




Prediction Type of Codon Effect in Each Disease Based on Intelligent Data Analysis Techniques

Zena A. Kadhuim¹ and Samaher Al-Janabi²(✉) 

¹ Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

² Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq

Samaher@itnet.uobabylon.edu.iq

Abstract. To determine the codon usage effect on protein expression genome-wide, we performed whole-proteome quantitative analyses of FFST and LSTM whole-cell extract by mass spectrometry experiments. These analyses led to the identification and quantification proteins. Five human diseases are due to an excessive number of “cytosine (C), adenine (A), guanine(G)” as (i.e., CAG)repeats in the coding regions of five different genes. We have analyzed the repeat regions in four of these genes from nonhuman primates, which are not known to suffer from the diseases. These primates have CAG repeats at the same sites as in human alleles, and there is similar polymorphism of repeat number, but this number is smaller than in the human genes. In some of the genes, the segment of poly (CAG) has expanded in nonhuman primates, but the process has advanced further in the human lineage than in other primate lineages, thereby predisposing to diseases of CAG reiteration. Adjacent to stretches of homogeneous present-day codon repeats, previously existing codons of the same kind have undergone nucleotide substitutions with high frequency. Where these lead to amino acid substitutions, the effect will be to reduce the length of the original homopolymer stretch in the protein. In addition, RNA-sequencing (seq) analysis of the mRNA was performed to determine correlations between mRNA levels with codon usage biases. To determine the codon usage bias of genes, the codon bias index (CBI) for every protein-coding gene in the genome was calculated. CBI ranges from -1 , indicating that all codons within a gene are nonpreferred, to $+1$, indicating that all codons are the most preferred, with a value of 0 indicative of random use. Because CBI estimates the codon bias for each gene rather than for individual codons, the relative codon biases of different genes can be compared.

Keywords: Codon DNA · Protein · Information Gain · LSTM

1 Introduction

A Disease is an extraneous condition that affects the body’s organs with sporadic damage, and its functions stop working either temporarily or for a long time [1]. Recently many disease present like (COVID19, hemorrhagic fever ([2]. It is even estimated that

the number of patients with the aforementioned diseases is three million patient daily. The reason behind the tendency of effected disease is people not apply protective equipment and also not eat an healthy food [3]. The human need to protect their body against this deathly disease. All human has mRNA (Ribonucleic Acid) It is a complex compound with a high molecular weight that is involved in the protein synthesis process inside the cell [4]. Each mRNA sequence contain number of limited codon that effect directly on indirectly to disease caused to human [5]. Intelligent Data Analysis (IDA) is an interdisciplinary of study that focuses on extracting meaningful knowledge from data using techniques from artificial intelligence, high-performance computing, pattern recognition, and statistics [6]. The IDA process include three primary step, first must work on problem from real word and understand both problem and parameter of these problem second build model for this problem (clustering, classification, prediction, optimization, etc.) and evaluate the result. Finally Interprets the results so that they are understandable by all specialists and no specialists [7]. Data analysis is divided into four types Descriptive Analysis, Diagnostic Analysis, Predictive Analysis and Prescriptive Analysis. Descriptive analysis is a sort of data analysis that helps to explain, show, or summarize data points in a constructive way so, that patterns can develop that satisfy all of data conditions. While Diagnostic Analysis is the technique of using data to determine the origins of trends and correlations between variables is known as diagnostic analytics. After using descriptive analytics to find trends, it might be seen as a logical next step. Predictive analytics is a group of statistical approaches that evaluate current and historical data to, generate predictions about future or otherwise unknown events. It includes data mining, predictive modelling, and machine learning. Finally Prescriptive analytics is after data analytics maturity stage, for better decision in appropriate time. Different type of data analysis has been introduced with their used in different field and advantage of it. Main aim of intelligent Data analysis is to extract knowledge from data. Prediction is a data analysis task for predicting a value not known of target feature, we know the prediction techniques split based on the scientific field into two fields; prediction techniques related to data mining and prediction related to deep learning (i.e., Neuron computing Techniques) [8]. Different type of data analysis has been introduced with their used in different field and advantage of it. Aim of prediction is the process of making estimates for the future on the basis of past and present data and its impact on analysing trends. Bioinformatics is a sub discipline of biology and computer science, concerned with the extract, storage, analysis, and dissemination of biological data, for manage data in such a way that it, allows easy access to the existing information and to, submit new entries as they are produced and developing technological tools that help analyse data biology. Bioinformatics encompasses a wide range of disciplines, including Drug Designing, Genomics, Proteomics, System Biology, Machine Learning, Advanced Algorithms for Bioinformatics, Structural Biology, Computational Biology, and many others. Bioinformatics is consisted of complex DNA and amino acid sequences called protein after extracting from DNA. Bioinformatics is itself a great area of research at present [9].

