International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 09 Issue 11 ǁ 2021 ǁ PP. 58-65

Comparative Analysis of Different Gene Prediction Algorithms

Sukhpreet Kumar
Research Scholar
DBFGOI
Moga

Deepak Sharma
Assistant Professor
DBFGOI
Moga

Abstract:
Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. The science of Bioinformatics, which is the melding of molecular biology with computer science, is essential to the use of genomic information in understanding human diseases and in the identification of new molecular targets for drug discovery(e.g. DNA computing, neural computing, evolutionary computing, immuno-computing, swarm-computing, cellular-computing). In this research author has put his efforts to compare various existing gene prediction algorithms.
Keywords: *gene prediction, bioinformatics, DNA.*

Date of Submission: 12-11-2021 Date of acceptance: 28-11-2021

I. Introduction:

Bioinformatics is the application of computer technology to the management of biological information. It deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory for generating new knowledge of biology and medicine, and improving & discovering new models of computation. With the help of computer tools, biological information is gathered and analyzed. It is the science of managing, mining and interpreting information from biological sequences and structures. It deals with algorithms, databases and information systems, data mining, image processing and improving & discovering new models of computation. This scientific field deals with the computational management of all kinds of biological information. This information can be on genes and their products, whole organisms or even ecological systems. Mainly bioinformatics involves merger of different applications of mathematical, statistical, computational or molecular biological tools to gather different types of information and by analyzing them, researches can be carried out. Over the past few decades, major advancements in this field have led to an explosive growth in the biological information. The computerized databases are used to organize, store and index the data. Java, XML, Perl, C, C++, SQL and MATLAB are the programming languages popularly used in this field. The tools of bioinformatics include computer programs that help to reveal fundamental mechanisms. The biological problems related to the structure and function of macromolecules, disease processes, and evolution are contained in the tools. The applications of the tools is being categorised into sequence analysis, structure analysis, and function analysis. These three aspects of bioinformatics often interact to produce integrated and good results. Bioinformatics includes the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins. Extract DNA and Protein Sequences from Database. Whole of the database is being searched to compare the DNA's in a pair-wise fashion. DNA is transcribed to RNA which is further translated to proteins. This make possible to analyze the behaviour of the cell. After the alignments, a structure occurs in form of a tree.

Introduction to DNA and RNA:
DNA : Basic structure of DNA is shown in the following figure 1:

Figure1 DNA Structure[29]

figure1 shows DNA is short for deoxyribonucleic acid. Two chains of four chemical bases (abbreviated A, T, C and G) make up DNA and act as a cell's recipe book to make proteins. The particular sequence of a DNA chain – meaning the precise order of the four chemical bases – determines what protein will be made. A DNA segment beginning with ATTCGC would produce a very different protein than one that starts with CCGTAT. This can be likened to adjusting the order of letters in a word. Though the letters are the same, the meaning changes. For example, actmeans something very different than cat.

Not all DNA is destined to become a protein. Just as a recipe might contain more information than just a list of ingredients, only some regions of your DNA – called genes – are directly translated into proteins. Cellular machinery follows the instructions written in a gene's recipe to create a corresponding sequence of messenger ribonucleic acid (mRNA), which is chemically similar to DNA but acts as a messenger, carrying the recipe from the nucleus. Out in the cell's cytosol, the mRNA sequence is read by more machines, called ribosomes. Following the mRNA instructions, ribosomes string together amino acids, the building blocks that make up proteins. Proteins do most of the work in the cell.As cells divide, producing two cells where there was once only one, the parent cell's DNA is duplicated and the same protein-making recipe is passed on to the daughter cell.

DNA has following points:-

a) Just a 4 letter alphabet (GATC)
b) Encodes proteins with 3 letter codons
c) Punctuation determines transcription starts and stops
d) Information in DNA has stored in code made up of 4 chemical basis.
e) Determines its own replication.

b) RNA: basic structure of RNA is shown in the following figure 2:

