Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 4, Number 3 (2011) pp. 273-282
© Research India Publications
http://www.ripublication.com/jcib.htm

Clustering Algorithm Approach for Prokaryotic and Eukaryotic Gene Prediction

Sandeep Kaur[1]*, Vipan Kumar Sohpal[2] and Preetkanwal Singh[2]

[1]Department of Computer Science & Engineering
Swami Sarvanand Institute of Engg. & Technology, Gurdaspur, India
[2]Department of Chemical Engineering & Bio Technology
Beant College of Engineering and Technology, Gurdaspur, Punjab, India
**Corresponding Author E-mail: sandeep.kaur18@gmail.com*

Abstract

In this paper we have design a computational model for prokaryotic and eukaryotic gene prediction by using the clustering algorithm. The input DNA (Deoxyribonucleic Acid) sequence is spliced and the open reading frames are identified. For identification of consensus sequences various data for the consensus sequence is collected and data mining algorithm is applied for creation of clusters. The gene density is calculated at the end. GC (Guanine and cytosine) content is usually expressed as a percentage value. This model saves the implementation time, as whole of the database is present online so the sequence to be predicted is just taken from any one of the available database. Several experiments have been done where the parameters of gene prediction are changed manually. The performance has been tested on different unknown DNA sequences found on the internet. The basis for making clusters is local alignment between the sequences, as larger the score more the sequences will be similar. So the sequences having score greater than or equal to the threshold value are entered into one cluster and rest of the sequences having score less than the given threshold are entered into second cluster and GC-content percentage is calculated.

Keywords: DNA, RNA, GC, FASTA.

Introduction

Bioinformatics is the science of managing, mining and interpreting information from biological sequences and structures using mathematical & information technology tool. In this area of science, biology, computer science and information technology, all the three merge into a single discipline. Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. During the last few years, bioinformatics has been overwhelmed with increasing floods of data, both in terms of volume and in terms of new databases and new types of data. The aims of bioinformatics are threefold. First, at its simplest bioinformatics organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced, for example the Protein Data Bank for 3D macromolecular structures. While data-creation is an essential task, the information stored in these databases is essentially useless until analyzed. The second aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences. This needs more than just a simple text-based search and programs such as FASTA. The third aim is to use these tools to analyze the data and interpret the results in a biologically meaningful manner. Basic types of data that can be analyzed in bioinformatics are (a) Raw Deoxyribonucleic Acid sequences (b) Protein sequence (c) Macromolecular structure (d) Genome (e) Gene Expression [1].For researchers to benefit from the data stored in a database, two additional requirements must be met first is easy access to the information and second is a method for extracting only the information needed to answer a specific biological question.

Literature Review

Lakshmi et al. (2004) discussed that the biological research in the twenty-first century is primarily driven by high precision instrumentation for exploring the complexity of biological systems in greater detail. Very large datasets are generated from these instruments that require efficient computational tools for data mining and analysis. The definition of the term ―high-throughput has had to be redefined at regular intervals because of the exponential growth in the volume of data generated with each technological advance. For addressing the needs of modeling, simulation and visualization of large and diverse biological datasets from sequence, gene expression and proteomics datasets, systems biology approaches are being developed for construction of gene regulatory networks. Among the various biomolecules present inside the cell, ribonucleic acid plays important roles in executing inherited genetic instructions. Myburgh (2005) proposed an algorithm using matlab for detecting the euokaryotic *RNA* Polymerase II start site using artifical neural networks. Author has discussed the genetic principles involved in DNA, RNA and promoter sequences. The data acquisition process is implemented. And finally the detection program is tested and compared with other promoter detection methods.Ruskin (2007), proposed the clustering technique of gene expression data obtained from microarray experiments. Their survey states the art of applications which recognizes implements the procedures to overcome them. It provides a framework for the evaluation of clustering in gene expression analyses. The nature of microarray data is discussed briefly with suitable examples.

Valdimir (2001) suggested new promoter finding method, known as Dragon Promoter Finder (DPF) which locates RNA polymerase II promoter in DNA sequences of vertebrates by predicting transcriptional start site (TSS) positions. DPF uses sensors for three functional regions (promoters, introns, exons) and an artificial neural network. Results are better as compared to previous promoter-finding algorithms.

