Journal of Theoretical and Applied Information Technology
15th December 2024. Vol.102. No. 23
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN APPLICATION OF CLUSTERING TECHNIQUE FOR SELECTING SNP MOLECULAR MARKERS IN RICE GENOME

LUXSANAN PLOYWATTANAWONG [1], SUCHA SMANCHAT [1], SISSADES TONGSIMA[2]

[1] Department of Information Technology, Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
[2] National Biobank of Thailand, National Science and Technology Development Agency,
Pathum Thani, Thailand

E-mail: [1]luxsanun.p@rmutsb.ac.th., [1]sucha.smanchat@itd.kmutnb.ac.th., [2]sissades.ton@nstda.or.th.

ABSTRACT

A selection of Single Nucleotide Polymorphism (SNP) molecular markers whose unique genotypic combinations represent individual rice breeds is a critical consideration in rice breeding programs. Due to the complexity of SNP data with unknown target phenotypes, identifying trait-associated markers presents a significant challenge. Existing research has applied both supervised and unsupervised techniques to genetic and protein datasets; however, little effort has been directed toward SNP analysis. To mitigate the time and difficulties involved in exploring and identifying important markers for biologists, we employ a clustering technique for selecting significant SNPs from the rice genome. The experimental dataset comprises genome-wide SNPs from 88 rice breeds, each containing 50,172 SNPs. We propose an iterative application of the K-means clustering method to cluster these rice breeds into an increasing number of clusters. To identify potentially important SNP markers, the frequency with which each SNP is closest to the centroid of its group is counted. The SNPs are then ranked based on this frequency. The results demonstrate that the proposed method can distinguishes certain SNPs that are more frequently closest to the centroids, potentially indicating their importance as biomarkers. These SNPs among thousands can be recommended for further investigation for biologists in wet experiments.

Keywords: *Rice Genome; K-means Clustering; Molecular Markers; Single Nucleotide Polymorphism; Bioinformatics*

1. INTRODUCTION

Rice DNA and its derived genetic variation profile have the potential to identify rice breeds more precisely. By studying the genomic code of rice DNA, scientists can identify genome differences and diversity among rice breeds and understand their evolutionary relationships. Genomic research, specifically bioinformatics, utilizes these data for the analysis of the genomic code of living organisms allowing scientists to understand genome differences and diversity as well as evolutionary relationships in various aspects of living things [1].

The application of bioinformatics in the improvement of rice varieties has the potential to increase produce yield and improve resistance to harsh environments and diseases [2]. However, this process usually requires time-consuming wet experiments to identify the important molecular markers of plants that govern such traits. Applying machine learning techniques to predict results before commencing wet experiments may reduce budget and time. However, the molecular markers associated with rice varieties are difficult to identify, given a molecular marker dataset without known phenotypes (i.e., the final effect of gene expression).

To address this issue, we propose a technique that selectively distinguish SNP (pronounced as "snip" for single nucleotide polymorphism) molecular markers for the benefit of understanding genetic traits from DNA genotype to phenotypic expression. We employ the K-means clustering technique to select important SNP markers from rice genome from a database of 88 rice cultivars with 50,172 SNPs. It will assist biologists in reducing the number of SNPs that are to be investigated further reducing the time and effort needed for the experiments in wet labs.

The rest of the paper is organized as follows. Section 2 explains the concept and work related to molecular markers in rice species. Section 3 and 4 describe the proposed application of machine learning to identify important markers along with experiment results and limitation. Section 5 concludes this paper with the direction of future work.

2. RELATED WORK

A genetic marker is a unit that controls the genetic characteristics of an organism [3]. A genetic marker is a short strand of DNA that contains a sequence of base pairs. There are several types of DNA markers, namely, Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism (AFLP), Random Amplified Polymorphic DNA (RAPD), Simple Sequence Repeat (SSR) [4] and Single Nucleotide Polymorphism (SNP) [5], which is the focus of this paper.

