O-MB12

Computer Aided Simulation of DNA Fingerprint Amplified Fragment Length Polymophism (AFLP) Using Suffix Tree Indexing and Data Mining

Kestrilia Rega P[1], Sulistyo Emantoko[1], Bhinawan Whendy[2], Agung Budiman[1]

[1]Biotechnology Faculty, University of Surabaya
[2]Departement of Informatics, Engineering Faculty , University of Surabaya
kestrilia@ubaya.ac.id, emantoko@ubaya.ac.id

Abstract

AFLP is one of the DNA Fingerprinting techniques which have broad application as genetic marker in various fields. Begin with the DNA sequence digestion using one or more particular restriction enzyme, ligation of the adapters to the overhanging sticky ends followed by DNA fragments amplification using PCR. The PCR reaction uses primers that match the adapter sequence and have some (1 to 3) additional "selective" bases which could be any bases, this reduces the number of bands that will be amplified. Such technique intended to increase the amplified fragments peculiarity so the polymorphism of the organism being studied could be well visualized by gel electrophoresis. The computer aided of AFLP simulation developed in this research was aimed to predict this electrophoresis result by simulate the digestion, ligation and PCR process using some pattern recognition algorithm applied to the DNA sequence from online databases. Through this simulation the researcher could determine the best combination of restriction enzyme and selective bases for their laboratory experiment. Suffix tree indexing was conducted during the exploration process of the genome sequence (in FASTA format) to find the restriction sites rapidly and create fragments of it. Data modeling enable the system draws the fragments into virtual DNA's electrophoresis pattern. Data mining accomplish the simulation by exploring overall possible virtual DNA's electrophoresis pattern and determine the best restriction enzyme and selective bases combination by calculating certain quantitative criteria.

Keywords : DNA Fingerprint, AFLP, PCR, Suffix Tree Indexing, Data Mining

I. INTRODUCTION

Since its first development in the mid-1980's, technique for DNA fingerprinting has rapidly evolved. In the field of agriculture, this technology assisted seed selection in order to acquire high quality plant such as cereals [1] and tea [2]. Many researcher suggested that Amplified Fragment Length Polymorphism (AFLP) is the best genetic marker nowadays in term of it's information quantity, reproducibility and resolution of genetic polymorphism. With this technique, DNA treated with restriction enzymes is amplified with PCR. It also allows selective amplification of restriction fragments, giving rise to large numbers of useful markers which can be located on the genome relatively quickly and reliably. Users can determine the specificity level of genetic marker by altering the restriction enzyme and sequence of bases in primer's selective bases. Unfortunately, due to the operation cost, it is

not an easy task to conduct trial and error attempt to find the best combination of restriction enzyme and selective bases. Therefore AFLP simulation program (in silico experiment) was developed in this research to help researchers simulate combinations of restriction enzymes and selective bases on virtual AFLP procedure by computational method so that they can determine the combinations that can be used to produce the desired genetic marker through in vitro experiment.

II. MATERIALS AND METHODS

Input of this computational method is DNA sequence from the online database. *Vitis Vinifera* genome sequence was taken from GenBank NCBI as an example and as much as 145 type II restriction enzymes were downloaded from the online Restriction Enzyme Database (rebase.neb.com). In order to make the simulation operational in the wet laboratory these 145 restriction enzymes were selected based on following criteria : (1) palindromic; (2) sticky end; (3) cut the DNA precisely on the restriction site; (4) no ambiguous and methylated bases on the restriction site; (5) at least one supplier available. Virtual restriction digestion then conducted by applying suffix tree algorithm as string pattern matching technique on the genome sequence. This algorithm will rapidly seek the string pattern which is match the restriction site of the enzymes being studied and then separate the genome sequence into subsequences. Hence, virtual PCR is done by exploring the compatibility between sub sequences and the primer-selective bases being studied. At the end of the simulation, exponential regression data modeling would enable the system draws the subsequences into virtual DNA's electrophoresis pattern. Data mining accomplish the simulation by exploring overall possible virtual DNA's electrophoresis pattern and determine the best possible restriction enzyme and selective bases combination by calculating certain quantitative criteria and conduct cluster analysis.