It is found from the literature survey that yet there is a requirement to design new model which predicts the gene in lesser time and with less complexity both for prokaryotes and eukaryotes.

Gene Expression or Evolutionary Algorithms

While the specific sequence of nucleotides in a deoxyribonucleic acid molecule can have important information content for a cell, it is actually proteins that do the work of altering a cell’s chemistry by acting as biological catalysts called enzymes. In chemistry catalysts are molecules that allow specific chemical reaction to proceed more quickly than they would have otherwise occurred. Catalysts are neither consumed nor altered in the course of such a chemical process and can be used to catalyze the same reaction many times. The term gene is used in many different ways, but one of its narrowest and simplest definitions is that genes spell out the instruction needed to make the enzyme catalysts produced by cells. Few basic terms related with gene expression are Deoxyribonucleic Acid, m-ribonucleic acid, t-ribonucleic acid, Chromosomes, consensus sequences, promoter sequences, open reading frames, introns and exons. In the world of cells there are two major groups of cells: the prokaryotes and eukaryotes. The major similarities between prokaryotes and eukaryotes are that they both have deoxyribonucleic acid as their genetic material, both have ribosome’s and open reading frames. The major differences between gene expression of prokaryotes and eukaryotes are that prokaryotes have single circular chromosome, whereas eukaryotes have many bar shaped chromosomes. The gene expression for eukaryotes completes in two different steps, one with transcription than with translation. In eukaryotes the open reading frames are longer as compared to open reading frames of prokaryotes.

Genetic or evolutionary algorithms borrow their inspiration from the process of evolution by natural selection found in nature. They start with a population of possible hypotheses, and evaluate them on some training data. The best hypotheses are kept and used to create a new generation of hypotheses. New hypotheses can be obtained from old by two different operations-crossover and mutation. Neural Networks is used for eukaryotic gene prediction. Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics. Gene prediction algorithms are based on the apparently simple rules that the transcriptional machinery uses: strong, easily recognizable signals within the genome such as open reading frames, consensus splice sites and GC content. There are two basic problems in gene prediction: prediction of protein coding regions and prediction of the functional sites of genes. Two classes of methods are generally adopted: similarity based searches and *ab initio* prediction. Various gene prediction algorithms were developed both for prokaryotic and eukaryotic genomes, few of them are HMM, Glimmer, GenScan, BLAST, FASTA. These algorithms are compared on the availability and accuracy as the primary criteria. Accuracy means consistent performance of a program on a set of standard test cases. An evaluation is made on the number of true positives (TP), where the length and end sequence positions are correctly predicted and the number of over predicted positive predictions or false positives (FP), true negatives (TN) and unpredicted residues as false negative (FN) predictions. Based upon this (TP, TN, FP, FN) specificity and sensitivity is measured [9].

Concept of Gene Prediction

The main focus of this work is to design and implement a new model for gene prediction. To achieve this objective, the proposed work will involve the use of data mining techniques to analyze the given database and retrieve the desired information. The data will be taken from online resources and web servers. A new model is created which computes the given DNA sequence based upon its GC content, and open reading frames. Connectivity is done between front-end and back-end and an interface is created. DNA sequence is entered to this interface. Algorithm searches all the possible reading frames from six reading frames. It calculates the resultant vector for each of the open reading frames. Than classify open reading frames as genes and non-genes. After this the GC content is computed. Based on parameters if ratio count>= 6, than it is eukaryotes otherwise prokaryotes. Algorithm includes both transcription and translation for predicting the gene [4, 12].The process by which information is extracted from the nucleotide sequence of a gene and then used to make a protein is essentially the same for all living things on Earth and is described by the grandly named central dogma of molecular biology shown in figure 1. Quite simply, information stored in DNA (Deoxyribonucleic Acid) is used to make a more transient, single stranded polynucleotide called RNA (Ribonucleic Acid) that is in turn used to make proteins. The process of making a ribosomal ribonucleic acid copy of a gene is called transcription and is accomplished through the enzymatic activity of an RNA polymerase.

