Identification of Mutations in Long Non-Coding RNA Sequence of Covid-19 using Machine Learning Approach

Revathi Annem [1], Jyothi Singaraju [2]

[1]Researchscholar, Department of Computer Science, SPMVV, Tirupati
arevathi20@gmail.com
[2]Professor, Department of Computer Science, SPMVV, Tirupati.
jyothi.spmvv@gmail.com

Abstract:

The transcripts which do not produce proteins and longer than 200 nucleotides are known to be long non- coding RNAs (lncRNAs). Mutations associated with disease is often studied to know about disease and its prevention. The mutations also help to diagnosis the diseases and develop new drug for the treatment of the diseases. Various computational methods have been developed to study about the lncRNAs functions and mutations associated with diseases but still, it is an unknown task. Machine Learning is one of the methods, which used to study about errors in the RNA sequence. As a lncRNA is novel class of RNAs the mutations of it are not yet studied. The mutations in lncRNA sequence play an important role in the disease development, so which can also be used as a strong biomarker of the diseases. Previous studies identified the mutations using high throughput DNA sequencing technologies. This proposed method focused on the mutation identification in Covid-19 long non-coding RNA sequence using Machine Learning Approach. The proposed system is a novel Machine Learning method for identifying the possible point mutations in the long non-coding RNA sequence. The results shown that this novel method has high accuracy in identifying point mutations.

Keywords: lncRNAs, long non coding RNAs, Covid-19, Machine Learning, DNA, RNA, non-coding RNA, Mutations, Point Mutations

1. INTRODUCTION

It has been a global pandemic due to the novel virus which is known as coronavirus or Covid-19. The first whole genomic sequence coronavirus is placed in the GenBank of NCBI, which was found in the China laboratory [1]. It is identified that the Covid-19 virus is transmitted from one human to other human through direct contact or droplets [2,3]. The mutation of this virus has an effect in the human body, it is a RNA virus [1].

The DNA (deoxyribonucleic acid) is a hereditary component in humans. Every cell in the body has same DNA. It consists of four bases of chemical, adenine(A), Thymine(T), guanine(G), cytosine(C). Human DNA has 3 billion bases, almost 99 % people has same bases. The sequence of these bases helps in maintaining and building of an organism. The DNA molecule is a double helix consist of two strands which are twisted around each other. RNA is a single strand, which plays an important role in producing proteins required by the human body. DNA and RNA are important components of viruses.

Previous research about Covid-19 is on the analysis of coding RNAs and shown that only 2% of human genome has coding regions, the remaining is junk non-coding regions. The non-coding is the portion where proteins are not produced. When compared to non-coding regions, the coding portion is very less in the human genome.

Studies shown that non-coding RNAs are more related to human diseases. The long non-coding RNAs (lncRNAs) are the one of the groups of non-coding RNAs which is grater then 200 nucleotides in length. Recently research shown that mutations in non-coding RNAs alter their function which leads to cause of disease. The newly identified long non-coding RNAs mutations more associated with disease.

1.1 About lncRNA mutations

The changes in the DNA/RNA sequence of nucleotides are known as mutations.The mutations in the sequence can change function of protein in the body, which leads to cause of disease. The viruses related to RNA are different when compare to DNA related viruses i.e. they mutation rates are higher[4]. The evolution of mutation continuously leads to loos immunity and become even more harmful[4]. Point mutation is one of the RNA mutations which effect only one or some nucleotides in the RNA sequence[4].Recently the scientists have proved that changes in the regions of RNA that do not produce proteins can also lead to many diseases. previous studies shown that the gene activity is controlled by long non coding RNA , i.e. it decides where and when the gene is to be turn on and turn off. The changes or mutations of long non coding RNA can turn on the gene and produce the protein in the wrong place at the wrong time and also the mutations can eliminate or reduce the important protein production, when it is required to the human body. The normal development or cause of a disease is due to disturb of actual protein production.

ln many biological and cellular process, the lncRNAs are involved. The studies shown that the cause of disease is due to dysfunction, misregulation or mutation of lncRNA [9, 10] [b]. LncRNA disease database contains the information of association of disease with lncRNAs [11].So identification of the mutations related to lncRNA helps to diagnosis several disease, by which disease prevention or new drug can be developed for the treatment.So, all the mutations in the lncRNA have impact on health.