III. SYSTEM'S DESIGN

III.1 Input

DNA's genome sequence in FASTA format is required as system's raw material as well as the information of enzyme's restriction site pattern. The sequence could be store in several files (one file for each chromosome) in txt format. This FASTA sequence then considered as a text. Hence, all algorithm used in the consecutive processes should be string based algorithms.

III.2 Suffix Tree Algorithm

The first process is tracing the whole text (whole genome sequence) to find the short text (sub sequences) which is match the restriction sites of the restriction enzymes being

studied. The major computational problem when dealing with genome scale sequence is execution time due to computer's processor and memory performance limitation. It can take time up to one hour to find one short sequence along the whole genome [3]. Therefore, an effective string matching technique should be implemented to speed up the process. Hence, more restriction enzyme combination could be simulated. One popular technique to run fast string matching is suffix tree algorithm. Suffix tree are versatile data structures that can help execute short subsequences (queries) very efficiently. In fact, suffix trees are useful for solving a wide variety of string based problems [4]. For instance, the exact substring matching problem can be solved in time proportional to the length of the query, once the suffix tree is built on the database string. The example of suffix tree construction is shown in Figure 1 [5].

.

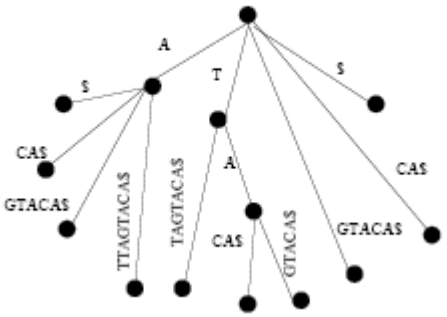


Figure 1. Suffix Tree Representation

The tree will inform every possible subsequence from a sequence as a pattern. One pattern is considered as particular path from the top node (root) to the most bottom node (leaf), for the example on the figure there are 10 possible sub sequences for the ATTAGTACA$ sequence. The $ character is added to inform the end of the sequence. There are three main function in this exploration process :

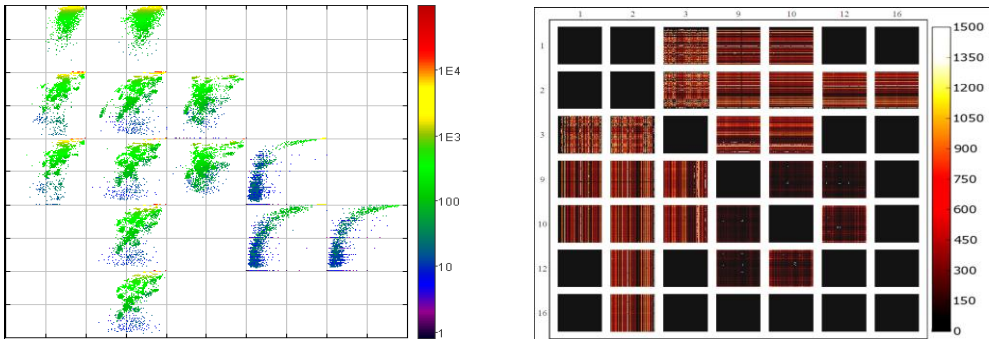
1. Build tree , construct suffix tree on the database. Every sequence (in FASTA format) subjected to the exploration should be transformed to the tree structure. Once it build, the FASTA format no longer needed so that it can be deleted and provide more space on the computer's memory.
2. Node searching, explore the tree for the queries, begin from the root (the top node) and end up at the leaf which is the most bottom node. If the query doesn't exist the system will report as "nothing". Each subsequence being found is indexed by number, represent its location on the sequence and its length (the number of the string).
3. Dispose, automatically erase the tree from the memory after it is stored on the database.

There will be 53.248 search on the *Vitis Vinifera* sequence's tree, the detail is explained in the following paragraph.