A SNP is a variation of a single base pair (i.e. one of adenine (A), cytosine (C), guanine (G), and thymine (T)) in a genome that appears in at least 1% of a population (otherwise it is considered as a point mutation) [6]. In a dataset, SNPs may be represented as codes from the paternal and maternal DNA strands as shown in Table 1. For example, “R” represents the SNP with one of the strands having the base pair “A” and the other strand having the base pair “G”. A part of the rice SNP dataset used in this research is shown in Figure 1 (where “NA” represents a missing value).

A SNP could lead to a different phenotype in individuals in a population. Therefore, the identification of SNPs is valuable in the improvement of rice varieties or cultivars and is usually relied upon the expertise of biologists.

Usually, datasets can be analysed to predict the important SNPs given known phenotypes, which are the final effect of gene expression such as produce yield and resistances to diseases. To facilitate the analysis, machine learning techniques have been applied, especially classification approaches. However, for a dataset without known phenotypes, classification and feature selection approaches cannot be effectively utilized. In addition, while the influences of some SNPs may have already been known to biologists, many others are still unknown and may be of value. Therefore, it is beneficial to explore SNP datasets to identify potentially important SNPs whose phenotypes have not been identified.

In a genome that may contains thousands of SNPs, identifying important SNPs is comparable to selecting important features in a dataset. Feature selection is a technique of dimensionality reduction, which aims to select a small subset of relevant features from the original features [7] using three approaches. The supervised approach reduces the redundant and irrelevant features based on feature relevance [8]. The unsupervised approach exploits a target prototype and performs the selection by dividing a dataset into training and testing sets [9]. The semi-supervised approach performs a supervised selection on the labelled data in a dataset to infer the characteristics of the dataset and then applies to unlabelled data evaluation [10].

Exploring unknown SNPs is akin to an unsupervised feature selection. However, without clear target phenotypes, it is hard to define training and testing data. To assist biologists in exploring SNPs, clustering techniques can be valuable. Clustering techniques divide a large set of data into smaller clusters based on their similarity and are useful in analysing data whose details are unknown. Some existing research works have proposed using clustering techniques on biological data such as clustering gene expression [11].

Xu et al. [12]. proposed an unsupervised gene selection using a filter-based evaluation framework to solve the problem of multi-dimensional system in the original dataset, effectively representing the geometric description of the data. Then, the optimal feature subset is obtained from clustering with neural networks and fuzzy ART.

Kim and Gao [13], in their research, extracts a subset of physically meaningful genes based on their ability to create a projection of the sample onto the principal components (PCs) using the Least-Square-Estimation (LSE). Furthermore, they used the boost-expectation-maximization (BEM) clustering to improve the partitioning quality.

An application of K-means clustering is utilized to predict protein-protein interaction [14]. Their approach uses an increasing value of K in iterative clustering until there are no changes in cluster centroids.

Most of the existing research has focused on genetic data with known phenotypes as target classes. Without target phenotypes, unsupervised techniques can be applied. Although unsupervised techniques have been proposed before, the focus has been on clustering gene selection. The application of unsupervised techniques on SNP selection has received little attention. Unlike existing work, we need to explore a SNP marker dataset of rice genome without any known phenotypes. To achieve this, we adapt an iterative approach of the K-means clustering [14] to identify potentially important SNP markers.

3. RESEARCH METHOD

The genome database of Thailand's indigenous rice cultivars under the supervision of National Biobank of Thailand (NBT), National Science and Technology Development Agency, was used in this research. The database contains 50,172 SNPs of 88 rice cultivars. Each SNP is a character code as explained earlier. The database contains a lot of missing and invalid values denoted by “N” and data cleansing is required under two removal conditions. The markers (columns) that contain the value of “N” in more than 20% of the rice cultivars are to be removed. The rice cultivars (rows) that contain the value of “N” in more than 20% of the markers are also to be removed. With these removal conditions, no cultivar was removed, and there were 30,901 SNPs remaining after cleansing. These SNPs were then encoded into integers for K-means clustering [15].

Because the names of the SNPs in the database are just codes, it is not possible to effectively select markers based on their linguistic meaning. In order to identify potential markers from a database containing thousands of unknown markers, K-means clustering, an unsupervised machine learning technique, is employed [16 - 18].

