IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 4, Aug-Sept, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

An Innovative Gene Identification and Classification Method by Bidirectional String Matching Algorithm and SVM Classifier

Debashis Ghosh[1], Ankur Mondal[2]

[1]Department of CSE, AITS,
Haldwani,Naninital,Uttrakhand,India

[2] Department of CSE, GNIT,
Kolkata, India

Abstract

The major research efforts in Bioinformatics include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution. Identification of genes is an important problem in Bioinformatics. But since biological data include massive information regarding genomic sequences of different species, changes due to evolution, and changes in their protein sequences, it requires systematic sieving of the data to categorize and catalogue them. Automatic identification of genes has been an actively researched area of Bioinformatics. Compared to earlier attempts for finding genes, the recent techniques are significantly more accurate and reliable. Many of the current gene finding methods employs computational intelligence techniques that are known to be more robust when dealing with uncertainty and imprecision. Here the Bidirectional String Matching Algorithm and SVM classifier classifies a number sample DNA sequences into different groups.

Keywords: Bidirectional String Matching Algorithm, DNA, Gene Identification, Support Vector Machine (SVM) Classifier.

1. Introduction

An overview of some of the important aspects of the DNA molecule from the signal processing view point can be found in the introductory magazine-article by Anastasia. The four bases or nucleotides attached to the sugar phosphate backbone are denoted with the usual letters A, C, G, and T (respectively, adenine, cytosine, guanine, and thymine). The forward genome sequence corresponds to the upper strand of the DNA molecule, and in the example shown this is ATTCATAGT. The ordering is from the so-called 5/ to the 3/ end (left to right). The complementary sequence corresponds to the bottom strand, again read from 5/ to 3/ (right to left). This is ACTATGAAT in our example. DNA sequences are always listed from the 5/ to the 3/ end because, they are scanned in that direction when triplets of bases (codons) are used to signal the generation of amino acids. Typically, in any given region of the DNA molecule, at most one of the two strands is active in protein synthesis (multiple coding areas, where both strands are separately active, are rare). The genes are responsible for protein synthesis. Even though all the cells in an organism have identical genes, only a selected subset is active in any particular family of cells. A gene can be treated as a sequence made up from the four bases, can be divided into two sub-regions called the exons and introns. (Procaryotes, which are cells without a nucleus, do not have introns). Only the exons are involved in protein-coding. The bases in the exon region can be imagined to be divided into groups of three adjacent bases. Each triplet is called a codon. Evidently there are 64 possible codons. Scanning the gene from left to right, a codon sequence can be defined by concatenation of the codons in all the exons. Each codon (except the so-called stop codon) instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein.

2. Gene Identifying Classical Approaches:

There are several approaches built over the decades to identify the Gene. The recent development in Bi-informatics most used classical approaches are as follows;

2.1 HMM Algorithms:

The Viterbi and Expectation Maximization (EM) algorithms are used for computing with HMM during its training and testing.

2.2 Dynamic Programming:

The use of dynamic programming in gene finding is briefly reviewed in. The dynamic programming algorithm is a well -established recursive procedure for finding the optimal

(e.g., minimal cost or top scoring) pathway among a series of weighted steps.

2.3 Bayesian Networks:

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis.

3. Gene Identifying Computational Intelligence Approaches:

Different types of computational approaches are also very popular research methodology used in gene identification. Those approaches are as follows:

3.1 Case Based Reasoning (CBR):

Case-based reasoning (CBR) is a model of reasoning where the systems 'expertise is embodied in a library of past cases (stored as a case base) already experienced by the system, rather than being encoded explicitly as rules, or implicitly as decision boundaries.

3.2 Neural Networks:

Artificial Neural Networks (ANNs) are computer algorithms based loosely on modeling the neuronal structure of natural organisms. They are stimulus-response transfer functions that accept some input and yield some output.

3.3 Decision Trees:

Decision tree algorithms are important, well -established machine learning techniques that have been used for a wide range of applications, especially for classification problems. Decision trees have been found to accurately distinguish between coding and non-coding DNA for sequences as short as 54 bp.

3.3 Genetic Algorithms:

A genetic or evolutionary algorithm, first proposed by J. H. Holland, applies the principles of evolution found in nature to finding an optimal solution to an optimization problem. In a genetic algorithm (GA) the problem is encoded in a series of bit strings that are manipulated by the algorithm; in an evolutionary algorithm the decision variables and problem functions are used directly.

