GENE PREDICTION: A REVIEW

Renu Sharma* Ajay Kaushik**

*Lecturer, Electronics and communication Engineering
R.P.Inderaprastha Institute of Technology, Karnal, India
** Lecturer, Electronics and communication Engineering, MMU, Mullana, India

ABSTRACT

Millions of bases of genomic DNA are sequenced daily in genome centres worldwide and the list of completely sequenced genomes from different organisms is growing rapidly, tools for interpreting the content of these genomes are more important than ever. The task of gene prediction is to find sub sequences of bases that encode proteins. Intrinsic method use statistical features to differentiate in between exons and introns. Extrinsic method is used to find similarities in between genomics sequence and proteins. Many software are available that predict gene sequences perfectly with more than 80% accuracy.

Keywords: **DNA, Extrinsic Method, Gene Prediction, Intrinsic Method, Protein.**

1. INTRODUCTION

Bioinformatics means applying computer science methods on biological problems. It is divided into different sub areas, whereas the area concerning this paper is Gene prediction. DNA molecules constitute the genetic blueprint of living organisms. DNA sequence can be divided into genes and intergenic spaces. The genes are responsible for protein synthesis. A gene can be divided into two sub-regions called the exons and the introns. There are two basic problems in gene prediction: prediction of protein coding regions and prediction of the functional sites of genes. Thus DNA sequence does not contain any relevant information, but only a small part does contain the genes. Gene prediction deals with the problem of finding these genes, which is still not solved satisfyingly.

2. PREDICTION USING SEVERAL GENE FINDING SOFTWARE

A large amount of literature on the subject of gene prediction as well as number of developed gene-finding algorithms further illustrates the importance of this area of research. Burge and Karlin [1] proposed a new computer program (GENSCAN), which captures potentially important dependencies between signal positions. It identified complete exon/intron structures of genes in genomic DNA. Fourier spectrum is good discriminator of coding potential and this feature used by GENSCAN and 75 to 80% of exons identified exactly. Claverie proposed [2] Computational methods, the best program currently available perfectly locates more than 80% of the internal coding exons, and only 5% of the predictions do not overlap a real exon. If the performances are satisfactory for the identification of the coding moiety of genes (internal coding exons), the determination of the full extent of the transcript (54 and 34 extremities of the gene) and the location of promoter regions are still unreliable. As the human and mouse genome sequencing projects enter a production mode, the fully automated annotation of megabase-long anonymous genomic sequences is the next big challenge in bioinformatics. Culotta et al. [3] proposed that instead the use of a discriminatively trained sequence model, the conditional random field (CRF). CRFs perform better than HMM-based models at incorporating homology evidence from protein databases, achieving a 10% reduction in base-level error. Salzberg et al. [4] developed GLIMMERM to find genes in the malaria parasite Plasmodium falciparum. Because the gene density in P. falciparum is relatively high, the system design was based on a successful bacterial gene finder, GLIMMER. GLIMMERM predict 87% exons exactly. A program CRITICA (Coding Region Identification Tool Invoking Comparative Analysis) introduced by Badger and Olsen [5] for identifying likely protein coding sequences in DNA by combining comparative analysis of DNA sequences with more common non-comparative methods. CRITICA is not dependent upon the existence or accuracy of coding sequence annotations in the databases. This independence makes the method particularly well-suited for the analysis of novel genome. Milaesi et al. [6] analyze the full gene structure in different organism, necessary to combine statistical properties and potential function signal of coding sequence. Its sensitivity is 89% and it approximately predicts 91% exon correctly. Conserved Exon Method(CEM) introduced by Bafa and Huson [7] based on the idea of looking for conserved protein sequences by comparing pairs of DNA sequences, identifying putative exon pairs based on conserved regions and splice junction signals then chaining pairs of putative exons together. A method GAZE introduced by Howe et al. [8] uses assembling arbitrary evidence for individual gene components (features) into predictions of complete gene structures. GAZE uses a dynamic programming algorithm to obtain the highest scoring gene structure according to the model and posterior probabilities that each input feature is part of a gene. A novel pruning strategy ensures that the algorithm has a run-time effectively linear in sequence length. It doesn't work directly with genomic DNA sequence. It predict gene have result of any signal or content sensor. Taher et al. [9] purposed www server for homology-based gene prediction. The user enters a pair of evolutionary related genomic sequences, for example from human and mouse. Alignment of the input sequence is calculated using CHAOS and DIALIGN and then searches for conserved splicing signals and start/stop codons around regions of local sequence similarity. Stanke and Waack [10] proposed a method AUGUSTUS, used for the prediction of protein coding genes in eukaryotic genomes. The program is based on a Hidden Markov Model and integrates a number of known methods and submodels. It employs a new way of modelling intron lengths. AUGUSTUS have achieved relatively high accuracy on short genomic sequences but do not perform well on longer sequences with an unknown number of genes in them. JIGSAW a new gene finding system designed by Allen and Salzberg [11] to automate the process of predicting gene structure from multiple sources of evidence, with results that often match the performance of human curators. JIGSAW computes the relative weight of different lines of evidence using statistics generated from a training set, and then combines the evidence using dynamic programming. Its sensitivity and specificity are 92% and 72% respectively. Gross and Brent [12] proposed N-SCAN used to model the phylogenetic relationships between the aligned genome sequences, context dependent substitution rates, and insertions and deletions. An implementation of N-SCAN was created and used to generate predictions for the entire human genome and the genome of the fruit fly Drosophila melanogaster. Bernal et al. [13] introduced a program CRAIG for intrinsic gene prediction based on conditional random field with a semi-markov structure. It uses three benchmarks test sets BGHM53, TIGR251, ENCODE294. It distinguish two different type of introns short 980bp and long greater than 980 bp. Vinson et al.[14] said that CRF directly model the conditional Probability of a vector of hidden states conditioned on set of observation. Cryptococcus neoformans strain JEC21 used for this method. Accuracy using EST is 89.0-91.7% and on adding gap feature it increases 93.6-95.4%. Akhter at al. [15] said that DNA symbolic-to-numeric representations are presented and compared with existing techniques in terms of relative accuracy for the gene and exon prediction problem. Novel signal processing-based gene and exon prediction methods are then evaluated together with existing approaches at a nucleotide level using the Burset/Guigo1996, HMR195, and GENSCAN standard genomic datasets. Cai et al. [16] said that computational method is the best prediction method and the prediction accuracy ranges from 84.16% & 90.06% for basic and testing data. The accuracies of various gene predictor software is mentioned in table 1.