K-means clustering is a very popular partition clustering by dividing data into K groups [19, 20], where the value of K is set by the user. The K-means algorithm first selects K random groups, starting with the centre of each group. Then each data point is added to the group with the least distance to itself, i.e., the most similar group to itself. The new centre, called centroid, of the group is then recalculated using the average of the data in that group. The process is repeated until all data points are grouped [21].

The K-means is iteratively applied to cluster the rice cultivars based on the 30,901 SNPs remaining from the data cleansing to identify the SNPs that are most important in each clustering. SNPs that are closest to the centroids of their respective clusters repeatedly should potentially be of value for further study as they may represent certain traits in the phenotypes in rice cultivars.

The iteration of K-means clustering is applied to rice cultivars with increasing numbers of K target clusters. Our initial experiment sets the K target clusters from 2 to 20 clusters, totalling 19 rounds. In the first round, the value of K is set to 2 (i.e. two clusters). Once the first clustering round finishes, the SNPs that are closest to the centroids of their respective clusters are identified; there could be more than one SNP being closest to the centroid of a cluster. The K value is then increased to 3 and the process is repeated until the K value reaches 20.

After 19 rounds, the frequency of each SNP being closest to the centroid of its group is determined. The SNPs are then sorted based on their frequencies. The whole process is depicted in Figure 2.

4. RESULTS AND DISCUSSION

After performing the iterative clustering, the frequencies of 30,901 SNPs being closest to their centroids are determined and sorted. We report the number of SNPs being closest to their centroid in Table 2. The majority of the SNPs are closest to their centroids approximately 7 – 13 times from the total 19 clustering rounds. Less SNPs are closest to their centroids at higher frequencies (e.g. from 14 to 17 times) and at lower frequencies (e.g. from 2 to 6 times). There is no SNP being closest to its centroid in every clustering round (19 out of 19 rounds). Also, there is no SNP being closest to its centroid 18 of out 19 rounds.

It can be seen that it is possible to use this method to distinguish certain SNPs that are closest to their centroid most often, potentially representing certain traits in rice cultivars. Five SNPs are closest to their centroids 17 times out of 19 rounds, namely, rs17921738, rs19665236, rs54178375, TBGI258369, and rs20124503. These five SNPs could potentially be important in rice cultivars for further investigation in web experiments as they are closest to the centroid with the highest frequency. The additional 106 SNPs that are closest to their centroids 16 times may also be suggested to scientists as necessary. The frequencies may also be used to prioritize the SNPs needing the attention of scientists.

In the experiment, the number of clusters was set to 2 to 20 clusters. However, a limit to the increment of the number of clusters needs to be defined. In an attempt to identify this iteration limit, the numbers of cultivar members in the clusters in every iteration are reported in Table 3 sorted in a non-increasing order.

The minimum size of a cluster in K-means clustering is 1, meaning that the initial random member is alone in the cluster. Therefore, the iterative application of K-means clustering should stop when there is at least one cluster with the minimum size of 1. According to Table 3, the first time a cluster having the size of 1 occurs when K is set to 12. We use this as the iteration limit and revise the process as shown in Figure 3. The result of the iterative clustering up to 11 rounds is reported in Table 4.

From Table 4, the iterative clustering result also reveals a similar frequency distribution of SNPs being closest to their centroids. Particularly, only three SNPs are most frequently closest to their centroids: vcZ25GIOP, vcZ2IRUB9, and S12_14772751. However, these SNPs are not among the best rank when reporting the result up to the K value of 20. This suggests that excessive iterations may lead to inconsistent results. Additionally, there are 111 SNPs closest to their centroids in the second rank, as opposed to 106 in Table 3. This indicates that not only should the SNPs in the best rank be recommended, but also those in the second rank as well.

However, the limitation of this work lies in the inherent lack of known phenotypes. Without target phenotypes, the effectiveness and the efficiency of the proposed iterative clustering technique could not be fully evaluated. In addition, the missing and the invalid values that were removed from the dataset may influence the outcome of the iterative clustering.