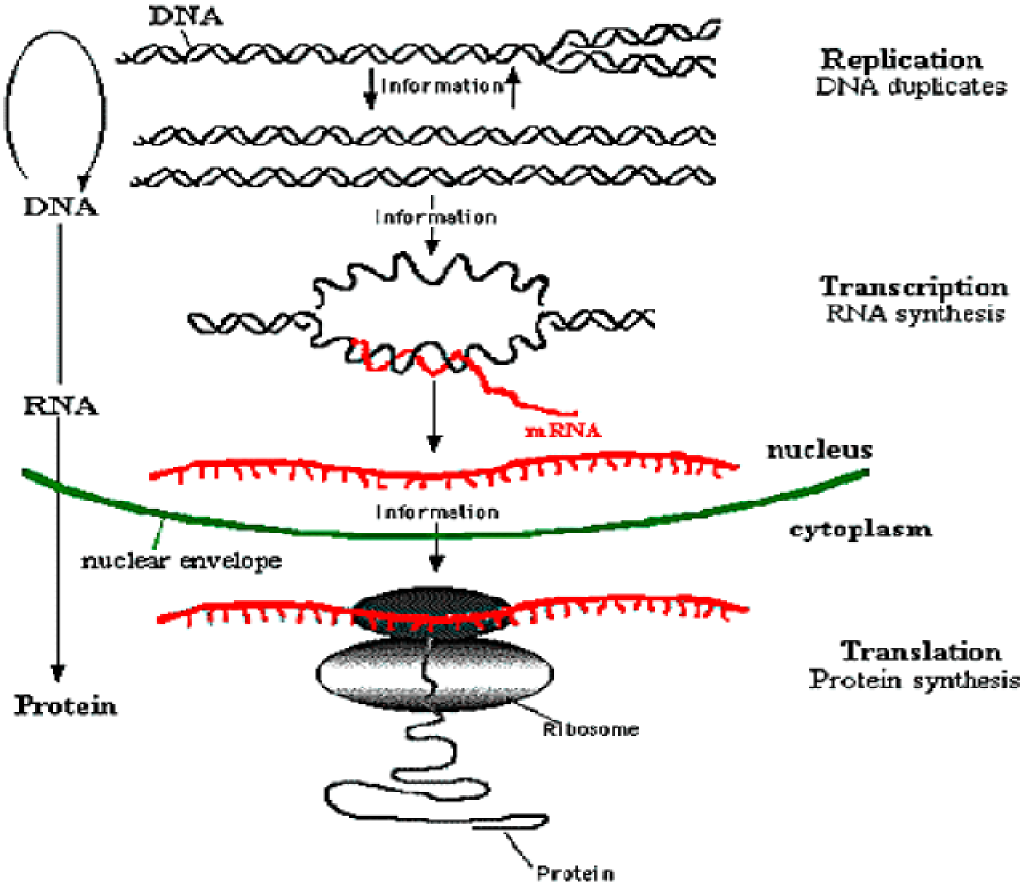


Figure 1: Gene expression.

There is one-to-one correspondence between the nucleotides used to make ribonucleic acid (G, A, U and C where ―U is an abbreviation for Uracil) and the nucleotide sequences in deoxyribonucleic acid (G, A, T, and C, respectively). The process of converting that information from nucleotide sequences in ribonucleic acid to the amino acid sequences that make a *protein* is called translation and is performed by a complex of proteins and called ribonucleic acid ribosomes. Finding the particular beginning of genes for transcription is done by ribonucleic acid polymerase and that beginning sequence is known as Promoter Sequences. In case of prokaryotic genomes the promoter sequences are easy to find as compared to those in eukaryotic genomes. The problem of recognizing eukaryotic genes in genomic sequences data is a major challenge for bioinformatics. The best methods used are neural network and dynamic programming techniques.After finding the longest reading frames, GC content is computed. Most prokaryotic genes are represented only once in the genome. This is not true for eukaryotic genes are present in multiple copies. Eukaryotes have large genomes but low gene density. Some genes have strong and others have weak promoters. Strong promoters have sequences close to the ideal consensus sequences

TTGACA (-35 box) or TATAAT (-10 box) shown in figure 2. So at the least there must be one promoter region upstream of TSS for the polymerase to bind [14].

