International Journal of Biotechnology and Microbiology
www.biotechnologyjournals.com
Online ISSN: 2664-7680, Print ISSN: 2664-7672
Received: 25-07-2021; Accepted: 11-08-2021; Published: 26-08-2021
Volume 3, Issue 2, 2021, Page No. 29-33

Computational genome analysis for identifying biliary atresia genes

Sudheer Menon
Department of Bioinformatics, Bharathiar University, Coimbatore, Tamil Nadu, India

Abstract
Biliary atresia considered to be a very uncommon birth defect. Out of every 10,000 births 0.5 to 0.8 birth posses this defect. The Kasai portoenterostomy operation has shown good results with biliary atresia treatment, however age factor highly influences the results of the operation. Many unrevealed genetic and associated factors are the cause of biliary atresia, hence considered as multifactorial etiologies. There are three types of biliary atresia have been reported, classified based on biliary obstruction. Type I is bile duct blockage, with approximate prevalence of 5% cases. Type II is the blockage of hepatic duct with prevalence of around 2% cases. Type III is the blockage of porta hepatis. In this article different computational and bioinformatics tools are mentioned that could be utilized for *In Silico* analysis of genes causing biliary atresia. Like Gene ID, FGENESH, Gene Parser, Genie, AUGUSTUS, GENSCAN and MZEF. Two types of gene analysis and gene predictions approaches can be used in this case (I) Sequence similarity searches and (II) *Ab initio* gene prediction methods. Both approaches use different set of tools working on different algorithms. However, GENSCAN tool based on GHMM algorithm resulted to be the most accurate in gene prediction based on the accuracy parameters Burset and Guigo's sequence set.

Keywords: biliary atresia, FGENESH, gene parser, genie, AUGUSTUS, GENSCAN and MZEF

Introduction
Biliary atresia (BA) is a complex liver disease of infancy disease characterized by the extra hepatic destruction of biliary system by T-cells [1-3]. The patients with biliary atresia lack large bile ducts which can lead to cholestasis and can be life threatening if left untreated. At present Kasai portoenterostomy is the only treatment available for biliary atresia to re-establish bile flow from the effected liver. Kasai portoenterostomy has a success rate of less than 80% and almost 50% patients with successful results from this procedure will eventually need liver transplant. Biliary atresia is the primary reason for liver transplant in infancy around the globe which accounts for 40-50% of liver transplants of infants [4].

The etiology of BA is yet to be explored and in clinical settings it is divided into two types, first is the syndromic BA which occur during developmental stage of embryo and characterized by defective morphology of bile duct. Second type is non-syndromic BA which occurs during the perinatalperiod and affects bile duct by sclerosing inflammatory process. Non-syndromic BA is linked to both environmental and genetic factors [5]. Several environmental factors including viral infections and plant toxins have been reported to be associated with BA pathogenesis in human [6–8]. Genetic factors are also studied extensively and found to play role in biliary atresia [9]. Genome wide association studies (GWAS) on Chinese population has associated chromosome 11q24.4 to BA which maps intergenic region of *ADD 3* and *XPNPEP1* [10]. Genes associated to BA via GWAS are *ADD 3, EFEMP 1, GPC1* and *ARF 6* [10-13].

The process of identification of gene involved in complex disease is known as gene finding. It is very important process of analyzing the genome of an organism. In past process of gene finding use to be costly and time taking as well as it required a lot of *in vivo* experimentation. But due to the recent advancement in bioinformatics and statistical tools, gene finding has become easier. Here we discuss the bioinformatics tools used to identify the pathogenic genes responsible for BA.

