A study on complex decision tree construction for getting the rules of contig binding in DNA double assembly

Ayako Ohshiro[1,a)] Takeo OKAZAKI[2] Morikazu NAKAMURA[2]

Abstract: To derive the restored sequence called contig without reference, some assembly approaches have been proposed. Decision of binding sequences depends on accuracy of contigs. Decision tree learning algorithms have been proposed for decision of classifications, and also used for contig binding. We proposed complex decision tree for getting the rules of contig binding in DNA double assembly, by use of multiple objective variables, accuracy of binded contigs, minimum coverage value among contigs of before binding, overlap length of binded contig. We evaluated its performance from two points of the view that assembly quality and classification ability of rules from complex decision tree. As evaluation indices, we used correct ratio *corR*, coverage ratio *covR*, N50, $M_{correct}$, $M_{incorrect}$, of output contigs, by applying complex decision tree to double assembly and compared assembly quality to traditional assembly method.

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

1. Introduction

Thanks to the development of giga sequencer platforms with parallel processing[1][2] and cost reduction of them, the research area of DNA assembly method such as ABySS[3], Velvet[4] and SSAKE[5] has been actioned. And yet the same time, read error correction from sequencer has been a challenge such as Trimmomatic[6] and approaches by use of *k*-mer such as EDAR[7], ECHO[8], BLESS[9], and Quake[10]. In addition, Bayesian Genome Assembly[11] and [12], are proposed. Because assembly result depends on assembly algorithm and *k* value, it is difficult to obtain assembly results robustly and hybrid assembly algorithms by integrating the results of traditional assembly method have been proposed such as IDBA[13], MAIA[14], GAA[15] and CISA[16]. Boisvert S.el [17] proposed Ray that is a hybrid of sequence technology. Marcel.el[18] proposed Oases multiple *k*-mer assembly method for mRNA sequence data sets. We proposed a double assembly method merging different k-mers and applied it to binding rules with a characteristic distribution of *k*-mer's coverage value for contig named DAwCC[19].

In the process of DAwCC, we used C4.5[20] to construct contig binding rule, the research area of the decision tree algorithm continues to develop over the years. Recent years, ensemble machine learning that integrating traditional classifier have been proposed. Breimen proposed Bagging[21] that dividing training data to subdata and generating classifiers for each of them and finally determined decision by majority from multiple classifier, and proposed Random Forest[22] that was randomly explanatory variable selected for each divided sub-data, and finally determined decision by majority from multiple classifier as Bagging. Boosting was proposed by updating explanatory variable by weighting on each misclassified data. Wei-Yin Loh[23], J. R. Quinlan[24] had comparative performance studies about ensemble machine learning algorithms. In this paper, we proposed complex decision tree by use of multiple objective parameters, in order to get the contig binding rules for DNA double assembly.

2. Possibility of multiple objective variable

In this section, we report about a pre-experiment for the possibility of multiple objective variable. Generally, the traditional decision tree is constructed by multiple explanatory variable and one objective variable. We applied accuracy of combined contigs as objective variables named *accu* and distribution of *k*-mers on each binded contig as explanatory variables as Table.1, on the process of DAwCC. In order to improve the classification ability of contig binding rules, we tried to verify the possibility of the use of multiple explanatory variable that they were related to each other. We can expect that multiple explanatory variable can classify test data that *accu* rule couldn't classify. As the regions with small *k*-mers coverage value are removed in traditional assembly method, we can consider that binded contigs with small *k*-mer's coverage value are likely incorrect contigs. In tha same way, we can consider that binded contigs with large overlap region are likely correct contigs. Firstly, We had pre-experiments to verify them by correct contigs obtained from traditional assembly methods by use of *k*-mer. Fig.1 shows the relation about the accuracy of binded contigs and overlap length, named *ovlp* of them. Fig.2 shows the relation about the accuracy of binded contigs and minimum coverage value, named *mincov* of contigs before binded.