The reset of paper summarized as: related works are explained in Sect. 2. In Sect. 3 present the theoretical background on techniques use, Sect. 4 show the methodology

of this work. Section 5 shows the results and discussion. Finally, Sect. 6 states the conclusions and future work of this model.

2 Related Work

Protein prediction is one of the most important concerns that directly affect people's lives and the continuance of a healthy lifestyle in general. The goal of this method is to establish a prediction of a mount of disease method for dealing with multiple type of disease and stop it later. Therefore, many researchers work on this filed summarized below.

In [10], Ahmed et al., implement a new model based on artificial intelligence to perform genome sequence analysis of human that infected by COVID-19 and other viruses that like COVID-19 example SARS and MERS and Ebola and middle east respiratory syndrome. The system helps to get important information from the genome sequences of different viruses. This done by extracting information of COVID-19 and perform comparative data analysis to original RNA sequences to detect gene continue virus and their frequency by count of amino acids.at end of method, classifier-based machine learning called support vector machine used to classify different genome sequences. The proposed work uses (accuracy) for measuring performance of algorithm. This study implements high accuracy level (97%) for COVID-19.

In [11], Narmadha and Pravin introduce a method called graph coloring-deep neural network to predict influence protein in infectious diseases. The method starts by coloring the protein that have more interaction in the network represent disease. The main aim of this method is to development of drug and early diagnosis and treatment of the disease. They used various datasets for different diseases (cancer, diabetes, asthma and HPV viral infection). The result show that for predicting cancer 92.328% accuracy, 93.121% precision, 92.874% recall and f measurement 91.102%.

In [12], Asad Khan et al. proposed a new method to predict the existence of m6A in RNA sequences this method used statistical and chemical properties of nucleotides and called (m6A-pred predictor) and uses random forest classifier to predict m6A by identify features that was discriminative. The proposed work uses (accuracy) and (Mathew correlation coefficient values) for measuring performance of algorithm. This study show high accuracy level (78.58%) with Mathew correlation coefficient values (79.65%) of 0.5717. Our work similarity with this work in evaluation measurement but differ from method used to discover protein based on intelligent data analysis and techniques used.

The authors in [13] predict protein position of S-sulfenylation by a new method called SulSite-GTB. This protein involved in a different biological processes important for life like (signaling of cell, increasing stress). The methods summarize by four steps: combine amino acid composition, dipeptide composition, grouped weight encoding, K nearest neighbors, position-specific amino acid propensity, position-weighted amino acid composition, and pseudo-position specific score matrix feature extraction. Secondly, to process the data on class imbalance, To remove the redundant and unnecessary features, the least absolute shrinkage and selection operator (LASSO) is used. Finally, to predict sulfenylation sites, the best feature subset is fed into a gradient tree boosting classifier. Prediction accuracy is 92.86%.

As for the work in [14], Athilakshmi et al. design a method using deep learning to discover anomaly causing genes in mRNA sequences cause brain disorders such as Alzheimer's disease and Parkinson's disease (Table 1).

3 Theoretical Background

The following section show the main concepts used in this paper.

3.1 Fast Frequency Sub-graph Mining Algorithm (FFSMA)

Frequency Sub-graph mining algorithm is an algorithm-based pattern growth for extracting all frequent sub-graph from data and then accepting the most frequent sub-graph according to some minimum support [15]. Many algorithms of FSM Work with graph, FFSM is a Fast Frequent Sub-graph mining algorithm its outperform all FSM algorithms include (gSpan, CloGraMi and Hybrid tree miner) because of two reason, first its contain incidence matrix normalize that compute each node and its connected edge second for each sub-matrix add all possible edge that have not found in it [16, 25].

3.2 Feature Selection

The act of selecting a subset of pertinent features, or variables, from a larger data collection in order to build models is known as feature selection. Other names for feature selection include variable selection, attribute selection, and variable subset selection [17]. It makes the machine learning algorithm less complex, allowing faster training, and is simpler to understand. If the proper subset is selected, a model's accuracy is increased. Finally, feature selection minimizes over fitting [17] 18. Entropy is a metric for a data-generating function's diversity or randomness. Full entropy data is utterly random, and no discernible patterns can be discovered. Data with low entropy offers the ability or potential to forecast newly created values [19]. On the other side Information gain, which is the decrease in entropy or surprise caused by altering a dataset. By comparing the entropy of the dataset before and after a transformation, information gain is computed, Entropy can be used to determine how a change to the dataset, such as the distribution of classes, affects the dataset's purity using information gain. A lower entropy indicates greater purity or decreased surprise [20, 21].

The connection between two variables is known as a correlation. Using the features of the input data, we can forecast our target variable using these variables. Based on their association, many variables are put together in this metric [22].