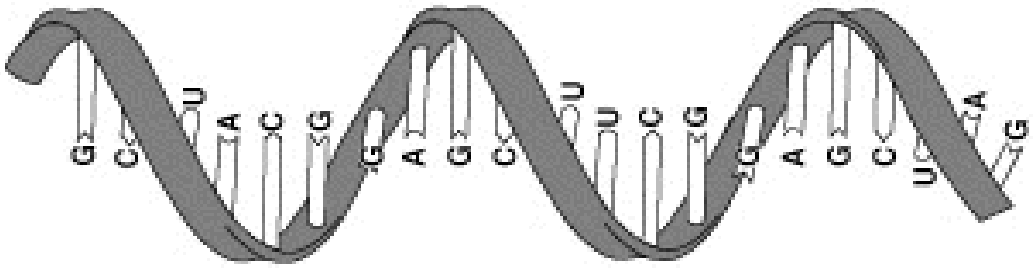


Figure 2: Structure of RNA [30]

Ribonucleic acid or RNA, is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life. Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase (sometimes called a nitrogenous base), a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. All cellular organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins. In addition, some viruses use RNA instead of DNA as their genetic material; perhaps a reflection of the suggested key role of RNA in the evolutionary

The chemical structure of RNA is very similar to that of DNA, with two differences: (a) RNA contains the sugar ribose, while DNA contains the slightly different sugar deoxyribose (a) type of ribose that lacks one oxygen atom), and (b) RNA has the nucleobase uracil while DNA contains thymine. Uracil and thymine have similar base-pairing properties.

Fundamental of Gene

The entire nucleic acid sequence necessary for the synthesis of a functional protein (or functional RNA)

A gene includes the nucleotides encoding information amino acid sequence (coding sequence), sequences for controlling synthesis (enhancers, promoters, polyadenylation site, splice sites, transcriptional termination site, ribosome binding site as well as intervening regions that contain no information. Basic gene structure is shown in following figure 3:

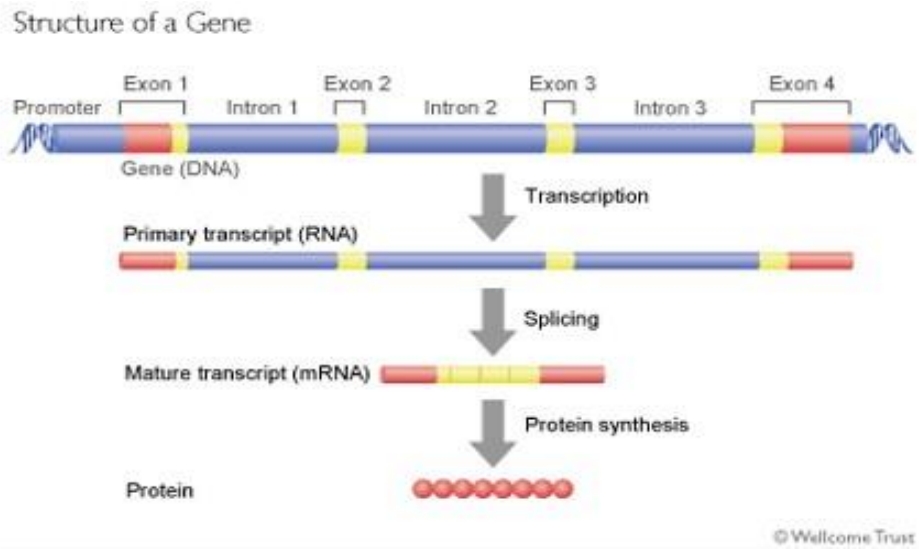


Figure 3 GENE Structure[6]

As shown in above figure 3, shows that there are two general types of gene in the human genome i.e. Non-coding RNA genes and Protein-coding genes. Non-coding RNA genes represent 2-5 per cent of the total and encode functional RNA molecules. Many of these RNAs are involved in the control of gene expression, particularly protein synthesis. They have no overall conserved structure. Protein-coding genes represent the majority of the total and are expressed in two stages: transcription and translation (Gene expression). They show incredible diversity in size and organisation and have no typical structure.

Existing study:

Zhou et al. (2010) discussed a method to found the nearest neighbors in biological databases using less distance computations that involved the similarity comparison between two objects. A substantial speedup technique for the well-studied k-nearest neighbor (k-nn) search is used, which is based on novel concepts of virtual pivots and partial pivots, such that a significant number of the expensive distance computations can be avoided. Some methods are included for k-nn searching and that are M-tree, OMNI, SA-tree, LAESA.