1.2 Types of mutations

A change in the RNA sequence is known as a mutation. Mutations may be in small or large. Small mutations are the changes in the single nucleotide or more nucleotides, whereas large mutations are changes that effect the genes on a chromosome. The virus mutation at micro-level which infects the immune system.

1.2.1 Mutations are of different types: point mutation, insertion mutation, deletion mutation and substitution mutation and frame shift mutation [13]. Point mutation changes only one nucleotide. Substitution mutation one or more nucleotides or substituted or replaced with same nucleotide. When one new nucleotide is inserted in the sequence, then it is known as insertion mutation and when any nucleotide is deleted from the sequence then it is called as deletion mutation.

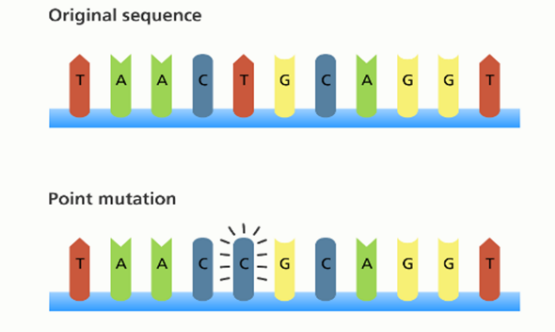

Fig 1: Representation of original sequence and point mutation sequence

Again, the point mutation is divided into silent, nonsense and missense. Silent mutation is the one which changes codon codes of the amnio acid. The nonsense mutation is one which change the stop codon. Missense Mutation is new Condon codes after mutation for different amino acid

.

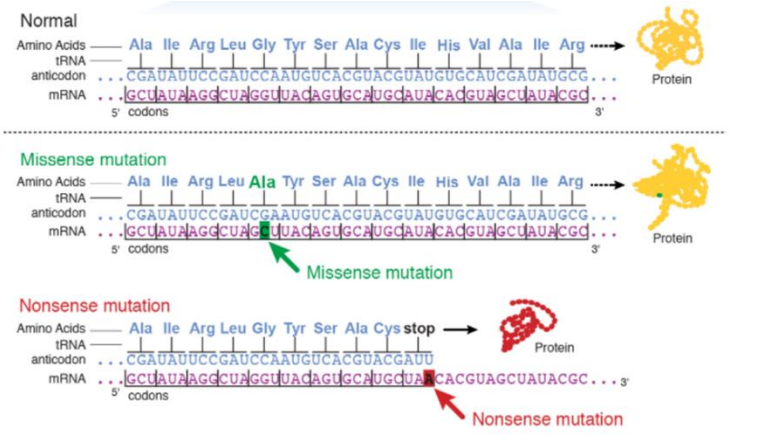

Fig 2: Representation of different types of point Mutations

The paper is organized as follows: Section 2 is background of Covid-19, DNA and RNA mutations. The proposed algorithm to find point mutations in the Covid-19 lncRNA sequence and its classification using Machine learning algorithm is in the section 3. Section 4 provides results and performance evaluation of the algorithm. Then section 5 concludes with future work.

2. BACK GROUND

In past the researchers had found 12,706 mutations in Covid-19 genome, the major changes is due to singe nucleotide and also discovered that from sequencing data that there is two single nucleotide mutations in a month in Covid-19 Genome [5]. It is also identified that this virus carries the protein name spike mutation which is very major in the world [5].

Researchers are observed the genetic mutations in coronavirus genome to detect novel mutations, for understanding the biological potential changes. The rate of mutation in RNA viruses is faster than their hosts [6,7]. Many mutations will not affect the transmission of viral, because they will not change the protein structure [8,9]. The collection of mutations is marker for the viral strength.

The studies related to the mutations are helpful in developing vaccines and drugs related to an antiviral [5].

