Int. J. Nonlinear Anal. Appl. (xx) xx xx, xxx-xxx
ISSN: 2008-6822 (electronic)
http://dx.doi.org/11

Roughsets For Identifying Rice Protein Component from DNA Sequence

Rossy Nurhasah[a], Ade Sarah Hufaizah[a], Reza Taqyuddin[a], Erna Budhiarti Nababan[a,*]

[a] *Department of Information Technology, Universitas Sumatera Utara, Indonesia*

(Communicated by Erna Budhiarti Nababan)

Abstract

Text of the abstract. Good quality rice may be identified on DNA level by extracting protein

.

.

.

Keywords: (DNA Sequence, Rough Sets, Identification)
2021 MSC: Primary xxxxx (mandatory); Secondary xxxxx, xxxxx (optionally)

1. Introduction

Rice is a first-class commodity in Asia that commonly consumed as primary food source because it has carbohydrate as energy source for people to help them activity. Some Asian countries made rice as primary indicator to maintain Food Tenacity Index. Demands of high-quality rice keeps increased every year, it helped the rapid development of agricultural and bio science. The expansion of genomic research combining with genetics and molecular biology and advanced technological modelling used to conduct data analysis to gain another perspective [1]. Design a DNA sequence project from human genome was set by the 1000 Genomes Project [2] to get the maximum output of human.

Genetic factors contribute significantly to determine property of high-quality rice, using cooking and eating quality indicator, a rice grain must consist of following features such as quality of the starch, total protein in grain, flavor, and grain size [4]. An analyzer used in evaluating rice quality which can analyze starch consistency because it called Rapid Visco-Analyzer (RVA) because it only needs small number of samples to conduct data analysis, the identification of new quantitative trait loci (QTLs) for RVA was profiling of great significance to improve rice quality [5].

[*] Corresponding author

Email addresses: rossynurhasanah@usu.ac.id (Rossy Nurhasah), adesarah@usu.ac.id (Ade Sarah Hufaizah), rezataqyuddin@usu.ac.id (Reza Taqyuddin), ernabrn@usu.ac.id (Erna Budhiarti Nababan)

Received: xxx yyyy *Accepted:* xxx yyyy

Research by Li [6] represent a statistical framework to Single nucleotide polymorphisms (SNP) calling using Japonica Rice variant of Oryza Sativa to map Apparent amylose content (AAC) and Protein Component (PC) as main factor of rice quality [7]. In this study, we will identify the quality of rice based on amylose protein from DNA sequence data using Rough Sets Algorithm to determine weight of feature of each Protein.

2. Methodology

2.1. Data Preparation

First, we collect the rice DNA data from site rice.uga.edu. The site itself has consist of collection that provides sequence and annotation data for rice gnome from National Science Foundation [8]. We construct an automatic data scraping tools using python programming using locus category as reference from funrice.github.io/categories. Web scraping is an alternate method to gather data from website other than conventional API by automate a program to query and parse the webpage to be extracted [9]. In a scraping process there are several steps to follow:

- Retrieve DOM structure of target Website.
- Parse data target.
- Save the data.
- Continue move to next page

These steps constructed as pseudocode below :

```
def scrape_data:
   init starting_page
   set target_selector
   parse dom for every target_selector :

   while target next not null or empty :
       data <- extracted from target
       contain data{}
       continue next
```

2.2. DNA Sequence Reading

DNA, or deoxyribonucleic acid are material in almost any organism. It constructed by a pair of nucleotides making a genome. To read DNA information, scientist store it as code of 4 chemical base (A, C, G, and T) sequentially and carries information to assemble protein and RNA molecules [10]. Research by Heather lays out the history of DNA sequencing where researchers for years have been figuring out how to read DNA to get nucleotide information quickly. The first Generation of DNA Sequence on 1977 introduced a technology with the development of Sanger's chain-termination that makes use deoxyribonucleotides chemical analogues (dNTPs) which is the monomer of DNA strands as shown in figure 1 [11] A new method using ATP sulfurylase to convert pyrophosphate into ATP to infer sequence nucleotide in order turn over the template DNA, The pyrosequencing method then evolved to successful as Next-Generation sequencing technology.