3.3 Deep Prediction Neuro Computing Techniques

Prediction is a method that used to predict some value or features according to founded once [17] prediction techniques, either related to data mining (SVM [23], LR [24], RF [25]) or related to Deep prediction neuro computing techniques (LSTM [26], BiLSTM [27], MLSTM [28], RNN [29], GRU [30]). Prediction in deep neuro techniques is outperform data mining techniques in term of accuracy but on the other hand take long time for predict an accurate result [30].

Table 1. Summarized on literate survey

Author	Data set/Database	Preprocessing	Method	Evaluation
Wang [13]	Independent test set (protein sequence) https://github.com/QUEST-AIBDRC/SulSite-GTB/ .	Feature en-coding	SulSite-GTB	Accuracy
Ahliakshmi et al. [14]	Gene Sets of Alzheimer's and Parkinson's http://www.genecards.org	Feature en-coding	DL based Anomaly Detection	MSE
Khan et. al. [12]	RNA sequences	Feature ex-traction	m6A-predction	Accuracy and Mathew correlation coeffi-cient
Narmada and Pravin [11]	Protein sequence Collection of PPI (string DB, IntAct, DIP) database	Segmentation	graph coloring-deep neu-ral network	Confusion matrix
Imran Ahmed et al. [10]	DNA sequences	Feature ex-traction	ML	Accuracy

4 Proposed Method

The data set used in our work is Codon usage Data Set published in machine learning respiratory at <https://archive.ics.uci.edu/ml/datasets/Codon+usage#>. There is 64 codon of Amino Acid related to 13028 disease. Each one of these Amino Acid have a percentage of bias in each disease [31]. In this paper we show how grouping codons effect positively in term of reduce computing and time for entering to farther stages.

4.1 Step of Proposed Method

Amino acids produce the taste of food and keep us healthy. For example, they are used for sports nutrition, medicine, beauty products and to reduce calorie intake. In this proposed method. We implement some of intelligent data analysis techniques for reduce the computation of working on dataset in [31]. All work summarized under the following main points:

- For all CTUG dataset, data pre-processing is performed to group feature by feature selection. We calculate information gain for all feature after convert all description feature to numeric value feature.
- Calculate Minkonisky distance for result from one for creating group between features.
- Grouping features into sixteen group four nodes for each sub-group.
- Enter all group to FFSMA for delete duplication subgroup for each group.
- The confidence of relation is computed to see how correct rule of result extracted.
- Finally, the results enter to long short-term memory to see how result valid

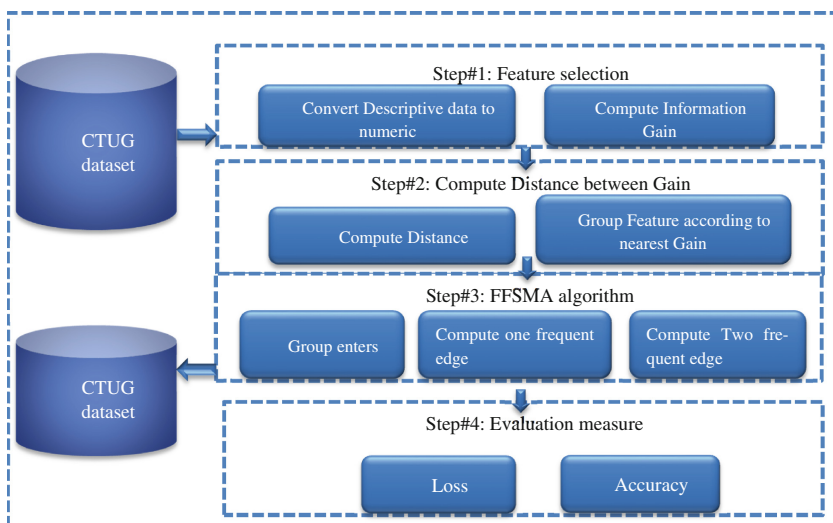


Fig. 1. Proposed New MVA-FFSM Method

- Our step of work is implemented via several stages to finally effect of each codon the details of proposed algorithm explain in algorithm #1 and Fig. 1.