Objectives
1. Enlisting Sequence similarity searches method for gene prediction
2. Enlisting *Ab initio* gene prediction methods

Materials and Methods
1. Gene ID
This programs used for gene prediction of unspecified genome sequences planned with a various leveled structure. First step in this program involves the prediction of start/stop codons and splice sites and the prediction is scored along the succession utilizing Position weight arrays (PWAs). Second step involves the exons which are constructed from predicted sites. Exons are scored as the amount of the scores of the characterizing locales, in addition to the log-probability proportion of a Markov Model for coding DNA. In the last step, a full fledge gene structure is constructed from the group of predicted exons. GeneID provides kind of a support to assimilate predictions coming from different means by utilizing gff files and the reconstruction of universal gene model or structure is likewise practical. The precision of this bioinformatics approach analyzes well to that of other existing

approaches, however GeneID is possibly more productive as far as speed and memory utilization. At present, on a processor Intel(R) Xeon CPU 2.80 Ghz v1.2 version of GeneID examines the entire human genome in as few as 3 hours. [14, 15]

GeneID yield can be altered to various degrees of detail, including thorough posting of likely signals and exons. Besides, a few yield designs as gff or XML are accessible. There are accessible parameters documents in geneid v 1.2 for Tetraodon nigroviridis, Dictyosteliumdiscoideum, human and Drosophila Melanogaster among numerous others for species traversing the four Kingdoms. [15, 16]

2. Fgenesh

The main developer of programming approaches for genomic examination is Softberry Inc. Softberry Inc. has focused in on computational techniques which include programming tools for SNP detection, transcriptome examination, and next generation sequencing methods and techniques for the detection of disease causing SNP subsets. Research oriented programs of softberry have been utilized in hundreds of research projects i: e FGENESH. FGENESH research tool is cited in thousands of research articles used for the identification of eukaryotic genes. In year 2015 only this tool was used in 2000 publications [17, 18]. Another programming application named Hundred's can be downloaded or used online at Softberry website for academic usage. It has collaborated in many academic projects globally [16, 18].

3. Geneparser

GeneParser is another computer based tool which is used to identify protein structure genes in DNA sequences. This program identify subinterval in a sequence and score them which indicate exons, introns and the locations or sites that indicate the boundaries of introns and exons. This data is than measured by a neural system to surmise the log-probability for each subinterval precisely addresses an exon or intron. Then by applying a vibrant computational algorithm to this information, a perfect blend of introns and exons that has the maximum probable function is discovered. This program has been tested on number of human genes and a correlation coefficient for GC rich genes and exon nucleotide prediction correlation was achieved which was 0/94 and 0.89 respectively [19]. In a study the program was trained on more than 50 human genes and when tested against it GeneParser accurately finds 75% exons and precisely anticipated 85% of coding sequence with correlation coefficient of 0.85 [16, 20].

4. Génie

In biomedical literature is a traditional source used by scientist to be aware of relevant genes to their research topics. However, a number of genes are not found in literature particularly genes from inadequately studied organisms. In addition it is impossible to summarize the literature related to genes. To overcome these above mentioned problems an algorithm Génie was introduced. It evaluates literature according to the topic to find out related genes and its orthologs in a genome. As compare to other algorithms Génie has very high performance, sensitivity and precision rates.

Génie is supported with thousands of species and genes. It is capable of analyzing genome in less than one minute even when examining human genome by utilizing records in literature. These features make it suitable for its use in lab work. Currently more than 4000 species are accessible in Génie [16, 21, 22].

5. Augustus

In a latest ENCODE genome annotation project (EGASP), few of the newly developed and widely used gene prediction tools have been tested and compared on information from human genome. One of the tested tools was AUGUSTUS. It can be used as an ab initio program. An ab initio program utilizes just one genomic sequence for input data [23]. Also, it can join data from the genomic arrangement under investigation with outer clues from different data sources. In EGASP, protein sequences, expressed sequence tags alignment and genomic sequence alignments were used along with other additional informational sources. In the class of ab initio programs AUGUSTUS altogether anticipated a genes more precisely than any other of the programs in this category. Moreover simultaneously it also predicted lowest number of false positive exons and genes which any other of the program failed to do. Among all ab initio tested tools in EGASP, AUGUSTED resulted as the finest and most precise. Also it is entirely adaptable in light of the fact that it can take data from a few sources all the while into consideration [24, 25, 26].

