ISSN No. 0976-5697

Volume 4, No. 5, May 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

Genetic Algorithm – A Case Study in Gene Identification

V. Bhaskara Murthy
Sr. Asst. Professor, Padmasri Dr. BVRICE
Vishnupur, Bhimavaram, W.G.Dt. A.P.
Email: murthyvb@gmail.com

Dr. G. Pardha Saradhi Varma
Professor & Director of PG Courses
Head, Department of IT, S.R.K.R. Engineering College,
Chinamiram, Bhimavaram., W.G.Dt. A.P.
Email: gpsvarma@yahoo.com

Abstract: Gene prediction has been an interesting area of research in Bioinformatics. Many of the recent gene identification methods adopt different approaches which are more robust when dealing with uncertainty and ambiguity. In this paper a detailed survey of methods are discussed.

Index Terms: *Content-Based Reasoning, Decision Tree, Neural Networks, Genetic Algorithms*

I. INTRODUCTION

A large amount of raw sequence data generated because of advancement in sequencing technology requires biological interpretation in an effective and optimal way. This is known as annotation. Although the Human Genome Project was completed in April, 2003, the exact number of genes encoded by the human genome is still unknown and estimated genes are in the range of 20,000-25,000. The steps involved in genome annotation are classified as three categories.

1. Gene Identification (Nucleotide level)
2. Structure determination of proteins. (Protein level)
3. Mechanism of biochemical reactions. (Process level)

Among three, nucleotide-level annotation is a primary step in molecular biology[1]. Only 80% of genes are accurately predicted at the nucleotide level , 45% are predicted at the exon level and nearly 20% at the whole genome level. Even if all human genes are experimentally determined, it would still be important to know how the structures of genes are organized and defined, and how they can be recognized.

II. PROBLEM DEFINITION

In this section, some basic terminology related to the problem of gene prediction is given. The problem is then stated formally.

Basic Terminology:

Gene: Gene is defined as a segment of DNA that contains the necessary information to produce a functional product, usually a protein.

Promoter: Promoter is the regulatory region of DNA located upstream of a gene. It provides a control point for regulated gene transcription.

Core Promoter: It is the minimal portion of the promoter required to initiate transcription properly. It serves as a binding site for RNA polymerase and general transcription factors.

Proximal Promoter: The proximal sequence upstream of the gene that tends to contain regulatory elements is known as proximal promoter. It serves as a binding site for specific transcription factors.

OPEN READING FRAME(ORF): An ORF is a sequence of DNA that starts with a start codon normally 'ATG' and ends with one of three codons TAA, TAG, TGA.

Coding sequence: Coding sequence(CDS) is the actual region of DNA that is translated to form proteins[2].

Gene Prediction: The characterization of genomic features using computational and experimental methods is called gene prediction.

The following information is required:

1. Coding region for a protein
2. The DNA strand is used to encode the gene
3. The start and end positions of the gene
4. The exon-intron boundaries in eukaryotes
5. The regulatory sequences for that gene

It is still difficult to predict genes accurately due to:

- DNA sequence have low information content
- Difficult to discriminate real signals.
- Contain sequencing errors.
- Short genes are found in prokaryotes that have little information
- The presence of alternate splicing mechanism in eukaryotic genes makes its detection difficult.

Computational gene finding is a process of the following:

- Identifying common phenomena in known genes.
- Building computational model that can accurately describe the common phenomena.
- Using the model to scan an uncharacterized sequence to identify regions that match the model.
- Test and validate the predictions.

III. NOVEL INTELLIGENCE COMPUTER TECHNIQUES FOR GENE PREDICTION

Novel Intelligence computational techniques like artificial intelligence techniques genetic algorithms, neural networks, which are top-to-bottom. These techniques differ from traditional and logic-based methods.

CONFERENCE PAPER
"National Conference on Networks and Soft Computing"
On 25-26 March 2013
Organized by
Vignan University, India

Case-Based Reasoning

Case-based reasoning has been formalized for purposes of computer reasoning as a four-step process[3] 4R's.

