Scientific Research Publishing

International Journal of Intelligence Science, 2017, 7, 9-23
http://www.scirp.org/journal/ijis
ISSN Online: 2163-0356
ISSN Print: 2163-0283

Identifying Cancer Disease through Deoxyribonucleic Acid (DNA) Sequential Pattern Mining

Lailil Muflikhah, Ilham Yuliantoro

Computer Science Department, Brawijaya University, Malang, Indonesia
Email: laililmf@gmail.com, ilhamyuliantoro@gmail.com

How to cite this paper: Muflikhah, L. and Yuliantoro, I. (2017) Identifying Cancer Disease through Deoxyribonucleic Acid (DNA) Sequential Pattern Mining. *International Journal of Intelligence Science*, **7**, 9-23.
http://dx.doi.org/10.4236/ijis.2017.71002

Received: October 1, 2016
Accepted: January 10, 2017
Published: January 13, 2017

Copyright © 2017 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

This paper aims to propose the sequential pattern discovery method of Deoxyribonucleic Acid (DNA) sequence database in order to identify cancer disease. The DNA which is composed of amino acids of gene *P*53 is mutated. It effects to change of *P*53 formation. Sequential pattern discovery is a process of extracting data to generate knowledge about the series of events that has the sequences in a certain frequency so that it creates a pattern. PrefixSpan is to propose method to find a pattern of DNA sequence database. As a result, there are various selected patterns of DNA sequence. The pattem which has high similarity is used as biomarker to identify the breast cancer disease. The performance measure of support value average is 0.8. It means that the frequent sequence pattern is high. Another measure is confidence. All of the confidence values are 1. Then, the last performance measure is lift ratio at average more than 1. It means that the composed sequence items in the pattern has high dependency and relatedness. Futhermore, the selected patterns are applied as biomarker with accuracy as 100%.

Keywords

Sequential Pattern, Breast Cancer, DNA, PrefixSpan, Lift Ratio

1. Introduction

Cancer is classified as malignant and deadly disease. The cancer disease is effective to uncontrolled growth of cells and gene mutations, *i.e.* gene of *P*53. This disease changes *P*53 protein sequence [1]. This protein consists of a combination of 20 amino acids which are synthesized by ribosomes and are performed based on the genetic code of the Deoxyribonucleic Acid (DNA). If the DNA is mutated, then the protein composition will be incorrect. Continuously, it is effective

to a variety of diseases and disorders such as cancer. Therefore, early detection can be conducted by analyzing of protein sequence through the blood test. The most frequently altered gene of *P*53 or *TP*53 mutations is found in human cancers. There are 30,000 somatic mutations of various cancer types of *TP*53 database which is collected over 20 years. Recently, the methodology of sequencing cancer genome impacts on the healing and data management [2]. According to Soussi, *P*53 mutation analysis of the pattern has become essential to investigate the cause of cancer. His test result shows that infrequent of gene mutation is associated to the normal activity of the *P*53 protein [2].

A field that is expected to provide his role is bioinformatics. This field is a knowledge discipline that combines the study of molecular biology, mathematics and information technology (IT). It is defined as the application of computational tools and analysis to capture and interpret molecular biology data. Molecular biology itself is also an interdisciplinary field in molecular level of life sciences [3]. Bioinformatics has a very important role, including the data management for molecular biology, especially DNA sequences and has huge volume of genetic information.

Many researches are conducted to find genome mutation using DNA sequence approach, identifying cancer patients through DNA micro-array to classify class of cancer using Genetic algorithm and K-means [4]. Another research is to find Ebola disease through DNA sequence pattern discovery [5]. Therefore, this research is to propose method of sequence pattern mining to identify cancer disease through DNA sequence discovery. Sequential pattern discovery is a part of data mining task that generates knowledge about the series of events that have frequent occur with specified threshold value [6]. The pattern is expected to use as biomarker of cancer disease. In this research is used PrefixSpan algorithm to discovery its pattern of DNA sequence database. It is a method of Sequential Pattern Discovery which has high performance of computational time [7]. This paper is organized as follows. In Section 2, it consists of literature studies including sequential genome, Sequential Pattern Discovery and performance measure evaluation. The design and implementation method of Prefix span is presented in this section. Then experimental result and analysis is provided in Section 3. The last section contains conclusion of this paper.

2. Research Method

In this research is started on data collection which is taken from NCBI. The related literature is sequential pattern discovery, prefix span algorithm and evaluation of the performance measure.

2.1. Sequencing Genome

DNA sequencing is a process of determining three million nucleotide bases order which consist of adenine, guanine, cytosine and thymine (A, T, G, C) in a DNA molecule. However, sequencing genome is the determination of the nucleotide sequence of DNA bases in the genome or in the body of an organism.

Sequencing results are expressed in the form of a sequence of letter nucleotide bases in specific DNA sequence, for example AGTCCGCAGGCTCGGT. Sequencing genome is always compared to coding process, whereas the sequencing process is not only defining a code, but also analogous genome sequence of letter from a mysterious language. It has an important and specific meaning.

There is a few mutations of DNA sequence that can effect to disorder gene and then it can be diseases. Hence, in medicine, sequencing genome is used to diagnose cancer diseases. A cancer is caused of abnormal DNA sequence construction in body cell.

2.2. Sequential Pattern Discovery

Sequential pattern mining is a data mining process that generates knowledge about the sequence of events that have frequent occur exceeds the threshold value [6]. Sequential pattern is a derived pattern from the association rules, because of the both indicate events relationship. The difference is that the sequential pattern of events focused on finding patterns of event that appear after another event. It is consider to time sequence. However, association rules is a pattern of events that appear with other events together.