4. Problem definitions:

In the post-genome era, efforts are focused on biomarker discover and the early diagnosis of cancer through the application of various omics technologies - transcriptomics, proteomics, metabonomics, peptidomics, glycomics, phosphor proteomics or lipidomics - on tissue samples and body fluids. No matter which omics technology is used in biomarker development, bioinformatics tools are required to extract the diagnostic or prognostic information from the complex data. Based on pattern recognition technologies, discriminatory patterns (a panel of gene, protein or peptide patterns) can be identified for the diagnosis of persons with and without cancer. In Efficient Technique for Gene Identification problem, we are given the data sets of samples for both of normal human DNA sequence and cancer affected DNA sequence. The DNA sequence consists of manly four bases or nucleotides attached to the sugar phosphate backbone are denoted with the usual letters A, C, G, and T (respectively, adenine, cytosine, guanine, and thymine). DNAs are differed with the variation of the sequences of the A, C, G, and T. Now it is here to analyze and pictorially represent the unknown sample of DNA sequence. If any unknown sequence is similar with the normal human DNA sequence then we consider that this sequence is not cancer affected. Otherwise the known sequence is similar with cancer affected DNA sequence, then this sequence can be considered as cancer affected.

4.1 Proposed method:

The goal of this work is to use supervised learning to classify and predict cancer, based on the gene expressions sequence databases. Known sets of data will be used to train the machine learning protocols to categorize the DNA according to their sequence analysis result. The outcome of this study will provide information regarding the efficiency of the machine learning techniques, in particular a Support Vector Machine (SVM) method. Before using SVM, different codons (i.e. defined combinations of A, C, G, T) the number of occurrences of each codons in a particular DNA sequence, must have to determine. In this work the chromosome sequences are identified which are cancer affected and which are not cancer affected.

4.2 Bidirectional String Matching Algorithm:

```
Begin
m:=|x|; n=|y|;
s:=0; r:=n; i:=1; j:=m; p=0;
for s:=0 to (r-m+1)
do
    while y[s+i-1]=x[i]
    i=i+1;
    end while
    while y[r+j-m]=x[j]
    j=j-1;
    end while
if(i>m) then
    pattern matches occur at text position 's'
    p=p+1;
end if
if(j<0) then
    pattern matches occur at text position 'r-m+1'
    p=p+1;
end if
```

4.3 Complexity Analysis of the Algorithm:

In the worst case, the algorithm requires n numbers of searches in the given text for the patterns (codons) from both end, where the length of the text is equal to n. Thus the time taken in the worst case is proportional to n and m, where m is the length of the pattern.

Let Pi is the probability of finding a pattern in a text in average case. The average length of search for a text of n elements of character is given by:

t(n)=1*P(1)+2*P(2)+3*P(3)+….+n*P(n)

where, the length of pattern is m and the searching is done for m times.

Hence the average searching efficiency function becomes:

t(n)*m=[1*P(1)+2*P(2)+3*P(3)+….+n*P(n)]*m

=(1+2+3+4+….+n)*Pa *m,

where, P=1=P(1)+P(2)+P(3)+……+P(n).

[Considering Pa =P/n average probability]

=m**[n(n+1)/2]

=m(n+1)/2

➔O(m)*O(n/2)

Implies Time complexity of order O(mn).

The average case time complexity depends on the probability of finding of the pattern in text characters. In case of unsuccessful matching, a mismatch will be detected much earlier than all m comparisons, which will reduce the average case complexity of this algorithm. The average performance of the algorithm will be much better if the alphabet set Σ contains the large number of symbols. In average case, the time complexity order is linear.

In the best case, the both while loops in the algorithm will execute in the order of O(m). It requires nearly n/2 times for loop execution in the next of length m. In the best case the pattern may be found at the start position as well as at the end position of the text, and the pattern present without being overlapped. Thus, it requires n/2 and O(m) for the loop comparisons.

In this case, the efficiency function is as follows:

m+1*P(1)+2*P(2)+3*P(3)+……+n/2*P(n/2)

=m+(1+2+3+….+n/2)* Pa

[Consider Pa =P/(n/2) average probability]

=m+[(n/2*(n/2+1)/2]*1/(n/2)

=m+(n+2)/4

➔ O(m)+O(n/4)

Implies Time complexity of order O(m+n)

Using the above algorithm we write a code for create a sample data matrix of numeric values of order n*64 (n is the number of sequences) in a text file for a number of sample DNA sequence. This matrix can be used further.

