text{freq}} \leq 7.6, -0.10 < L^l_{\text{freq}}$
Trd11	$1 < D^f, 3 < H, 0.1 < Q^l, 0.291 < \rho, U^f_{\text{freq}} \leq 18.2$
Trd12	$0.1 < R, \Phi_{\text{freq}} \leq 2908.5, 0.291 < \rho, 19.155 < W^l$
Trd13	$0.1 < R, \Phi_{\text{freq}} \leq 2908.5, 1 < D^l, 0.291 < \rho$
Trd14	$D^l \leq 1, 5 < H, Q^l \leq 0.3, 401.4 < U^f_{\text{freq}}, -0.10 < L^l_{\text{freq}}$
Trd15	$1 < D^l, Q^f \leq 0.1, \rho \leq 0.291, 4.21 < W^f$
Trd16	$0.3 < Q^l, \rho \leq 0.291, 4.214 < W^f$
Trd17	$1 < D^l, \rho_{\text{freq}} \leq -0.17, 4.21 < W^f, -0.101 < L^l_{\text{freq}}$

The comparative results showed that the ability of the hybrid assembly method can generate longer correct overlapped contigs than traditional assembly methods. Furthermore, these results suggest that the application of discriminant rules generated by a CDT to the task of combining obtained contigs has good potential to generate a higher number of correct contigs and improve the accuracy of the resulting combinations of contigs compared to a traditional decision tree.

4. Conclusion

In order to improve the accuracy of the combination of over-

Table 9 The list of negative rules that could remove a 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

Table 10 Results of performance comparisons between traditionals and hybrid methods.

Method	#.Output	#.Correct	%.Correct	N50	%.Mapped	ML.Incorr	ML.Corr	%.Correct$_{1500}$	%.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

lapping contigs for double assembly, we proposed a complex decision tree with multiple objective variables. In the process of generating the complex decision tree, the combination of discriminant rules for both correct and incorrect combinations was utilized. Comparisons with traditional assembly methods showed improvements of the quality indices, length of correct overlapping contigs, correct ratio, and coverage ratio of the large correct contig group. Furthermore, the ability of discriminant rules was improved compared to those simply generated with the traditional method. Thus, these results indicate that to achieve an error-free read dataset generated artificially, our proposed CDT approach could generate more adequate discriminant rules than a traditional decision tree algorithm. Consequently, the application of discriminant rules obtained with our proposed method to a double assembly could achieve more accurate combinations of contigs.

References

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

[2] Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I.: ABySS: A parallel assembler for short read sequence data *Genome Research*, Vol.19, No.4, pp.1117–1123 (2009).

[3] Warren, R.L., Sutton, G.G., Jones, S.J.M. and Holt, R.A.: Assembling millions of short DNA sequences using SSAKE, *Bioinformatics*, Vol.23, No.4, pp.500–501 (2006).

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

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

[6] Kelley, D.R., Schatz, M.C., Salzberg, S.L., et al.: Quake: quality-aware detection and correction of sequencing errors, *Genome Biol.*, Vol.11 (2010).

[7] Miller, J.R., Koren, S. and Sutton, G.: Assembly Algorithms for Next-Generation Sequencing Data, *Genomics*, Vol.95, pp.315–327 (2010).

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

[9] Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marais, G., Pop, M. and Yorke, J.A.: GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res.*, Vol.22, pp.557–567 (2012).

[10] Nijkamp, J., Winterbach, W., van den Broek, M., Daran, J.-M., Reinders, M. and de Ridder, D.: Integrating genome assemblies with MAIA, *ECCB*, Vol.26, pp.433–439 (2010).

[11] Yao, G., Ye, L., Gao, H., Minx, P., Warren, W.C. and Weinstock, G.M.: Graph accordance of next-generation sequence assemblies, Vol.28, No.1, pp.13–16 (2012).

[12] Lin, S.-H. and Liao, Y.-C.: CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes, Vol.8, No.3, e60843 (2013).

[13] Peng, Y., Leung, H., Yiu, S.M. and Chin, F.Y.L.: 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 (2010).

[14] Peng, Y., Leung, H.C.M., Yiu, S.M. and Chin, F.Y.L.: 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 (2012).

[15] Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E.: Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels, *BIOINFORMATICS*, Vol.28, No.8 (2012).

[16] Choi, J.-H., Kim, S., Tang, H., Andrews, J., Gilbert, D.G. and Colbourne, J.K.: A machine-learning approach to combined evidence validation of genome assemblies, *BIOINFORMATICS*, Vol.24, No.6, pp.744–750 (2008).

[17] Breiman, L.: Random Forests, *Machine-Learning*, Vol.45, pp.5–32 (2001).

[18] Ohshiro, A., Okazaki, T. and Nakamura, M.: 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 (2014).

[19] Breiman, L.: Bagging Predictors, *Machine-Learning*, Vol.24, pp.123–140 (1996).

[20] Freud, Y. and Schapire, R.E.: A decision- theoretic generalization of on-line learing and an application to boosting, *Journal of Computer and System Sciences*, Vol.55, pp.119–139 (1995).

[21] Loh, W.-Y.: Classification and regression trees, Vol.1, WIREs Data Mining and Knowledge Discovery (2011).

[22] Quinlan, J.R.: Bagging, boosting, and C4.5, *Proc. 13th National Conference on Artificial Intelligence*, pp.725–730 (1996).

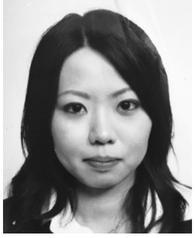
[23] Palmer, L.E., Dejori, M., Bolanos, R. and Fasulo, D.: Improving *de novo* sequence assembly using machine-learning and comparative genomics for overlap correction, *BMC Bioinformatics*, Vol.11, No.33, DOI: 10.1186/1471-2105-11-33 (2011).

[24] Angeleri, E., Apolloni, B., de Falco, D. and Grandi, L.: DNA fragment assembly using neural prediction techniques, *Int. J. Neural. Syst*, Vol.9, No.6, pp.523–544 (1999).

[25] Quinlan, J.R.: *C4.5: Programs for machine-learning*, Morgan Kaufmann Publishers (1993).

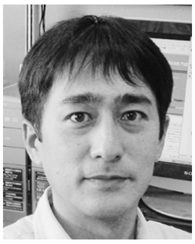
[26] Ma, Y., Lao, S., Takikawa, E. and Kawade, M.: Discriminant Analysis in Correlation Similarity Measure Space, ICML'07 Proceeding of the 24th International Conference on Machine Learning, pp.577–584 (2007).

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



Ayako Ohshiro took her B.Sc. and M.Sc. degrees in Information Engineering from University of the Ryukyus. She belongs to doctoral course in same university. Her research area is Bioinformatics. Especially, she is interested in development of DNA assembly algorithm.

Hitoshi Afuso took Ph.D. from University of the Ryukyus in 2013 and worked as post-doctoral researcher in Hokkaido University from 2013 to 2015 and he is studying bioinformatics in University of the Ryukyus.



Takeo Okazaki took his B.Sc. and M.Sc. degrees from Kyushu University in 1987 and 1989, respectively. He took Ph.D. from University of the Ryukyus in 2014. He has been a associate professor at University of the Ryukyus. His research interests are statistical data normalization for analysis, and causal relationship analysis.

Morikazu Nakamura took his B.E. and M.E. degrees from University of the Ryukyus in 1989 and 1991, respectively. He took Ph.D. from Osaka University in 1995. He has been a professor at University of the Ryukyus. His research interest includes design and analysis of parallel and distributed algorithms.

(Communicated by *Masakazu Sekijima*)