Fig.1 shows that most *mincov* of incorrect contigs are lower

[1] Interdisciplinary Intelligent Systems Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus, Okinawa 903-0213, Japan
[2] Department of Information Engineering, Faculty of Engineering, University of the Ryukyus, Okinawa 903-0213, Japan
[a)] ayami@ms.ie.u-ryukyu.ac.jp

Table 1 Explanatory variables with p-value's distribution of k-mer's coverage

Fluctuation	$Coef_{wav}^{f,l}$ Gradient of waveform
	$R_{inc}^{f,l}$ Rate of increase value
	$F_{high}^{f,l}$ High-frequency component of Fourier transform
	$Sum_{F}^{f,l}$ Powered value of Fourier transform
	$F_{low}^{f,l}$ Low-frequency component of Fourier transform
Distribution	$p_{null}^{f,l}$ p-value with null frequency value
	$Sum_{Ffreq}^{f,l}$ Powered value of Fourier transform for frequency distribution
	CC Correlation coefficient
Correlation	CC_{freq} Correlation coefficient of frequency distribution
	CCF_{freq} Maximum cross-correlation function for Fourier transform of frequency distribution
	M_{ccf}^{F} Maximum cross-correlation function for Fourier transform
	M_{ccf} Maximum cross-correlation function
	D_{ham} Hamming distance of frequency distribution
	M_{ccf}^{freq} Maximum cross-correlation function for frequency distribution
	$\|p_{f,l}\|$ Norm value of end point of former and start point of latter

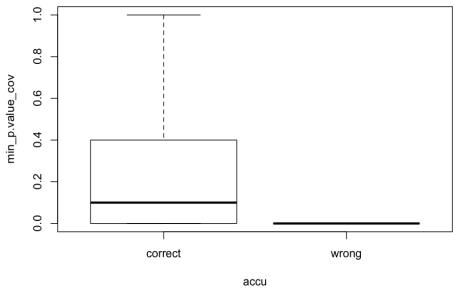

Fig. 1 Distribution of *mincov* for correct and incorrect binded contigs

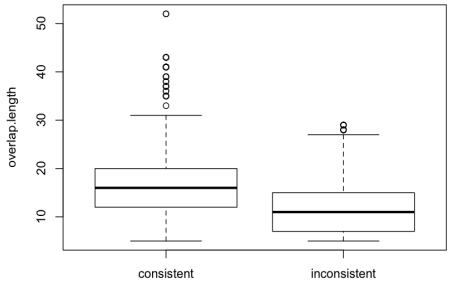

Fig. 2 Distribution of *ovlp* for correct and incorrect binded contigs

than 0.1, and Fig.2 shows most *ovlp* of incorrect contigs are lower than 10 bases. We can find that the larger *mincov* and larger *ovlp*, more correct binded contigs. Secondly, we observed about the classification distribution about multiple objective variable, by applying *mincov*, *ovlp*, *accu* to same test data. Fig.3-4 shows classification distribution by positive and negative rules correctly.

Fig.3 shows that 4 binded contigs were correctly obtained by *mincov* rules, but *accu* and *ovlp* rules couldn't obtain and 289 binded contigs were correctly obtained by *accu* rules but *mincov* and *ovlp* rules couldn't obtain. 53 binded contigs were correctly obtained by *accu*, *mincov*, *ovlp* rules. In a similar way, Fig.4 shows that 180 binded contigs were eliminated as incorrect binded contigs, with *accu*, *mincov*, *ovlp* rules and 160 binded contig were able to eliminate with *mincov* rules but *accu* and *ovlp* rules couldn't. We can find that decision tree will be improved using multiple objective variables because they are in a complementary style each other. We considered about the addition of

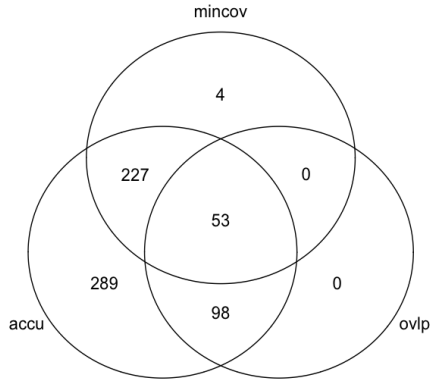

Fig. 3 Classification distribution of positive rules

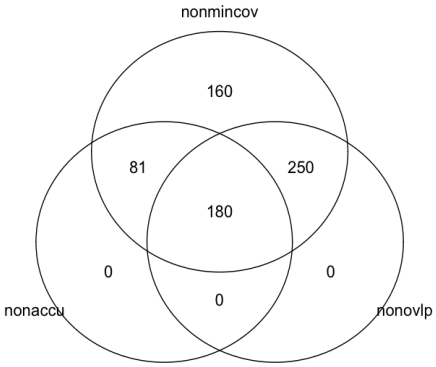

Fig. 4 Classification distribution of negative rules

discriminator as *mincov*, *ovlp* to traditional decision tree, *accu*.