Results and Discussion

Sequence similarity searches

The basis of sequence similarity search is to find resemblance among genomes, proteins or ESTs to the input sequence. It is very simple approach which is based on a theory that exons or functional regions are more evolutionary conserved than that of intronic or non-functional regions. If the resemblance is found between a particular genomic sequence and protein, DNA or ESTs than this resemblance information can be utilized to deduce the structure of gene or function of that particular region. Similarity based on EST has certain disadvantages as ESTs is only defines small segments of gene sequence. Hence it is not easy to predict structure of gene for a given region. Two methods dependant on similarity searches are global alignment and local alignment. BLAST family of programs is a local alignment tool which is used most commonly to identify the resemblance of sequence to known ESTs or proteins. Global alignment method is used by two programs, GeneWise [28] and PROCRUSTES [27] for gene prediction. Another software CST finder is dependent on pairwise genome comparison [29]. The major drawback of these approaches is that 50% of the discovered genes are considerably homologous to genes in databases.

Ab initio gene prediction programs

Ab initio gene prediction is the next group of techniques for the computational detection of genes which employs the structure of gene as a template for gene identification. This prediction method is dependent on two kinds of sequence information, signal and content sensors. The first one, Signal sensors are concerned with small sequence motifs, like start/stop codons, polypyrimidine tracts, branch points and splice sites. The identification of exon must dependent on

content sensors, which allow coding region to stand out from the non- coding by statistical identification and it also refers to the unique codon patterns of particular specie. Various algorithms such as Meural Network, Hidden MarkovModel, Linguist methods, linear discriminant analysis and Dynamic Programming are used for constructing gene structure models. Many ab initio programs for gene prediction have been designed on based on these models. Few of the most widely used are shown in Table 1. Among them Hiden Markov Model (HMM) is the most successful [30]. HMM is mainly describes here to learn about other programs reader can find in references. In HMM, conversions among two sub-models of specific gene constituents are modeled as unnoticed or hidden Markov procedure which is capable of determining the possibility of forming specific observable nucleotides. In a real gene a more general model is needed to precisely describe the lengths of intron and exon, as the lengths of intron and exon are apparently limited by the pre-mRNA splicing factors. Consequently a Generalized Hidden Markov Model (GHMM) is created in which ensuing states are produced by a Markov chain however instead of fixed unit they have subjective length disseminations. The illustration of state transition in genomic sequences of eukaryotes is shown if Figure 1 [16].

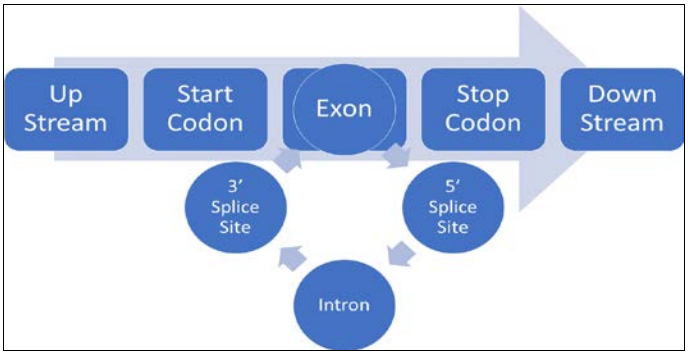

Fig 1: HMM modeling

Table 1: *ab initio* programs for Gene Prediction

Program	Species or organism	Algorithm	Website
GeneID	Plants and Vertebrates	Dynamic programming	https://genome.crg.es/geneid.html
FGENESH	Drosophila, Human and rat,	Hidden Markov Model (HMM)	http://www.softberry.com/berry.phtml?topic=fgenes_plus&group=programs&subgroup=gfs
GeneParser	Vertebrates	Neural Networks	http://stormo.wustl.edu/src/GenParser/
Genie	Drosophila,human, other	Generalized HMM	http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=6
AUGUSTUS	Human,Arabidopsis	Interpolated Markov Model	http://augustus.gobics.de/

Assessment of Programs used for predicting gene

Gene prediction programs are present in abundance which causes inadequate evaluation of their quality. Quality of algorithms in terms of reliability and accuracy must be taken into consideration and also the type of algorithm used such as evaluation method, HMM, neural networks or others. The ranking of a prediction tool is not only based on a single feature. Specificity and sensitivity are likely to be the most features corresponding to an prediction which are demonstrated by Burset and Guigos [31]. The prediction is measured for its accuracy at three various levels which are protein product, exon structure and coding sequence.