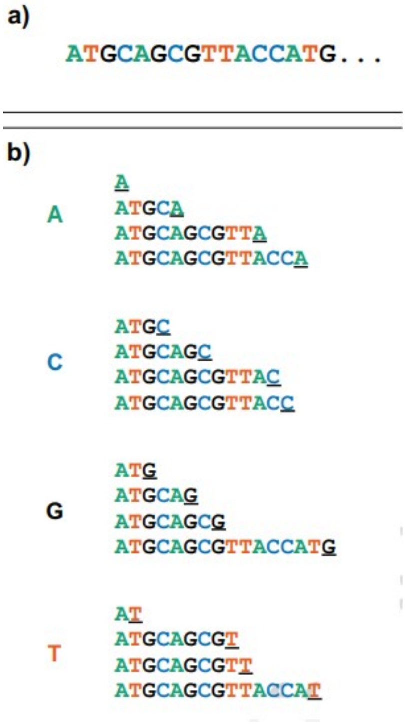


Figure 1: DNA Sequencing Model [1]

2.3. *Rough Sets*

Since DNA data contains collections of attributes that has correlation to each other's we need to define which attribute has greater impact for identification. In Mathematics, sets must be exact to make reasoning possible. Pawlak proposed Rough set theory to overcome vagueness of sets mathematically with the assumption every object in data has association with others. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated by characterizing similar information. The given set of functional attributes make the indiscernibility In the rough set approach is defined relative.

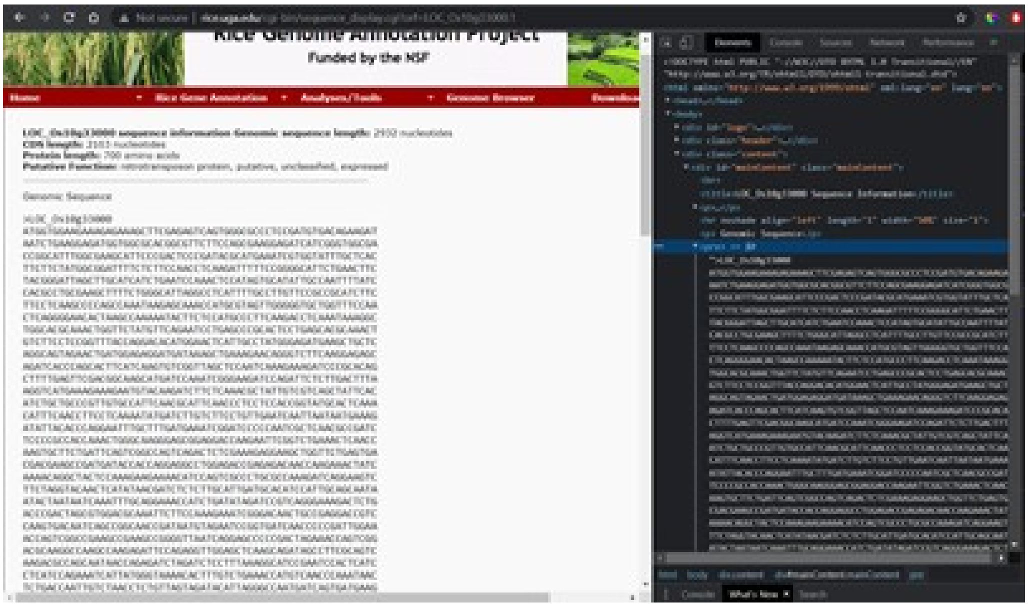
2.4. *Feature Selection using RoughSets*

Imperfect knowledge commonly faced by researcher because too many irrelevant features, noise, and even misleading feature. We need to know and understand the features of the data. Feature and subset usefulness determined by relevancy and redundancy. Features that highly correlated with decision features(s) but least correlated with each other called a good feature. To discover data dependencies in a dataset, a Rough set theory (RST) can be used to reduce the number of attributes without require additional information. In perspective of dimensional reduction, predictive class attributes are those that has most informative feature [11] [12]. Research by Riza [13] has result on released packages for Rough Sets in R programming with implementation of Rough Set Theory (RST) and Fuzzy Rough Set Theory (FRST).