According to its restriction site, the restriction enzymes were classified into 44 groups and the simulation was conducted on 13 combinations among it. The combinations were determined as follow : (1) Three group having 4 bases of restriction site were paired with 3 group having 6 bases of restriction site, all with the most frequent match on the genome sequence; (2) The *Eco*RI and *Mse*I pair also included in the combinations although *Eco*RI do not fulfil the criteria because of there are facts that thus pair was used frequently for AFLP experiment [6,7,8,9]; (3) The restriction site of each pair do not overlap because such condition could lead bad and unpredicted restriction result. Three nucleotide selective bases were used for each subsequence's right and left hand end. Because there are 4 possible base (A,T,C,G), the total combination for selective bases should be 4^6 = 4.096. Therefore the total run for searching process on the tree is 13 x 4.096 = 53.248.

III.3 Cluster Analysis

The exploration result from the suffix tree then analyse by regarding on some criteria, which are : (1) Fragment (subsequence) length; (2) Percent of "in range" fragment, the number of fragment with the length does not exceed the polyacrylamide gel range criteria divided by the total fragment; (3) Percent of redundancy, the number of fragment with same length but different sequence divided by the total fragment. The analysis was done using multi dimension cluster analysis. The example of cluster representation is shown in Figure 2.



(a) (b)

Figure 2. Cluster Respresentation, (a) restriction enzyme and selective bases combination with their percent of "in range" fragment and percent of redundancy; (b) restriction enzyme and selective bases combination with their fragment length

III.4 The Selection Criteria

In order to find the best ten combinations of restriction enzyme and selective bases, the selection criteria should be well define. The combination will be considered good if : (1) Percent of redundancy less than 25%, too many different subsequence which have same length will reduce the polymorphism information; (2) Percent of fragment "in range" more

than 75%, too many fragment “out range” cause electrophoresis failure due to most of the fragment can not well visualize; (3) The average difference of the fragment length should be large enough so that it could be nicely separate on electrophoresis process. The selection is done by applying IF THEN rules.

III.5 Exponential Regression Model

To simulate the electrophoresis process, the system provide 1 Kbp DNA ladder from which the exponential model was developed. The exponential model between fragment size (bp) and its distance (cm) from the well is as follow:

$$\ln(\text{size}) = 10.81 - 0.736 * \text{distance}$$

IV. RESULT AND DISCUSSION

IV.1 Genome Description

The FASTA format of *Vitis Vinifera* genome sequence was separated in 19 different txt file, one file for one chromosome. Table 1 contains the description of each chromosome sequence component :

Table 1. The *Vitis Vinifera* Chromosome Sequence Description

Chromosome	Ambiguous bases per 1000 bases	GC Content (%)	Size of FASTA file (kb)
1	28	34,45	15.701
2	83	34,48	17.682
3	48	34,42	10.233
4	54	34,40	19.380
5	47	34,85	23.533
6	55	34,45	24.257
7	70	34,46	15.302
8	36	34,47	21.654
9	31	33,69	16.607
10	55	34,53	9.691
11	21	34,46	13.999
12	49	34,51	18.624
13	38	34,19	15.260
14	25	34,57	19.568
15	25	33,72	7.728
16	32	34,14	8.196
17	33	34,89	13.118
18	23	34,70	18.780
19	33	34,05	14.135
Total	**42**	**34,43**	**300.211**

IV.2 Exploration Process Performance

The main problem when facing with simulation of genome scale sequence is the operation time, but it is proven that by conducting suffix tree algorithm the operation time could be reduced significantly. Table 2 describes the time needed for suffix tree construction based on the size of the genome sequence. It is shown that the time needed increase in

linear form with the size of genome sequence, however the system could still operate in reasonable time (less than 3 minute) to handle genome sequence up to 12.1 Mbp long.. Once the suffix tree is constructed, all short pattern searching could be done in no time.