In the future, when the result of the phenotypic study becomes available, it will be possible to adjust our technique further to reflect the proper selection of SNPs. This includes the refinement of the condition for the iteration limit of K-means clustering and the effect of missing data values. It is also possible to employ and evaluate other clustering techniques for SNP selection.

5. CONCLUSION

Due to the difficulty of identifying and selecting potential Single Nucleotide Polymorphism or SNP markers that may influence the phenotypes of rice cultivars, scientists usually need to spend valuable time in wet labs studying them. With unknown target phenotypes of SNPs, classification techniques could not be effectively applied. This paper aims to alleviate this problem by applying K-means clustering iteratively with increasing number of clusters to identify the SNPs that are closest to the centroid of their clusters most frequently. These SNPs may potentially represent certain traits in the phenotypes in rice cultivars. The experiment shows that our application of iterative K-means clustering can distinguish certain SNPs based on the frequencies of being closest to their centroids. The frequencies of being closest to the centroid can also be used to rank SNPs as necessary.

Using this technique, scientists can identify potentially important SNPs among thousands to prioritize the selection of SNPs for further investigation in the wet lab to minimize their valuable effort and budget.

Table 1: SNP data code

Genotype	AA	CC	GG	TT	AG	CT	CG	AT	GT	AC
Code	A	C	G	T	R	Y	S	W	K	M

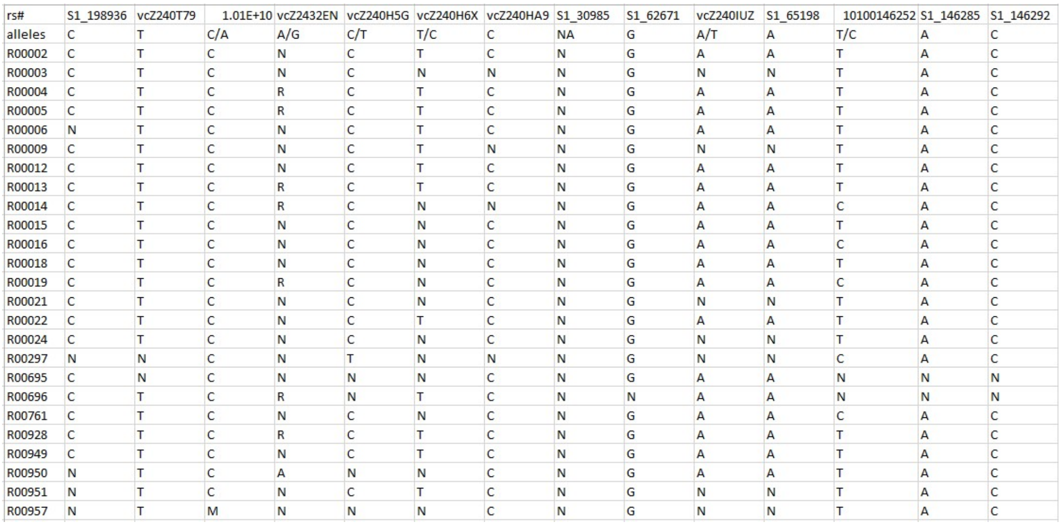
Table 2: Number of SNPs being closest to their centroids sorted by frequency

Number of SNPs closest to their centroids	Frequency of SNPs being closest to their centroids out of 19
0	19 / 19
0	18 / 19
5	17 / 19
106	16 / 19
197	15 / 19
553	14 / 19
1926	13 / 19
4245	12 / 19
5350	11 / 19
5200	10 / 19
4473	9 / 19
4053	8 / 19
2922	7 / 19
1029	6 / 19
374	5 / 19
371	4 / 19
93	3 / 19
4	2 / 19
0	1 / 19