4.4 Support Vector Machine Classification:

Classification of data into two groups: Group = svmclassify (SVMStruct, Sample) classifies each row of the data in Sample using the information in a support vector machine classifier structure SVMStruct, created using the svmtrain function. Sample must have the same number of columns as the data used to train the classifier in svmtrain. Group indicates the group to which each row of Sample has been assigned. Support vector machine classifier structure SVMStruct, can be created in this way:

SVMStruct = svmtrain (chr_train, type)

where _chr_train' is matrix of training data, where each row corresponds to an observation or replicate, and each column corresponds to a feature or variable and 'type' is column vector, character array, or cell array of strings for classifying data in 'chr_train' into two groups. It has the same number of elements as there are rows in 'chr_train'. Each element specifies the group to which the corresponding row in 'chr_train' belongs. Here the 'type' column contain to group one is 'Cancer' and another is 'Normal'. Each row of 'Group' specify the corresponding sequence of the 'Sample' data (i.e. test data).

4.5 Plotting of svmClassify data: For this purpose we follow the following steps.

1. Load the Sample data, which includes DNA sequence data of 64 measurements on a sample of 70 sequences.

load Sample; [Here a matrix Sample of 70x64 is create]

2. Create data, a two-column matrix containing sepal length and sepal width measurements for 70 sequences.
Data1 = [Sample (:, 1), Sample(:,2)];

3. From the type vector, create a new column vector, groups, to classify data1 into two groups: Cancer and Normal. groups = ismember (type,'Cancer‘);

4. Randomly select training and test sets.
[train, test] = crossvalind ('holdOut', groups);
cp = classperf (groups);

5. Use the svmtrain function to train an SVM classifier using a linear kernel function and plot the grouped data.
svmStruct = svmtrain (data1 (train,:), groups (train), 'showplot', true);

6. Classify the test set using a support vector machine.
classes = svmclassify (svmStruct, data1(test,:), 'showplot', true);

7. Evaluate the performance of the classifier.
classperf(cp,classes,test);

Here performance is measured by CorrectRate. In this experiment the value of the CorrctRate is 0.8285714285.