Mutation analysis is very important, so one of the advanced computational approach machine learning is used for it. Some of the machine learning methods such as support vector machine, decision trees and neural networks are used for analysing mutations based on protein behaviour [6] and is used to find the secondary structure formation based on sequence primary structure [7-9]. Machine learning techniques are used to predict the mutations that affect the behaviour of protein and also used to help to know the new mutations which influence the certain drugs with no effect [6]. Markov chain is machine learning tool developed to find the relative rate of changes in the different nucleotides and also used to search the nucleotide replacements in the sequence of RNA.

The virus mutation in the genome plays a very important role in disease development and also reduce the life of the vaccine [30]. Pervious research identified the mutation in the protein structure and analysed many biochemical properties related to protein to know the mutations evolution.

The life of the vaccine can be determined by the mutations [30].The vaccine is not effective when a mutations are very high in rate [30].Previously they studied point mutation and its properties related to biochemical using bioinformatic tools[30].In depth study of origin of mutations of protein is analysed to understand the changes of virus spread[30].Covid-19 mutations in the protein are analysed to know its pattern growth and also helps to have to further study of proteins[30].Mutation analysis is going to help in both non-structural and structural proteins of the virus and it is very important to determine the origin of the virus. The recent studies on proteins and sequences identified that S-protein and N-protein are helpful for drug design [30]. So, analysing the mutations in the proteins and sequences are very important to fight against Covid-19[30]. The infection is mild and severe in human due to Covid-19 [30].

Attributes of protein is parameter which specifies the important information of sequence of protein [31]. To replace composition of amino acid to reduce losing order of sequence information, the composition of pseudo amino acid is found [2001 32,33]. It is used extensively to know protein related attributes, which helped to identify virulent bacterial protein [34], Super secondary structure prediction [35], protein location of subcellular [35-38] types protein membrane prediction [39], finding proteins allergenic [40], protein structural class prediction [41], amnio acids classification [42] and so on

2.1 Mutations in the lncRNAs related to many diseases

lncRNAs are present in the complete life cycle of cell and plays very important roles in the biological processes[10]. The research evidences shows that the dysregulations and the mutations of lncRNA involved in various human complex diseases development.Analyzing the disease association with lncRNA is become very important task in the bioinformatics [10]. Understanding the disease mechanism at level of lncRNAS helps to identify disease biomarker, diagnosis of disease, prevention, treatment and prognosis [10]. lncRNAs are playing very important roles in many complex diseases of humans. The dysregulation and mutations of lncRNAs are related to various diseases of human [11,12,13,14,15-17], such as cancer related to thyroid [18], lung [19,20], breast [21,22], bladder [23], ovarian [24] and colon [24]. The other diseases such as

diabetes [25, 26], AIDS [27] and AD [28] like so many more associated with many lncRNAS. Recently many lncRNA databases have been developed such as LncRNADisease (http://www.cuilab.cn/lncrnadisease) [29]. Many models related to machine learning are developed to predict the novel lncRNAs associations with disease. All those models have its own advantages and disadvantages.

3. PROPOSED METHODOLOGY

It is the new method, which uses lncRNA sequence to find the mutations related to Covid-19 diseases. The mutations of RNA sequence analysis help to know the mutation process, cause of the mutations that directs to develop drugs related to it. In this proposed system we used novel algorithm idlncRNAseq [43] to extract long non coding RNA sequence from the Covid-19 DNA sequence. The extracted long non coding RNA is used as input to this proposed system to find the point mutations in the long non coding sequence. By using proposed algorithm, the system is trained by giving set of Covid-19 long non coding sequences as input. In the input every feature is a nucleotide in the long non coding RNA sequence which matches to the output feature of the sequence.

The data of long non coding RNA is collected from that novel algorithm idlncRNAseq [43] by giving Covid-19 DNA sequences as input. The Covid-19 DNA sequences which are aligned are collected from NIH Genomic data. The learning is performed to find the virus mutation rules to predict the point mutations. The mutations in lncRNA sequence can be used for disease diagnosis and also useful for developing new drug for the Covid-19

This proposed system is focused on the point mutations that occur in the long non coding sequence of Covid-19 virus. This analysis of detecting point mutation is done by monitoring the changes in the nucleotides sequence. The mutation of the lncRNA sequence of Covid-19 virus is predicted by applying the techniques of machine learning to extract their patterns and rules. The proposed algorithm is applied on an aligned lncRNA sequence which is extracted by the novel algorithm recently.