Algorithm #1	
Input:	CUTG dataset //set of all codon related to each disease of kingdom entitled in dataset.
Output:	// non-frequent records
1:	Begin
2:	For all element in CUTG dataset
3:	Perform a feature selection
4:	Compute information Gain(feature, Target)
5:	Remove features have less information gain
6:	End for
7:	For all features in CUTG dataset
8:	Compute minkonisky distance for between inf.G
9:	Group CUTG base on minkonisky distance (nearest Gain)
10:	End for
11:	For all features in group of dataset perform FFSMA
12:	While F(k) > 0 \\ for candidate generation set to ← Φ
13:	For i= 1 to n
14:	TFG(Gi,K+1)= Φ
15:	for each k-edge frequent sub graph gn(Gi,)∈FG(Gi,)
16:	N ← list of all edge (e) not connected to Gi
17:	For each e in N do
18:	gn(G,k+1) ← gn(G,k+1) U e
19:	If gn(G,k+1) not belong (Gi,K+1)
20:	TFG(Gi,K+1)= TFG(Gi,K+1) U gn(G,k+1)
21:	End
	End
22:	End
23:	For each gn(G,k+1) in TFG(Gi,k+1)
24:	If gn(G,k+1) not belong TFGC(G,k+1) then
25:	TFGC(G,k+1)= TFGC(G,k+1)U n(G,k+1)
26:	Freq_gn(G,k+1)=1
27:	Else
28:	Freq_gn(G,k+1)= Freq_gn(G,k+1)+1
29:	End
30:	F(K+1)=0
31:	For each gn(G, k+1) in TFGC(k+1)
32:	if Freq_gn(G, k+1)> minisup then
33:	FGS(GDB)= FGS(GDB) U {gn(G, k+1)}
34:	F(k+1)= F(k+1)+1
35:	End
36:	End
37:	For i= 1 to length of N
38:	FG(G, k+1)=Φ
39:	For each gn(Gi, k+1) in TFGC(k+1) do
40:	if gn(Gi, k+1)∈FGS(GDB) then
41:	FG(G, k+1)= FG(G, k+1) U {gn(G, k+1)}
42:	End
43:	End
44:	End
45:	K=k+1
46:	End for
47:	Delete non maximum frequent sub graph

4.2 Data Preprocessing (Feature Selection)

All CTUG dataset is entered to the system. For our work we need only specific filed for work on it, this field is name of disease and all 64 amino acid to compute there effect to disease. So, for this we show the 64 codon feature by compute information gain for each feature with target that represent Disease (SpieseName).

First, we calculate entropy for each 64 features:

$$\text{Entropy} = -\sum_{n=1}^l p_i * \log_2 p_i \tag{1}$$

Then compute information gain for each features with Target (features, Disease):

$$\text{Information Gain} = \text{Target entropy} - \text{Entropy child} \tag{2}$$

$$\text{Entropy child} = \sum E \text{ splits} * W \tag{3}$$

where:

- Pi probability of element in column
- E splits calculate Entropy for splits depend on feature selected
- W how many element found in splits
- Target Entropy Entropy of target feature

4.3 Selecting Different Records

After preparing data, we have 13028 record and each one of them have 64 codon effect to it. In this work we work on more frequent disease to see the effect of 64 codon. Before FFSMA work we group features according to nearest value of information gain by minkonisky distance.

$$M(D) = \left(\sum i = 1n|X_i - Y_i|p \right) 1/p \tag{4}$$

where:

- xi vector one of value
- y1 vector one of value
- p default value of squirt.

Then group 64 feature into 16 group of 4 nodes according to value of information gain from lowest to highest:

[['UAG', 'UAA', 'UGA', 'CGG'], ['CGA', 'UGC', 'AGG', 'UGU'], ['UCG', 'CGU', 'ACG', 'CCG'], ['AGU', 'AGC', 'CAU', 'GGG'], ['UGG', 'CGC', 'AGA', 'CAC'], ['CCU', 'UAC', 'UCC', 'GCG'], ['CCC', 'UCU', 'GUA', 'AUG'], ['ACU', 'UUG', 'CUA', 'UCA'], ['CCA', 'GUC', 'GCA', 'GGA'], ['GGU', 'CUU', 'AAC', 'GUU'], ['GCU', 'CAA', 'CAG', 'UAU'], ['GUG', 'UUC', 'ACA', 'GGC'], ['AUA',

'CUC', 'ACC', 'CUG'], ['UUA', 'GAC', 'AUC', 'AAG'], ['GAU', 'UUU', 'AAU', 'GAG'], ['GAA', 'GCC', 'AAA', 'AUU']]

Then after group it, each group is a sub graph enter to FFSMA and dealing with each column as value of node and compute frequent edge as follow:

Enter **Sub-graph 1** that have 4 node to FFSMA algorithm:

- $N = AUU, GCC, AAA, GAA$, whereas:

First compute One frequent edge:

- $E1 = \{AUU, GCC\}$ at Attrition T1,
- $E2 = \{GCC, AAA\}$ at Attrition T1,
- $E3 = \{AAA, GAA\}$ at Attrition T1,

Second compute Two frequent edge:

- $E1E2 = \{AUU, GCC, AAA\}$ at Attrition T2,
- $E2E3 = \{GCC, AAA, GAA\}$ at Attrition T2,

Third compute Three frequent edge:

- $E1E2E3 = \{AUU, GCC, AAA, GAA\}$ at Attrition T3,

Results of Remove Duplication Sup-graph FFSMA.