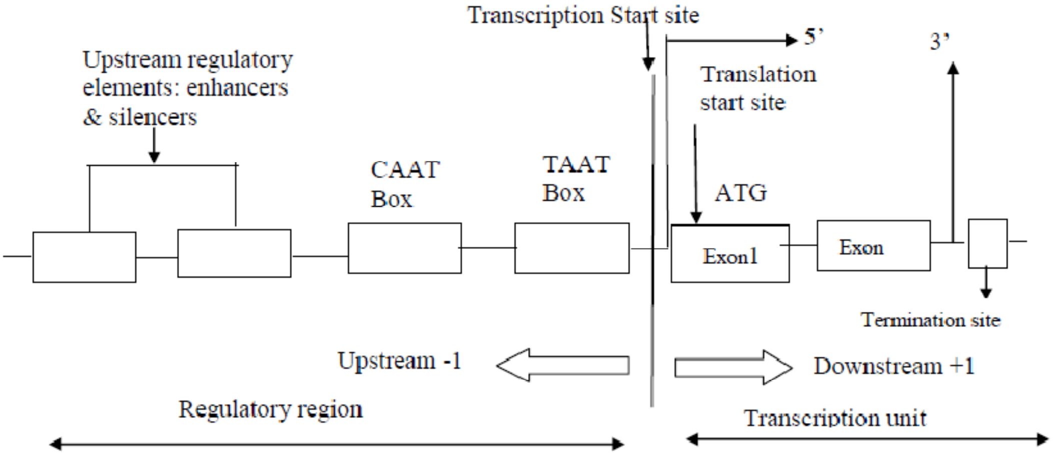
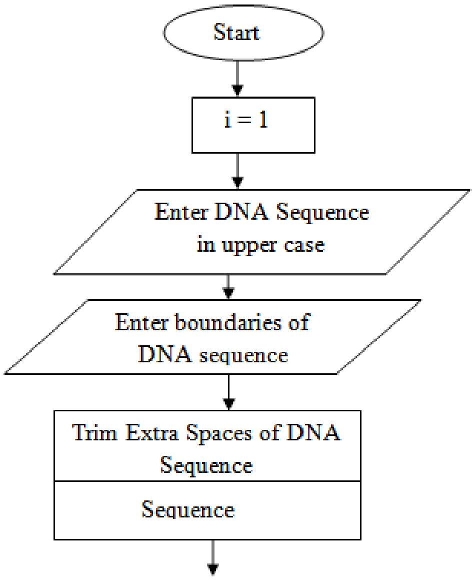


Figure 2: A generalized structure of genes transcribed by RNA polymerase II displaying various structural and functional domains [14].

Design of Model

This computational model implements in two main modules (1) applying the algorithm for ORF Prediction and (2) applying the model for finding the GC content.


Start
i = 1
Enter DNA Sequence in upper case
Enter boundaries of DNA sequence
Trim Extra Spaces of DNA Sequence
Sequence