The technique of the proposed system is used to find point mutations in the Covid-19 lncRNA sequence. The first step of technique is to provide both input and output. The input and output of the system is defined by 4 times of the number of nucleotides in the lncRNA sequence. hence it is represented by 4 binary bits i.e., the nucleotide A is represented by 0 and the nucleotide C is represented by 2 and the nucleotide G is represented by 2 and nucleotide U is represented by 3. The distance of each nucleotide is same, but the distance between 1st and 3rd are not same as between 1st and 2nd nucleotide. The frequent nucleotides bases are identified by applying sequential rules between the nucleotides. Next finally the mutations in the sequence are identified with the help of codon table and all the features of nucleotides.

Let consider nucleotide sequence N with set of nucleotides

N=(n1,n2……nn)

Let take target class M to find mutation in the Covid-19 lncRNA sequence.

The target class maximizes the likelihood Of M

P(M|N) = P(n1,n2,…nn|B) (1)

since the present work is to classify the lncRNA sequence into mutation and non-mutation sequence, a binary class B ∈ {0,1} was considered where 1 is the mutation in the sequence and 0 denotes non mutation in the sequence. For Binary classification the class for the sequence sample could be determined by comparing two classes such as

$$= \frac{p(B=1)\prod_i^n = pi\,(ni \mid B=1)}{p(B=0)\prod_i^n i = Pi\,(ni \mid B=0)} \quad (2)$$

$$\frac{p(B=1 \mid N=n1,n2,n3.....n_n)}{p(B=0 \mid N=n1,n2,n3,....n_n)}$$

The equation (2) can be represented as

$$= \log \frac{p(B=1)}{p(B=0)} + \sum_{i=1}^{n} \log \frac{Pi\,(ni \mid B=1)}{Pi\,(ni \mid B=0).} \quad (3)$$

Hence the mutation will be identified as

$$\frac{p(B=1 \mid N=n1,n2,n3.....n_n)}{p(B=0 \mid N=n1,n2,n3,....n_n)} >= 0 \quad (4)$$

The classifier performance depends on the equation (4)

3.1 The sequence of lncRNA is expressed as

lncRNA = n1, n2, n3, … ,n_i, … , n_n

Where s €{A(adenine), C (cytosine), G(guanine), U(uracil) }

3.1.1 Algorithm: Pseudocode of IdmutlncRNAseq

Input: Long non coding RNA Sequence each of N nucleotides, each nucleotide have A,C,G,U

Output: nucleotide n1,n2,,,,n_n,where nm1,nm2…nm_n are the mutations

1.Begin

2.finding frequent nucleotide base(s)

3.Applying sequential rules between nucleotides bases(s)

4.def lncRNAseqMuation(lncRNA):

5.Comparing nucleotides with codon table

6.def idMutiation(lncRNA,codon)

7.for pos in range(len(lncRNA)):

7.1 lncRAN[pos]! =codon[po]

7,2 display the Mutation

7.3 return MutationSeq

8.Displaying the mutations

9.End

According to my review and knowledge, for classification of lncRNA sequence into mutation and non-mutation sequence there is no computational approach developed until now. An important task of machine learning approach is the classification, here the proposed system classifies the lncRNAs sequence with point mutations and the sequences without point mutations.

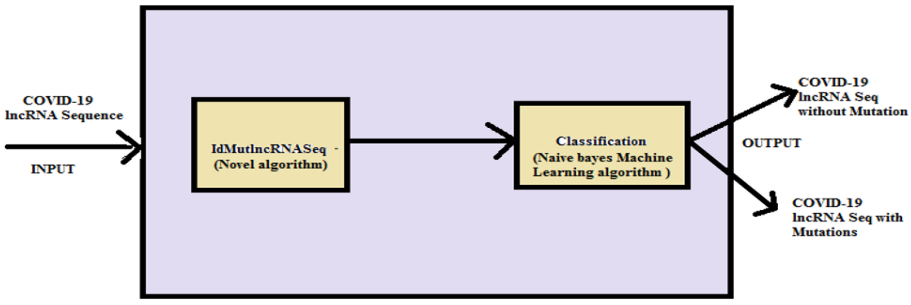

Fig 3: proposed system with novel approach of identification of point mutations and its classification.