Table 1: Accuracy of gene predictor software

Software	Accuracy
HMM Gene	64.87%
CRITICA	67%
AUGUSTUS	71.58%
JIGSAW	72%
GENSCAN	75-80%
GLIMMERM	87%
GAZE	85-90%
GeneBuilder	91%

3. PREDICTION USING DIGITAL SIGNAL PROCESSING

The field of signal processing deals with numerical sequences rather than character strings. However, if a character string is properly mapped into one or more numerical sequences, then digital signal processing provides a set of novel and useful tools for solving gene prediction problem.

Sahu and Panda [17] discussed two new methods based on sliding DFT (SDFT) and adaptive autoregressive modeling for efficient and cost effective prediction of the exons in the gene. Receiver operating characteristic curve (ROC) analysis is used to predict gene. Anastassiou [18],

[19] proposed a mapping technique to optimize gene prediction using Fourier analysis and introduced color spectrogram for exon prediction. Although this mapping technique gives comparatively good results than DFT but it is DNA sequence dependent and thus requires computation of the mapping scheme before processing for gene prediction. Vaidyanarthan and Yoon [20] proposed digital resonator (antinotch filter). This scheme can be implemented with only one multiplier per output sample. Multisatge filter is used to improve gene prediction. Vaidyanathan and Yoon [21] said that gene location can be predicted using period-3 property of codon structure, and in fact allows the prediction of specific exons within the genes of eukaryotic cells. They introduce a simple and efficient scheme for identifying the period-3 regions of DNA sequences based on antinotch IIR filters instead of the DFT. These filters can be implemented very efficiently using the one-multiplier Gray and Markel lattice structure. Vaidyanathan [22] said that the protein-coding regions of DNA sequences exhibit period-3 behaviour due to codon structure. Identification of the period-3 regions helps in predicting the gene locations, and in fact allows the prediction of specific exons within the genes of eukaryotic cells. Traditionally these regions are identified with the help of techniques such as the windowed DFT purposed. A new technique (a single digital filter operation followed by a quadratic window operation) was introduced by Fox and Carrerira [23] that suppresses nearly all of the non-coding regions. The proposed method therefore improves the likelihood of correctly identifying coding regions in such genes. Various methods that have been used previously to automatically identify the coding regions, have been predominantly ‘frequency’ domain techniques. Numerous ‘time’ domain techniques are available from the signal processing literature. Rao and Shephard, in [24], assumed the DNA sequence to be generated from a white random process through an all pole system and thus used Auto-Regressive modeling to replace Fourier analysis for exon prediction. Two techniques new to this application are introduced by Epps and Akhtar [25], are Time Domain Periodogram (TDP) and the Average Magnitude Difference Function (AMDF). They also present an indicative comparison of time domain and existing frequency domain techniques, from which the AMDF appears to be the most promising technique. Previously binary indicator sequence and electron-ion interaction pseudo potentials (EIIP) indicator sequence has been used for the identification of the coding regions. . Nair and Mahalakshmi [26] used Cumulative Categorical Periodogram (CCP) and done spectral analysis. In CCP there is no longer overload to handle subsequences. Hota and Srivastava [27] observed that complex indicator sequence provides strong spectral component compared to EIIP indicator sequence. They observed that windowed DFT taking complex indicator sequence provides better exon prediction compared to windowed DFT taking EIIP indicator sequence and digital filters methods. Computational overhead is reduced by 75% in complex indicator sequence compared to binary indicator sequence. There is maximum discrimination between coding and non-coding regions in complex indicator sequence. The effect of window lengths on selected signal processing-based gene and exon prediction methods was firstly investigated by Akhtar et al. [28] and these methods were then optimized to improve their prediction accuracy by employing the best DNA representation, a suitable window length, and boosting the output signals to enhance protein coding and suppress the non-coding regions. It is shown therein that the optimized method outperforms major existing time-domain, frequency domain, and combined time-frequency approaches. By comparison with the existing DFT-based methods, the proposed method not only requires 50% less processing but also exhibits relative improvements of 53.3%, 46.7%, and 24.2% respectively over spectral content, spectral rotation and paired and weighted spectral rotation measures in terms of prediction accuracy of exonic nucleotides at a 5% false positive rate using the GENSCAN test set. Tomar et al. [29] said that filtering techniques are able to detect smaller exon region and adaptive MV filter minimize power in introns. MV filter suppress introns, so make exons peak more visible. Background noise is almost negligible in this case. Ahmad et al. [30] represented DNA symbolic-to-numeric sequence and compared with existing techniques in terms of relative accuracy for the gene and exon prediction problem. Ahmad et al. [31] incorporates denoising DNA signal with discrete wavelet transforms and indicator sequence. 1/f nose was greatly reduces by Upsampling and downsampling of signalAhmad et al. [32] proposed that discrete wavelet transform which greatly reduces the background noise and visible peak observed in power spectral estimation. The computational overhead reduces 75% than ordinary binary indicator sequence. They predict S.Cerevisiae Chromosomes gene sequence.