As considering that each of two train data is constructed by *mincov* or *ovlp* as objective variable and distribution of k-mer's coverage value as explanatory variable, we used multiple regression analysis to get discriminant function. Multiple regression analysis outputs discriminant function by parameter selection with *AIC*, multiple correlation coefficient and determination coefficient, which represents the fitness to the discriminant functions. Table.2 shows the selected variables for a regression function named *variables*, partial regression coefficient named $Coef_{reg}$, multiple correlation coefficient named $Multiple_{coef}^{cor}$ and determination coefficient named $Coef_{det}$, *AIC* for the discriminant function of *mincov* and *ovlp*.

Table 2 Discriminant function of *variables*, $Coef_{reg}$, $Multiple_{coef}^{cor}$, $Coef_{det}$ for *mincov* and *ovlp*

	variables	$Coef_{reg}$	$Multiple_{coef}^{cor}$	$Coef_{det}$	AIC
mincov	M_{ccf}^{F}	-6.230e-06	0.01161	0.008545	-1829.74
	CC	5.141e-06			
	Intercept	1.191e+01			
ovlp	M_{ccf}^{F}	-2.144e-07			
	CCF_{freq}	5.558e-08			
	$Coef_{wav}^{l}$	-4.299e-03			
	R_{inc}^{l}	-3.893e-01			
	R_{inc}^{l}	-3.242e-01	0.2143	0.2032	
	Sum_{F}^{f}	6.061e-05			
	Sum_{F}^{l}	7.211e-05			
	F_{high}^{f}	-1.409e-04			
	F_{high}^{l}	-1.700e-04			
	Intercept	4.461e-01			

The $Multiple_{coef}^{cor}$ and $Coef_{det}$ were closer to 1.0 and AIC is larger, the fitness of discriminant function were higher. From Table.2, we could find that each $Multiple_{coef}^{cor}$ of *mincov* and *ovlp*

were 0.01161 and 0.2143, each $Coef_{det}$ of *mincov* and *ovlp* were 0.008545 and 0.2032, adequacy of discriminant functions for *mincov* and *ovlp* were not high. We can consider that the reason of it is high variance of objective variable, and it is difficult to represent discriminator by linear classifier. Therefore, we need to reconsider about the treatment of objective variables by converting to qualitative variables. Decision tree requires objective variables as qualitative variable, we convert the qualitative variable to quantitative variable. We generated decision tree composed of *accu* as objective variables, *mincov*, *ovlp* as explanatory variable. Detail of generating process is described in section 3. To check the fitness of rules, in similar to $Multiple^{cor}_{coef}$ or AIC about multiple regression analysis, we observed learning ability for rule about decision tree from *accu*, *mincov* and *ovlp*. Table.3 shows discriminant result form and we defined learning ability as formula (1). Table.4 shows the learning ability of contig binding rule by C4.5 for training data about decision tree of *accu*, *mincov* and *ovlp*.

Table 3 Discriminant result form

	consistent	inconsistent
correct	*num*1 contigs judged as correct for "consistent"	*num*2 contigs judged as correct for "consistent"
incorrect	*num*3 contigs judged as incorrect for "consistent"	*num*4 contigs judged as incorrect for "inconsistent"

$$Le-R = 1 - \frac{num2 + num3}{num1 + num2 + num3 + num4} \quad (1)$$

Table 4 $Le-R$ of each decision tree from *mincov*,*accu*,*ovlp*

$Le-R$ of decision tree for *mincov*			
	consistent	inconsistent	Le-R
correct	394	3	0.937
incorrect	14	236	
$Le-R$ of decision tree for *accu*			
	consistent	inconsistent	Le-R
correct	394	3	0.983
incorrect	8	242	
$Le-R$ of decision tree for *ovlp*			
	consistent	inconsistent	Le-R
correct	400	8	0.951
incorrect	24	215	

We could find that each $Le-R$ of decision trees for *accu*, *mincov*, *ovlp* was larger than 0.93, so it has high fitness for training data. When considered in term of fitness, we decided to apply of decision tree with multiple objective variable named complex decision tree.

3. Complex Decision Tree with Multiple objective variables for double assembly :(CDTwM)

In this section, we discuss about construction of CDTwM (Complex Decision Tree with Multiple objective variables in double assembly). In order to construct complex decision tree by C4.5, it is necessary to generate training data composed with *accu*, *mincov*, *ovlp* and distribution with *k*-mer's coverage value of binded contigs. We obtained training data from binded contigs with overlap region more than 5 bases, and determined *accu*, *mincov*, *ovlp* of them. After getting rules from the complex decision tree with train data, we applied them to target binded contigs. The actual process for generating complex decision trees is described as follows and Fig.5.

step1 Prepare the whole sequence and read dataset whose base allocation is known.

step2 Obtain contigs from traditional assembly methods for some *k*-mers.