Table 2. Time for Suffix Tree Construction Based on Genome Sequence Size

Size of Genome Sequence (Mbp)	Time (second)
2,43	17
4,87	32
7,3	52
9,74	66
12,1	80

IV.2 Restriction Result Description

By conducting the data mining technique, there are several information that could be infer about the restriction result. It is known that a lot of small fragments were formed using a pair of restriction enzyme with 4 nucleotide restriction site, in the other hand just few bigger fragments were formed using a pair of restriction enzyme with 6 nucleotide restriction site. This facts were inline with the restriction digestion theory, restriction site with many nucleotide will have less probability to match the genome sequence. Therefore, the combination of restriction enzyme with 4 and 6 nucleotide of restriction site seems to be the better choice. These combinations will produce moderate number of fragments with moderate length as well.

IV.3 The Best Ten Combinations

Regarding to the selection criteria, the best ten combinations of restriction enzyme and selective bases were found. Table 3 describes thus combinations.

Table 3. The Best Ten Combinations Description

Rank	Restriction Enzyme		Selective Base		Range	Fragment in range		Redundancy	% of Restriction	% of Amplified Fragment
	1	2	1	2		Total	%			
1	AATT	ATGCAT	GCA	TAA	25-150 (126)	48	80,00%	22,92%	1,24%	0,04%
2	AATT	ATGCAT	GCA	CCA	25-150 (126)	44	80,00%	25,00%	1,24%	0,03%
3	AATT	AAGCTT	GCA	CTC	25-150 (126)	43	82,69%	23,26%	1,23%	0,03%
4	AATT	ATGCAT	GCA	TCT	25-150 (126)	42	76,36%	19,05%	1,24%	0,03%
5	AATT	ATGCAT	GCC	AAA	25-150 (126)	42	80,77%	23,81%	1,24%	0,04%
6	AATT	AAGCTT	GCA	ACA	25-150 (126)	41	77,36%	24,39%	1,23%	0,04%
7	AATT	ATGCAT	GCA	GAA	25-150 (126)	40	75,47%	20,00%	1,24%	0,03%
8	AATT	ATGCAT	GCC	TAA	25-150 (126)	38	77,55%	15,79%	1,24%	0,03%
9	AATT	ATGCAT	GCA	CGA	60-400 (341)	38	77,55%	21,05%	1,24%	0,03%
10	AATT	AAGCTT	GCA	TTT	25-150 (126)	38	77,55%	21,05%	1,23%	0,04%

Figure 3 depicts the visualization of virtual electrophoresis pattern based on the exponential regression model using 1 Kbp DNA Ladder. The blue line indicate that there is only one kind of subsequence with particular size, the green line indicate that there are two kind of subsequences with the same size, the red line indicate that there are three kind of subsequences with the same size and finally the black line indicate that there are more than three kind of subsequences with the same size. The black line should appears as the most thick and bright band in real gel electrophoresis result.