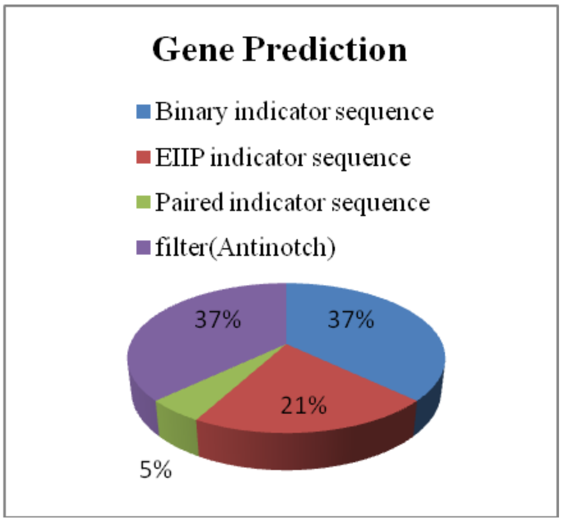

Fig: 1 Gene prediction with various indicator sequences

From Fig 1 it is found that a little work is to be done on paired indicator sequence.

4. CONCLUSION

a) When gene prediction is to be carried out using specialized softwares then the accuracy of GeneBuilder is found out to be maximum.
b) Lots of work has been done on antinotch filters and binary indicator sequence.
c) A few publications are found out on converting the character string of gene into paired indicator sequence and then passing through filter.
d) There is future scope of converting character string into paired number indicator sequence and also in Real number indicator sequence and then passing through the filter.

REFERENCES

[1] Chris Burge and Samuel Karlin, “Prediction of Complete Gene Structures in Human Genomic DNA”, J. Mol. Biol., pp 78-94, 1997.

[2] Jean-Michel Claverie,” Computational methods for the identification of genes in vertebrate genomic sequences”, Human Molecular Genetics, Vol. 6, No. 10 pp 1735–1744, 1997.

[3] Aron Culotta, David Kulp and Andrew McCallum, “Gene Prediction with Conditional Random Fields”, pp 1-14, 1998

[4] Steven L. Salzberg, Mihaela Pertea , Arthur L. Delcher , Malcolm J. Gardner, and Herve Tettelin “Interpolated Markov Models for Eukaryotic Gene Finding” Genomics 59, pp 24-31, 1999.

[5] Jonathan H. Badger and Gary J. Olsen, “CRITICA: Coding Region Identification Tool Invoking Comparative Analysis”, Molecular Biology and Evolution 16(4): pp 512-524, 1999.

[6]Luciano Milanesi, Dino D’Angelo and Lgor B. Rogozin, “GeneBuilder: Interactive in silico prediction of gene structure”, pp 612-621, 1999.