step3 Extract all the pairs of contigs with more than 5 bases overlap region.

step4 Distinguish binded contigs as accuracy named *accu* to correct or incorrect by comparing to the original sequence

step5 Generate training data constructed explanatory variable as Table.1, and overlap length named *ovlp* and minimum coverage value named *mincov* among each contigs.

step6 Obtain rules of decision tree from C4.5 constructed by *accu* as objective variable and *mincov*, *ovlp* as explanatory variables, they are quantitative variables.

step7 Convert *mincov*, *ovlp* to qualitative variable accordance with decision tree obtained at the step6.

step8 Derive contig binding rules by C4.5 with training data that consists of explanatory variables as step4 about *accu*, *ovlp*, and *mincov*.

step9 Output rules from decision tree about *accu*, *ovlp*, and *mincov* as complex decision tree.

step10 Obtain binded contigs from the result of judgement by complex decision tree.

We used E.coli data as Table.5 and generated a different short read set with the same combination of *k*-value and traditional assembly method as Table.6.

Table 5 Detail of reference sequence

Species	Length	read length	number of reads
Escherichia coli	30000	50	30000

	Train and test data	
Combination of Method	ABySS	Velvet
Combination of *k*-mer	16,18	15,17

Table 6 Detail of experimental datasets from different short reads with the same reference

It the process of step6, we got the decision tree constructed by *accu* as objective variable and *ovlp*, *mincov* as explanatory variables, as Table.7. For example, rule1 means that when the *mincov* is more than 0, binded contig is correct with a probability of 99.6 %.

Table 7 Decision tree composed of *ovlp* and *mincov* as explanatory variable, and *accu* as objective variable

rule1	*mincov* > 0 -> class correct [0.996]
rule2	*ovlp* > 14 -> class correct [0.933]
rule3	*ovlp* <= 14 and *mincov* <= 0 -> class incorrect [0.728]

We converted *mincov* larger than 0 and *ovlp* larger than 14 to the positive rule as *correct*, and *mincov* lower than 0 and *ovlp*