2.3. PrefixSpan Algorithm

This research is conducted to find the pattern of DNA sequence using Prefix Span for patient having breast cancer. PrefixSpan (Prefix-Projected Sequential Pattern Growth) is a development approach that uses an algorithm to search for sequential pattern sequences. PrefixSpan will seek frequent sequences of the elements and then develop the sequences by adding elements one by one. As a result, the additional sequence is still the previous sequence. This way is not necessary to generate and test for candidates. There are two steps to find the candidates, including Prefix and projected database.

2.3.1. Prefix

If the sequence $A=\{a_1,a_2,\cdots,a_n\}$, sequence $B=\{a'_1,a'_2,\cdots,a'_m\}$, ($m <= n$) is known as prefix of A iff:

1) $a_1 = a'_1$
2) a'_m is set of a_n
3) all items in $(a_n - a'_m)$ as alphabetic is appear after a'_m

2.3.2. Projected Database

Given the sequence A and B so that B is subsequence of A. A sequence A' is subsequence of A is projected of prefix B, iff:

1) A' has refix B

2) There is no supersequence A'' of A' so that A'' is sub-sequence of A and has prefix B

The pseudocode of PrefixSpan algorithm is shown in **Figure 1** [7]

The pseudocode of Prefix Span algorithm in **Figure 1** is designed and applied to the system as shown in **Figure 2**. The initial step is input for DNA sequence

```
Input: sequence database, minimum support
Output: A complete set of sequential pattern
Method: Call prefixspan({},(),S).
Subroutine: Prefixspan(α,l,S|α)
Parameter: α: a sequential pattern, l: length α,S|α: projected
database, if α ≠(); else if then sequence database is S

Method:
1. Scan S|α once, find set frequent itemset, so that:
   a. b can be combined to last element from α to construct
      sequential pattern; or
   b. <b> can be added to α for constructing squential
      pattern
2. For each item b is appear, add to α for constructing a
   sequential pattern α', and output α';
3. For each α', construct α' projected database S|α', and call
   PrefixSpan (α', l+1, S|α').
```

Figure 1. Pseudocode of PrefixSpan algorithm.

database and minimum support as constraint. The second step is sequence numbering. It is based on transaction identifier to be sorted by length of string. The next step is scanning sequence database. It is to get the DNA sequential patterns and to count their frequency based on minimum support. The next step is divide search space to find the projected database and sequence pattern. In this step, it is retrieved a subsequence of their patterns and frequencies. If the number of frequencies is greater than or equal to the minimum support, then it is added as a postfix. These steps are repeated recursively until there is no subsequence of pattern.

2.4. Evaluation

There are several measurements to know the performance of system including support, confidence, lift ratio and accuration rate.

- Support

It is probability of frequent itemset in whole transaction. It is ratio of itemset transaction as in Equation (1) [6].

$$\begin{aligned} \text{support} &= P\left(A \cap B\right) \\ &= \frac{\text{number of transactions containing both A and B}}{\text{total number of transactions}} \end{aligned} \tag{1}$$

- Confidence

It is measurement that shows relation between two items conditionally as in Equation (2) [6].

$$\text{Confidence}, C\left(A \to B\right) = \frac{s\sigma\left(A \cup B\right)}{\sigma\left(A\right)} \tag{2}$$

where:

$\sigma\left(A \cup B\right)$ = the number of itemset in all transactions

$\sigma\left(A\right)$ = the number of antecedent in transaction

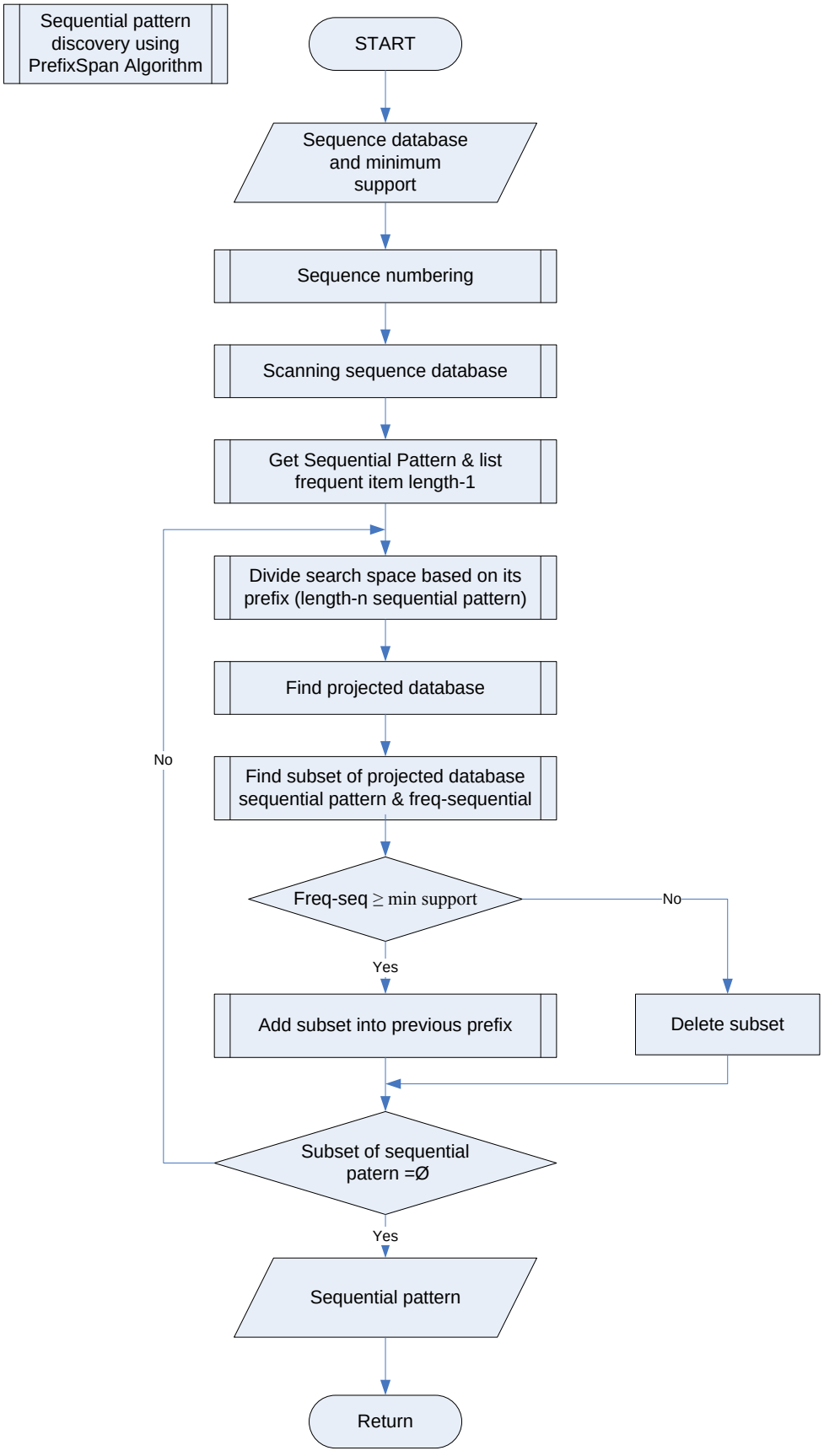

Figure 2. Flowchsart of sequential pattern using PreffixSpan algorithm.

- Lift ratio

Lift ratio is a measure to know the strength of constructed rule of sequential pattern mining algorithm. The value of lift ratio is between 0 and unlimited. It has definition as below: [8]

- If lift ratio value is less than 1, then it means that rule antecedent will be effect to negative of rule consequence.
- If lift ratio value is equal to 1, then the rule is frequent but it is independent.
- The others, if lift ratio value is more than 1, then it means that the rule antecedent will be effect to positive of rule consequent. It is recommended value.

The formula of lift ratio is stated in Equation (3) and Equation (4).

$$\text{Expected Confidence, } EC(A \to B) = \frac{\sigma(B)}{m} \tag{3}$$

$$\text{Lift} = \frac{\text{Confidence}}{\text{Expected Confidence}} \tag{4}$$

where:

- $\sigma(B)$ = the number of *consequence* in transaction
- m = the number of transaction

- Accuracy Rate

The accuracy rate in this research is purposed to find a motif (pattern) of DNA sequences. If the pattern (without consider to contiguous sequence) is match then it indicates that the accuration rate is 100%.

In this research, it is used modified method of string matching algorithm approach. The pattern of string is not exactly the same but it is considered to the same sequence pattern. It is evaluated using threshold minimum support. As illustration, the algorithm can be described as the below example:

SequenceID1: gttggctctg
The pattern: attca
Output: NotMatching
SequenceID2: **a**c**t**g**t**a**cca**c
The pattern: attca
Output: Matching (accuracy rate = 100%)

3. Results and Analysis

The data set of this experiment is taken from DNA sequence of patient's breast cancer. The DNA sequence database which is taken from 7th exon of ten human as is shown in Table 1. The Exon is a subset of DNA sequence to construct the code of protein.

As ilustration, the interface of application is shown at Figure 3. The DNA sequence database as input is taken from NCBI with minimum support and sequence length as threshold. Then the sequence data is performed as transaction data before applying Prefix Span method.

Table 1. Illustration of DNA sequence database (Homo sapiens breast cancer anti-estrogen resistance 3 (BCAR3), transcript variant 4, mRNA.

Human #	DNA Sequences
1	gttggctctgactgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
2	gttggctcaggactgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
3	gttggctctgactgtaccaccatcaactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
4	gttggctcggactgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
5	gttggctctgactgccaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
6	gttggctctgactgtaccaccatcaactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
7	gttggctcggactgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
8	gttggctctgacttgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
9	gttggctctgagctgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag
10	gttggctctgacgtgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggcccatcctcaccatcatcacactggaagactccag

