© 2015 IJSRST | Volume 1 | Issue 4 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science

Bioinformatics Approaches for Gene Finding

Mahin Ghorbani [1]*, Hamed Karimi [2]
[1]Department of Biotechnology, Fergusson College, F.C. Road, Pune, Maharashtra, India
[2]Department of Information Technology, Payam noor university of Farokh-shahr , Farokh-shahr, Chaharmahl va bakhtiari, Iran

ABSTRACT

Gene finding as process of identification of genomic DNA regions encoding proteins , is one of the important scientific research programs and has vast application in structural genomics ,functional genomics ,metabolomics, transcriptomics, proteomics, genome studies and other genetic related studies including genetics disorders detection, treatment and prevention .It is prominent that for study of all above mentioned research programs , identification of fundamental and essential elements of genome such as functional genes, intron, exon, splicing sites, regulatory sites, gene encoding known proteins, motifs, EST, ACR, etc are formed principle basis of the studies and these functions are employed by gene prediction or finding process. So gene finding process plays significant role in the study of genome related programs. Several methods are available for gene finding such as laboratory -based approaches, feature- based approaches homology based approaches, statistical and HMM –based approaches. In this paper, we aim to discuss Insilco approaches for gene prediction in order to make scientist familiar with available bioinformatics tools for gene finding to take benefit from their advantages including low in cost, rapid in time, high in accuracy and large in scale.

Keywords: Gene finding, Gene prediction, bioinformatics tools, genome study, Insilco approaches.

I. INTRODUCTION

The process of identification of genomic DNA regions encoding proteins is defined as gene prediction or gene finding. Gene finding is one of the most significant process in understanding and analysis of an organism's genome after its sequencing. Previously the process of gene finding was relied on effortful experimentations on organisms and living cells which required high expenditure in cost and time but recently with advancement in statistical and bioinformatics tools, these troubles have been decreased. Statistical approaches in combination with computational techniques provide analysis of homologous recombination rates of various genes which lead in determination of their order on a particular chromosome and obtained information from such experiments are helpful in creation a genetic map which help in specification the rough locus of known genes related to each other . Availability of vigorous computational tools aided gene finding to demonstrate significant position is genome studies. Gene finding process is able to distinguish determination of function of gene or its product from a functional gene, although prediction of a gene function and its confirmation still demand *in vivo*, gene knockout experimentations and other assessments, but bioinformatics approaches have great ability to predict the gene function based on its sequence alone. Additionally, gene finding process is able to predict structural genes which are fundamental basis for understanding biochemical process within the cells including transcription. Protein-protein interactions and regulatory system process which contributed in Omics fields' research such as structural genomics, functional genomics, metabolomics, transcriptomics a proteomics. It is clear that for each research process, knowledge of fundamental elements are required , in application gene finding for research studies knowledge of some bases are essential such as , gene definition, gene structure types of genes, differences between types of genes, gene feature and DNA characteristics. In this part we briefly discuss such issues for better understanding gene finding process. A DNA segment expressed for production of a functional product like a protein or RNA is called as a gene. Generally genes structure consist of following parts : *upstream* (intergenic region) , *promoter* (for example , TATA box with consensus sequence TATA(A/T)A(A/T), *first exon*(transcriptional start,5'-UTR), *intron(s)* (frequent stop codons), *exon(s)*(CDS/ORF and enhancer sites), *intron(s)* (frequent stop codons)), *last exon* Transcriptional stop, Poly A insertion sites , *downstream (intergenic region).* Generally there are two types of genes based on organism: prokaryotic and eukaryotic genes which show following features: *prokaryotic genome:* small in size, high gene density, terminator important, no introns (or splicing), no RNA processing, similar promoters, and overlapping genes. *Eukaryotic genome*: large in size, low gene density, terminator not important, presence of introns (or splicing), presence of RNA processing, heterogeneous promoters, polyadenylation. Knowledge of pattern recognition including gene feature and DNA characteristics are also important and prior to applying gene finding process, these are such as coding sequences (open reading frames (ORFs), GC-rich , CpG-content), PolyA-signals ,(consensus sequences), translational start and stop sites(start codons (ATG), stop one(TAA,TAG,TGA), splice sites,(consensus sequences) promoter regions(TATA, shine Dalgarno, Kozak consensus, CpG content, Prinbnow). Totally gene finding methods can be divided into two types: laboratory based approaches and *web* based approaches which itself consist of three types namely: feature –based, homology based and statistical and HHM based approaches. Laboratory based approaches are such as southern blotting, northern blotting methods, Zoo blots, S1 nuclease mapping, Primer extension, Exon trapping, Reverse transcriptase polymerase chain reaction (RT-PCR) and *In situ* hybridization (*ISH*) In this paper we aim to introduce several online database and tools related to feature –based, homology based and statistical and HHM based approaches in order to utilize their benefits in gene prediction processes. [1-4].