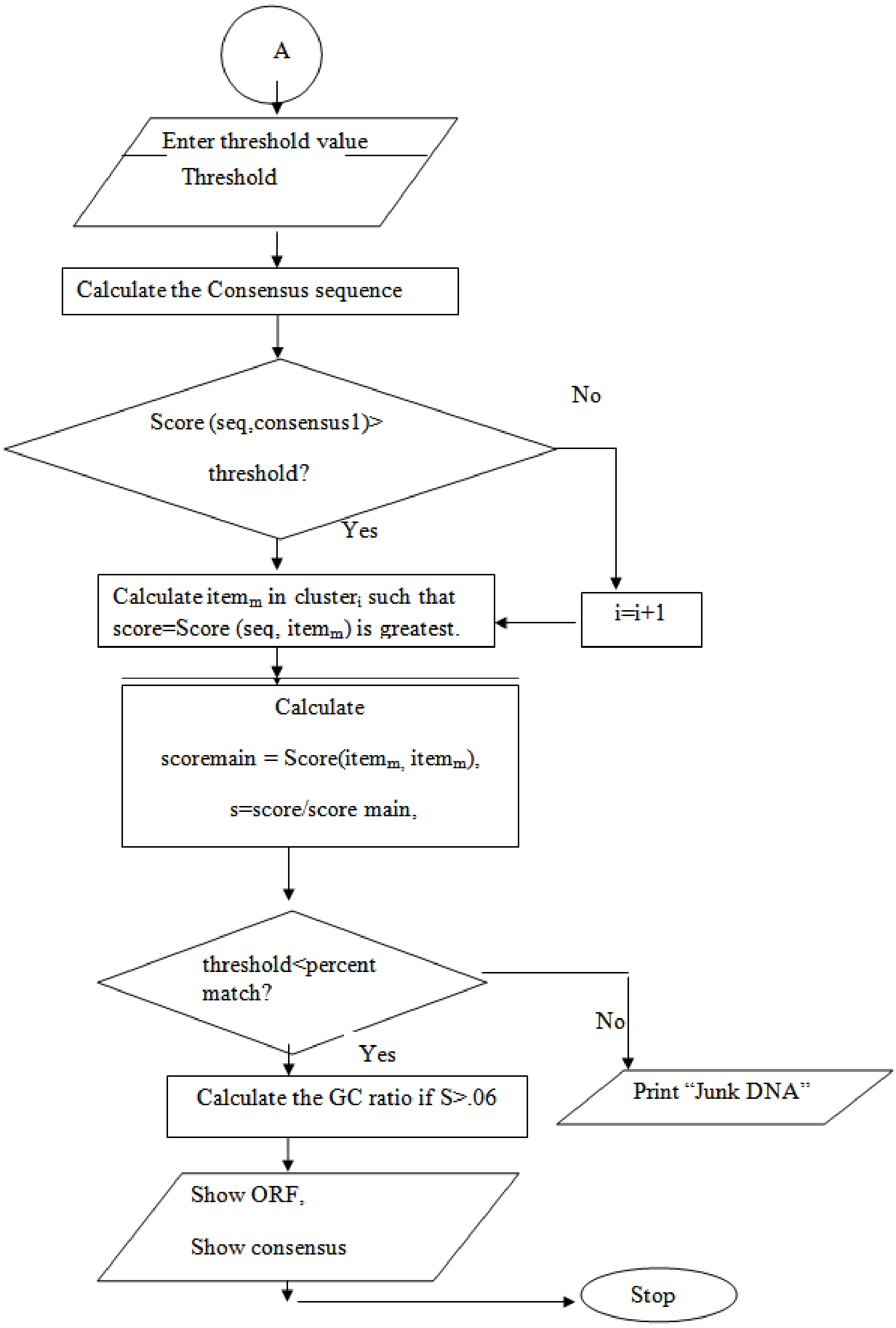

A
Enter threshold value
Threshold
Calculate the Consensus sequence
Score (seq,consensus1)>
threshold?
No
Yes
Calculate item$_m$ in cluster$_i$ such that score=Score (seq, item$_m$) is greatest.
i=i+1
Calculate
scoremain = Score(item$_m$, item$_m$),
s=score/score main,
threshold<percent match?
No
Yes
Calculate the GC ratio if S>.06
Print "Junk DNA"
Show ORF,
Show consensus
Stop

Figure 3: Flowchart of applying clusters for ORF Prediction & GC content

Results and Discussion

We used the two assumptions for computational model mentioned in figure 3, these are inputs data in Genomic sequences and extra spaces present in the input sequence trimmed. Local Alignment score is calculated with the sequences having greatest similarity calculated though the data mining algorithm applied on the consensus sequences. If score >=8 then score of entered DNA sequence is matched with the items of the cluster having similarity>=8 and consensus sequence having greatest score is found out otherwise best similarity with the second cluster is calculated. Then the percentage match of the entered DNA sequence with the consensus sequence is calculated and if that match is greater than or equal to the entered threshold value then various outputs are displayed.

