ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Rice Genomes Classification based on Efficient Distance Measures Classifiers

S S Patil*, Kiran S K
University of Agricultural Sciences, Bangalore, India

Abstract

The structure and composition of genomes is swiftly systematic in pace with their sequencing. The promising data show that a significant portion of Rice genomes is composed of transposable elements. Given the profusion and diversity of TEs and the hustle at which large quantities of sequence data are rising, detection and annotation of TEs presents a significant confront. Here we propose integrated classification system, designed on the basis of the transposition mechanism; sequence similarities and structural relationships can be easily applied by amateur. We used machine learning technique based on different classifier algorithms with four distance measures.

Keywords: Machine learning technique, Euclidean algorithm, Hamming algorithm, Mahalanobis algorithm, Minimum algorithm, classification accuracy.

Introduction

Classification of genome sequence data for identifying the homology and properties of genome sequences is a difficult task is due to large data. Rice genome one of the important gene which is staple food across the world, genome having smallest genome size among all cereal crops. Recently, DNA sequencing is a important to determine exact order of nucleotide in a DNA molecule. Genome sequence dataset available in public domain (databanks) and the classification large datasets is a challenging task in data mining. [13] Uncovered the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. [3] Rice, one of the most important food crops for humans, is the first crop plant to have its genome sequenced. Rice whole-genome microarrays, genome tiling arrays and genome-wide gene-indexed mutant collections have recently been generated. With the availability of these resources, discovering the function of the estimated 41,000 rice genes is now within reach. Such discoveries have broad practical implications for understanding the biological processes of rice and other economically important grasses such as cereals and bioenergy crops.

Sequence classifying is very important to study genome data [5], [7],[8], we propose schemes using motif based clustering to analyze grass genome sequences data [1], [2]. Classification is performed using similarity or distance measures [7]. The evaluation of classifiers is assessed on CA: Based on training and test data sequences can be classified using sequence similarity to the known class/family of the sequences [4]. This helps in predicting the structure and function of unknown sequences to save the expenses on the biological experiments. Classification of multiple sequences is important to determine the molecular evolution of the gene family and to understand their current status of evolution in biological systems. The distance measure requires neither homologous sequence nor prior sequence an alignment is used to search for similar sequence from a database [15] Comparing the whole length of sequences with each other using distance measure is very difficult. [6] A sub string of sequence motifs can be used to generate genome profiles. Evolution wise, the Rice species is fascinating due to the degree of variation found in genome size; ploidy level and chromosome number [11], [14]. Most of the comparative genomic studies have been initiated in barley, wheat, maize, rice, and sorghum to understand the diversity and structure function relationship of the genomes [12], [11]. This paper reports on the distance measure classifier algorithm based on machine learning technique, we identified the time complexity and classification accuracy of the classifiers.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Materials and methods

The genome sequences dataset available in a public domain like NCBI database. The rice genome sequence extracted through NCBI database (www.ncbi.nlm.nic), sequence compiled in a FASTA format. This FASTA format file gives a clear description of the particular sequence, which begins with the '>' symbol. The format is helpful to find similarity between the sequence and it also adopted for searching through a pattern database instead of sequence database. The 5600 sequences are collected as whole sequence dataset (OSD1) among these 1400 sequences are selected as benchmark sequence dataset (OSD2). We made comparison between the sequence data set and based on machine learning technique divided the dataset in to two third of training data and one third of testing data. This data set will achieve time complexity and classification accuracy with the four distance measure classifier (algorithms) such as Euclidean algorithm, Hamming algorithm, Mahalanobis algorithm and Minimum algorithm. Among these distance classifiers, the minimum distance classifier is achieved time complexity least which as noticed classification accuracy highest

A scalar function, d(x, y), of the ordered pair of vectors x, y, is a distance function it satisfies the following axioms of a distance measure

Algorithm	Data set	Total	Threshold	Training Data	Test Data	Training Time	Test Time	Class	CA in %
Euclidean	OSD1	5600	0.6	3733	1867	20	218	13	88.56
	OSD2	1400	0.6	933	467	12	157	14	89.65
Hamming	OSD1	5600	0.6	3733	1867	18	30	14	90.12
	OSD2	1400	0.6	933	467	11	20	13	91.26
Minimum	OSD1	5600	0.6	3733	1867	17	11	14	92.58
	OSD2	1400	0.6	933	467	11	9	13	92.65
Mahalanobis	OSD1	5600	0.6	3733	1867	28	157	14	92.15
	OSD2	1400	0.6	933	467	22	117	14	92.25

d(x, y) ≥ 0
d(x, y) = 0 if x = y,
d(x, y) =d(y, x) and
d(x, y) ≤ d(x, z) + d(z, y) for any z.