Origin Sub-graph—Results Sub-graph = Deleted Duplication Sub-graph.

5 Results and Discussion

The result of our work represents how each feature selection can reduce a computation of our work according to Gain of each feature and scale it, Table 2 shows the result of our features with normalized disease.

In Table 2, there are five columns: in the first column are the main characteristics in the dataset, which represent the codons of each disease, which are 64 codons found in all creatures that are associated with 13,028 diseases, in the second column the entropy values for each codon from among the 64 codons, and the third column It represents the value of the information gain in relation to the codon and its association with each disease, the next column is the conversion of the entropy values with the scaling function between (1 and -1) and the last column is the normalization of the information gain value to be between (1 and 0). Figure 2 represent important codon related to each disease that in range (1, 0).

Table 2. Illustrate the relation between feature and disease in term of Gain, Correlation

Feature	Entropy	Gain	Correlation	GN	GS
UUU	11.86638	5.14777	0.148125	0.748361	0.496721
UUC	11.65497	4.94355	0.292521	0.718672	0.437344
UUA	11.65547	5.02011	0.194589	0.729802	0.459604
UUG	11.17456	4.64629	-0.11831	0.675458	0.350915
CUU	11.46138	4.75966	0.255333	0.691939	0.383878
CUC	11.65372	4.97407	0.226646	0.723109	0.446218
CUA	11.3578	4.6904	0.410942	0.68187	0.36374
CUG	11.63117	4.9805	-0.17578	0.724044	0.448087
AUU	12.04675	5.32029	0.221926	0.773441	0.546881
AUC	11.83587	5.11225	0.243239	0.743197	0.486394
AUA	11.64205	4.97307	0.325031	0.722963	0.445927
AUG	11.26315	4.60613	-0.29516	0.669619	0.339238
GUU	11.46679	4.79023	-0.14321	0.696383	0.392766
GUC	11.37616	4.7118	-0.11657	0.684981	0.369962
GUA	11.27364	4.60142	0.167697	0.668935	0.337869
GUG	11.45692	4.92349	-0.28198	0.715756	0.431511
GCU	11.49003	4.79992	-0.08557	0.697792	0.395583
GCC	11.94552	5.24025	0.021014	0.761805	0.52361
GCA	11.40672	4.71649	0.067029	0.685663	0.371326
GCG	10.83358	4.50338	-0.3017	0.654682	0.309364
CCU	11.04018	4.37547	0.029213	0.636087	0.272174
CCC	11.24163	4.57031	0.185779	0.664412	0.328824
CCA	11.37799	4.69991	0.241677	0.683253	0.366505
CCG	10.56902	4.21152	-0.27587	0.612253	0.224505
UGG	10.77833	4.2783	-0.26737	0.621961	0.243921
GGU	11.40249	4.75339	-0.22887	0.691027	0.382055
GGC	11.4311	4.97036	-0.17221	0.722569	0.445139
GGA	11.62821	4.73219	0.231896	0.687945	0.375891
GGG	10.88869	4.26292	-0.05329	0.619725	0.23945
UCU	11.27058	4.5941	0.107305	0.66787	0.335741
UCC	11.17281	4.48813	0.265926	0.652465	0.30493
UCA	11.372	4.69187	0.274799	0.682084	0.364168

(continued)

Table 2. (continued)

Feature	Entropy	Gain	Correlation	GN	GS
UCG	10.36185	4.02861	-0.20983	0.585662	0.171324
AGU	10.68159	4.23582	-0.15994	0.615785	0.23157
AGC	10.8407	4.25646	-0.12671	0.618786	0.237571
ACU	11.29789	4.61877	0.031038	0.671457	0.342914
ACC	11.68092	4.97473	0.132544	0.723205	0.446409
ACA	11.66511	4.96873	0.226249	0.722332	0.444665
ACG	10.57329	4.1723	-0.29004	0.606551	0.213102
UAU	11.57677	4.90166	-0.02509	0.712582	0.425164
UAC	11.15769	4.48105	0.029102	0.651436	0.302871
CAA	11.48509	4.81417	0.015902	0.699863	0.399726
CAG	11.34212	4.84099	-0.30663	0.703762	0.407524
AAU	11.86419	5.16431	-0.09713	0.750765	0.50153
AAC	11.47363	4.77758	0.038796	0.694544	0.389088
UGU	10.30101	3.96058	-0.09584	0.575772	0.151544
UGC	10.40001	3.9367	-0.04899	0.5723	0.144601
CAU	10.89653	4.2591	0.002728	0.61917	0.238339
CAC	11.01853	4.36198	0.186279	0.634126	0.268252
AAA	11.99542	5.31061	-0.08887	0.772034	0.544067
AAG	11.63019	5.11527	-0.23254	0.743636	0.487272
CGU	10.49915	4.09973	-0.25308	0.596001	0.192002
CGC	10.73903	4.29636	-0.27613	0.624586	0.249172
CGA	10.37923	3.82518	0.301776	0.556088	0.112176
CGG	9.7358	3.65667	-0.1815	0.531591	0.063182
AGA	10.12659	4.2995	-0.1104	0.625043	0.250085
AGG	9.52235	3.9369	-0.11753	0.572329	0.144659
GAU	11.75633	5.14641	-0.34814	0.748163	0.496326
GAC	11.7045	5.04355	-0.25787	0.733209	0.466419
GAA	11.87286	5.19904	-0.22795	0.755814	0.511628
GAG	11.69344	5.16498	-0.27195	0.750862	0.501725
UAA	8.00838	2.22431	0.111846	0.323361	-0.35328
UAG	5.96591	1.36074	0.045477	0.197818	-0.60436
UGA	8.51548	2.7087	0.442073	0.393779	-0.21244