The accuracy at nucleotide level which determine specificity (Sp), sensitivity (Sn) and approximate coefficient (AC) provides a good judgment of how close are actual coding and predicted regions are when aligned but it does not precisely identify exon boundaries. Accuracy evaluation at exon level gives an idea how similar sequence signals like start/stop codon and splice sites are. Actual and anticipated exons with test sequences are compared to measure the accuracy (Figure 2). The accuracy at protein level is evaluated by the comparison of proteins products encoded predicted and actual gene in the test sequence. Only Fields and Sonderlunds [32] have evaluated the algorithm at protein level in literatures of gene prediction, which is why it is not used widely. The accuracy of the prediction of few widely used algorithms has been examined on sequence set of Burset and guigo [31] and the results are shown in Table 2 [33]. The result illustrated that GHMM based GENSCAN is distinctively more precise than other programs.

Table 2: Gene Prediction accuracy comparison among different programs

Program	Sp	Sn	MR	WR
Genie	0.46	0.54	0.18	0.30
FGENESH	0.65	0.60	0.15	0.13
GeneID	0.45	0.45	0.27	0.22
GeneParser	0.39	0.36	0.33	0.15
GENSCAN	0.80	0.77	0.10	0.07

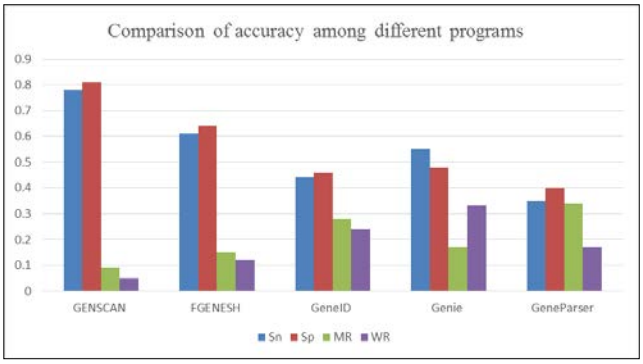

Fig 2: Accuracy Comparisons of Gene Prediction Programs

Conclusion

Associated chromosome 11q24.4 has been reported after the Genome wide association studies (GWAS) on Chinese population, which maps intergenic region of *ADD3* and *XPNPEP1*. Genes associated to BA via GWAS are *ADD3*, *EFEMP1*, *GPC1* and *ARF6*. To evaluate and analyze these genes in human genome for their pathogenic mutations which cause biliary atresia can be done by the mentioned bioinformatics tools. These tools have been classified according to the type of gene prediction method. Sequence similarity searches is the simple method is based on finding similarity in biliary atresia gene sequences between ESTs

(expressed sequence tags), proteins, or other genomes to the input genome. *Ab initio* gene prediction methods which use wild biliary atresia gene structure as atemplate to detect gene mutation and pathogenicity. Prediction of gene requires great efforts by both experimental and computational biologists to make gene prediction more accurate, which can definitely speed up gene discovery and knowledge mining.