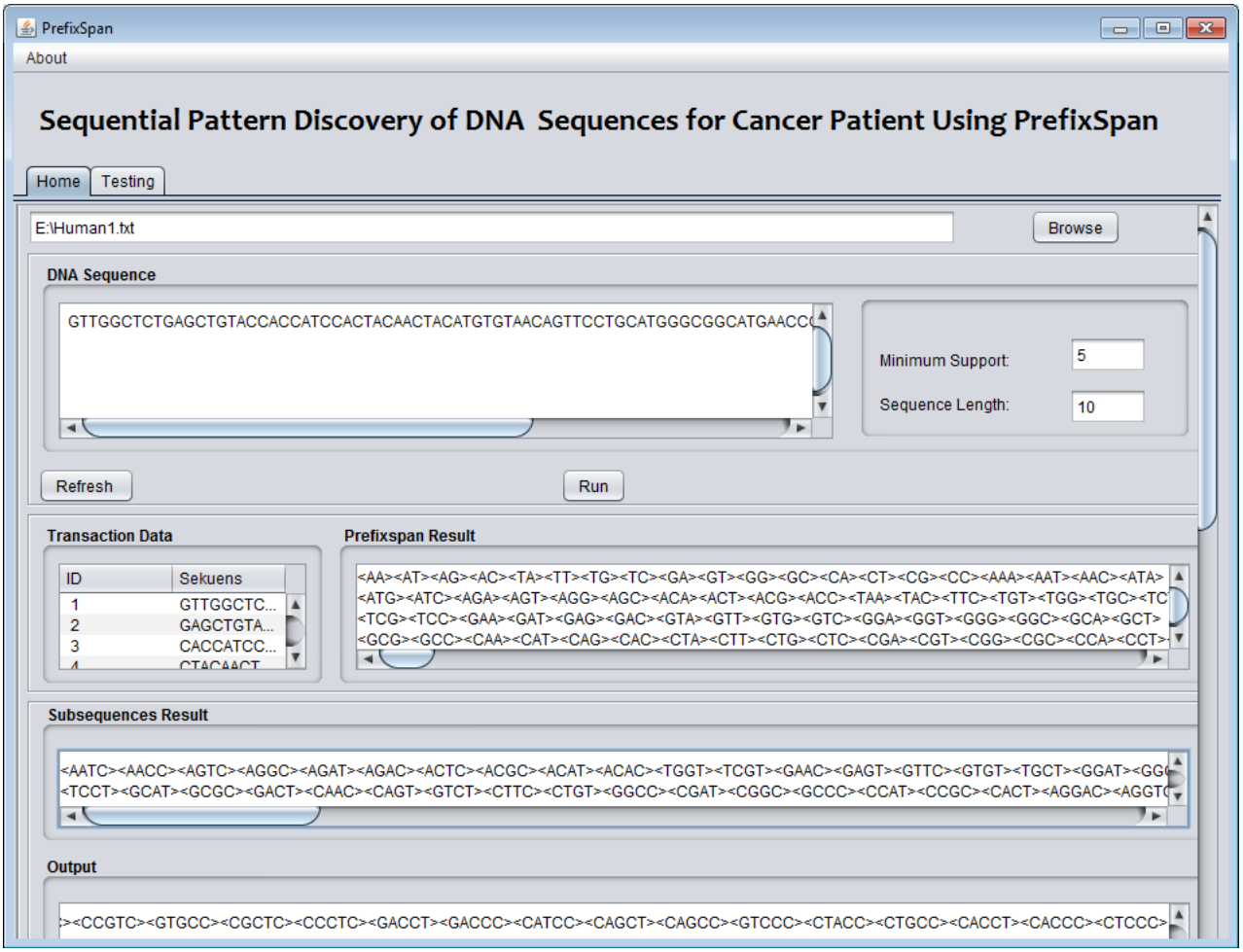

Figure 3. Interface of sequential pattern mining for cancer patient's DNA sequence.

3.1. Experimental Result

Several thresholds such as minimum support, maximum sequence length is given to the system and as a result, the sequence pattern of DNA is constructed. As

illustration, it is given minimum support count = 7 and sequence length = 10. It is applied to the 7th exon of human#1, then the patterns are found as in Figure 4.

A pattern, <aac> or a → a → c, which is shown in Figure 4 means that it is composed from sequence a is followed by a and the latest is c. This pattern occurred minimum seven times per ten sequences.

After finding several patterns in each human, it is retrieved a certain pattern of DNA sequence database which has high similarity. In this experiment, the data sets are the 7th of exon five humans(human #1, human #2, human #3, human #4, and human #5) for pattern retrieval. By giving various minimum support, the selected patterns are applied to five others DNA sequences database of human (human #6, human #7, human #8, human #9, human #10). Then, the selected pattern and accuracy rate are shown in Table 2.

3.2. Analysis

In order to know performance of the system, it is evaluated based on performance measure of support, confidence, lift ratio and accuracy. First, the experiment is applied to the training data with various minimum support and sequence length of 7th exon for human #1, human #2, human #3, human #4, and human #5.

The five graphs at Figure 5 up to Figure 9 show that the performance measurement including lift ratio and support of the constructed pattern tends to be

<aac><gtt><ggt><ggc><gct><ctt><cgt><cgc><cct><gtac><agtc><gtgc><ctac>
<actc><ctgc><atcc><gatc><agcc><accc><catc><gacc><gtcc><cacc><ctcc>

Figure 4. The resultedDNA sequential pattern of human # 1.

Table 2. The selected sequential pattern of DNA.

Min_sup	Pattern
2	<agggt><agcgt><aggct><acggt><agcct><accgt><acgct><ttgtg><ttgtc><accct> <ttctg><ttctc><ttggg><ttcgg><ttgcg><ttccg><gggtc><ggctc><gcgtc><gcctc> <gggct><ggcct><cggtc><gcgct><cgctc><gccct><ccgtc><ccctc><cggct><cgcct> <agggac><ccgct><cccct><agcgac><aggcac><acggac><agccac><accgac><acgcac> <acccac><gggatc><ggcatc><gcgatc><gccatc><ggacat><gcacat><cggatc><cgcatc> <ccgatc><cccatc><cgacat><ccacat><ggacct><gcacct><cgacct><ccacct>
3	<aaat><aagt><tttc><aaact><aagac><atacc><agtac><actac><ggact><gcact> <aggatc><agcatc><cgact><acgatc><ccact><accatc>
4	<tgtc><tggg><tctc><tcgg><tgcg><tccg><aggat><agatc><acgat><agcat> <accat><acatc><agacc><acacc>
5	<aatc><agat><acat><aggtc><aggac><acgtc><acgac><agctc><agcac><accac><acctc>
6	<aat><ggt><cgt><aaac><aacc><atac><agac><agct><acac><acct><gacc><cacc>
7	<agtc><actc>
8	<agt><act><agcc><accc>
9	<aac>
10	<tcc>