2.5. *Rough Set Theory (RST)*

Research from Pawlak [12] consider the relation to model indiscernibility, an information system $(\cup, A)$ for any relation of $B \subseteq A$ the equivalence relation R_B is defined by :

$$R_b(X, Y) = (X, Y) \in U^2 | \forall a \in B, a(X) = a(Y)$$



LOC_Os10g33000 sequence information Genomic sequence length: 2932 nucleotides
CDS length: 2103 nucleotides
Protein length: 700 amino acids
Putative Function: retrotransposon protein, putative, unclassified, expressed

Genomic Sequence

>LOC_Os10g33000
[illegible]

Figure 2: Document Object Model of target Data

2.6. Fuzzy Rough Set Theory (FRST)

In FRST, it is assumed that R is at least a fuzzy tolerance relation. To construct a fuzzy B-indiscernibility relation for $B \in A$ proceeds by considering a fuzzy tolerance relation R_a for each quantitative attribute a, such as the following equations considered by Jensen and Shen in [11]:

$$R_a(X,Y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|}$$

$$R_a(X,Y) = exp(-\frac{(a(x) - a(y))^2}{2\sigma_a^2})$$

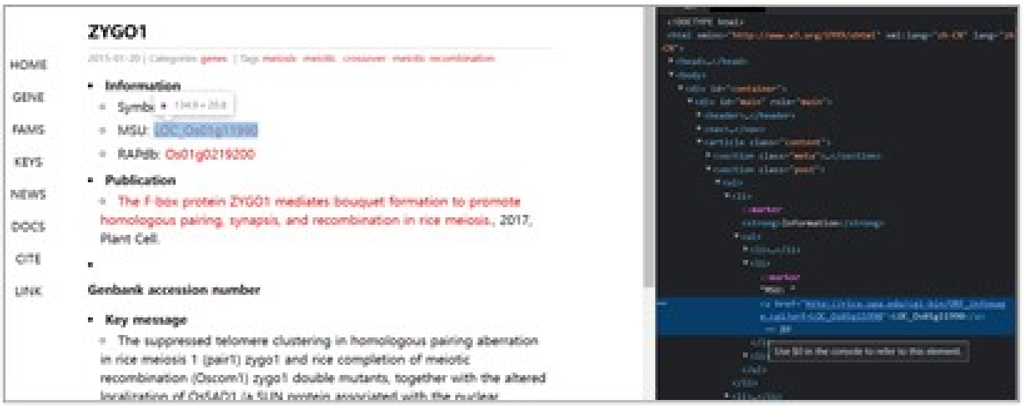
$$R_a(X,Y) = max(min(\frac{(a(x) - a(y))^2}{2\sigma_a^2}))$$

3. Result and Discussion

3.1. Data Collection

We collect data from the rice.uga.edu website by extracting information on the DOM elements on the website as shown in the figure 1. This is the target of our data so we will determine whether it can be automatically scraped or not. Since the URL require locus name as query parameter, then we try to find list of rice locus as reference. We find a GitHub repository which has curated list of rice genes by locus name from funricegene.github.io, see figure 2.

After finding the exact DOM containing our data, we can construct a scraping script. Scraping uses the Document Object Model to select a target then extract the value from html tag. HTML tag contains many information stored like innerHtml, attribute value or even data-* and other object. The scraping continue working in the background by crawling next hyperlink to another then extract the information and save it into csv file. After running the script, we successfully gather about 544 locus data with different length of DNA strand.



ZYGO1

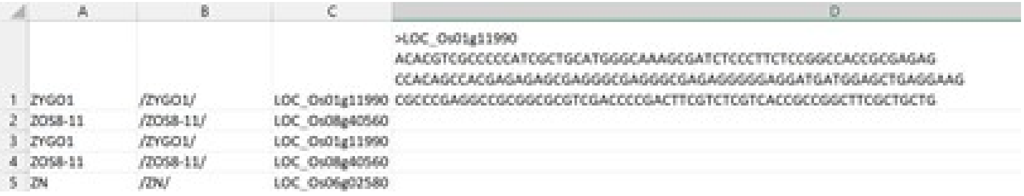
HOME
GENE
FAMS
KEYS
NEWS
DOCS
CITE
LINK

- **Information**
  - Symbol:
  - MSU: LOC_Os01g11990
  - RAPdb: Os01g0219200
- **Publication**
  - The F-box protein ZYGO1 mediates bouquet formation to promote homologous pairing, synapsis, and recombination in rice meiosis., 2017, Plant Cell.
-