References

1. Mieli-Vergani G, Vergani D. Biliary atresia. Semin. Immunopathol,2009:31:371-381. Doi:10.1007/s00281-009-0171-6.
2. Bezerra JA. Potential etiologies of biliary atresia. *Pediatr. Transplant*,2005:9:646-651. Doi:10.1111/j.1399-3046. 2005.00350.x.
3. Hartley JL, Davenport M, Kelly DA. Biliary atresia. *Lancet (London, England)*,2009:374:1704-1713, Doi:10.1016/S0140-6736(09)60946-6.
4. Sokol RJ, Shepherd RW, Superina R, Bezerra JA, Robuck P, Hoofnagle JH *et al*. Screening and outcomes in biliary atresia: summary of a National Institutes of Health workshop. *Hepatology*,2007:46:566-581, Doi:10.1002/ hep.21790.
5. Davenport M. A challenge on the use of the words embryonic and perinatal in the context of biliary atresia. *Hepatology*,2005:41:403-405.
6. Harper P, Plant JW, Unger DB. Congenital biliary atresia and jaundice in lambs and calves. Aust. Vet. J,1990:67:18-22. Doi:10.1111/j.1751-0813.1990.tb07385. x.
7. Drut R, Drut RM, Gómez MA, Cueto Rúa E, Lojo MM. Presence of human papillomavirus in extrahepatic biliary atresia. J. Pediatr. Gastroenterol. Nutr,1998:27:530-535, Doi:10.1097/00005176-199811000-00007.
8. Domiati-Saad R, Dawson DB, Margraf LR, Finegold MJ, Weinberg AG, Rogers BB. Cytomegalovirus and human herpesvirus 6, but not human papillomavirus, are present in neonatal giant cell hepatitis and extrahepatic biliary atresia. Pediatr. Dev. Pathol. Off. J. Soc. Pediatr. Pathol. Paediatr. Pathol. Soc,2000:3:367-373. Doi:10.1007/s100240010045.
9. Leyva-Vega M, Gerfen J, Thiel BD, Jurkiewicz D, Rand EB, Pawlowska J, *et al*. Genomic alterations in biliary atresia suggest region of potential disease susceptibility in 2q37.3. Am. J. Med. Genet. A,2010:152A:886-895, Doi:10.1002/ajmg.a.33332.
10. Garcia-Barceló MM, Yeung MY, Miao XP, Tang CSM, Cheng G, So MT *et al*. Genome-wide association study identifies a susceptibility locus for biliary atresia on 10q24.2. Hum Mol Genet,2010:19:2917-2925, Doi:10.1093/hmg/ddq196.
11. So J, Ningappa M, Glessner J, Min J, Ashokkumar C, Ranganathan S *et al.* Biliary-Atresia-Associated Mannosidase-1-Alpha-2 Gene Regulates Biliary and Ciliary Morphogenesis and Laterality. *Front. Physiol*, 2020:11:538701. Doi:10.3389/fphys.2020.538701 .
12. Asai A, Miethke A, Bezerra JA. Pathogenesis of biliary atresia: defining biology to understand clinical phenotypes. *Nat Rev Gastroenterol Hepatol*,2015:12:342-352. Doi:10.1038/nrgastro.2015.74.
13. Chen Y, Gilbert MA, Grochowski CM, McEldrew D, Llewellyn J, Waisbourd-Zinman O *et al*. A genome-wide association study identifies a susceptibility locus for biliary atresia on 2p16.1 within the gene EFEMP1. *PLOS Genet*,2018:14:e1007532.
14. Parra G, Blanco E, Guigó R. GeneId in Drosophila. *Genome Res*,2000:10:511-515. Doi:10.1101/gr.10.4.511.
15. Sudheer Menon. "Preparation and computational analysis of Bisulphite sequencing in Germfree Mice" International Journal for Science and Advance Research In Technology,2020:6(9):557-565.
16. Sudheer Menon, Shanmughavel Piramanayakam, Gopal Agarwal. "Computational identification of promoter regions in prokaryotes and Eukaryotes" EPRA International Journal of Agriculture and Rural Economic Research (ARER),2021:9(7):21-28.
17. Sudheer Menon. "Bioinformatics approaches to understand gene looping in human genome" EPRA International Journal of Research & Development (IJRD),2021:6(7):170-173.
18. Sudheer Menon. "Insilico analysis of terpenoids in Saccharomyces Cerevisiae"international Journal of Engineering Applied Sciences and Technology,2021:6(1):43-52. ISSN No. 2455-2143.
19. Sudheer Menon. "Computational analysis of Histone modification and TFBs that mediates gene looping" Bioinformatics, Pharmaceutical, and Chemical Sciences (RJLBPCS),2021:7(3):53-70.
20. Sudheer Menon, Shanmughavel piramanayakam, Gopal Prasad Agarwal. "FPMD-Fungal promoter motif database: A database for the Promoter motifs regions in fungal genomes" EPRA International Journal of Multidisciplinary research,2021:7(7):620-623.
21. Sudheer Menon, Shanmughavel Piramanayakam, Gopal Agarwal. Computational Identification of promoter regions in fungal genomes, International Journal of Advance Research, Ideas and Innovations in Technology,2021:7(4):908-914.
22. Sudheer Menon, Vincent Chi Hang Lui and Paul Kwong Hang Tam. Bioinformatics methods for identifying hirschsprung disease genes, International Journal for Research in Applied Science & Engineering Technology (IJRASET),2021:9(7):2974-2978.
23. Sudheer Menon. Bioinformatics approaches to understand the role of African genetic diversity in disease, International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET), 2021:4(8):1707-1713.
24. Guigó R, Knudsen S, Drake N, Smith T. Prediction of gene structure. J Mol Biol,1992:226:141-157. Doi:10.1016/0022-2836(92)90130-C.16.
25. Wang Z, Chen Y, Li Y. A brief review of computational gene prediction methods. *Genomics, proteomics Bioinforma./Beijing Genomics Inst.*2004:2:216-221, Doi:10.1016/S1672-0229(04)02028-5.
26. Zhang S, Li D, Zhang G, Wang J, Niu N. The Prediction of Rice Gene by Fgenesh. Agric Sci China,2008:7:387-394. Doi: https://doi.org/10.1016/S1671-927(08)60081-4.
27. Rinaldi AJ, Lund PE, Blanco MR, Walter NG. The Shine-