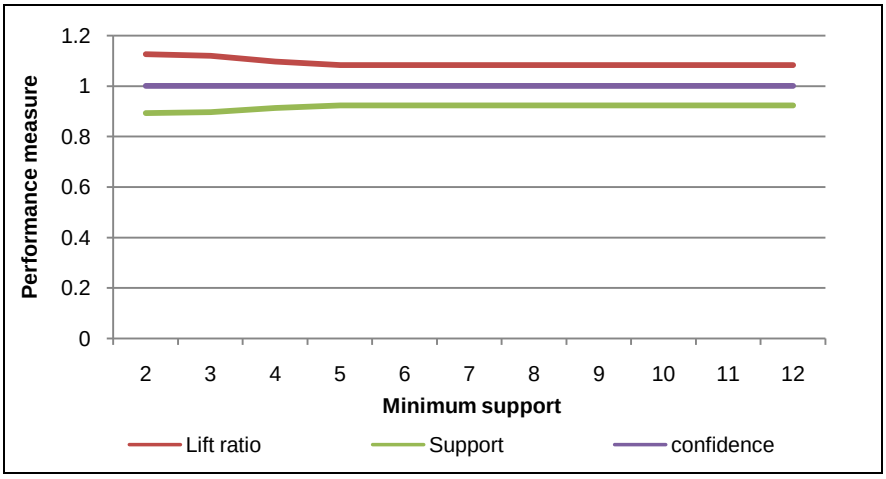

Figure 5. Performance measure of DNA sequence for human #1 with various minimum supports.

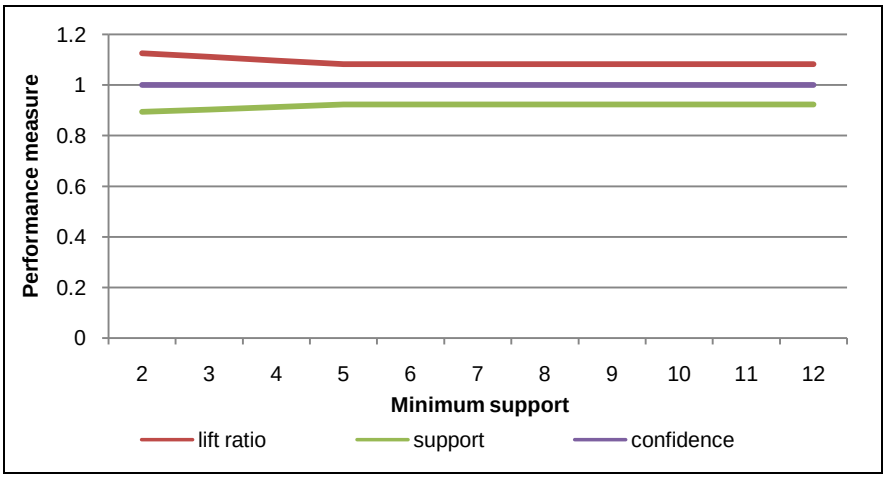

Figure 6. Performance measure of DNA sequence for human #2 with various minimum supports.

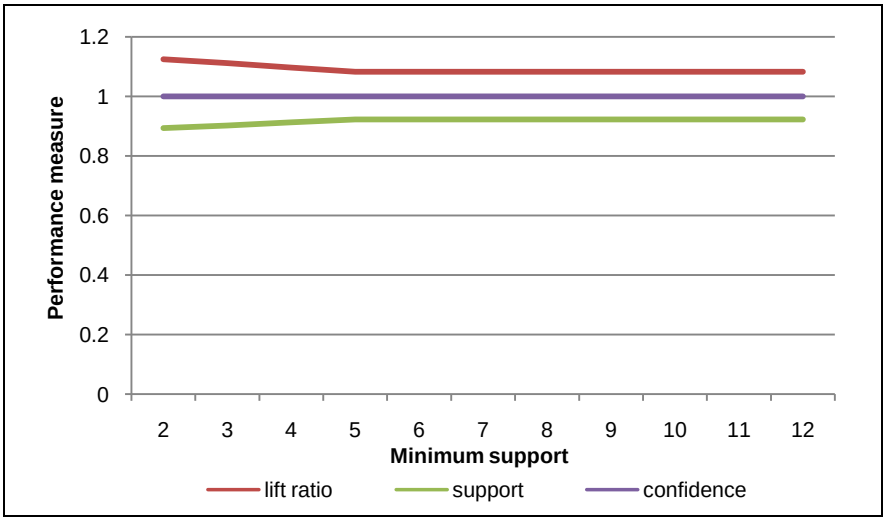

Figure 7. Performance measure of DNA sequence for human #3 with various minimum supports.

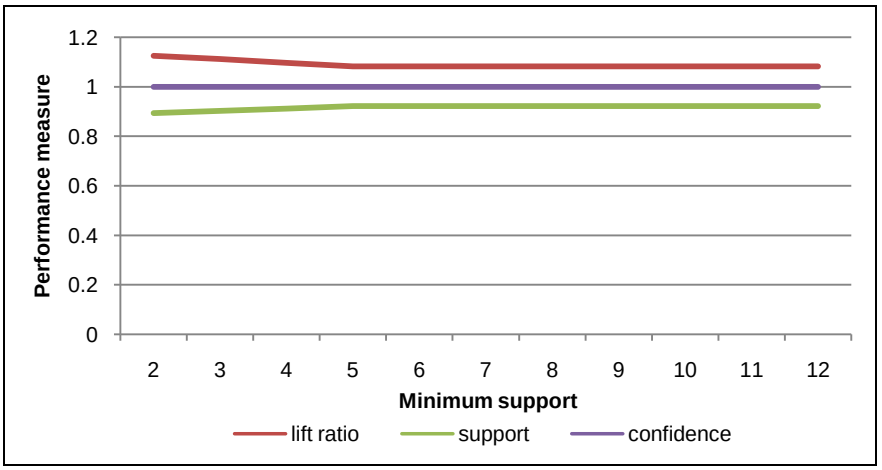

Figure 8. Performance measure of DNA sequence for human #4 with various minimum supports.