Genbank accession number

- **Key message**
  - The suppressed telomere clustering in homologous pairing aberration in rice meiosis 1 (pair1) zygo1 and rice completion of meiotic recombination (Oscom1) zygo1 double mutants, together with the altered localization of OsSAD1 (a SUN protein associated with the nuclear

Figure 3: Locus Name target on funricegene.github.io



	A	B	C	D
1	ZYGO1	/ZYGO1/	LOC_Os01g11990	>LOC_Os01g11990 ACACGTCGCCCCCATCGCTGCATGGGCAAAGCGATCTCCCTTCTCCGGCCACCGCGAGAG CCACAGCCACGAGAGAGCGAGGGCGAGGGCGAGAGGGGGAGGATGATGGAGCTGAGGAAG CGCCCGAGGCCGCGGCGCGTCGACCCCGACTTCGTCTCGTCACCGCCGGCTTCGCTGCTG
2	ZOS8-11	/ZOS8-11/	LOC_Os08g40560	
3	ZYGO1	/ZYGO1/	LOC_Os01g11990	
4	ZOS8-11	/ZOS8-11/	LOC_Os08g40560	
5	ZN	/ZN/	LOC_Os06g02580	

Figure 4: Generated Locus DNA Data

3.2. Data Analysis

We continue our research by conducting analysis from collected data using BioPython library to parse the DNA sequence from fasta file format into computer readable. Since the library only can read from a fasta file format, we dump each DNA into separate dataset based on locus name e.g ZYG01.fasta and so on. After separating our dataset, we parse each fasta file to statistical data so we could extract the sequence information. The read.fasta function parse the strand into represent objectIDs, and Sequence string. From our 544 dataset, we make a histogram of each sequence length as seen on figure 5

At this point, we only can see the distribution of DNA strand length and GC-content (Guanine-Cytosine) values from each locus. For further exploratory data analysis, we read the DNA strand to get the possible sequence that has effectiveness from each others. For example, if we have a sequence seq(”GATCGATGGGCCTATATAGGATCGAAAATCGC”), we may count any combination of 4 main characters : A, T, C and G so we can get another attribute describing each locus as shown in Table 1.

Table 1: DNA information extracted model

Locus Name	Length	GC%	AA	AG	AT	AC	GG	GA	GT	GC	CC	CA	CG	CT	TT	TA	TC	TG
LOC_Os01g11990	3049	54.83	97	140	157	128	158	190	184	258	158	190	184	258	183	167	215	243
LOC_Os08g40560	925	57.08	32	55	51	42	74	62	51	84	48	58	90	42	47	28	64	52
LOC_Os01g11990	2660	48.83	161	168	194	119	144	177	114	209	116	178	160	164	157	126	174	172
LOC_Os08g40560	3771	50.99	154	218	220	183	178	233	221	285	195	270	171	299	223	118	272	350
...																		
LOC_Os06g02580	6261	40.07	387	341	524	324	198	327	346	271	240	408	128	495	491	455	436	474

====== Next Explanation on weighted attribute
Grouping attribute and Feature Selection using roughset
========