This proposed system extracted features using novel algorithm idmutlncRNAseq and identified point mutations in the lncRNA sequence and used Naive Bayes classification algorithm to classify the covid-19 lncRNA sequence into mutation and non-mutation sequence. Naïve Bayes is a statistical algorithm for classification which is used successfully in bioinformatics. The naïve bayes is same that of Covariance Determinant (CD). This algorithm assumes the features are independent of each other for the outcome. Naive bayes classification algorithm finds the results with maximum probability of set of observed features

4. Performance comparison

The algorithm performance is evaluated using specificity, sensitivity and accuracy.TN TP FN and FP represents the number of correctly not identified mutations, the number correctly identified mutations and correctly not identified non mutations and correctly identified non mutations of the nucleotides respectively. The present classifier dependents on independent attribute, and ROC (receiver operating characteristic curve also represented. The classifier quality can be evaluated by measuring the area under the ROC. The ROC score ranges from 0 to 1, with 0.5 is random classification and score 1.0 is correct classification.

4.1 Results:

ACC **=** accuracy = (TP+TN)/(TP+TN+FP+FN)

Sp =specificity = TN /(FP + TN)

TPR = sensitivity = TP /(TP+FN)

FPR = (1-specificity) = FP/(FP/TN)

Results

Sensitivity	98.000%	89.353% to 99.949%
Specificity	95.000%	83.080% to 99.389%
AUC	0.965	0.903 to 0.992
Positive Likelihood Ratio	19.600	5.074 to 75.713
Negative Likelihood Ratio	0.021	0.003 to 0.147
Disease prevalence	55.556%	44.700% to 66.036%
Positive Predictive Value	96.078%	86.380% to 98.954%
Negative Predictive Value	97.436%	84.501% to 99.624%
Accuracy	96.667%	90.566% to 99.307%

Fig 4: performance of the proposed system

The identification of Mutations in the lncRNA sequence and the Naive bayes classifier is used to classify the mutation and non-mutation lncRNA sequence with the accuracy of 96.66% with average specificity 95 % and sensitivity 98% were obtained for the classification of mutation and non-mutation by using the above novel algorithm features.

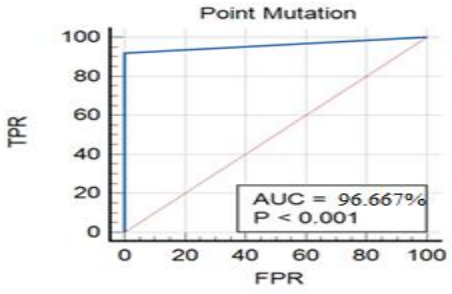

Fig 5: proposed system maximum Accuracy of point Mutation

A twofold cross-validation approach is done on the sequence dataset and found that the proposed algorithm achieved maximum accuracy of 96.6%. Present method yielded a best ROC score of and predictive accuracy with to the best of my knowledge, there is no method to classify the lncRNA

sequence into mutation and non-mutation sequence for any other disease, so with the published result this algorithm is confirmed that it is best method.

5. Conclusion:

This proposed system identified point mutations in the Covid-19 lncRNA sequence and classified the lncRNA sequence into mutation sequence and non-mutation sequence. The accuracy of the classification of mutation sequence and non-mutation sequence is 96.6%. This classification is more useful for diseases analysis and development of new drug. In future, it can be extended to find severity of diseases with the help of mutations in the lncRNA sequence.