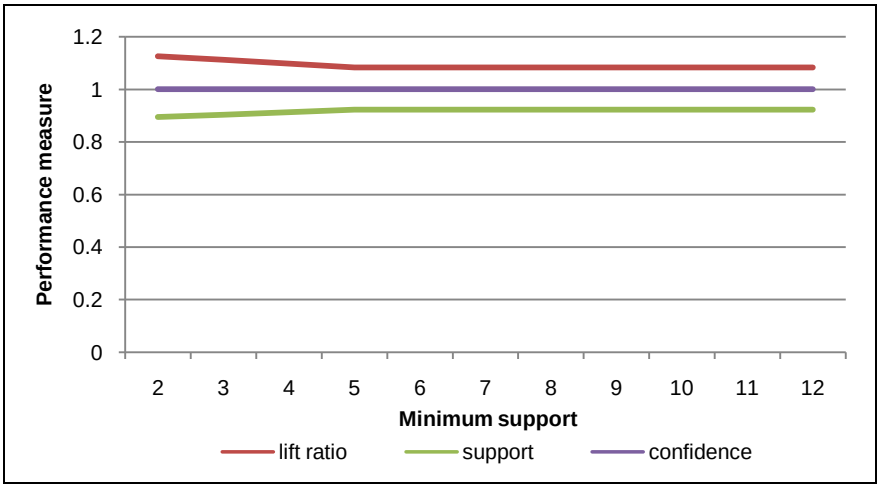

Figure 9. Performance measure of DNA sequence for human #5 with various minimum supports.

stable. The best performance is at minimum support in range 6 and 11 at the sequence length 10. The both measurements almost close to 1. Also, for the confidence is always at value 1. It means that the selected patterns have high strengthness and correctness.

Then, the second experiment is used various sequence length with minimum support 7. As a result, the patterns are constructed with performance measure as shown in Figure 10 up to Figure 14. Almost all graphs on the Figure show that lift ratio and support value are stable. They have the best performance within interval range 6 - 12 of sequence length. However, the confidence value is not depend on the sequence length. It is always at 1.

In general, the experimental result is presented that the more minimum support, the less number of the selected pattern is as shown in Table 2. Also, the maximum minimum support is ten for selecting pattern of five human. All accuration rates are 100% (match within pattern sequence) as in Table 3. This shows that the selected patterns are appear more frequently than and equal to

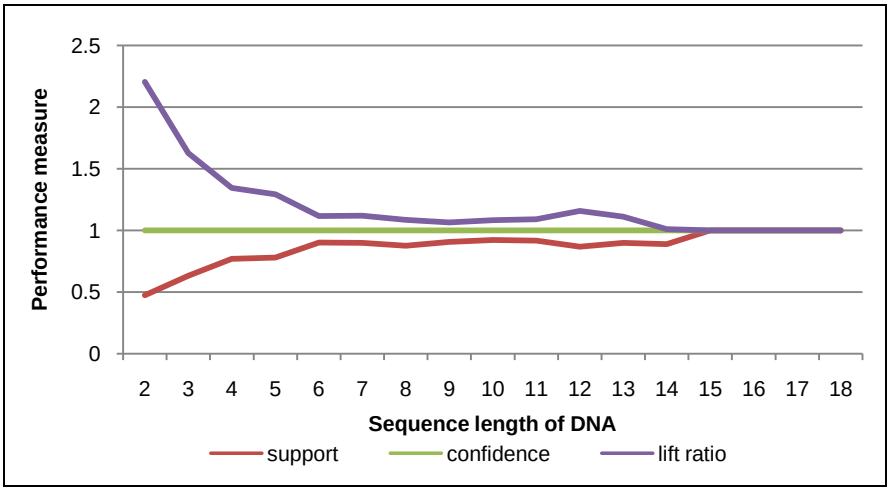

Figure 10. Performance measure of DNA sequence for human #1 (by various sequence length).

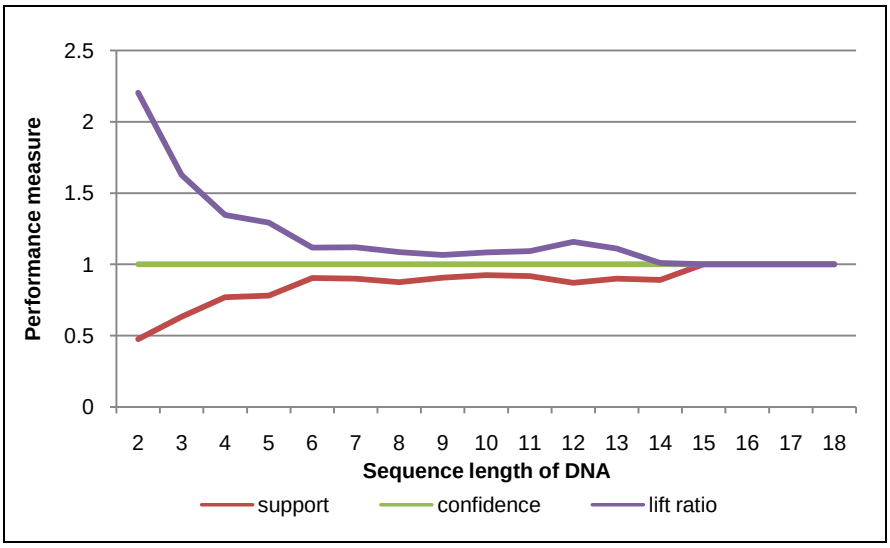

Figure 11. Performance measure of DNA sequence for human #2.