5. Result Analysis:

5.1 Graphical representation of known DNA sequence:

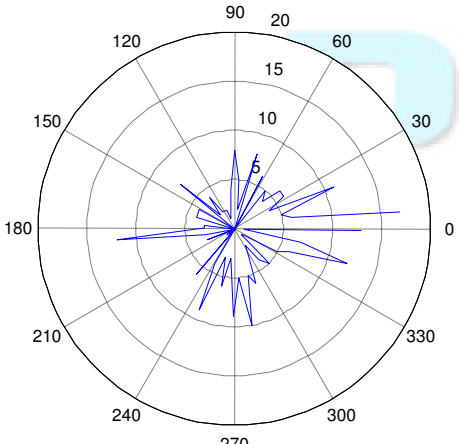

Figure 1: Cancer affected DNA structure

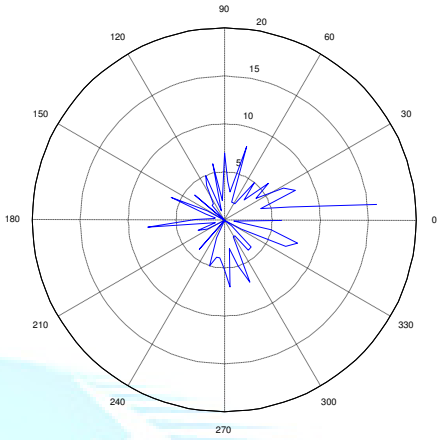

Figure 2:Cancer affected DNA structure

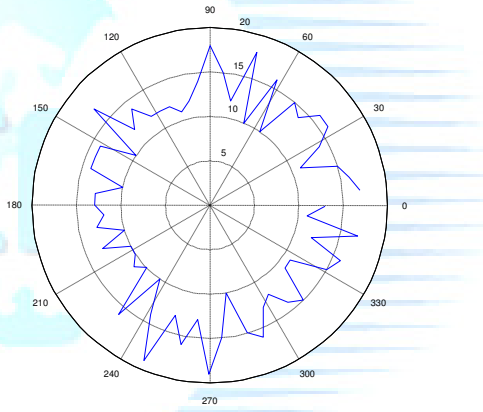

Figure 3: Normal DNA structure

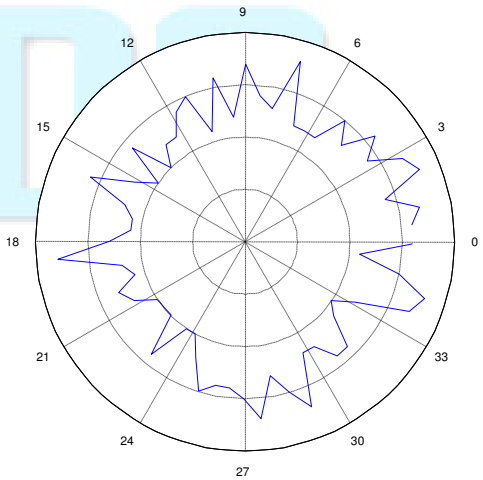

Figure 4: Normal DNA structure

5.2 Result of svmclassify of Sample data:

Group	Group	Group
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Normal'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Normal'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Normal'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Normal'
'Normal'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Normal'	'Cancer'	'Normal'
'Cancer'	'Cancer'	'Cancer'
'Normal'	'Normal'	'Cancer'
'Cancer'	'Cancer'	'Cancer'
'Normal'	'Cancer'	'Cancer'
'Normal'	'Cancer'	'Cancer'
'Cancer'	'Cancer'	
'Cancer'	'Cancer'	
'Cancer'	'Normal'	
'Cancer'	'Cancer'	
'Cancer'	'Cancer'	

5.3 Resulting plot of svmclassify of Sample data:

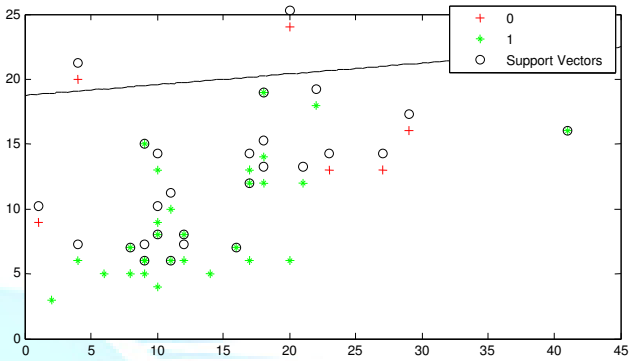

Figure 5: svmclassify plot

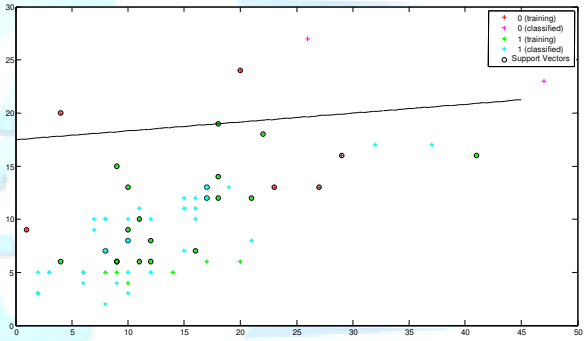

Figure 6: svmclassify plot with training and classified data

6. Conclusions and future work:

The proposed method is able to deal with thousands of examples while combining hundreds of kernels within reasonable time, and reliably identifies a few statistically significant positions. Support Vector Machines (SVMs) – using a variety of string kernels – have been successfully applied to biological sequence classification problems. While SVMs achieve high classification accuracy they lack interpretability. In many applications, it does not suffice that an algorithm just detects a biological signal in the sequence, but it should also provide means to interpret its solution in order to gain biological insight.