REFERENCES

1. Refat Khan Pathan , Munmun Biswas ,Mayeen Uddin Khandaker"Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model",2020 Sep;138:110018. doi: 10.1016/j.chaos.2020.110018. Epub 2020 Jun 13.
2. Wang C, Horby PW, Hayden FG, Gao GF "A novel coronavirus outbreak of globalhealth concern". Lancet North Am Ed 2020;395(10223):470–3.
3. Cucinotta D, Vanelli M. "WHO declares COVID-19 a pandemic". Acta Bio-medica2020;91(1):157–60.
4. Mostafa A. Salama, Aboul Ella Hassanien and Ahmad Mostafa1 "The prediction of virus mutation using neural networks and rough set techniques",EURASIP J Bioinform Syst Biol. 2016 Dec; 2016: 10. Published online 2016 May 13. doi: 10.1186/s13637-016-0042-0
5. AkkizH,"Implications of the Novel Mutations in the SARS-CoV-2 Genome for Transmission, Disease Severity, and the Vaccine Development", 2021 May 7;8:636532. doi: 10.3389/fmed.2021.636532. PMID: 34026780; PMCID: PMC8137987.
6. Duffy S. "Why RNA virus mutation so damn high?"PLoS Biol. (2018) 16:e3000003. 10.1371/journal.pbio.3000003 .
7. FitzsimmondsWj, Woods RJ, McCrone JT, Woodman A, Amold JJ, Yennawar M, et al.. A speed-fidelity trade-off determines the mutation rate and virulence of an RNA virus. PLoS Biol. (2018) 16:e2006459. 10.1371/journal.pbio.2006459
8. Korber B, Fischer WM, Gnanakaran S, Yoon H, Thelier J, Abfalterer W, et al.."Tracking changes in SARS-CoV-2 spike : evidence that D614G increases infectivity of the COVID-19 virus". Cell. (2020) 182:1–16. 10.1016/j.cell.2020.06.043
9. Callaway E. "Making sense of coronavirus mutations. Nature. (2020)" 585:174–7. 10.1038/d41586-020-02544-6
10. Xing Chen, Chenggang Clarence Yan, Xu Zhang, Zhu-Hong You ,"Long non-coding RNAs and complex diseases: from experimental results to computational models",Briefings in Bioinformatics, Volume 18, Issue 4, July 2017, Pages 558–576, https://doi.org/10.1093/bib/bbw060

11. Taft RJ, Pang KC, Mercer TR, et al. “Non-coding RNAs: regulators of disease”. J Pathol2010;220:126–39.

12. Li J, Xuan Z, Liu C. “Long non-coding RNAs and complexhuman diseases”. Int J Mol Sci 2013;14:18790–808.

13. Ponting CP, Oliver PL, Reik W.” Evolution and functions oflong noncoding RNAs”. Cell 2009;136:629–41

14. Mercer TR, Dinger ME, Mattick JS. “Long non-coding RNAs: insights into functions”. Nat Rev Genet 2009;10:155–9

15. Spizzo R, Almeida MI, Colombatti A, et al. “Long non-codingRNAs and cancer”: a new frontier of translational research&quest. Oncogene 2012;31:4577–87.

16. Cheetham S, Gruhl F, Mattick J, et al. “Long noncoding RNAsand the genetics of cancer”, Br J Cancer 2013;108:2419–25.

17. Gutschner T, Diederichs S. “The hallmarks of cancer: a longnon-coding RNA point of view. RNA” Biol 2012;9:703–19.

18. Jendrzejewski J, He H, Radomska HS, et al. “The polymorphism rs944289 predisposes to papillary thyroid carcinomathrough a large intergenic noncoding RNA gene of tumorsuppressor type”. Proc Natl Acad Sci 2012;109:8646–51.

19. Zhang X, Zhou Y, Mehta KR, et al. “A pituitary-derived MEG3isoform functions as a growth suppressor in tumor cells”. JClin Endocrinol Metab2003;88:5119–26

20. . Ji P, Diederichs S, Wang W, et al. MALAT-1, a novel noncoding RNA, and thymosin b4 predict metastasis and survival inearly-stage non-small cell lung cancer. Oncogene 2003;22:8031–41.

21. . Gupta RA, Shah N, Wang KC, et al. Long non-coding RNAHOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010;464:1071–6.

22. Guffanti A, Iacono M, Pelucchi P, et al. A transcriptionalsketch of a primary human breast cancer by 454 deepsequencing. BMC Genomics 2009;10:163.

23. Zhang Z, Hao H, Zhang C, et al. Evaluation of novel geneUCA1 as a tumor biomarker for the detection of bladder cancer. Zhonghua Yi Xue Za Zhi2012;92:384–7.

24. Pibouin L, Villaudy J, Ferbus D, et al. Cloning of the mRNA ofoverexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. Cancer GenetCytogenet2002;133:55–60.