Zimek et al. (2010) studied the hierarchical and flat classification of proteins. The problem in classification of proteins was received significant attention. One feature of this problem is that expert-defined hierarchies of protein classes exist and can potentially be exploited to improve classification performance. They compared multiclass classification techniques that exploited the information in those class hierarchies and those that do not, using logistic regression, decision trees, bagged decision trees, and support vector machines as the underlying base learners.

Chen et al. (2011) discussed the classification trees that included nonparametric statistical learning methods having incorporated feature such as selection and interactions, possess intuitive interpretability, efficiently, and have high prediction accuracy when used in ensembles. However, it provided a brief introduction to the classification tree-based methods, a review of the recent developments, and a survey of the applications in bioinformatics and statistical genetics.

Kakiuchi et al. (2011) gave a characteristics of a new neighborhood that determined by three parameters. The neighborhood is generated from a special capacity that determined by three parameters. Neighborhood has a certain combination of contamination and gap from the model. Various new neighborhoods are obtained from changing the values of the three parameters. One of the parameters expressed the size of contamination and the others determined the size of gap from the model.

Vijan et al. (2011) defined a biological sequence alignment for bioinformatics applications by using MATLAB. The biological sequence alignment is widely used operation in the field of bioinformatics and computational biology as it is used to determine the similarity between the biological sequences. The proposed method described the two basic alignment algorithms i.e. Smith Waterman for local alignment and Needleman Wunsch for global alignment. The algorithms have been developed and simulated using MATLAB for genome analysis and sequence alignment. The local and global alignment has been presented and the results are shown in the form of dot plots and local and global scores for the sequences. The goal is to develop a tool that can aided in the exploration, interpretation and visualization of data in the field of molecular biology.

Chris Burge* and Samuel Karlin(1997) discussed and introduce a general probabilistic model of the gene structure of human genomic sequences which incorporates descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic Regions. Distinct sets of model parameters are derived to account for the many substantial differences in gene density and structure observed in distinct C. G compositional regions of the human genome. Vladimir Pavlovi´c, Ashutosh Garg(1997) provide Gene identification and gene discoveryin new genomic sequences is one of the most timely computationalquestions addressed by bioinformatics scientists.This computational research has resulted in severalsystems that have been used successfully in manywhole-genome analysis projects. As the number of suchsystems grows the need for a rigorous way to combinethe predictions becomes more essential.In this paper we provide a Bayesian networkframework for combining gene predictions from multiplesystems. The framework allows us to treat the problemas combining the advice of multiple experts. Previouswork in the area used relatively simple ideas suchas majority voting. We introduce, for the first time, theuse of Hidden Input/Output Markov models for combininggene predictions.

Aron Culotta, David Kulp and Andrew McCallum Department of Computer Science University of Massachusetts(2001)provide this paper Given a sequence of DNA nucleotide bases, the task of gene prediction is to find subsequences of bases that encode proteins. Reasonable performance on this task has been achieved using generatively trained sequence models,such as hidden Markov models. We propose instead the use of a discrimini-tively trained sequence model, the conditional random field (CRF).

M. Sohel Rahman1,Costas S. Iliopoulos1,And L. Mouchard,(2001) provide this paper, we consider the pattern matching problem in DNA and RNA sequences where either the pattern or the text can be degenerate i.e. contain sets of characters. We present an asymptotically faster algorithm for the above problem that works in O (n log m) time, where n and m is the length of the text and the pattern respectively. We also suggest an efficient implementation of our algorithm, which works in linear time when the pattern size is small. Finally, we also describe how our approach can be used to solve the distributed pattern matching Problem.

Yang weng Dept. of Math., Sichuan Univ., Chengdu yunmin zhu(2006) provide Although the computational gene-finding programs have been greatly improved in recent years, our ultimate goal is far to be met. Even the best program cannot be used automatically to identify genes and other genomic elements. Fortunately, there are only a tiny number of exons completely missed by all programs. Therefore the combination of predictions from the gene-finding programs is a convenient way to improve gene prediction. In this paper we motivate the use of the Dempster-Shafer theory of evidence as an appropriate theory for modelling

combination of gene predictions, and give the mathematical framework for combining gene predictions of gene-finding programs by using Dempster-Shafer combination rule.

Bandyopadhyay, S.Maulik, U. Roy ,D. Indian Stat. Inst., Kolkata (2008) provideAutomatic 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 employ computational intelligence techniques that are known to be more robust when dealing with uncertainty and imprecision. In this paper, a detailed survey on the existing classical and computational intelligence based methods for gene identification is carried out.