Table 3: The number of cultivar members in each cluster

k	Number of cultivar members																			
2	61	27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	35	28	25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	34	23	23	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	24	21	20	15	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	24	21	20	12	8	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	20	20	16	13	9	8	2	-	-	-	-	-	-	-	-	-	-	-	-	-
8	20	17	14	13	9	8	5	2	-	-	-	-	-	-	-	-	-	-	-	-
9	22	14	14	11	8	8	6	3	2	-	-	-	-	-	-	-	-	-	-	-
10	13	12	11	10	10	10	8	7	4	3	-	-	-	-	-	-	-	-	-	-
11	14	13	12	9	8	8	8	5	5	4	2	-	-	-	-	-	-	-	-	-
12	14	14	12	12	8	7	6	5	4	3	2	1	-	-	-	-	-	-	-	-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



rs#	S1_198936	vcZ240T79	1.01E+10	vcZ2432EN	vcZ240H5G	vcZ240H6X	vcZ240HA9	S1_30985	S1_62671	vcZ240IUZ	S1_65198	10100146252	S1_146285	S1_146292
alleles	C	T	C/A	A/G	C/T	T/C	C	NA	G	A/T	A	T/C	A	C
R00002	C	T	C	N	C	T	C	N	G	A	A	T	A	C
R00003	C	T	C	N	C	N	N	N	G	N	N	T	A	C
R00004	C	T	C	R	C	T	C	N	G	A	A	T	A	C
R00005	C	T	C	R	C	T	C	N	G	A	A	T	A	C
R00006	N	T	C	N	C	T	C	N	G	A	A	T	A	C
R00009	C	T	C	N	C	T	N	N	G	N	N	T	A	C
R00012	C	T	C	N	C	T	C	N	G	A	A	T	A	C
R00013	C	T	C	R	C	T	C	N	G	A	A	T	A	C
R00014	C	T	C	R	C	N	N	N	G	A	A	C	A	C
R00015	C	T	C	N	C	N	C	N	G	A	A	T	A	C
R00016	C	T	C	N	C	N	C	N	G	A	A	C	A	C
R00018	C	T	C	N	C	N	C	N	G	A	A	T	A	C
R00019	C	T	C	R	C	N	C	N	G	A	A	C	A	C
R00021	C	T	C	N	C	N	C	N	G	N	N	T	A	C
R00022	C	T	C	N	C	T	C	N	G	A	A	T	A	C
R00024	C	T	C	N	C	N	C	N	G	N	N	T	A	C
R00297	N	N	C	N	T	N	N	N	G	N	N	C	A	C
R00695	C	N	C	N	N	N	C	N	G	A	A	N	N	N
R00696	C	T	C	R	N	T	C	N	N	A	A	N	N	N
R00761	C	T	C	N	C	N	C	N	G	A	A	C	A	C
R00928	C	T	C	R	C	T	C	N	G	A	A	T	A	C
R00949	C	T	C	N	C	T	C	N	G	A	A	T	A	C
R00950	N	T	C	A	N	N	C	N	G	A	A	T	A	C
R00951	N	T	C	N	C	T	C	N	G	N	N	T	A	C
R00957	N	T	M	N	N	N	C	N	G	N	N	T	A	C