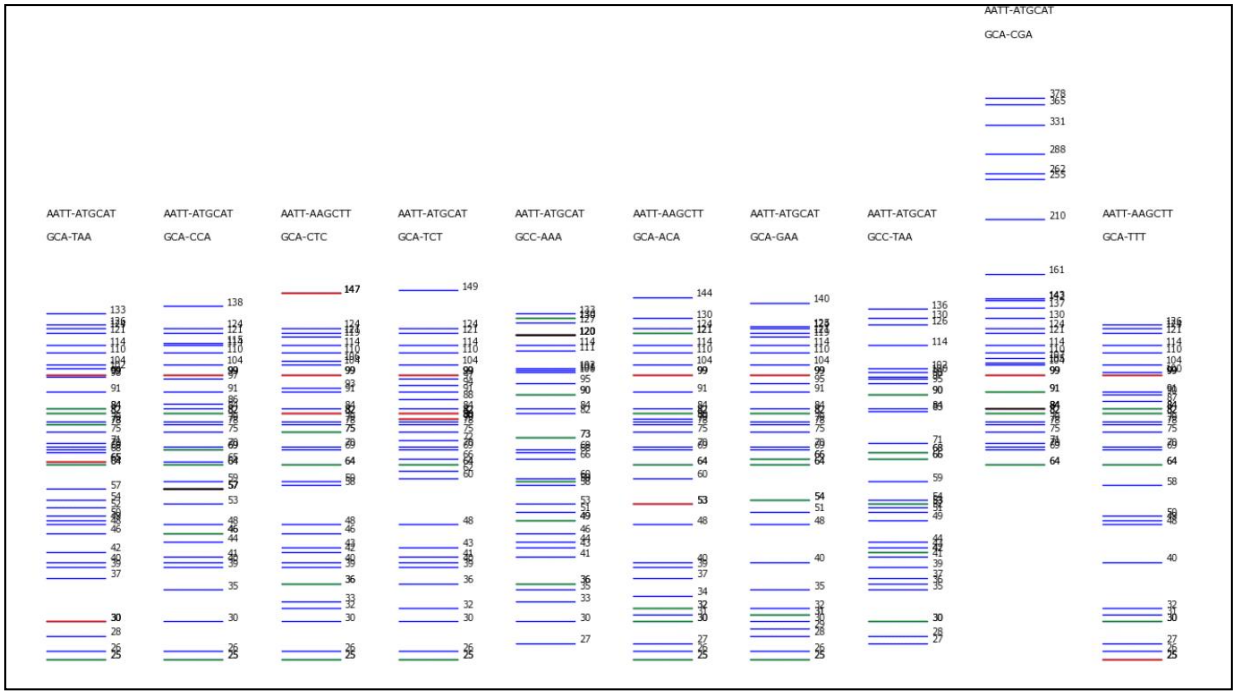


Figure 3. Visualization of The Virtual Electrophoresis Pattern

V. CONCLUSION

Like other simulation software, many factors embedded in laboratory experiment could not completely cover in this system, so that the result should be considered as recommendation (certainly with its probability of failure). However, so far the simulation result of AFLP with suffix tree indexing and data mining shows quite promising guidance for the laboratory experiment. The system developed in this research is a prototype from which more automatic and integrated system could be easily constructed. Machine learning technique such as genetic algorithm could be implemented to automate the optimization of selection criteria. At the end, laboratory conformation for this research result still could not leave behind. Therefore in the short incoming time such laboratory experiment should be conducted.

REFERENCES

[1] Korzun, V. 2003. *Molecular Markers and Their Applications in Cereals Breeding.A paper presented during the FAO international workshop on "Marker assisted selection: A fast track to increase genetic gain in plant and animal breeding?"*. Turin, Italy.

[2] Sui, N., Yao, M., Chen, L., Zhao, L., & Wang, X. 2008. *Germplasm and Breeding Research of Tea Plant Based on DNA Marker Approaches*. Front. Agric. China 2 (2): 200-207.

[3] Budiman, A. 2010. *Pembuatan Program Simulasi AFLP In Silico (Studi Kasus pada Vitis Vinivera)*.Thesis. Surabaya.

[4] D. Gusfield. 1997. *Algorithms on Strings, Trees and Sequences:Computer Science and Computational Biology*. Cambridge University Press.

[5] Tata, Sandeep., Hankins, Richard A., Patel Jignesh M. 2004. *Practical Suffix Tree Construction*. Proceedings of the 30th VLDB Conference, Toronto, Canada.

[6] Cervera, M. T., Gusmao, J., Steenackers, M., Van Gysel, A., Van Montagu, M., & Boerjan, W. 1996. *Application of AFLPTM-Based Molecular Markers to Breeding of Populus spp*. Plant Growth Regulation 20: 47-52.

[7] Koopman, W. J. M. & Gort, G. 2004. *Significance Tests and Weighted Values for AFLP Similarities, Based on Arabidopsis in Silico AFLP Fragment Length Distributions*. Genetics 167: 1915-1928.

[8] Mueller, U. G. & Wolfenbarger, L. L. 1999. *AFLP Genotyping and Fingerprinting*. TREE 14: 389-394.

[9] Stefenon, V. M., Gailing, O., & Finkeldey, R. 2006 *Phylogenetic Relationship Within Genus Araucaria (Araucariaceae) Assessed by Means of AFLPFingerprints*. Silvae Genetica 55 (2): 45-52.