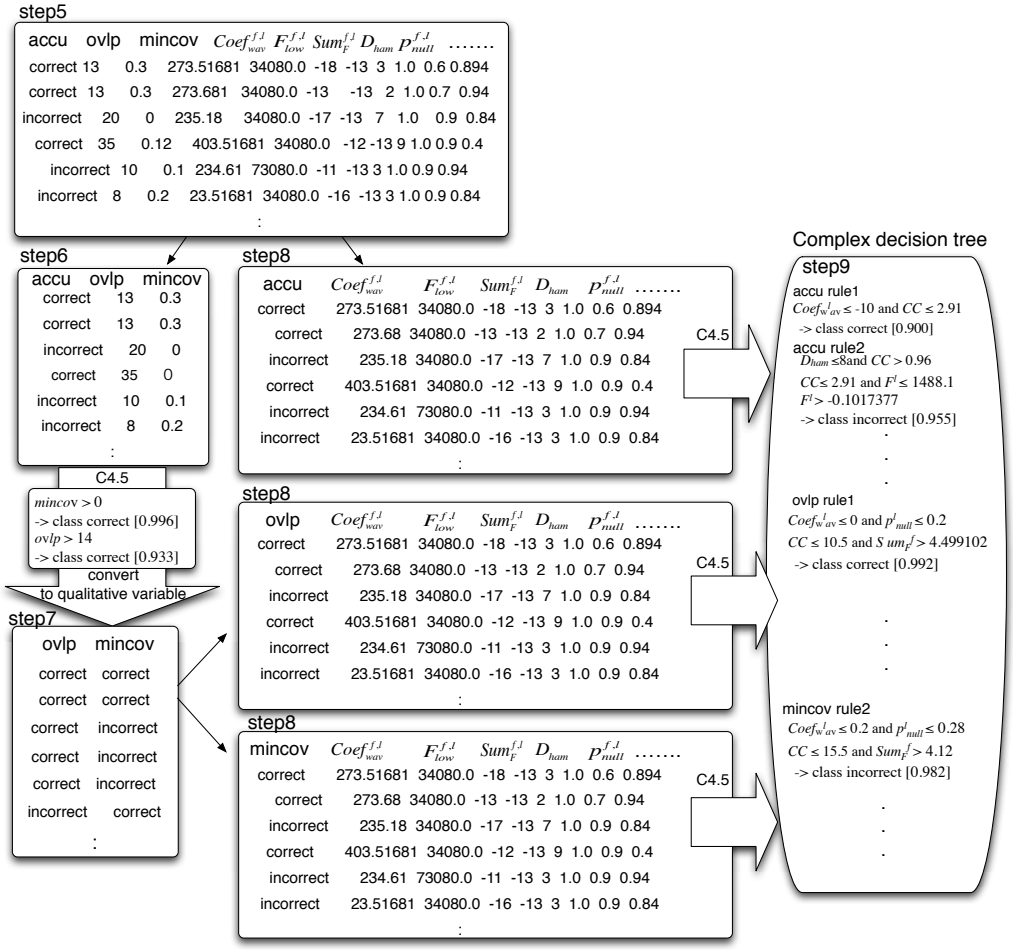

Fig. 5 Flow of generating complex decision tree after obtaining of training data

lower than 14 as a negative rule as *incorrect* from Table.5. In this experiment, we got 37 positive rules from 9 *accu* rules, 10 *ovlp* rules and 18 *mincov* rules and 29 negative rules from 14 *accu* rules, 6 *mincov* rules and 9 *ovlp* rules and Table.8 shows some of rules from a complex decision tree with *accu*, *ovlp* and *mincov*. We defined *accu*, *ovlp*, and *ovlp* correct rules as positive rules. In a similar way, we defined *accu*, *ovlp*, and *ovlp* incorrect rules as negative rules.

Table 8 Some of rules from complex decision tree with *ovlp* and *mincov*, and *accu*

Positive	ovlp rule1	$Coef_{wav}^{l} \leq$ -10 and $CC \leq 2.91$ -> class correct [0.900]
	mincov rule1	$Coef_{wav}^{l} \leq 0$ and $p_{null}^{l} \leq 0.2$ $CC \leq 10.5$ and $Sum_F^f > 4.499102$ -> class correct [0.992]
	accu rule1	$\|p_{f,l}\| \leq 0.1$ and $F_{high}^{f} >$708.6 -> class correct [0.976]
Negative	ovlp rule2	$D_{ham} \leq 8$ and $CC > 0.96$ $CC \leq 2.91$ and $F_{high}^{l} \leq 1488.1$ $F_{low}^{l} >$ -0.1017377 -> class incorrect [0.955]
	mincov rule1	$Coef_{wav}^{l} >0$ and $Sum_F^{l} \leq$4.658966 -> class incorrect [0.989]
	accu rule2	$p_{null}^{f,l} > 0.1$ and $CCF_{freq} > 6787.5$ $CC > 2.91$ and $CC \leq 182.32$ $R_{inc}^{l} \leq 0.2365952$ -> class incorrect [0.986]

Next, in order to evaluate the suitability of complex decision trees to double assembly, we had a comparative experiment about the performance of double assembly method with rules by traditional decision tree, and without rules. We defined double assembly method without rules as *DA*, applying traditional rules as DA_{cor}^{DT} and DA_{incor}^{DT}, complex decision tree rules as $DA_{positive}^{CDT}$, $DA_{negative}^{CDT}$.

We defined 5 indices to evaluate the performance of double assembly. *corR* means the rate of correct binded contigs for the output binded contigs and *covR* means the ratio of mapped binded contigs to the reference. When all output binded contigs are correct, *corR* become 1.0 and output contigs mapped entire of reference, *covR* become 1.0. In addition, large correct binded contigs are most ideal output, and large incorrect binded contigs are not most ideal. M_{cor} and M_{incor} means the longest correct or incorrect binded contigs among output. Table.9 shows the comparative result of double assembly methods.

Table 9 Comparative result of *DA* without rules, *DA* with rules by traditional C4.5, *DA* with rules by complex decision tree

Method	*DA*	DA_{cor}^{DT}	DA_{incor}^{DT}	$DA_{positive}^{CDT}$	$DA_{negative}^{CDT}$
Output	671(c412)	338(c234)	444(c315)	512(c350)	20
corR	0.614	0.69	0.71	0.68	1.0
N50	7849	7849	7906	7356	
covR	0.99	0.96	0.99	0.96	0.62
M_{incor}	15767	15767	15767	15767	
M_{cor}	15767	15767	15767	15767	

From Table.9, we can find the effect of the discriminator for double assembly method with contig binding rules. For double assembly method with negative rules, especially $DA^{CDT}_{negative}$, the output is all correct binding contigs. However, 392 correct binded contigs are removed by comparing *DA* and *covR* was decreased by 0.37 points. On the other hand, for double assembly method with positive discriminator, *corR* of $DA^{CDT}_{positive}$ was improved by comparing to *DA* but worse to DA^{DT}_{cor}.