Figure 1: A part of rice SNP dataset

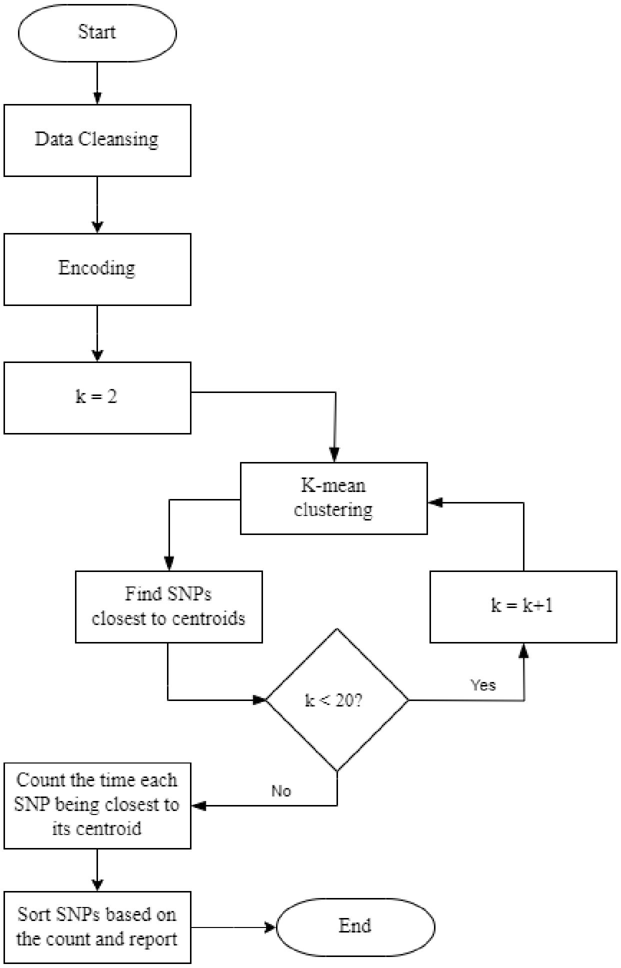

Start
Data Cleansing
Encoding
k = 2
K-mean clustering
Find SNPs closest to centroids
k = k+1
k < 20?
Yes
No
Count the time each SNP being closest to its centroid
Sort SNPs based on the count and report
End

Figure 2: Flowchart of the initial clustering process

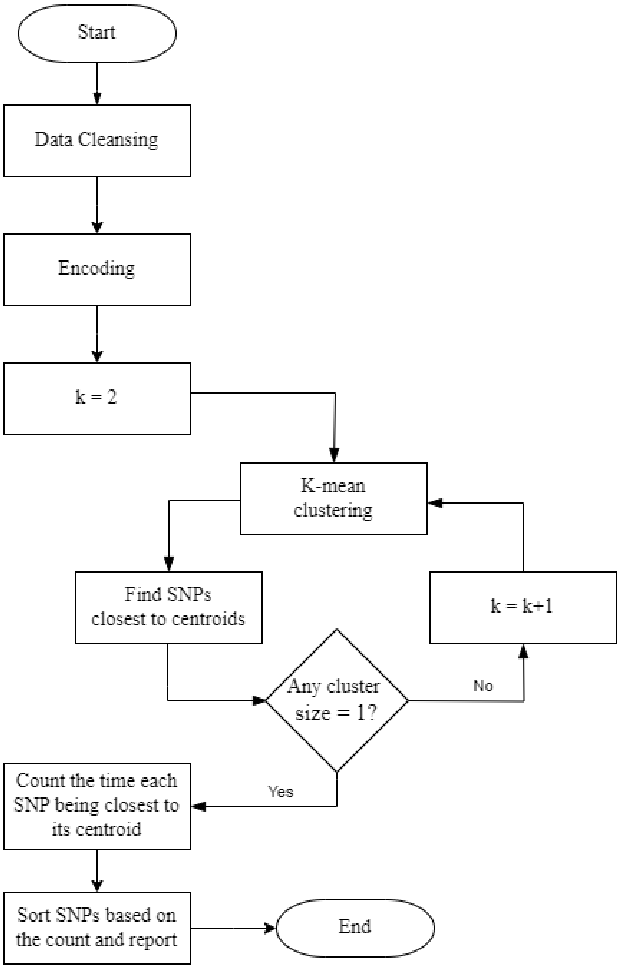

Start
Data Cleansing
Encoding
k = 2
K-mean clustering
Find SNPs closest to centroids
k = k+1
Any cluster size = 1?
No
Yes
Count the time each SNP being closest to its centroid
Sort SNPs based on the count and report
End

Figure 3: Flowchart of the iterative clustering process

Table 4: Number of SNPs being closest to their centroids up to K=12

Number of SNPs closest to their centroids	Frequency of SNPs being closest to their centroids out of 11
0	12 / 12
3	11 / 12
111	10 / 12
1129	9 / 12
3536	8 / 12
8259	7 / 12
8272	6 / 12
6601	5 / 12
2081	4 / 12
713	3 / 12
179	2 / 12
17	1 / 12

ACKNOWLEDGEMENTS

The authors would like to express sincere gratitude to the National Biobank of Thailand (NBT), National Science and Technology Development Agency for providing the SNP molecular marker database for our experiment.