1. **Retrieve:** Given a target problem, retrieve from memory cases relevant to solving it.
2. **Reuse:** Map the solution from the previous case to the target problem.
3. **Revise:** Having mapped the previous solution to the target situation, test the new solution and, if necessary, revise.
4. **Retain:** After the solution has been successfully adapted to the target problem, store it as a new case in memory.

```
... G T|A G C C G A A T C|G ...      Target Sequence
       |A C G A A G A T C|           Case Exon

... G T|A - G C C G A A T C|G ...
       |A C G A A G - A T C|
          d   s s   i
          1   1 1   1        = 4    Edit Distance
```

Fig. 1. Example edit distance computation.

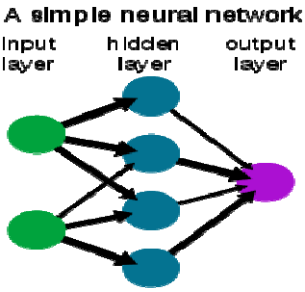
Figure 1 shows an example edit distance computation between a case exon and a target DNA sequence segment, with which the case is aligned. The minimal transformation cost between the exon case and the target segment requires a deletion(d) from the target, two substitutions (s), and an insertion (i) in the case, giving an edit distance of 4. This type of similarity is also conceptually appealing, as it computes similarity using adaptability.

CBR employs a case library of nucleotide segments that have been categorized as coding (exon) or noncoding (intron), in order to locate the coding regions of a new DNA strand. Costello and Wilson [4] have investigated a number of possible approaches to similarity including longest common subsequence and sequence alignment methods. The CBR framework has been applied to the problem of annotating genes and the regulatory elements their proximal promoter regions.

A database EpoDB[5] is a big database consists information for the genes that are expressed in vertebrate red blood cells. A detailed survey of the applications of CBR in molecular biology, including gene identification, is provided in [6].

Neural Networks:

Artificial neural networks (ANNs) [7] are computer algorithms based on modeling the neuronal structure of natural organisms. In general if given sufficient complexity, there exists an ANN that will map every input pattern to its appropriate output pattern, as mapping is not one-to-many.
Figure2:



If the input data are not linearly separable, a least means square solution is generated to minimize the means square error between the calculated output of the network and the actual desired output.

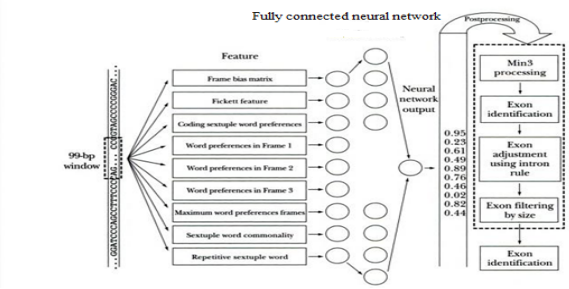


Figure 3: Gene identification.

A 99-base pair window is interrogated for a prediction of coding/noncoding on the central nucleotide. Input features are fed to the neural network, which provides an output value between -1 (noncoding) and +1(coding). Postprocessing features are used to make identification of a coding region (exon) boundary.

Well known GRAIL software used an ANN to combine a number of coding indicators calculated within a fixed sequence window[8] . The classification process using evolved ANNs proceeded as follows:

A sequence of DNA was interrogated using a window of 99 nucleotides. The ANN was used to classify the nucleotide in the center of the window as either coding or noncoding. For this analysis, the neural network architecture was fixed and consisted of 9 input nodes corresponding to 9 features as shown in fig:3, 14 hidden nodes and one output node. The output decision was normalized on from -1 to +1 for each position in the sequence. If the output was less than -0.5 it was classified as coding, if it was less than +0.5 classified it as noncoding.

In evolved ANN, genetic algorithms have been used for determining the appropriate network architecture [9]. ANNs are combined with a rule-based system has been used for splice site prediction in human using a joint prediction scheme for local splice site assignment.[10]

Decision Tree

Multi-frame Optimal Rule-based Gene Analyzer(MORGAN) is a approach based on decision tree classifiers, signal recognition algorithms and dynamic programming. It is highly modular allowing improvements in any one aspect of the gene-finding task to be incorporated relatively into the system.