II. METHODS AND MATERIAL

Computational tools for gene finding:

CRAIL (Gene Recognition and Analysis Internet Link) (http://compbio.ornl.gov/grailexp/). It is one the most known computational tools mostly used for ORF identification. This tools analyses potential of a DNA sequence for protein coding .The scheme provided by CRAIL is variable –length windows tailoring to each possible ORF by a pair of start, acceptor and stop sites. This scheme provides more genomic context information such as translation starts, splice junctions, non-coding scores of 60 base regions on both sides of putative exon. [5].

FindPatterns (http://www.accelrys.com/products/gcg_wisconsin_package/programlist.html#FindPatterns) Is another computational tools used to scan patterns of ORFs.[1].

Sequencher https://www.genecodes.com/. It is a computational tools used for different analysis purpose such as ORFs analysis, restriction enzyme mapping, contig assembly,cDNA to Genomic DNA large map alignment ,SNP analysis, motif analysis and heterozygous detection.[6]

Mac Vector 6.5 (http://www.sxst.it/oxm_mcv.htm).Used for detection of ORFs based on Fickett's statistical method.[7]

TestCode (http://Www.accelrys.com/products/gcg_wisconsin_package/) is another computational tools used for verification of potential of ORFs to encode a proteins. For example this tool helps in determination of correspondence of the codons in ORFs to those used in other genes of the same organism , possibility of ORFs translation into amino acid sequences and so on.[8].

Procrustes software program (http://hto13.usc.edu/software/procrustes). This program is based on homology and its entity is one genomic DNA sequence and one or more protein sequences. The targets are similar to the protein encoded in the genomic DNA sequence and exon chain is found by this homology based finding .[1

GeneMark http://www.ebi.ac.uk/genemark. Is another computational program based on statistical and HMM approaches used for gene identification.[1,9]

HMMgene (http://www.cbs.dtu.dk/services/HMMgene/.).It is a computational approach based on HMM used for gene prediction in an anonymous DNA.[1,10]

Glimmer {http://www.tigr.org/software/glimmerm.).Prediction of genes in microbial DNA can be performed computationally using this tool. This system also uses IMMs approach for identification of coding regions and their distinguishing from non-coding regions .[11,12]

Veil system (www.tigr.org/~salzberg/veil.html). This program used to identify genes in Eukaryotic DNA based on HMM approach.[1]

GENSCAN (http://genes.mit.edu/GENSCAN.html). It is a computational tool used for prediction of complete gene structures . It also used for identification of exons, introns, polyA signals, promoter sites , etc.[1]

Genie (http://www.fruitfly.org/seq_tools/genie.html). Is a computational program based on generalized HMMs and neural networks.[1]

Signal Scan
(http://www.cbs.umn.edu/software/sigscan.htm). It is another computational program used for finding potential transcription binding sites in DNA sequences using transcription factor sequences database.[1]

GenLang
(http://www.cbil.upenn.edu/genlang/genlang_home.html) It is a linguistic pattern recognition system using tools of computational syntactic for identification of genes in sequence data .[13]