Because the complex decision tree is constructed from multiple positive and negative rules as Table.8、we can expect that the combination of each of positive and negative rules will improve them. Furthermore, incorrect large contigs represented as M_{incor} are treated as maximum noisy data in the assembly process, it is necessary to apply negative rules that can remove such as M_{incor}.

4. Complex Decision Tree by positive and negative Rules with Heuristic for double assembly :(CDTwRH)

In this section, we discuss about the possibility of a complex decision tree with the combination of each of positive and negative rules, in order to improve assembly evaluation value. Firstly, in order to confirm the effect about each of negative rule, we applied them to the result of $DA^{CDT}_{positive}$ as Fig.6 and evaluated performance of them.

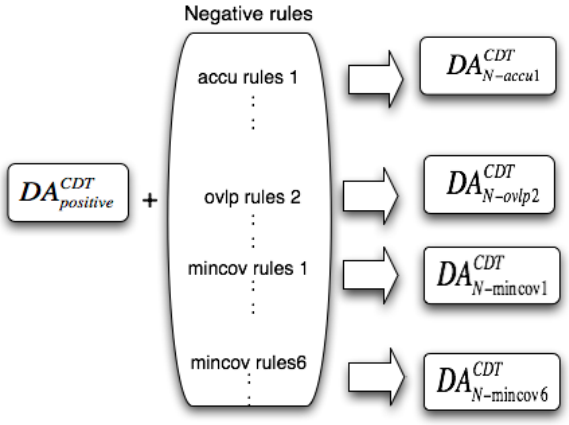

Fig. 6 Flow of generating complex decision tree after obtaining of training data

Table.10 shows the effect of each of the negative rule by complex decision tree.

Table 10 Results of complex decision tree with combination of positive rules and negative rules

Method	Outout	cor	*corR*	N50	*covR*	M_{inco}	M_{cor}
$DA^{CDT}_{negative}$	20	20	1.0		0.62		10854
$DA^{CDT}_{N-ovlp5}$	236	194	0.82	7891	0.96	10860	10855
$DA^{CDT}_{N-ovlp7}$	498	350	0.70	5658	0.96	15767	18729
$DA^{CDT}_{N-ovlp8}$	478	330	0.69	6585	0.96	15767	18729
DA^{CDT}_{N-cov2}	494	350	0.708	7357	0.96	15767	18729
DA^{CDT}_{N-cov6}	59	50	0.847	7899	0.81	10860	10855

We could find that *corR* for CDTwRH decreased by comparing to $DA^{CDT}_{negative}$, but *covR* of that improved as 0.81 to 0.96 by Table.10. About M_{inco} of CDTwRH, $DA^{CDT}_{N-ovlp5}$ and DA^{CDT}_{N-cov6} were improved but M_{cor}. And *corR* of DA^{CDT}_{N-cov6} was higher, but it removed too much correct binded contigs, so the suitability of double assembly is not high. Therefore, as high *corR* and high *covR*, the result of $DA^{CDT}_{N-ovlp5}$ has high performance better than that of $DA^{CDT}_{negative}$.

Secondly, we extracted negative rules these could remove large incorrect binded contigs, treated as maximum noise in the assembly. By considering the length of the reference length as 30000base, we extracted rules that could remove more than 10000 bases incorrect binded contigs as Table.11.

Table 11 The list of rules that could remove large incorrect binding contigs

Length of removed incorrect contigs	Negative rules
15767	ovlp5,ovlp9,cov6
15767	ovlp4,ovlp5,ovlp6,ovlp7,ovlp9,cov6
12695	ovlp4,ovlp5,ovlp6,ovlp7,ovlp8,ovlp9,cov6
10872	ovlp2,ovlp4,ovlp5,ovlp7,ovlp9,cov6
10860	ovlp2,ovlp7,ovlp9,cov1
10859	ovlp9,cov1

We considered the rules that could remove multiple large incorrect binded contigs as high usability. From Table.11, we decided to apply ovlp5, ovlp9 and mincov.1 rule, to $DA^{CDT}_{positive}$. According to the experimental result, we defined the combination of negative rules for complex decision tree as $DA^{CDT}_{N-ovlp5,2}$, $DA^{CDT}_{N-ovlp5,9}$, and $DA^{CDT}_{N-ovlp5,cov1}$.

Furthermore, because getting of large correct binded contigs is ideal output, and most binded contig's length of double assembly is more than 1500 as Fig.7.