II. Methodology

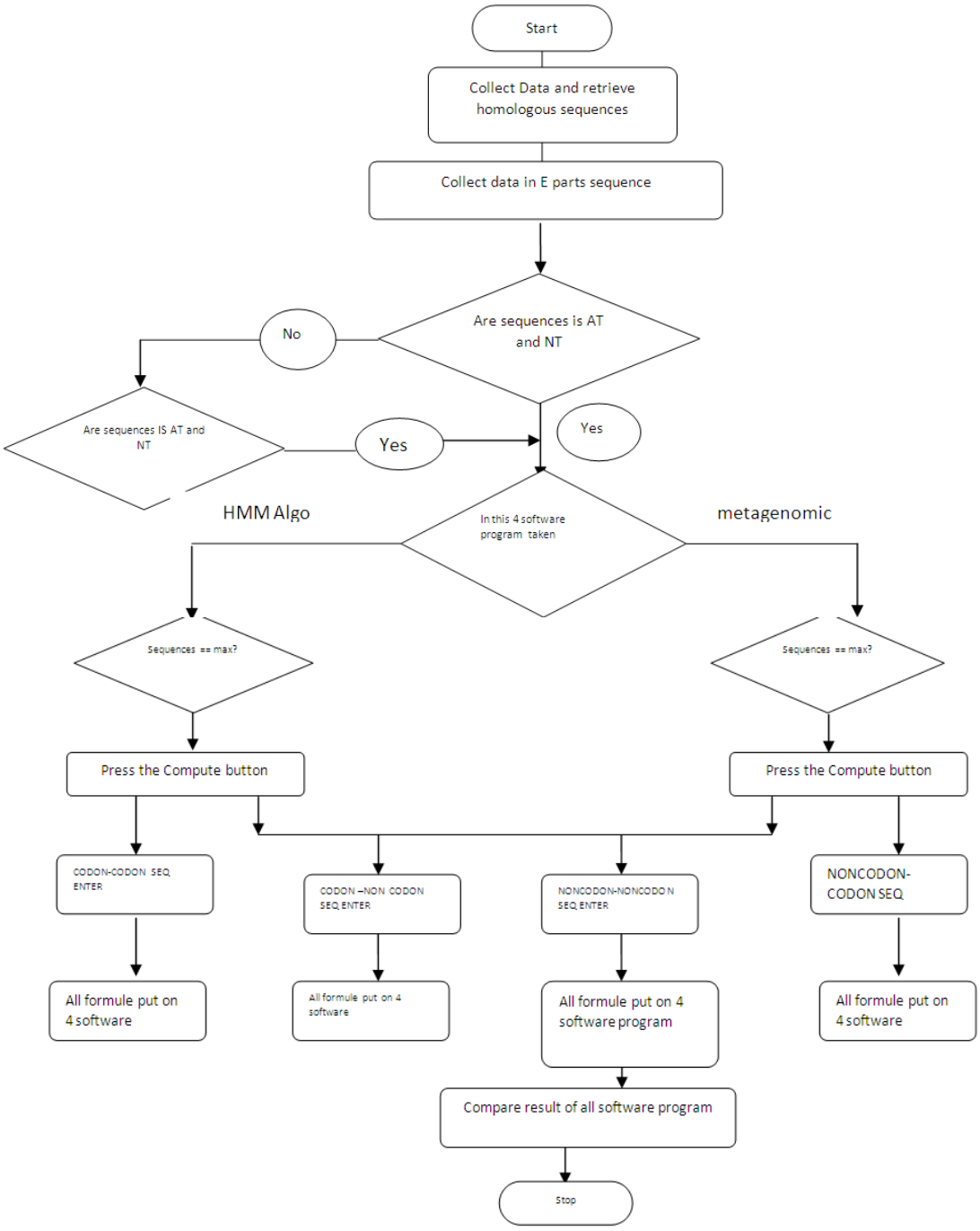
The methodology for the work will involve the use of METAGENOMIC DATA technique to analyze the given database and retrieve the desired information.In this me do manually calculation to take out result of sequences and here we have some formulas to do calaulations to find out the results of various gene prediction program of various sequences and then me calculate average results of all the program then me differentiate all result. In this work on various gene prediction program and analyzing them by using various calculation method formulas:-