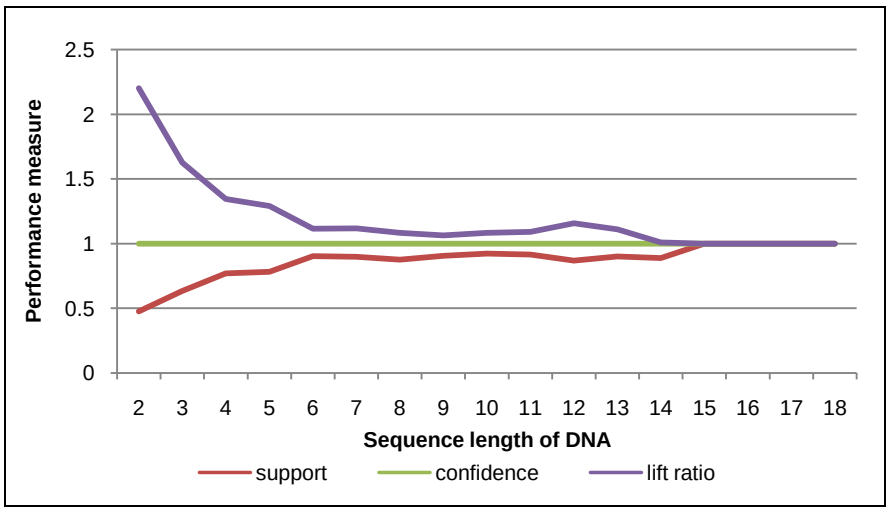

Figure 12. Performance measure of DNA sequence for human #3.

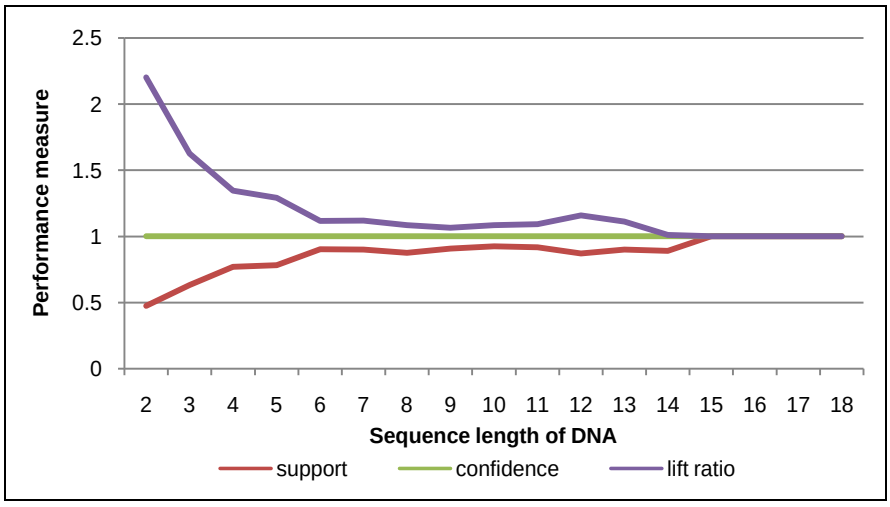

Figure 13. Performance measure of DNA sequence for human #4.

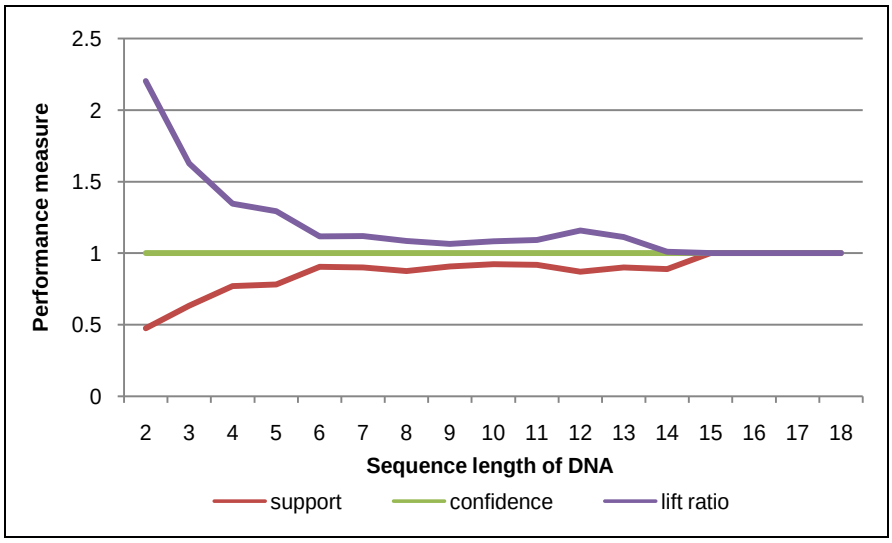

Figure 14. Performance measure of DNA sequence for human #5.

the threshold as minimum support and sequence length. Also, the maximum selected pattern is ten times.

As illustration, one of the threshold is minimum support = 7 and the sequence length = 10. Then, there are two resulted patterns, such as <agtc> and <actc>. The both patterns, a → g → t → c and a → c → t → c, are found minimum seven times per length 10 of DNA sequences.

Furthermore, the computational time of proposed method for DNA sequential pattern mining is evaluated as at Table 4. The trend of this performance is related to the number of constructed rule (pattern) as in Table 5. It has similar characteristic to the selected pattern. The more minimum support value is given, the less time computation is required.