Though this work successfully classify the Sample DNA sequence into two groups one is cancer and another is normal, the accuracy can be improved by a large number of training data. This work can be implemented in different applications like separating coding and non-coding regions, identification of introns, exons and promoter regions for annotating genomic DNA, gene product prediction, analysis of genomic sequences of simple

organisms simplifies the problem of understanding complex genomes (its applications are mainly in gene product prediction and forensic analysis), species identification, different diseases identification, specify different type of cancer. Evaluating the accuracy of a microbial gene finder is difficult, because the genes annotated in GenBank do not always have biological evidence to back up their existence. As the annotation becomes more stable, more accurate estimates of accuracy will be possible. At the same time, better gene finders should result because the available training data will improve. Although GLIMMER'S sensitivity is nearing 100% already, there are several important areas of future improvements. One is to improve its specificity by reducing the number of false positives (after first confirming that the unanimated genes found by the system are in fact false).We will be able to apply in the cancer research:

(1) To study the change in tissue specific protein expression

(2) Change in immune response of the proteins in cancerous conditions,

(3) To conduct translational research in different organ confined carcinoma: translating results from the laboratory bench to bed side delivery of patient care & communicating the lessons learned back to the bench.

(4) to develop an ANN (artificial neural network) to distinguish among members of a family of childhood tumors that include neuroblastoma rhabdomyosarcoma, non Hodgkins lymphoma etc.

(5) To develop and implement algorithms that help in differential diagnosis of organ confined tumor. This not only would eliminate a step in our current process, but in addition it would allow us to plot hybridization ratios along a chromosome. The peak of the hybridization ratio would identify the region with the greatest association, and hence the most likely mutant locus.

References

[1].Sanghamitra Bandyopadhyay, Ujjwal Maulik, and Debadyuti Roy, " Gene Identification: Classical and Computational Intelligence Approaches:" Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on (Volume:38 , Issue: 1)

[2].Dr. (Mrs.) Padma and R. Chavan, "Application of Bioinformatics in the Field of Cancer Research".

[3]. P. P. Vaidyanathan and Byung-Jun Yoon.," The role of signal-processing concepts in genomics and proteomics". Journal of Franklin Institute 341(1), 111-135

[4]. Claudio Anselmi, Gianfranco Bocchinfuso, Pasquale De Santis, Maria Savino, and Anita Scipioni, "A Theoretical Model for the Prediction of Sequence-Dependent Nucleosome Thermodynamic Stability:" Rome, Italy.

[5]. Xue-wen Chen, "Margin Based Wrapper Methods for Gene Identification Using Microarray".

[6]. Microbial gene identification using interpolated Markov models: Steven L. Salzberg1,2,*, Arthur L. Delcher3, Simon Kasif4 and Owen White1.1998 Oxford University Press. 544-548 Nucleic Acid Research, 1998, Vol.26,No.2 .

[7]. GISMO—gene identification using a support vector machine for ORF classification:Lutz Krause,* Alice C. McHardy,1 Tim W. Nattkemper, Alfred Pühler, Jens Stoye, and Folker Meyer2. Oxford Journals, Science & Mathematics, Nucleic Acids Research, Volume 35, Issue 2Pp. 540-549.

[8]. Sequence-Dependent Dynamics of Duplex DNA: The Applicability of aDinucleotide Model : T.M. Okonogi*, S.C. Alley*, †, A.W. Reese*, P.B.Hopkins* and B.H. Robinson, Biophysical Journal.

[9]. RAD marker microarrays enable rapid mapping of zebrafish mutations:Michael R Miller,1,2 Tressa S Atwood,1,3 B Frank Eames,2 Johann K Eberhart,2 Yi-Lin Yan,2 John H Postlethwait,2 and Eric A Johnson, *Genome Biology* 2007, **8**:R105 doi:10.1186/gb-2007-8-6-r105 Published: 6 June 2007

[10]. Microbial gene identification using interpolated Markov models: Steven L. Salzberg1,2,*, Arthur L. Delcher3, Simon Kasif4 and Owen White1. 1998 Oxford University Press. 544–548. Nucleic Acids Research, 1998, Vol. 26, No. 2.

[11]. An efficient Bi-directional String Matching Algorithm for Statistical Estimation: Chhanda Ray, Satish Tripathi, Arijit Chattejee, Anupam Das.

[12]. Gene Classification using Codon Usage and SVMs: Ma, J. Nguyen, M.N. Pang, G.W.L. Rajapakse, J.C. IEEE/ACM Transactions on Computational biology and Bioinformatics, vol. 6, no. 1, January-March 2009

[13]. Di-codon Usage for Gene Classification :Minh N. Nguyen, Jianmin Ma, Gary B. Fogel and Jagath C. Rajapakse. 4th IAPR International Conference, PRIB 2009, Sheffield, UK, September 7-9, 2009.