Sr. no.	Formula name	calculation
1	Predicted number of positives (PP)	TP+FP
2	Predicted number of negatives (PN)	TN+FN
3	Correlation-Coefficient (CC)	Gm/Gt
4	PredictionError	Gm/Gt
5	Prediction Accuracy	TP+TN/TP+FN+FP+TN
6	Positive predictive value(PPV)	PPV=TP/TP+FP
7	Annotation Error	|Lp−Lgb|+|Rp−Rgb||Fgb|
8	F−measure	2∗Sensitivity∗Specificity/Sensitivity+Specificity

n this research we used condons and non condons series of AT AND NT sequence to calaculate and differentiate all programs.
http://www.cbs.dtu.dk/biolinks/pserve2.php
http://www.ncbi.nlm.nih.gov/sites/gquery?itool=toolbar
http://www.genedb.org/Homepage
http://www.genome.jp/kegg/genes.html
www.hprd.org
and use server for findout
netglyserver.3.0
in this we use the concept of pattern matching for findout the gene sequences
In this me taken 5 results of each gene prediction program

a) In first me take codon and codon sequence
b) In this me taken codon and non-codon sequence
c) In this me taken non-codon and non-codon sequence
d) In this me taken non-codon and codon sequence.
e) In this me taken simple sequence

The work flow diagram as shown in Figure 4 represents the flow of my present work.



III. Results And Discussions:

Table 2 shows Performances of the four program metagenmark, Orphelia, glimmer and prodigal over case B fragments of the 1000 bp group .In this we taken coding and non-coding part both of 1000bp length. This fragment case is purely from gene region. The annotation errors onthis table are lower than those of case A fragments .

TABLE 2:- Performance of four programs

Measure	Metagenmark	Orphelia	Glimmer	Prodigal
Specificity	100%	100%	100%	100%
Sensitivity	100%	100%	80%	100%
Predicted number of positives(PP)	20	20	16	20
Predicted number of negative(PN)	0	0	4	0
Correlation-Coefficient (CC)	0	0	0	0

PredictionError	0	0	20	0
Prediction Accuracy	100%	100%	80%	100%
Positive predictive value(PPV)	100%	100%	100%	100%
F-measure	100%	100%	88%	100%

IV. Conclusion and future scope:

In this thesis the average result of all program has better than previous paper. In conclusion, gene-edge type fragments have a higher annotation error than intra-coding regions. The performancesof all algorithms worsen for shorter reads. Usageof all the four programs together in upper bound analysis enhance the accuracy of gene prediction as it is apparent. the Average sensitivity and specificity of all programs taken in this glimmer has low sensitivity and specificity.but other program has high sensitivity and specificity. We built an open source heuristic ab initio algorithm for metagenomic gene prediction using Prodigal. The program can analyze fragments independently and thereby achieve full speedup through utilization of multiple processors. Although we understand the problems posed by sequencing errors, we chose to focus instead on other problems that have received less attention,such as translation initiation site identification, handling of alternate genetic codes, and providing filtering mechanisms forscores based on confidence. In future versions, we hope to address sequencing errors in more detail, as well as provide further improvements to the program's performance at smaller fragment lengths.

References:

[1] S Mahony, T J Smith, J O Mclnerney, A Golden(2011),"A new approach to gene prediction using the self-organizing map"Computational Systems Bioinformatics CSB2003 Proceedings of the 2011 IEEE Bioinformatics.volume: 3,Page: 23-28. 2)

[2] Y Xu, E C Uberbacher(2011) Gene prediction by pattern recognition and homology search. Proceedings International Conference on Intelligent Systems for Molecular

[3] Xiaochuan Ai; Jingbo Xia(2009)," Functional gene prediction with vital reduced features Further topics for feature reduction and evaluation criteria for classifiers, Automation and Logistics (ICAL), 2009 IEEE International Conference on. Volume: 4, Pages: 254-267.

[4] Vikrant Tomar, Dipesh Gandhi, C. Vijaykumar, "Digital Signal Processing for Gene Prediction", pp 1-5, 2008.

[5] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah, "Signal Processing in Sequence Analysis:Advances in Eukaryotic Gene prediction",pp 310-321, 2008.