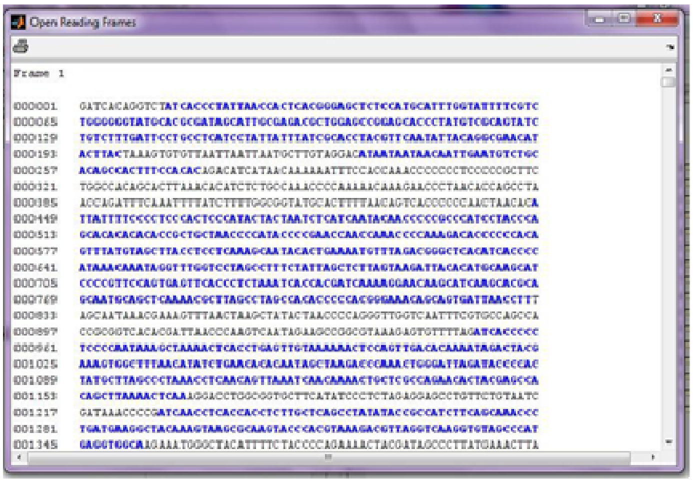
The clustering algorithm that works at back end is similar to the single link technique. With this serial algorithm, items are iteratively merged into existing clusters that are closest (in terms of similarity). In this algorithm a threshold, is used to determine if items will be added to existing clusters or if a new cluster is created. The basis for making clusters is local alignment between the sequences, as larger the score more the sequences will be similar. So the sequences having score greater than or equal to the threshold value are entered into one cluster and rest of the sequences having score less than the given threshold are entered into second cluster and GC-content percentage is calculated as:

$$((G+C) / (A+T+G+C))*100.$$

It is observed that GC content can vary so dramatically across prokaryotic species with values ranging from 25% to 75% GC. The proposed algorithm works with large data sets and the classification of the sequences is done on the basis of (1) Consensus Sequence, (2) Open Reading frame and (3) GC Content ratio shown in figure 4 & 5.The complexity of this algorithm actually depends on the number of items. For each loop, items must be compared to each item already in a cluster. This is n in the worst case. Thus, the time complexity is O (n^2). Space requirement is assumed to be also $O(n^2)$. That is same as nearest neighbor algorithm. Proposed algorithm is changed form of nearest neighbor algorithm. Changes are based on seeing the characteristics of input data are given in table 1.

Table 1: Comparison of clustering algorithms.

Nearest Neighbor	Partitional	$O(n^2)$	$O(n^2)$	Iterative
PAM	Partitional	O(n2)	O(tk(n-k)2)	Iterative; Adapted agglomerative; Outliers
K-means	Partitional	O(n)	O(tkn)	Iterative, not categorical
DBSCAN	Fixed	$O(n^2)$	$O(n^2)$	Sampling, outliners
MST	Hierarchical/ partitioned	$O(n^2)$	$O(n^2)$	Non incremental



```
Frame 1

000001  GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTC
000065  TGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATC
000129  TGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACAT
000193  ACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGC
000257  ACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTC
000321  TGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTA
000385  ACCAGATTTCAAATTTTATCTTTTGGCGGTATGCACTTTTAACAGTCACCCCCCAACTAACACA
000449  TTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAATACAACCCCCGCCCATCCTACCCA
000513  GCACACACACACCGCTGCTAACCCCATACCCCGAACCAACCAAACCCCAAAGACACCCCCCACA
000577  GTTTATGTAGCTTACCTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACCCC
000641  ATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCAT
000705  CCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCA
000769  GCAATGCAGCTCAAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTT
000833  AGCAATAAACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCA
000897  CCGCGGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
000961  TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGACTACG
001025  AAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGATACCCCAC
001089  TATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAACACTACGAGCCA
001153  CAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGGAGCCTGTTCTGTAATC
001217  GATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATACCGCCATCTTCAGCAAACCC
001281  TGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAGACGTTAGGTCAAGGTGTAGCCCAT
001345  GAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAGAAAACTACGATAGCCCTTATGAAACTTA
```

Figure 4: Open Reading Frame.