4. Conclusion

Sequential pattern discovery is a method in data mining task to find the pattern

Table 3. Accuracy rate of the selected pattern.

Min_sup	The selected pattern	Accuracy rate
2	<agggt><agcgt><aggct><acggt><agcct><accgt><acgct><ttgtg> <ttgtc><accct><ttctg><ttctc><ttggg><ttcgg><ttgcg><ttccg> <gggtc><ggctc><gcgtc><gcctc><gggct><ggcct><cggtc><gcgct> <cgctc><gccct><ccgtc><ccctc><cggct><cgcct><agggac><ccgct> <cccct><agcgac><aggcac><acggac><agccac><accgac><acgcac> <acccac><gggatc><ggcatc><gcgatc><gccatc><ggacat><gcacat> <cggatc><cgcatc><ccgatc><cccatc><cgacat><ccacat><ggacct> <gcacct><cgacct><ccacct>	100%
3	<aaat><aagt><tttc><aaact><aagac><atacc><agtac><actac><ggact> <gcact><aggatc><agcatc><cgact><acgatc><ccact><accatc>	100%
4	<tgtc><tggg><tctc><tcgg><tgcg><tccg><aggat><agatc> <acgat><agcat><accat><acatc><agacc><acacc>	100%
5	<aatc><agat><acat><aggtc><aggac><acgtc><acgac><agctc> <agcac><accac><acctc>	100%
6	<aat><ggt><cgt><aaac><aacc><atac><agac><agct><acac> <acct><gacc><cacc>	100%
7	<agtc><actc>	100%
8	<agt><act><agcc><accc>	100%
9	<aac>	100%
10	<tcc>	100%

Table 4. The running time of sequential pattern mining for five humans.

min_sup	Running Time (in second)				
	human #1	human #2	human #3	human #4	human #5
1	15.36530	16.33520	14.84030	15.97984	32.42403
2	3.18270	3.76670	3.55940	2.54936	15.15930
3	1.85060	1.48270	0.74580	0.40192	2.53526
4	0.73160	0.79890	0.31680	0.39710	1.85680
5	0.46220	0.43710	0.22890	0.27281	1.14540
6	0.06514	0.08399	0.09363	0.07259	0.99023
7	0.01747	0.01828	0.02560	0.02337	0.51263
8	0.01421	0.01970	0.00561	0.01601	0.43998
9	0.00300	0.00562	0.00186	0.00182	0.36660
10	0.00212	0.00164	0.00154	0.00146	0.29135
11	0.00109	0.00120	0.00101	0.00102	0.29135
12	0.00086	0.00075	0.00005	0.00037	0.14028

Table 5. The number of sequence pattern for five humans.

min_sup	Number of pattern				
	human #1	human #2	human #3	human #4	human #5
1	2384	2393	1570	1631	3220
2	554	588	430	410	1545
3	263	265	206	203	405
4	154	153	104	107	264
5	90	84	50	53	230
6	39	41	37	34	194
7	24	19	21	17	126
8	11	11	16	16	96
9	11	11	10	10	76
10	9	10	8	8	45
11	6	6	6	6	45
12	4	4	1	1	31

on DNA sequence database of patient's cancer disease. By giving the threshold such as various minimum support and sequence length, different sequence patterns are obtained. In order to know reliability of the system, it is evaluated using three performance measures, including support, confidence, and lift ratio. The experimental results show that almost all the support value is closed to 1. It means that the probability of co-occurrence for sequence-items in the pattern is high. Also, the confidence value is always 1. This shows that the pattern has believable value at 100%. Furthermore, the lift ratio is almost always more than 1. This means that the patterns which consist of items are dependent each other.

After applying to other dataset, *i.e.* five sequence DNA database from different human, it is achieved the accuration rate as 100%. It means that the pattern selected is valid with maximum appearance of items for ten times in each 10 sequential items. Therefore, the patterns can be used to define the DNA sequence characteristic (motif) of the patient's cancer diseases.

References

[1] Pustai L., Lewis, C. and Yap, E. (1996) Cell Proliferation in Cancer-Regulation Mechanisms of Neoplastic Cell Growth. Oxford University Press, Oxford.

[2] Soussi, T. (2011) TP53 Mutations in Human Cancer: Database Reassessment and Prospects for the Next Decade. *Advances in Cancer Research*, **110**, 107-139. https://doi.org/10.1016/B978-0-12-386469-7.00005-0

[3] Sander, C. (2001) Bioinformatics Challenges in 2001. *Bioinformatics*, **17**, 1-2. https://doi.org/10.1093/bioinformatics/17.1.1

[4] Zubi, Z.S. and Emsaed, M.A. (2013) Identifying Cancer Patients Using DNA Micro-Arrayy Data in Data Mining Environment. *Journal of Science and Engineering*, **3**, 63-75.

[5] Kalaiselvi, S. and Meena, A. (2016) Efficiency of Using Sequence Discovery for Po-

lymorphism in DNA Sequence. *International Journal of Scientific and Technical Advancements*, **2**, 95-100.

[6] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.

[7] Pei, J., Han, J.W., Mortazavi-Asl, B., Wang, J.Y., Pinto, H., Chen, Q.M., *et al.* (2004) Mining Sequential Patterns by Pattern Growth: The PrefixSpan Approach. *IEEE Transaction on Knowledge and Data Engineering*, **16**, 1424-1440.

[8] Fomby, T. (2011) Association Rules (Aka Affinity Analysis or Market Basket Analysis). Department of Economics, Southern Methodist University, Dallas, TX.

Scientific Research Publishing

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/
Or contact ijis@scirp.org