[6] Yang Weng; Yunmin Zhu(2006)," Combining Gene-Finding Programs by Using Dempster-Shafer Theory of Evidence for Gene Prediction," Computational Intelligence and Security,vol.16 no.1 2006 page 1-8.

[7] Intenational conference on Intelligent Systems in Molecular Biology (ISMB, 5:56-64.),

Blayo, P., Rouzé, P., and Sagot, M.-F. (2003). Orphan gene finding - an exon assembly approach. Theoretical Computer Science, 290:1407-1431.

[8] Chuang TJ, Lin WC, Lee HC, Wang CW, Hsiao KL, Wang ZH, Shieh D, Lin SC, Ch'ang LY. (2003) A complexity reduction algorithm for analysis and annotation of large genoomic sequences. Genome Res. 2003 13:313-22.

[9] Garg ashutosh,kasif simon, pavlovic Vladimir(2001)," A Bayesian framework for combining gene predictions",bioinformatics program boston university,vol.18 no.1 2001 page 19-27.

[10] Abdeddaim, S. and Morgenstern, B. (2001). Speeding up the DIALIGN multiple alignment program by using the `Greedy Alignment of BIOlogical Sequences LIBrary' (GABIOS-LIB). Lecture Notes in Computer Science 2066, 1 - 11.

[11] Biology ISMB International Conference on Intelligent Systems for Molecular Biology (2001)Volume: 4, Pages: 241-251.

[12] Castellano S., Morozova N., Morey M, Berry M.J., Serras F., Corominas M and Guigó R (2001) in silico identification of novel selenoproteins in the Drosophila melanogaster genome. EMBO Reports 2:697-702

[13] Abdeddaim, S. and Morgenstern, B. (2001). Speeding up the DIALIGN multiple alignment program by using the `Greedy Alignment of BIOlogical Sequences LIBrary' (GABIOS-LIB). Lecture Notes in Computer Science 2066, 1 - 11.

[14] Biology ISMB International Conference on Intelligent Systems for Molecular Biology (2001)Volume: 4, Pages: 241-251.

[15] Castellano S., Morozova N., Morey M, Berry M.J., Serras F., Corominas M and Guigó (2001) in silico identification of novel selenoproteins in the Drosophila melanogaster genome. EMBO Reports 2:697-702

[16] Guigó, R (1999) "DNA composition, codon usage and exon prediction." In M. Bishop, editor: Genetic Databases. Pp:53-80. Academic Press.

[17] Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., et al. (1999). The DNA sequence of human chromosome 22. Nature, 402(6761):489-495.

[18] Durbin, R., Eddy, S., Crogh, A., and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids. Cambridge University Press.

[19] Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Research, 8(9):967-974. 17

[20] Guigó, R. (1998). Assembling genes from predicted exons in linear time with dynamic programming. Journal of Computational Biology, 5:681-702.

[21] Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268:78-94

[22] Burge Chri and Samuel Karlin(1997)," Prediction of Complete Gene Structures in Human Genomic DNA, Department of Mathematics Stanford University, StanfordCA, 94305, USA. Volume: 4, Pages: 241-251.

[23] Burset M and Guigo R (1996) Evaluation of gene structure prediction programs. Genomics 34:353-367. 16

[24] Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. (1996). Gene recognitionvia spliced alignment. Proceedings National Academy Sciences USA, 93:9061-9066.

[25] Krogh A, Mian IS and Haussler D (1994) A hidden Markov model that finds genes in E . coli DNA. Nucl Acid Res. 22:4768-4778.

[26] Gelfand M and Roytberg M (1993) Prediction of exon-intron structure by a dynamic programming approach. BioSystems 30:173-182.

[27] Gish, W. and States, D. (1993). Identification of protein coding regions by database similarity search. Nature Genetics, 3:266-272.
[28] Brunak S Engelbrecht J and Knudsen S (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. J. Mol. Biol. 220:49–65.
[29] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215:403-410.
[30] http://www.abpischools.org.uk/res/coresourceimport/resources04/cancer/images/dna1.gif
[31] http://www.mysciencebox.org/files/images/RNA-codon.png
[32] http://www.rothamsted.ac.uk/notebook/images/pept.gif
[33] http://www.abpischools.org.uk/res/coresourceimport/resources04/cancer/images/dna1.gif
[34] http://www.mysciencebox.org/files/images/RNA-codon.png
[35] http://www.rothamsted.ac.uk/notebook/images/pept.gif