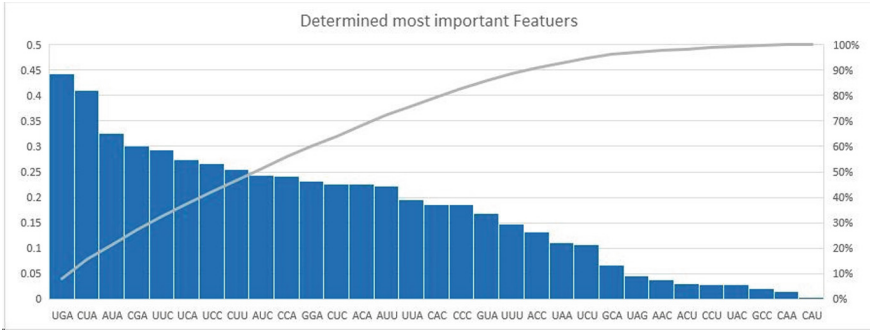


Fig. 2. Relation between codon to target disease

Then all feature must Grouped that represent a graph to enter to FFSMA algorithm, All 16 group of 64 codons from lowest to highs gain:

- G1** ['UAG', 1.36074, 'UAA', 2.22431, 'UGA', 2.7087, 'CGG', 3.65667]
- G2** ['CGA', 3.82518, 'UGC', 3.9367, 'AGG', 3.9369, 'UGU', 3.96058]
- G3** ['UCG', 4.02861, 'CGU', 4.09973, 'ACG', 4.1723, 'CCG', 4.21152]
- G4** ['AGU', 4.23582, 'AGC', 4.25646, 'CAU', 4.2591, 'GGG', 4.26292]
- G5** ['UGG', 4.2783, 'CGC', 4.29636, 'AGA', 4.2995, 'CAC', 4.36198]
- G6** ['CCU', 4.37547, 'UAC', 4.48105, 'UCC', 4.48813, 'GCG', 4.50338]
- G7** ['CCC', 4.57031, 'UCU', 4.5941, 'GUA', 4.60142, 'AUG', 4.60613]
- G8** ['ACU', 4.61877, 'UUG', 4.64629, 'CUA', 4.6904, 'UCA', 4.69187]
- G9** ['CCA', 4.69991, 'GUC', 4.7118, 'GCA', 4.71649, 'GGA', 4.73219]
- G10** ['GGU', 4.75339, 'CUU', 4.75966, 'AAC', 4.77758, 'GUU', 4.79023]
- G11** ['GCU', 4.79992, 'CAA', 4.81417, 'CAG', 4.84099, 'UAU', 4.90166]
- G12** ['GUG', 4.92349, 'UUC', 4.94355, 'ACA', 4.96873, 'GGC', 4.97036]
- G13** ['AUA', 4.97307, 'CUC', 4.97407, 'ACC', 4.97473, 'CUG', 4.9805]
- G14** ['UUA', 5.02011, 'GAC', 5.04355, 'AUC', 5.11225, 'AAG', 5.11527]
- G15** ['GAU', 5.14641, 'UUU', 5.14777, 'AAU', 5.16431, 'GAG', 5.16498]
- G16** ['GAA', 5.19904, 'GCC', 5.24025, 'AAA', 5.31061, 'AAU', 5.32029].

All sixteen-sub group enter to frequency sub graph mining algorithm (FFSMA) to remove duplication sub-graph. FFSMA results reduce the number of rows for whole dataset by remove frequent edge and save only different row that effect to each different disease and testing by association rule mining of dataset. Also the time of using techniques of preprocessing dataset is reduced compare with time of working to all dataset from 0.22 to 0.16 s.

Because of sensitive dataset, we select only rule that have high relation to feature. So in this case we select second rule and so forth.