The decision tree classifier by Quinlan [11] is one of machine learning techniques. A decision tree is made of decision nodes and leaf nodes. Each decision node corresponds to a test X over a single attribute of the input

data and has a number of branches, each of which handles an outcome of the test X. Each leaf node represents a class that is the result of decision for a case.

General Basic Tree Construction Procedure:

Let S = $\{(x_1,c_1), (x_2,c_2),\ldots\ldots\ldots\ldots (x_k,c_k)\}$ be a training sample. Constructing a decision tree from S can be done in a divide –and-conquer fashion as follows:

1: If all the examples in S are labeled with the same class, return a leaf labeled with that class.

2: Choose some test t that has two or more mutually exclusive outcomes $\{ o_1, o_2, \ldots.. o_r \}$.

3: Partition S into disjoint subsets S_1 , S_2 , …… S_r such that S_i consists of those examples Having outcome o_i for the test t. for i =1,2,…..r.

4: Call tree construction procedure recursively on each of the subset $S_1, S_2, \ldots. S_r$ and let the decision trees returned by these recursive calls be $T_1, T_2, \ldots\ldots. T_r$.

5: Return a decision tree T with a node labeled t as the root and the trees $T_1, T_2, \ldots\ldots. T_r$ as sub trees below that node.

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 noncoding DNA for sequences as short as 54bp. MORGAN, an integrated system for finding genes uses oblique decision tree system for solving the problem of discriminating coding and noncoding DNA. The general purpose of the entire procedure [12] consists on finding several partitions of the plane.

Genetic Algorithms

A genetic or evolutionary a The Genetic Algorithm (GA) was invented in the mid-1970s by John Holland[13]. It is based on Darwin's Evolution Theory. GA uses the concept of survival of the fittest and natural selection to evolve a population of individuals over many generations by using different operators: selection, crossover, and mutation. Genetic Algorithm can be used for optimization problems with multiple parameters and multiple objectives. It is commonly used to tackle NP-hard problems such as the DNA fragment assembly and the Travelling Salesman Problem (TSP). NP-hard problems require tremendous computational resources to solve exactly. Genetic Algorithms help to find good solutions in a reasonable amount of time.

The basic GA can be outlined as follows:

1. Generate random population of n chromosomes. i.e. suitable solutions for the problem.

2. Evaluate the fitness f(x) of each chromosome x in the population.

3. New population: Create a new population by repeating following:

(a) Selection: Select two parent chromosomes according to their fitness.

(b) Crossover: With a probability crossover, the parents to form new offspring (children).

(c) Mutation: With a probability mutate new offspring at each position.

(d) Fitness: Evaluate the fitness f(x) of each chromosome x in the new population.

4. If the end condition is satisfied, then stop and return the best solution in the current population .

5. Go to step 3.

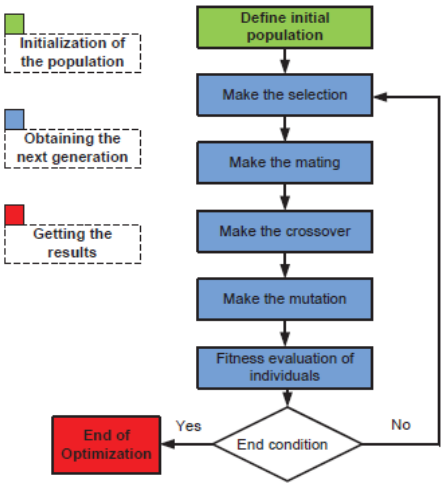


Fig. 4. Basic flow diagram of genetic algorithm

Population Representation

A permutation of integers represents a sequence of fragment numbers, where successive fragments overlap. The population in this representation requires a list of fragments assigned with a unique integer ID. In order to maintain a legal solution, the two conditions that must be satisfied are (1) all fragments must be presented in the ordering, and (2) no duplicate fragments are allowed in the ordering. For example, one possible ordering for 4 fragments is 3 0 2 1. It means that fragment 3 is at the first position and fragment 0 is at the second position, and so on. Use a fixed size population to initialize random permutations.