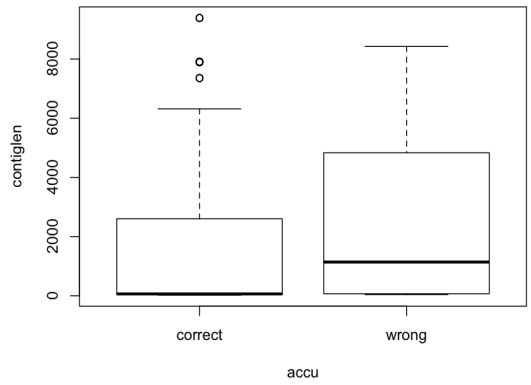

Fig. 7 Length distribution of binded contigs with double assembly

From Fig.7, we added evaluation indices in order to evaluate *corR* and *covR* in the large binded contig group. We considered the binded contigs with more than 1500 bases as a large binded contig group and defined their *corR* and *covR* as $corR_{1500}$ and $covR_{1500}$.

5. Comparative study

To confirm the effect of proposed method, we carried out comparative experiments of double assembly method with traditional C4.5, double assembly with a complex decision tree with multiple objective variables in section 3(CDTwM), double assembly with complex decision tree by positive and negative rules with heuristic in section4 (CDTwRH), traditional assembly method Velvet,

ABySS and CISA. From the comparative experiment, we defined the combination of negative rules for complex decision tree as $DA^{CDT}_{N-ovlp5,2}$, $DA^{CDT}_{N-ovlp5,9}$, and $DA^{CDT}_{N-ovlp5,cov1}$.

From Table.12, we could find that *corR* of *DA*, *DA* with C4.5 and complex decision tree were decreased, but M_{cor} are improved by comparison to traditional assembly method. *corR* of CDTwRH was increased by comparing to *DA* or *DA* with traditional C4.5 and we could verify that contig binding rules from complex decision trees removed incorrect binded contigs that traditional C4.5 couldn't. Because N50 and $corR_{1500}$ of $DA^{CDT}_{N-ovlp5,cov1}$ was increased by comparison to traditional assembly method that were Velvet, ABySS and CISA. $DA^{CDT}_{N-ovlp5,cov1}$ outputs many large correct binded contigs. *covR* of traditional assembly method depends on specific *k* value, but assembly method with hybrid or double, those were CISA, *DA* and DA_{incor} were improved. $corR_{1500}$ of $DA^{CDT}_{negative}$, $DA^{CDT}_{N-ovlp5,9}$, $DA^{CDT}_{N-ovlp5,cov1}$ were 1.0, that means they output large correct binded contigs and M_{cor} of them are improved by comparison to traditional double assembly methods. Furthermore, $covR_{1500}$ of DA^{DT}_{incor} were the highest and DA^{DT}_{cor}, $DA^{CDT}_{positive}$ were improved by comparison to traditional assembly method, that means double assembly method with complex decision tree output correct binded contigs.

6. Conclusion

In order to improve accuracy of double assembly by use of contig binding rules, we proposed complex discriminator with multiple objective variables. At the pre-experiment, we confirmed the possibility of using multiple objective variable by addition of *ovlp*, *mincov* to traditional objective variable *accu* by applying same data. We used multiple regression analysis and decision tree to obtain complex discriminator with multiple objective variable for applying candidate to double assembly. From comparative experiment result, we found that decision tree's leaning ability was higher than multiple regression analysis and proposed complex decision tree. In order to generate training data to apply decision tree, we converted *ovlp*, *mincov* those were quantitative variables to qualitative variables. In this way, we constructed complex decision tree with multiple objective variables named CDTwM. From experimental results, *corR* was improved, but *covR* was decreased. Therefore, because we can obtain positive rules and negative rules from each of decision trees of *ovlp*, *mincov*, *accu*, we constructed complex decision tree by a combination of positive and negative rules, named CDTwRH. Furthermore, we extracted rules that could longest incorrect binded contigs, that is treated large noise contig in the assembly process. From comparative experiments, we could obtain larger correct binded contigs M_{cor} and *covR* were improved than traditional assembly method, ABySS, Velvet, and CISA but *corR* were decreased. We confirmed the possibility of improvement of contig binding rules from decision tree by use of complex decision tree. And we found the possibility of improvement of accuracy of double assembly by a combination of positive and negative rules with heuristic approach.

References

[1] Staden, R. 'A new computer method for the storage and manip- ulation of DNA gel reading data.' Nucl. Acids Res. 8: pp3673-3694.(1980)

[2] Lincoln D Stein : The case for cloud computing in genome informatics. Genome Biology, vol.11 (2010).