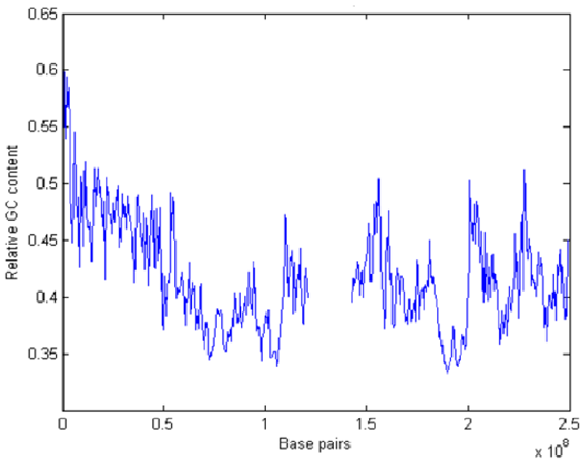


Figure 5: G-C Density.

Conclusions

In the field of Gene Prediction, the sequences of nucleotide in DNA molecules have important information contents of a cell. The information in DNA sequences is used to make single-stranded RNA sequence which in turn will further convert into Protein sequence. This designed computational model first find the open reading frames, based upon which the model distinguish between gene and non-genes and GC density is calculated, based upon these parameters the sequence is classified. This model saves the implementation time, as whole of the database is present online; the sequence to be predicted is just taken from any of the online available databases. Interface is opened and deoxyribonucleic acid sequence is entered in its FASTA format. All the complexities such as calculating GC content and locating open reading frames are computed by algorithm. Several experiments have been done where the parameters selected from classification changed manually. The global error was then estimated to about 10%. In general this error is too high. The performance has been tested on different unknown DNA sequences found on the internet.

References

[1] Ali Al.Shahib, Rainer B. and David R Gilbert. (2007), "Predicting protein function by machine learning on amino acid sequences – a critical Evaluation", BMC Genomics, 8:78, pp 1-10.

[2] Al-Shahib, A., Breitling, R. and Gilbert, D.(2005), "Feature selection and the class imbalance problem in predicting protein function from sequence", Applied Bioinformatics,.4, pp 195-203.

[3] Altschul, SF., Madden, TL., Schaffer, AA., Zhang, J., Zhang, Z., Miller W. and Lipman, DJ. (1997), vol.17, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research,. 25, pp.3389–3402.

[4] Au, W.H., Chan, K.C.C. and Yao, X. (2003), "A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction", IEEE Trans. Evolutionary Computation, 7(6), pp. 532-545.

[5] Baker, D. and Sali, A. (2001), "Protein structure prediction and structural genomics", Nucleic Acids Research, 30(1) pp. 93-96.

[6] Brunak, S., Engelbrecht, J. and Knudsen, S. (1991), "Prediction of human mRNA donor and acceptor sites from the DNA sequence", Journal of Molecular Biology, 220, pp. 49-65.

[7] Cadida, Ferreira. (2001), "Gene Expression Programming: A new adaptive algorithm for solving problems",Complex System Journal, 13(2), pp 87-129.

[8] Chris, B. and Samuel, K. (1997),"Prediction of Complete Gene Structures in Human Genomic DNA", Journal of Molecular Biology. 268, pp 78-94.

[9] Clare, A. (2003), "Machine learning and data mining for yeast functional genomics", Ph.D. thesis, University of Wales

[10] Claverie, J.M. (1997), "Computational Methods for the Identification of Genes in vertebrate Genomic Sequences", Journal Human Molecular Genetics, 6, pp 1735- 1744.

[11] Jiang, D., Tang, C. and Zhang, A. (2004), "Cluster Analysis for Gene Expression Data: A Survey", IEEE Transaction on Knowledge and Data Engineering, 16(11),pp. 1370-1386.

[12] Lakshmi, K. M and Steven, G. S. (2004), Department of Bioinformatics and Computational Biology George Manson university, Lecture- Bioinformatics tools and applications, book reference

[13] Myburgh, G. (2005), " Euokaryotic RNA Polymerase II start site detection using artifical neural networks", M.Tech thesis, University of Pretoria.

[14] Vladimir, Makarov. (2002), "Computer programs for eukaryotic gene prediction", Henary Stewart Publications 3(2), pp 195-199.