Fitness Function

A fitness function is used to evaluate how good a particular solution is. It is applied to each individual in the population and it should guide the genetic algorithm towards the optimal solution.

Mutation Operator

This operator is used for the modification of single individuals. The reason need a mutation operator is for the purpose of maintaining diversity in the population. Mutation is implemented by running through the whole population and for each individual, deciding whether to select it for mutation or not, based on a parameter called *mutation rate* (*Pm*).

Selection operator

The purpose of the selection is to weed out the bad solutions. It requires a population as a parameter, processes the population using the fitness function, and returns a new population. The level of the selection pressure is very important. If the pressure is too low, convergence becomes very slow. If the pressure is too high, convergence will be premature to a local optimum.

Crossover Operator

Two or more parents are recombined to produce one or more offspring. The purpose of this operator is to allow partial solutions to evolve in different individuals and then

combine them to produce a better solution. It is implemented by running through the population and for each individual, deciding whether it should be selected for crossover using a parameter called *crossover rate* (*Pc*). A crossover rate of 1.0 indicates that all the selected individuals are used in the crossover.

GA based method always converges to a good solution fast, since it is able to take advantage of the extra information to construct good local maps that can be then used to construct good global maps.

IV CONCLUSION

The identification of genes is an important problem in bioinformatics. Several methods of gene identification i.e. Hidden Markov Model, Dynamic Programming have been developed in the past. Given the difficulty of the problem, computational intelligence based methods have also been applied in recent times because of their robustness and ability to handle a noisy and incomplete data. This paper provides a comprehensive review of the different methods of gene identification, with special emphasis on computational intelligence techniques.

V ACKKNOWLEDGEMENT

The authors gratefully acknowledge the informative comments of the reviewers that will help in improving the quality and new sources of knowledge.

VI. REFERENCES

[1]. A.M. Campbell and L. J. Hyer, Discovering Genomics, Proteomics and Bioinformatics. Pearson Education, 2004

[2]. Role of 5_- and 3_-untranslated regions of mRNAs in human diseases Sangeeta Chatterjee and Jayanta K. Pal, University of Pune, Pune 411007, India Biol. Cell (2009) 101, 251–262 (Printed in Great Britain) doi:10.1042/BC20080104

[3]. Agnar Aamodt and Enric Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communications* 7 (1994): 1, 39-52.

[4]. Case-Based Approach to Gene Finding
Edwin Costello and David C. Wilson,Computer Science Department, University College Dublin, Dublin 4, Ireland edwin costello@hotmail.com,david.wilson@ucd.ie

[5]. Website: www.cbil.upenn.edu/**EpoDB**/

[6]. I. Jurisica and J.I. Glasgow "Applications of case-based reasoning in molecular biology" AI Mag., Vol 25, no. 1 pp85-96, 2004.

[7]. S. Haykin, Neural Networks:A comprehensive Foundation Pearson Education,2002.

[8]. E.C. Uberacher, Y.Xu, and R.J. Mural, "Discovering and understanding genes in human DNA sequence using GRAIL" Methods Enzymol, vol. 266, pp259-281, 1996.

[9]. G.B. Fogel, K. chellapilla and D.W. Corne, "Identification of coding regions in DNA sequences using evolved neural networks": Morgan Kaufmann, 2002, pp 195-218.

[10]. S.M. Hebsgaard, P.G.Kornig, N. Tolstrup, J. Engelbrecht, P. Rouze, and S. Brunak " Splice site prediction in Arabidopsis thaliana pre mRNA by combining local and global sequence Information." Nucleic Acids Res. Vol. 26, no. 1 pp 51-56, 2006.

[11]. Quinlan, J. R. (1993). *C4.5, Programs for Machine Learning*. Morgan Kaufmann San Mateo Ca, 1993.

[12]. Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk-Jean Gaudart**,** Belco Poudiougou**,** Stéphane Ranque and Ogobara Doumbo.

[13]. J. H. Holland. *Adaptation in Natural and Arti¯cial Systems*. The University of Michigan Press, Ann Arbor, Michigan, 1975.