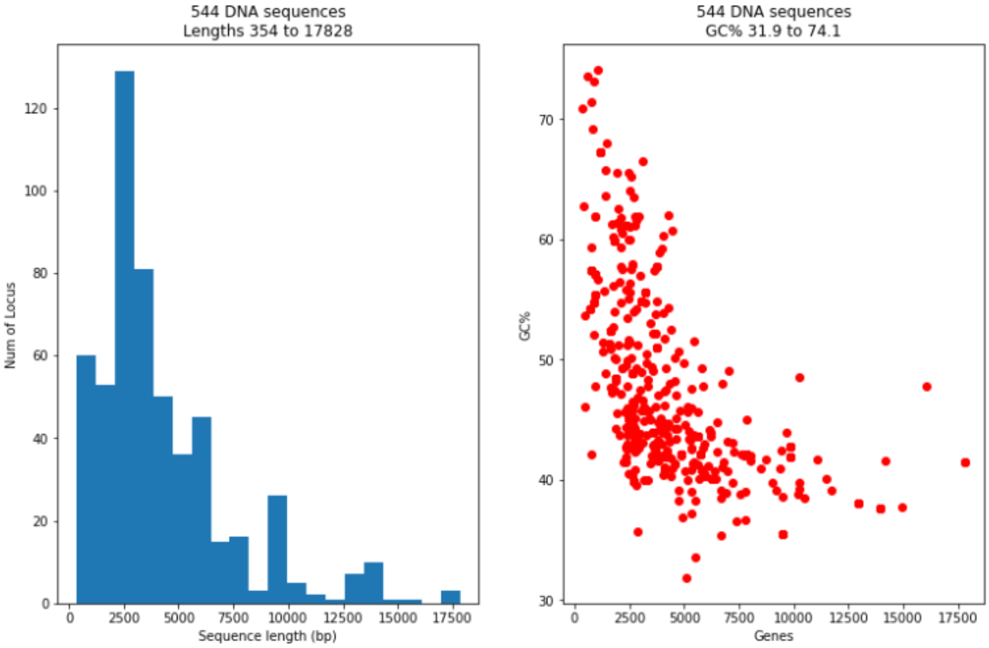


Figure 5: Data Distribution model

4. Conclusion and Future Work

Kita dapat memahami bagaimana informasi di dalam DNA tersusun melalui komponen sequence yang dibaca secara statistik. Setiap kombinasi komponen memiliki peran tersendiri pada untaian DNA. Penggunaan Roughset untuk menentukan komponen mana yang memiliki korelasi tinggi dengan yang lain dapat digunakan sehingga bisa mereduksi atribut yang banyak.

References

[1] Z. Zuo, C. Tang, Y. Xu, Y. Wang, Y. Wu, J. Qi and X. Shi *Gene Position Index Mutation Detection Algorithm Based on Feedback Fast Learcning Neural Network*, In *Computational Intelligence and Neuroscience*, 2021.

[2] L. Clarke, S. Fairley, X. Z. Bradley, I. Streeter, E. Perry, E. Lowy, A.M. Tasse and P. Flicek, *The International Genome Sample Resource (IGSR): A Worldwide Collection of Genome Variation Incorporating the 1000 Genomes Project Data* , In *Nucleic Acids Research*, 45 D1 (2016) D854-D859.

[3] FH. Clarke, YS. Ledyaev, RJ Stern and PR. Wolenski, *Nonsmooth analysis and control theory.* Springer Science and Business Media, 2008.

[4] T.-Q. Yu, W. Jiang, T.-H. Ham, S.-H. Chu and P. Lestari, *Comparison of Grain Quality Traits between Japonica Rice Cultivars from Korea and Yunan Province of China*, *Journal of Crop Science and Biotech,* 2008 135-140.

[5] Z. Changquan, B. Hu, K.-z. Zhu and H. Zhang, *QTL Mapping for Rice RVA Properties Using High-Throughput Re-sequenced Chromosome Segment Substitution Lines*, In *Rice Science*, 2013, 407-414.

[6] H. Li, *A statistical framework for SNP calling, mutation discovery association mapping and population genetical parameter estimation from sequencing data*, In *Bioinformatics*, 27, 2011, 2987-2993.

[7] Q. Liu, Y. Tao, S. Cheng, L. Zhou, J. Tian, Z. Xing, G. Liu, H. Wei and H. Zhang, *Relating Amylose and Protein Contents to Eating Quality in 105 Varieties of Japonica Rice*, In *Cereal Chemistry*, 97, 2020, 1303-1312.

[8] H. Sakai, S. S. Lee, T. Tanaka, H. Numa and J. Kim, *Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics*, In *Plant & cell physiology*, 2013.

[9] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*, California: O'Reilly, 2018.

[10] G. Chen and X. Xie, *Exon sequencing mutation detection algorithm based on PCR matching*, In *PLoS One*, 2020.

[11] R. Jensen and Q. Shen, *Rough Set-based Feature Selection: A Review*, In *Rough Computing : Theories, Technologies and Applications*, 2007.

[12] Z. Pawlak and A. Skowron, *Rough sets: Some extensions*, *Information Sciences*, 177, 1, 2007, 28-40.

[13] L. S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. S ´lezak and J. M. Benítez, *Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets"*, In Information Sciences, 287, 68-89, 2014.