[7] Vineet Bafna Daniel H. Huson,” The Conserved Exon Method for Gene Finding”, pp 3-12, 2000.

[8] Kevin L. Howe, Tom Chothia, and Richard Durbin,” A Generic Framework for the Integration of Gene-Prediction Data by Dynamic Programming”, pp 1418-1427, 2002.

[9] Leila Taher Oliver Rinner , Saurabh Garg1, Alexander Sczyrba, Michael Brudno, Serafim Batzoglou and Burkhard Morgenstern, “ homology-based gene prediction”, Vol. 19 no. 12, pp 1575–1577, 2003.

[10] Mario Stanke and Stephan Waack,” Gene prediction with a hidden Markov model and a new intron submodel”, Vol. 19 Suppl. 2, pp 215–225, 2003.

[11] Jonathan E. Allen and Steven L. Salzberg,” integration of multiple sources of evidence for gene prediction”, Vol. 21 no. 18, pp 3596–3603, 2005.

[12] Samuel S. Gross and Michael R. Brent, “Using Multiple Alignments to Improve Gene Prediction” ,Journal of Computational Biology Volume 13, Number 2, 2006.

[13] Axel Bernal, Koby crammer, Artemis Hatzigeorgiou, Fernando Pereira, “Global Discriminative Learning for Higher Accuracy Computational Gene Prediction”, pp 0488-0497, March 2007.

[14] Jade P. Vinson, David DeCaprio, Matthew D. Pearson, Stacey Luoma, James E. Galagan, “Comparative Gene Prediction using Conditional Random Field”, 2007.

[15] M. Akhtar, J. Epps, and E. Ambikairajah, “Time and frequency domain methods for gene and exon prediction in eukaryotes,” in Proc. IEEE ICASSP, pp. 573−576, 2007.

[16] Yudong Cai, Zhisong He, Lele Hu, Bing Li, Yi Zhou, Han Xiao, Zhiwen Wang, Kairui Feng, Lin Lu, Kaiyan Feng, Haipeng Li, "Gene finding by integrating gene finders" , pp 1061-1068, 2010.

[17] Sitanshu Sekhar Sahu and Ganapati Panda," An efficient signal processing approach in eukaryotic gene Prediction", pp 1-12, 1998.

[18] Anastassiou D., "Frequency-domain analysis of biomolecular sequences", Oxford University Press, Bioinformatics, vol. 16, pp. 1073-1081, 2000.

[19] Anastassiou, D., "Genomic Signal Processing", IEEE Signal Processing Magazine, pp. 8 – 20, July 2001.

[20] P. P. Vaidyanathan, and B. -J. Yoon, "Gene and exon prediction using allpass-based filters," in Proc. IEEE GENSIPS (Raleigh, NC, USA), 2002.

[21] P. P. Vaidyanathan and B.-J. Yoon, "Digital filters for gene prediction applications," in Proc. Asilomar Conference on Signals, Systems, and Computers, pp. 306–310, Pacific Grove, Calif, USA, November 2002.

[22] P.P. Vaidyanathan, "Genomics and Proteomics: A Signal Processor's Tour" IEEE Circuits and Systems Magazine, pp. 6-29, 2004.

[23] T.W.Fox and A.Carreira, "A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression" in EURASIP journal on Applied Signal processing, pp. 108-114, 2004.

[24] Rao N. and Shepherd S. J., "Detection of 3-periodicity for small genomic sequences based on AR technique", International Conference on communications, Circuits and Systems, ICCCAS, vol. 2, pp. 1032- 1036, June 2004.

[25] E. Ambikairajah, J. Epps, and M. Akhtar, "Gene and exon prediction using time-domain algorithms," IEEE 8th Int. Symp. On Sig. Proc. and its Appl., pp. 199-202, 2005.

[26] Achuthsankar S. Nair and T. Mahalakshmi, "Are Categorical Periodograms and Indicator Sequences of Genomes Spectrally Equivalent?", pp 215-222, 2006.

[27] M.K. Hota and V.K.Srivastava, "DSP technique for gene and exon prediction taking complex indicator sequence" in proc. IEEE TENCON, pp. 1-6, 2008.

[28] M. Akhtar, J. Epps, and E. Ambikairajah, "Optimizing period-3 methods for eukaryotic gene prediction," in Proc. IEEE ICASSP, pp. 621-624, 2008.

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

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

[31]Muneer Ahmad, Azween Abdullah and Khalid Burraga, "Optimal Nucleotides Range Estimation in Diffused Intron-exon Noise", pp 178-183, 2010

[32] Muneer Ahmad, Azween Abdullah and Khalid Buragga, "A Novel Optimized Approach for Gene Identification in DNA Sequences", pp 806-814, 2011.