The following distance measure or dissimilarity measure will be used to analyze the sequences (dataset) by using step by step procedure is called Algorithm.

Euclidean distance measure

The most common distance function is the Euclidean distance. It is the square root of the sum of the squared differences between all the dimensions of two elements.

$$d_{ij}=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ik}-x_{jk})^2}$$

This is useful in several applications where the input data consists of an incomplete set of distance and the output is a set of points in Euclidean space that realizes the given distance

Hamming distance measure

The Hamming distance has mostly seen in the categorical or nominal data. The Hamming distance measures the minimum number of substitutions are required to change one string into the other

$$D_H = \sum_{i=1}^{K} |x_i - y_i|$$
$$\text{where } x = y \Rightarrow D = 0,$$
$$x \neq y \Rightarrow D = 1$$

Hamming distance is mainly applicable for detecting the coding theory and biologically helpful to measure of genetic distance.

e.g.; "A**A**CT**G** and A**T**CA**G**" is 2.

Mahalanobis distance

The Mahalanobis distance is a unit less measure. It is a descriptive statistic that provides a relative measure of a data points distance (residual) from a common point.

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}.$$

The distance is applicable for Hierarchical classification; this distance tends to be a more robust to noisy data and predicting the protein or DNA structural class.

$$d_{ij}=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ik}-x_{jk})^2}$$

Minimum distance measure

The arithmetic mean of the maximum and minimum distances is (satellite or secondary star) from a primary distance. This is explained by following formula,

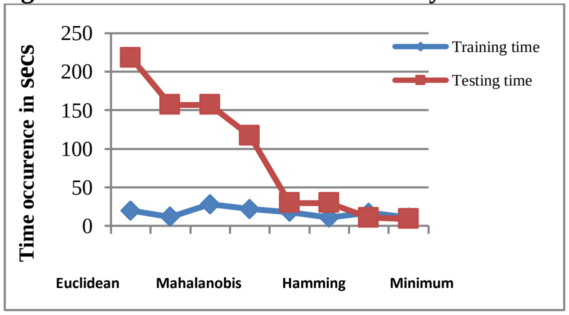
$$d_{ij} = \sqrt{\sum_{j=1}^{P}(x_{ij} - \bar{x}_{ij})^2}$$

The distance is a statistical method for fitting a model to data, the minimum length of a path between two points in a graph

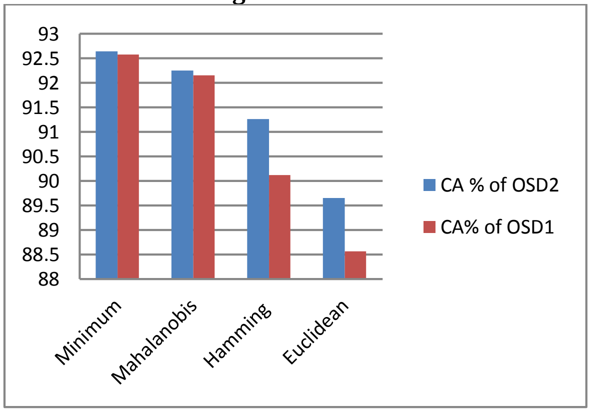
Results and discussion

The performance of the algorithms on set of 5600 Rice genome sequences dataset whose distinct classes were known. In the present investigation, genome sequence dataset carried out the four algorithm such as (1)Euclidean distance measure, (2)Hamming distance measure, (3) Minimum distance measure and (4)Mahalanobis distance measure. Similarity measure carried out based on above mentioned algorithms. Out of 5600 sequence dataset, arbitrarily 3733 sequences were as training data and 1867 as testing data. In the bench mark 933 sequences were as training and 467 sequences as testing. The bench mark dataset out performed better than whole sequence dataset in our context.