Dalgarno sequence of riboswitch-regulated single mRNAs shows ligand-dependent accessibility bursts. *Nat. Commun.*2016:7:8976. Doi:10.1038/ncomms9976.
28. Snyder EE, Stormo GD. Identification of protein coding regions in genomic DNA. J Mol Biol,1995:248:1-18. Doi:10.1006/jmbi.1995.0198.
29. Snyder EE, Stormo, GD. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids Res,1993:21:607-613. Doi:10.1093/nar/21.3.607.
30. Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Génie: literature-based gene prioritization at multi genomic scale. Nucleic Acids Res,2011:39:W455-W461. Doi:10.1093/nar/gkr246.
31. Reese MG, Kulp D, Tammana H, Haussler D. Genie--gene finding in Drosophila melanogaster. Genome Res,2000:10:529-538. Doi:10.1101/gr.10.4.529.
32. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol,*2006:7:S11. Doi:10.1186/gb-2006-7-s1-s11.
33. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics,* 2008:24:637-644. Doi:10.1093/ bioinformatics/btn013.
34. Zhang MQ. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci, 1997:94:565-568, Doi:10.1073/pnas.94.2.565.
35. Zhang MQ. Using MZEF to find internal coding exons. *Curr. Protoc. Bioinforma,* Chapter 4, Unit 4.2, 2002. Doi:10.1002/0471250953.bi0402s00.
36. Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced sequence alignment. Proc. Natl. Acad. Sci. U. S. A,1996:93:9061-9066. Doi:10.1073/pnas.93.17.9061.
37. Birney E, Durbin R. Using GeneWise in the Drosophila annotation experiment. *Genome Res,*2000:10:547-548. Doi:10.1101/gr.10.4.547.
38. Mignone F, Grillo G, Liuni S, Pesole G. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res,*2003:31:4639-4645. Doi:10.1093/nar/gkg483.
39. Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences. Genome Res,2000:10:1631-1642. Doi: 10.1101/gr.122800.
40. Burset M, Guigó R. Evaluation of gene structure prediction programs. *Genomics,*1996:34:353-367. Doi: 10.1006/geno.1996.0298.
41. Fields CA, Soderlund CA. gm: a practical tool for automating DNA sequence analysis. *Comput. Appl. Biosci.*1990:6:263-270. Doi:10.1093/bioinformatics/6.3.263.
42. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol,1997:268:78-94. Doi:10.1006/jmbi.1997.0951.