The original dataset is [13028 rows × 65 columns of features and normalized Disease]. The rule for each group is:

- G1 entered** [13028 rows × 4 columns] **out is:** [12037 rows × 4 columns]
- G2 entered** [13028 rows × 4 columns] **out is:** [12452 rows × 4 columns]

- G3 entered** [13028 rows × 4 columns] **out is:** [12615 rows × 4 columns]
- G4 entered** [13028 rows × 4 columns] **out is:** [12769 rows × 4 columns]
- G5 entered** [13028 rows × 4 columns] **out is:** [12701 rows × 4 columns]
- G6 entered** [13028 rows × 4 columns] **out is:** [12811 rows × 4 columns]
- G7 entered** [13028 rows × 4 columns] **out is:** [12826 rows × 4 columns]
- G8 entered** [13028 rows × 4 columns] **out is:** [12852 rows × 4 columns]
- G9 entered** [13028 rows × 4 columns] **out is:** [12814 rows × 4 columns]
- G10 entered** [13028 rows × 4 columns] **out is:** [12839 rows × 4 columns]
- G11 entered** [13028 rows × 4 columns] **out is:** [12818 rows × 4 columns]
- G12 entered** [13028 rows × 4 columns] **out is:** [12837 rows × 4 columns]
- G13 entered** [13028 rows × 4 columns] **out is:** [12814 rows × 4 columns]
- G14 entered** [13028 rows × 4 columns] **out is:** [12849 rows × 4 columns]
- G15 entered** [13028 rows × 4 columns] **out is:** [12863 rows × 4 columns]
- G16 entered** [13028 rows × 4 columns] **out is:** [12866 rows × 4 columns]

Finally, results of FFSMA must be trained by Long Short-Term Memory (LSTM) that produce a result according to different splitting (Table 3).

Table 3. Measurements criteria

Rate of Training and Testing Dataset	MSE (%)	ACCURACY (%)
50 train, 50 test	0.003	94.2431
70 train, 30 test	0.0019	94.678
90 train, 10 test	0.0005	96.162

We see how accuracy is increase according to splitting between multiple value of training and testing.

6 Conclusion

To determine the codon usage bias of genes, the codon bias index (CBI) for every protein-coding gene in the genome was calculated. CBI ranges from -1 , indicating that all codons within a gene are nonpreferred, to $+1$, indicating that all codons are the most preferred, with a value of 0 indicative of random use. Because CBI estimates the codon bias for each gene rather than for individual codons, the relative codon biases of different genes can be compared. The accuracy of proposed method is 96.162% while MSE is 0.0005 .

References

1. Al-Janabi, S.: Overcoming the main challenges of knowledge discovery through tendency to the intelligent data analysis. *Int. Conf. Data Anal. Bus. Ind. (ICDABI)* **2021**, 286–294 (2021)
2. Kadhuim, Z.A., Al-Janabi, S.: Intelligent deep analysis of DNA sequences based on FFGM to enhancement the performance and reduce the computation. *Egypt. Inform. J.* **24**(2), 173–190 (2023). <https://doi.org/10.1016/j.eij.2023.02.004>
3. Vitiello, A., Ferrara, F.: Brief review of the mRNA vaccines COVID-19. *Inflammopharmacology* **29**(3), 645–649 (2021). <https://doi.org/10.1007/s10787-021-00811-0>
4. Toor, R., Chana, I.: Exploring diet associations with Covid-19 and other diseases: a network analysis–based approach. *Med. Biol. Eng. Compu.* **60**(4), 991–1013 (2022). <https://doi.org/10.1007/s11517-022-02505-3>
5. Kadhuim, Z.A., Al-Janabi, S.: Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis. *Results Eng.* **17**, 100847 (2023). <https://doi.org/10.1016/j.rineng.2022.100847>
6. Nambou, K., Anakpa, M., Tong, Y.S.: Human genes with codon usage bias similar to that of the nonstructural protein 1 gene of influenza A viruses are conjointly involved in the infectious pathogenesis of influenza A viruses. *Genetica* 1–19 (2022). <https://doi.org/10.1007/s10709-022-00155-9>
7. Al-Janabi, S., Al-Janabi, Z.: Development of deep learning method for predicting DC power based on renewable solar energy and multi-parameters function. *Neural Comput. Appl.* (2023). <https://doi.org/10.1007/s00521-023-08480-6>
8. Al-Janabi, S., Al-Barmani, Z.: Intelligent multi-level analytics of soft computing approach to predict water quality index (IM12CP-WQI). *Soft Comput.* (2023). <https://doi.org/10.1007/s00500-023-07953-z>
9. Li, Q., Zhang, L., Xu, L., et al.: Identification and classification of promoters using the attention mechanism based on long short-term memory. *Front. Comput. Sci.* **16**, 164348 (2022)
10. Ahmed, I., Jeon, G.: Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. *Interdisc. Sci. Comput. Life Sci.* 1–16 (2021). <https://doi.org/10.1007/s12539-021-00465-0>
11. Narmadha, D., Pravin, A.: An intelligent computer-aided approach for target protein prediction in infectious diseases. *Soft. Comput.* **24**(19), 14707–14720 (2020). <https://doi.org/10.1007/s00500-020-04815-w>
12. Khan, A., Rehman, H.U., Habib, U., Ijaz, U.: Detecting N6-methyladenosine sites from RNA transcriptomes using random forest. *J. Comput. Sci.* **4**, (2020). <https://doi.org/10.1016/j.jocss.2020.101238>
13. Wang, M., Song, L., Zhang, Y., Gao, H., Yan, L., Yu, B.: Malsite-deep: prediction of protein malonylation sites through deep learning and multi-information fusion based on NearMiss-2 strategy. *Knowl. Based Syst.* **240**, 108191 (2022)
14. Athilakshmi, R., Jacob, S.G., Rajavel, R.: Protein sequence based anomaly detection for neuro-degenerative disorders through deep learning techniques. In: Peter, J.D., Alavi, A.H., Javadi, B. (eds.) *Advances in Big Data and Cloud Computing*. AISC, vol. 750, pp. 547–554. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1882-5_48
15. Cheng, H., Yu, J.X.: Graph mining. In: Liu, L., Özsu, M.T. (Eds.) *Encyclopedia of Database Systems*. Springer, New York, (2018)
16. Mohammed, G.S., Al-Janabi, S.: An innovative synthesis of optimization techniques (FDIRE GSK) for generation electrical renewable energy from natural resources. *Results Eng.* **16**, 100637 (2022). <https://doi.org/10.1016/j.rineng.2022.100637>
17. Kadhim, A.I.: Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. In: 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 124–128