REFERENCES:

[1] Sigala M, Beer A, Hodgson L, O'Connor A. Big data for measuring the impact of tourism economic development programmes: A process and quality criteria framework for using big data. in: Big Data and Innovation in Tourism, Travel, and Hospitality. Singapore: *Springer Singapore*; 2019. p. 57–73.

[2] Nguyen G, Dlugolinsky S, Bobák M, Tran V, López García Á, Heredia I, et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review* [Internet]. 2019; 52(1):77–124. Available from: http://dx.doi.org/10.1007/s10462-018-09679-z

[3] Paterson AH, Tanksley SD, Sorrells ME. DNA Markers in Plant Improvement. In: Advances in Agronomy. Elsevier; 1991. p. 39–90.

[4] Devos KM, Gale MD. The use of random amplified polymorphic DNA markers in wheat. *Theoretical and Applied Genetics* [Internet]. 1992; 84(5–6):567–72. Available from: http://dx.doi.org/10.1007/BF00224153

[5] Shastry BS. SNPs: impact on gene function and phenotype. *Single Nucleotide Polymorphisms* [Internet]. 2009;578:3–22. Available from: http://dx.doi.org/10.1007/978-1-60327-411-1_1

[6] Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion Plant Biology* [Internet]. 2002;5(2): 94–100. Available from: http://dx.doi.org/10.1016/s1369-5266(02)00240-6.

[7] Chandrashekar G, Sahin F. A survey on feature selection methods. *Computers & electrical engineering* [Internet]. 2014; 40(1):16–28. Available from: http://dx.doi.org/10.1016/j.compeleceng.2013.11.024

[8] El-Mageed A, Abohany AA, Elashry AA. Effective feature selection strategy for supervised classification based on an improved binary Aquila optimization algorithm. *Computers & Industrial Engineering*. 2023; 181.

[9] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing* [Internet]. 2018; 300:70–9. Available from: http://dx.doi.org/10.1016/j.neucom.2017.11.077.

[10] Li Z, Tang J. Semi-supervised local feature selection for data classification. *Science China Information Sciences* [Internet]. 2021; 64(9). Available from: http://dx.doi.org/10.1007/s11432-020-3063-0.

[11] Arima C, Hanai T, Okamoto M. *Genome Informatics* [Internet]. 2003; 14:334–5. Available from: http://dx.doi.org/10.11234/gi1990.14.334

[12] Xu R, Damelin S, Nadler B, Wunsch DC. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps," *Artificial intelligence in medicine*. 2010;48:91–8.

[13] Kim Y, Gao J. Unsupervised gene selection for high dimensional data. In: Sixth IEEE Symposium on BioInformatics and BioEngineering (BIBE'06). *IEEE*; 2006.

[14] Sun P, Ma Y, Wei Y, Ma Z, Lu L, Cui Y, et al. Application of improved K-mean clustering in predicting protein-protein interactions. In: 2008 International Conference on BioMedical Engineering and Informatics. *IEEE*; 2008. p. 83–6.

[15] Ridzuan F, Wan Zainon WMN. A review on data cleansing methods for big data. *Procedia Computer Science* [Internet]. 2019;161:731–8. Available from: http://dx.doi.org/10.1016/j.procs.2019.11.177

[16] Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*. 2013;1(6):90–5.

[17] Modha DS, Spangler WS. Feature Weighting in k-Means Clustering. *Machine learning*. 2003;52:217–37.

[18] Coates A, Ng AY. Learning feature representations with K-means. In: Lecture Notes in Computer Science. Berlin, Heidelberg: *Springer Berlin Heidelberg*; 2012. p. 561–80.

[19] Likas A, Vlassis N, J. Verbeek J. The global k-means clustering algorithm. *Pattern recognition* [Internet]. 2003; 36(2):451–61. Available from: http://dx.doi.org/10.1016/s0031-3203(02)00060-2

[20] Sinaga KP, Yang M-S. Unsupervised K-means clustering algorithm. *IEEE* [Internet]. 2020; 8:80716–27. Available from: http://dx.doi.org/10.1109/access.2020.2988796.

[21] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* [Internet]. 2023;622:178-210. Available from: http://dx.doi.org/10.1016/j.ins.2022.11.139.