Gene Parser
(http://beagle.colorado.edu/~eesnyder/GeneParser.html). This program used for prediction the most likely combined exons and introns in a genomic sequence . [1]

GeneId
(http://www1.imim.es/software/geneid/index.html). This programs predicts genes in unknown genomic sequences .[1]

HCpolyA
(http://125.itba.mi.cnr.it/~webgene/wwwHC_polya.html). It is another gene prediction tool used for prediction of poly-A sites using Hamming Clustering Methods.[1]

HCtata
(http://125.itba.mi.cnr.it/~webgene/wwwHC_tata.html). This program is used for prediction of TATA signal in Eukaryotic genes using Hamming Clustering Methods.[1]

MatInspector(www.genomatrix.de/cgi-bin/matinspector/ matinspector.pl). It is used for transcription factor identification.[14,15]

Tfsitescan (www.ifti.org/cgi-bin/ifti/Tfsitescan.pl) This tool used for analysis of promoter sequences and mostly work best with sequences of -500nt.[1]

FramePlot (http://watson.nih.go.jp/~jun/cgi-bin/ frameplot-3.0b.pl). This tool used for prediction of protein coding region in bacterial DNA –NIH-NET.
[16].

ORF Finder (www.ncbi.nlm.nih.gov/gorf/gorf.html) . This program used for identification of all open reading frames already available in its database.[1].

Exon Prediction Program – Perceval
(http://compbio.ornl.gov/grailexp/gxpfaq2.html) . It is another program used for prediction of protein coding exon and location of repetitive elements and CpG islands .[1].

Repetitive Elements Identification - RepeatMasker
(http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) This program used for analysis of repetitive elements in DNA sequences .[1].

tRNA gene Prediction
(www.genetics.wustl.edu/eddy/tRNAscan-SE/) It is a computational tools for identification of genomic tRNA.[1].

More tools are available for detection of other gene features such as single nucleotide polymorphism and so on.[17]

III. RESULT AND DISCUSSION

To understand the proteins encoded by the genes and their function and dysfunction, it is very important for scientists in order to analyse the cause of disturbance and find out treatment approaches. Of course for analysing and identification of causing agents for diseases, knowing the structures and specific genes encoding disease involving proteins are essential. One of the branches aids in this issue is gene finding. Gene finding provides applicable tools for researchers for searching genes encoding proteins, These tools are used for identification of different DNA features and corresponding DNA characteristics such as coding sequences (CDS) (ORFs, GC –rich,CpG-content), Splice sites (consensus sequences), Translational start and stop sites(codon start(ATG) and stop (TAA,TGA, TAG), promoter regions (TATA,Pibnow) and Poly A signals. Additionally the gene finding methods provides researchers with tools for identification of different gene structures such as protein binding sites in DNA sequences, exons, introns, combination of exons and introns, prediction of TATA signal, transcription factors, prediction of homologies of signal sequences, promoter sequence analysis, prediction of coding regions, analysis of repetitive elements in DNA sequences, prediction of tRNA and so on. SO understanding above characteristics and elements of encoding genes are essential and basis for the researcher for proper study of genomics, proteomics and other related studies. Although these gene finding approaches based on computational methods are very fast and economics as compared to lab based techniques which have limitation in time , cost and scale but they also have their own limitation in prediction type, sensitivity, specificity, sensitivity of exon and introns, sensitivity of exact exon and missed exons. Recently due to advancement in science and identification of many factors related to disease's emergence and treatment like biomarkers ,targets and drug targets such as ion channels, aquaporins, GPCRs, CDKs, and genes encoding them , these tools are applicable and can be used for identification of genes related to the above elements for better analysis of the factors involve in cause and treatment of diseases specially cancers and take benefits of their advantages such as rapid in time and save in cost , high in accuracy and large in scale. [18-23].

IV. REFERENCES

[1] Rastogi .S. C., Rastogi P and Mendiratta N., 2008. Bioinformatics Methods and Applications: Genomics .Proteomics And Drug Discovery PHI Learning Pvt. Ltd

[2] Korf I. (2004-05-14). "Gene finding in novel genomes". BMC Bioinformatics 5: 59–67.