25. Zhang X, Zhou Y, Mehta KR, et al. A pituitary-derived MEG3isoform functions as a growth suppressor in tumor cells. JClin Endocrinol Metab2003;88:5119–26.

26. Guan Y, Kuo W-L, Stilwell JL, et al. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. Clin Cancer Res 2007;13:5745–55.

27. Zhang Q, Chen C-Y, Yedavalli VS, et al. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. MBio 2013;4:e00596–12.

28. Faghihi MA, Modarresi F, Khalil AM, et al. Expression of anoncoding RNA is elevated in Alzheimer's disease anddrives rapid feed-forward regulation of b-secretase. Nat Med2008;14:723–30.

29. . Chen G, Wang Z, Wang D, et al. LncRNADisease: a databasefor long-non-coding RNA-associated diseases. Nucleic AcidsRes 2013;41:D983–6.

30. Tathagata Dey, Shreyans Chatterjee, Smarajit Manna, AsheshNandy, Subhas C. Basak,Identification and computational analysis of mutations in SARS-CoV-2,Computers in Biology and Medicine,Volume 129,2021,104166,ISSN 0010-4825,https://doi.org/10.1016/j.compbiomed.2020.104166.

31. Peng-Mian Feng,1 Hui Ding,2 Wei Chen,3 and Hao Lin,"Naive Bayes Classifier with Feature Selection toIdentify Phage Virion Proteins",Computational and Mathematical Methods in Medicine,Volume 2013, Article ID 530696, 6 pages,http://dx.doi.org/10.1155/2013/530696

32. K. C. Chou, "Prediction of protein cellular attributes usingpseudo-amino acid composition," Proteins, vol. 43, no. 3, pp.246–255, 2001.

33. K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," Bioinformatics, vol.21, no. 1, pp. 10–19, 2005

34. L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based onvariants of Chou's pseudo amino acid composition and onevolutionary information," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 2, pp. 467–475,2012.

35. D. Zou, Z. He, J. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," Journalof Computational Chemistry, vol. 32, no. 2, pp. 271–278, 2011

36. S. W. Zhang, Y. L. Zhang, H. F. Yang, C. H. Zhao, and Q. Pan,"Using the concept of Chou'spseudo amino acid compositionto predict protein subcellular localization: an approach byincorporating evolutionary information and von Neumannentropies," Amino Acids, vol. 34, no. 4, pp. 565–572, 2008.

37. K. K. Kandaswamy, G. Pugalenthi, S. Moller et al., "Prediction of ¨apoptosis protein locations with genetic algorithms and supportvector machines through a new mode of pseudo amino acidcomposition," Protein and Peptide Letters, vol. 17, no. 12, pp.1473–1479, 2010.

38. S. Mei, "Predicting plant protein subcellular multi-localizationby Chou's PseAAC formulation based multi-label homologknowledge transfer learning," Journal of Theoretical Biology, vol.310, pp. 80–87, 2012

39. Y. K. Chen and K. B. Li, “Predicting membrane protein typesby incorporating protein topology, domains, signal peptides,and physicochemical properties into the general form of Chou’spseudo amino acid composition,” Journal of Theoretical Biology,vol. 318, pp. 1–12, 2013

40. H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh,“Prediction of allergenic proteins by means of the concept ofChou’s pseudo amino acid composition and a machine learningapproach,” Medicinal Chemistry, vol. 9, no. 1, pp. 133–137, 2013

41. S. S. Sahu and G. Panda, “A novel feature representationmethod based on Chou’s pseudo amino acid composition forprotein structural class prediction,” Computational Biology andChemistry, vol. 34, no. 5-6, pp. 320–327, 2010.

42. D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, “Useof fuzzy clustering technique and matrices to classify aminoacids and its impact to Chou’s pseudo amino acid composition,”Journal of Theoretical Biology, vol. 257, no. 1, pp. 17–26, 2009

43. Revathi Annem , Jyothi Singaraju,” A Novel Machine Learning Algorithm to Identify Long Non-Coding RNA (lncRNA) Sequence of Covid-19”,Webology (ISSN: 1735-188X) Volume 18, Number 2, December, 2021