[3] Jared T. Simpson, kim Wong, Shaun D. Jackman, et al : ABySS: A parallel assembler for short read sequence data, Genome Research, vol.19, pp.1117- 1123, (2009)

[4] 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)

[5] 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)

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

[7] Xiaohong zhao, Lance E. Palmer, Randall Bolanos, Cristian Mircean, Dan Fasulo, and Gayle M. Wittenberg : EDAR: An Efficient Error Detection and Removal Algorithm for Next Generation Sequencing Data, vol.17, No 11, pp.1549—1560, (2010)

[8] Wei-Chun Kao, Andrew H. Chan and Yun S. Song : ECHO: A reference-free short-read error correc- tion algorithm Genome Res. 21,1181-1192 April 11, (2011)

[9] Yun Heo, Xiao-Long Wu, Deming Chen, Jian Ma and Wen-Mei Hwu, : BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads, pp1354-1362, vol.30, no.10, (2014)

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

[11] Mark Howison, Felipe Zapata, Erika J. Edwards, Casey W. Dunn : Bayesian Genome Assembly and Assessment by Markov Chain Monte Carlo Sampling, vol.9, Issue 6, e99497, (2014)

[12] Xiaohu Shen, Manohar Shamaiah, and Haris Vikalo : Iterative Learning for Reference- Guided DNA Sequence Assembly from Short Reads: Algorithms and Limits of Performance ,vol.22 (2014)

[13] Yu Peng, Henry Leung, S.M. Yiu, Francis Y.L. Chin : 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, 2010. Vol. 6044, pp.426-440, (2010)

[14] Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek,Jean-Marc Daran, Marcel Reinders, and Dick de Ridder : Integrating genome assemblies with MAIA, Vol.26 ECCB 2010, pp433—439

[15] Guohui Yao. Liang Ye, Hongyu Gao, Patrick Minx, Wesley C. Warren, George M. Weinstock : Graph accordance of next-generation sequence assemblies, Vol. 28 no. 1, pp13-16, (2012)

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

[17] Boisvert S, Laviolette F, Corbeil J : Ray simultane- ous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol. 17 (11): 1519—1533, (2010)

[18] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron and Ewan Birney : Oases: robust de novo RNA- seq assembly across the dynamic range of expression levels, BIOINFORMATICS, Vol.28, no.8, (2012)

[19] Ayako OHSHIRO, Takeo OKAZAKI, Morikazu NAKAMURA: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)

[20] J. Ross Quinlan Morgan Kaufmann, San Mateo,CA: C4.5:Programs for Machine Learning, (1993)

[21] LEO BREIMAN. : Bagging Predictors, Machine Learning, vol.24, pp.123-140, (1996)

[22] LEO BREIMAN : Random Forests, Machine Learning, vol.45, pp.5-32, (2001)

[23] Wei-Yin Loh : Classification and regression trees, Vol.1, WIREs Data Mining and Knowledge Discovery, (2011)

[24] J.R.Quinlan : Bagging, boosting, and C4.5, Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp.725-730, (1996)

Table 12 Comparative performance of proposal and traditional methods

Method		Output	cor	$corR$	N50	$covR$	M_{inc}	M_{cor}	$corR_{1500}$	$covR_{1500}$
Traditional assembly with specified k value	Velvet:k=15	20	19	0.95	2963	0.98		4815	1.0	0.81
	Velvet:k=17	12	12	1.0	7889	0.98		10850	1.0	0.90
	ABySS:k=16	54	54	1.0	3048	0.87		4817	1.0	0.72
	ABySS:k=18	40	40	1.0	7891	0.77		10852	1.0	0.72
Traditional hybrid assembly	CISA	38	38	1.0	4044	0.98		10852	1.0	0.90
Traditional double assembly with multiple k value	DA	671	412	0.614	7849	0.99	15767	18729	0.4648	0.9393
Double assembly with traditional C4.5	DA^{DT}_{cor}	338	234	0.69	7849	0.96	15767	18729	0.57	0.938
	DA^{DT}_{incor}	444	315	0.71	7906	0.99	15767	18729	0.61	0.94
Double assembly with CDTwM	$DA^{CDT}_{positive}$	512	350	0.68	7356	0.96	15767	18729	0.57	0.938
	$DA^{CDT}_{negative}$	20	20	1.0	7894	0.62		10854	1.0	
Double assembly with CDTwRH	$DA^{CDT}_{N-ovlp5,2}$	231	191	0.827	5631	0.96	10859	10854	0.89	0.9076
	$DA^{CDT}_{N-ovlp5,9}$	43	33	0.767	10854	0.628	63	10855	1.0	0.618
	$DA^{CDT}_{N-ovlp5,cov1}$	180	147	0.8167	3148	0.603	1122	7892	1.0	0.55