[3] Neelam Goel, Shailendra Singh, Trilok Chand Aseri (2013). "A comparative analysis of soft computing techniques for gene prediction". Analytical Biochemistry. doi:10.1016/j.ab.2013.03.015.

[4] Mathé C, Sagot MF, Schiex T, Rouzé P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 2002 Oct 1;30(19):4103-17.

[5] Uberbacher EC, Hyatt D, Shah M.GrailEXP and Genome Analysis Pipeline for genome annotation. Curr Protoc Hum Genet. 2004 Feb;Chapter 6:Unit 6.5. doi:10.1002/0471142905.hg0605s39.

[6] Sequencher® version 5.3 sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA http://www.genecodes.com

[7] Neena Haider. Biotech Software & Internet Report. November 2000, 1(5): 208-213. Doi: 10.1089/152791600750034730.

[8] Fickett JW.Recognition of protein coding regions in DNA sequences. TEST CODE Nucleic Acids Res. 1982 Sep 11; 10(17):5303-18.

[9] Lukashin A. and Borodovsky M. "GeneMark.hmm: new solutions for gene finding." Nucleic Acids Research (1998) 26 (4): 1107–1115.

[10] Anders Krogh.Using Database Matches with HMMGene for Automated Gene Detection in Drosophila. Genome Res. 2000 Apr; 10(4): 523–528.

[11] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER, Nucleic Acids Research 27:23 (1999), 4636-4641.

[12] Salzberg S, Delcher A, Kasif S, White O. Microbial gene identification using interpolated Markov models, Nucleic Acids Research 26:2 (1998), 544-548.

[13] Dong, S. and Searls, D.B."Gene Structure Prediction by Linguistic Methods" (1994) Genomics 23:540-551.

[14] Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T (2005)MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21, 2933-42.

[15] Cartharius K (2005) MatInspector: Analysing Promoters for Transcription Factor Binding Sites in "Analytical tools for DNA, genes and genomes: nuts & bolts", by Arseni Markoff ed., The nuts & bolts series, DNA Press, 2005, ISBN 0-9748765-1-8.

[16] Jun Ishikawa* and Kunimoto Hotta .FramePlot: a new implementation of the Frame analysis for predicting protein-coding regions in bacterial DNA with a high G+C content.FEMS Microbiology Letters Volume 174, Issue 2, pages 251–253, May 1999.

[17] Ghorbani M ,Karimi H ,Ten Bioinformatics Tools for Single Nucleotide Polymorphisms , American Journal of Bioinformatics ;2014:3(2):45-48

[18] Ghorbani M and Karimi H. Cyclin-Dependent Kinases as valid targets for cancer treatment. Journal of Pharmacy Research 2015,9(6),377-382

[19] Ghorbani M , Karimi H , 'Ion Channels Association with Diseases and their Role as Therapeutic Targets in Drug Discovery', International Journal of Scientific Research in Science and Technology(IJSRST), 1(3):65-69,July-August 2015.

[20] Ghorbani M , Karimi H , 'Role of Aquaporins in Diseases and Drug Discovery', International Journal of Scientific Research in Science and Technology(IJSRST),1(3):60-64, July-August 2015

[21] Ghorbani M, Karimi H, 'Role of Microarray Technology in Diagnosis and Classification of Malignant Tumours', International Journal of Scientific Research in Science and Technology(IJSRST), 1(3):117-121, July-August 2015

[22] Mahin Ghorbani, Hamed Karimi, 'Role of G-Protein Coupled Receptors in Cancer Research and Drug Discovery', International Journal of Scientific Research in Science and Technology (IJSRST),1(3), pp.122-126 , July-August 2015.

[23] Mahin Ghorbani, Hamed karimi, 'Role of Biomarkers in Cancer Research and Drug Development', International Journal of Scientific Research in Science and Technology(IJSRST),1(3), pp.127-132, July-August 2015