Comparison of various Distance measure algorithms with classification accuracy



Comparison of Training and Testing time in Distance measure algorithms



Classification accuracy (CA%) of Distance or Dissimilarity measure

Classification by pattern of similarity is an interesting and challenging problem. The computational time complexity was high to classify the genome sequences in tested data as trained data. Hamming distance algorithm performed well, compared to Euclidean distance algorithm and Minimum distance algorithm performed well, compared to Mahalanobis distance algorithm.

The study reported 1)EDC 2)MDC 3)HDC 4)MIDC Algorithms were used to perform classification of genome sequence data set, it is apparent from the table that the classification accuracy was almost same using the Euclidean with Hamming distance and Mahalanobis with Minimum distance. The employed scheme of classifier performance in the study showed that the algorithm derived from the pattern based motifs was good. The time complexity achieved was the best in MIDC and HDC. Classification accuracy achieved was the best in MIDC followed by MDC.

Conclusion

The study concluded, Classification of Rice genome data set of various distance similarity/ dissimilarity measures administered to achieve the improved performance of accuracy and tested results and found statistically significant. Among these four distance classifier, the Minimum distance algorithms obtained least time complexity and achieved best classification accuracy.

References

[1] Alison Abbot., "Bioinformatics institute plans public database for gene expression data", NATURE, 398:638-646, 1999.

[2] Gaut, B.S. "Evolutionary dynamics of grass genomes", NEW PHYTOL, vol. 154, 15–28, 2002.

[3] Jong-Seong Jeon, Ki-Hong Jung, Hyun-Bi Kim, Jung-Pil Suh, and Gurudev S Kush, "Genetic and Molecular Insights into the Enhancement of Rice Yield Potential", JOURNAL OF PLANT BIOLOGY, 54:1-9, JANUARY 2011.

[4] M. Fischer and M. Paterson.,"String-matching and other products". Proc. SIAM-AMS COMPLEXITY OF COMPUTATION, 113-125, 1974.

[5] Patil, S.S., Dhandra, B.V., and Angadi, U.B., "Efficient Scheme for Classifying Grass Genomes", Proceedings of the WORLD CONGRESS ON ENGINEERING AND COMPUTER SCIENCE, SAN FRANCISCO, USA, WCECS, **30**(1):20-22, 2009.

[6] Patil, S.S., and Angadi, U.B., 2014, "Motif based Clustering Techniques for

classification of Grass Genome Sequences", INTERNATIONAL JOURNAL OF ELECTRONICS COMMUNICATION AND COMPUTER ENGINEERING, 5(1):2278-2289.

[7] P.A. Vijaya, M.N. Murty and D.K. Subramanian. "An efficient incremental Clustering Algorithms for large data set," Proc. ICAAI, Kolhapur, India, 2005, pp79-85.

[8] R. Luo, Z. Feng, J. Liu, "Prediction of protein structural class by amino acid and polypeptide composition", EUROPEAN JOURNAL OF BIOCHEMISTRY, 269: 4219–4225, 2002.

[9] Saul B Needlemen and Christian D Wunch., "A General method Applicable to the search for similarities in the Amino acid sequence of Two proteins", J. MOL. BIOL. 48:443-453,1970,.

[10] Smith, and Waterman, "Identification of common molecular sub sequences", JOURNAL OF MOLECULAR BIOLOGY, 147:195-197, 1981.

[11] Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J., "Duplication and DNA segmental loss in the rice genome: implications for diploidization", NEW PHYTOL. 165: 937–946, 2005.

[12] Wei, F., et al. "Physical and genetic structure of the maize genome reflects its complex evolutionary history," PLOS GENET. Vol.3, 2007, p.123.

[13] Xuehui Huang, Xinghua Wei, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, Meng Li, Danlin Fan, Yunli Guo, Ahong Wang, Lu Wang, Liuwei Deng, Wenjun Li, Yiqi Lu, Qijun Weng, Kunyan Liu, Tao Huang, Taoying Zhou, Yufeng Jing, Wei Li, Zhang Lin, Edward S Buckler, Qian Qian, Qi-Fa Zhang, Jiayang Li & Bin Han ," Genome-wide association studies of 14 agronomic traits in rice landraces", NATURE GENETICS,42,961–967, 2010

[14] Yu, J., et al. "The genomes of Oryza sativa: A history of duplications", PLOS BIOL, 3: 266, 2005,

[15] Yusen Zhang and Wei Chen "A new Measure for similarity searching in DNA sequences", MATCH COMMUN. MATH. COMPUT. CHEM. 65:477-488 ,2011