18. Wang, S., Tang, J., Liu, H.: Feature selection. In: Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, MA (2017). https://doi.org/10.1007/978-1-4899-7687-1_101
19. Khan, M.A., Akram, T., Sharif, M., Javed, K., Raza, M., Saba, T.: An automated system for cucumber leaf diseased spot detection and classification using improved saliency method and deep features selection. *Multimedia Tools Appl.* **79**(25–26), 18627–18656 (2020). <https://doi.org/10.1007/s11042-020-08726-8>
20. Jia, W., Sun, M., Lian, J., Hou, S.: Feature dimensionality reduction: a review. *Complex Intell. Syst.* 1–31 (2022). <https://doi.org/10.1007/s40747-021-00637-x>
21. Rodriguez-Galiano, V., Luque-Espinar, J., Chica-Olmo, M., Mendes, M.P.: Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. *Sci. Total Environ.* **624**, 661–672 (2018)
22. Saqib, P., Qamar, U., Aslam, A., Ahmad, A.: Hybrid of filters and genetic algorithm-random forests based wrapper approach for feature selection and prediction. In: *Intelligent Computing- Proceedings of the Computing Conference*, vol. 998, pp. 190–199. Springer (2019)
23. Al-Janabi, S., Alkaim, A.: A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis. *Egypt. Informatics J.* **23**(2), 271–290 (2022). <https://doi.org/10.1016/j.eij.2022.01.004>
24. Liew, B.X.W., Kovacs, F.M., Rügamer, D., Royuela, A.: Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. *Eur. Spine J.* **1** (2022). <https://doi.org/10.1007/s00586-022-07188-w>
25. Hatwell, J., Gaber, M.M., Azad, R.M.A.: CHIRPS: Explaining random forest classification. *Artif. Intell. Rev.* **53**, 5747–5788 (2020)
26. Rodriguez-Galiano, V., Luque-Espinar, J., Chica-Olmo, M., Mendes, M.P.: Feature selection approaches for predictive modelling of foreseeing the principles of genome architecture. *Nat. Rev. Genet.* **23**, 2–3 (2022)
27. Liu, H., Zhou, M., Liu, Q.: An embedded feature selection method for imbalanced data classification. *IEEE/CAA J. Autom. Sin.* **6**, 703–715 (2019)
28. Lu, M.: Embedded feature selection accounting for unknown data heterogeneity. *Expert Syst. Appl.* **119** (2019)
29. Ansari, G., Ahmad, T., Doja, M.N.: Hybrid Filter-Wrapper feature selection method for sentiment classification. *Arab. J. Sci. Eng.* **44**, 9191–9208 (2019)
30. Jazayeri, A., Yang, C.: Frequent subgraph mining algorithms in static and temporal graph-transaction settings: a survey. *IEEE Trans. Big Data* (2021)
31. Khomtchouk, B.B.: Codon usage bias levels predict taxonomic identity and genetic composition (2020)