Survey and research proposal on Computational methods for gene prediction in Eukaryotes – A Report

Dhruv Jain

1. ABSTRACT

The rising popularity of genome sequencing in the field of Bioinformatics has resulted in the utilization of computational methods for gene finding in DNA sequences. Recently computer assisted gene prediction has gained impetus and tremendous amount of work has been carried out on this subject. Eukaryotic gene prediction is an important, long-standing problem in computational biology. This report (i) reviews the main mathematical models and computational algorithmic approaches to predicting genes in eukaryotic genomes and underlines their intrinsic advantages and limitations, (ii) Presents a method for a gene prediction using a three phase approach based on combination of comparison based (sequence alignment/database search) and ab-initio (standard HMM) methods, and (iii) Proposes methods for decreasing the computational time of the proposed tool using custom modifications in the various phases involved.

2. INTRODUCTION

Functionally, a eukaryotic gene can be defined as being composed of a transcribed region and of regions that cis-regulate the gene expression, such as the promoter region which controls both the site and the extent of transcription. The region between two transcribed regions is called intergenic. The promoter is in the intergenic region, immediately upstream of the gene and not overlapping with it. Genes sometimes overlap other genes creating a single transcriptional unit but these events are rare and complex enough that to my knowledge, no gene predictor has been able to explicitly model them [1].

For the automatic analysis and annotation of genomic sequences, computational gene prediction is becoming increasingly important [2]. Gene identification is for predicting the complete gene structure, particularly the accurate exon-intron structure of a gene in a eukaryotic genomic DNA sequence. After sequencing, finding the genes is one of the first and most significant steps in knowing the genome of a species [3]. Gene finding usually refers to the field of computational biology which is involved with algorithmically

recognizing the stretches of sequence, generally genomic DNA that are biologically functional. This specially not only involves protein-coding genes but may also include additional functional elements for instance RNA genes and regulatory regions as described above.

In general, most currently existing programs use two types of content sensors: one for coding sequences and one for non- coding sequences, i.e. introns, UTR (Universally Translated Regions) and intergenic regions.

3. REVIEW OF LITERATURE

Different Computational algorithms widely used in literature for gene prediction include, Support Vector Machine e.g. Kim et al.[5], which depends upon the SVMs for predicting the targets of a transcription factor by recognizing subtle relationships between their expression profiles.; Hidden Markov Model eg. Van Baren et al.[6], which can model the statistical dependencies between the adjacent bases; Software programs , Issac et al.[7], which can be trained for specific domain of organism or similar genes for greater accuracy; Machine Learning, Hoff et al.[8], where context-free grammars and other machine learning techniques are explored; Digital Signal Processing, Mabrouk et al. [9], where Discrete Fourier transforms (DFT) and filter based techniques are used and; Neural Networks algorithms based approach e.g. Mahony, et. Al [10]. The list is definitely not exhaustive.

3.1. Approaches

Computational gene prediction tools can be separated into two classes depending on how the content of exon/intron regions was assessed

3.1.1. Comparison based extrinsic approach: They exploit a sufficient similarity between a genomic sequence region and a protein, DNA or mRNA sequence present in a database in order to determine whether the region is transcribed and/or coding.
All the programs in this class may be seen as sophistications of the traditional Smith-Waterman local alignment algorithm where the existence of a signal allows for the opening (donor) or closure (acceptor) of a gap with an essentially free extension cost. Some software tools shown in table 1 are used for gene prediction in humans.

The obvious weakness of such extrinsic approaches is that nothing will be found if the

database does not contain a sufficiently similar sequence. Furthermore, even when a good similarity is found, the limits of the regions of similarity, which should indicate exons, are not always very precise and do not enable an accurate identification of the structure of the gene. Small exons are also easily missed.

3.1.2. *Ab initio* (Intrinsic Approach): Only about half of the genes can be found by homology to other known genes or proteins. In order to determine the remaining 50% of the genes, the only solution is to turn to predictive methods and to elaborate fast, accurate and reliable gene finders [11].
In *ab initio*(Latin: “from beginning”) approaches, Genomic DNA sequence alone is systematically searched for certain signals, specific sequences that indicate the presence of a gene nearby, or content, statistical properties of protein-coding sequence to identify protein-coding genes.

Ab initio gene finding in eukaryotes, especially complex organisms like humans, is considerably more challenging for several reasons. First, the promoter and other regulatory signals in these genomes are more complex and less well-understood than in prokaryotes, making them more difficult to reliably recognize. Second, since a typical protein-coding gene in humans might be divided into a dozen short exons, it is much more difficult to detect periodicities and other known content properties of protein-coding DNA in eukaryotes.

3.1.3. Integrated Approaches: To have the combined advantages of both, authors today are updating/developing tools to include evidence from both *ab-initio* and homology based programs. A pioneer in the area is GenomeScan[12] which is Burge's own extension of Genscan to incorporate similarity with a protein retrieved by BLASTX or BLASTP. Others include Twinscan[13], which use the FGENESH [16] and Genscan programs ; Eugene, which combines NetGene2[18] and SplicePredictor[19] for splice site prediction, NetStart [14] for translation initiation prediction, IMM-based content sensors and similarity information from protein, EST and cDNA matches ; GAZE[15] etc.

3.2. Problems with the current approaches

Several Issues like very long genes, very long introns, very conserved introns and very short exons, make the problem of eukaryotic gene finding extremely difficult [17].
So, there is a need to make the currently available tools more sophisticated to handle these and other boundary cases that might occur.

Also, there is a need for clean sequence databases which are not redundant, contain reliable and relevant annotations and provide all necessary links to further data [44].

4. METHODOLOGY FOR THE PROPOSED TOOL

There are three major phases involved in the tool:

4.1 Database search via sequence alignment

The first phase is to go through the genome and annotate genes that are high similarity matches to already known eukaryotes genes. The entire list of currently known human genes has been compiled and is referred to as RefSeq by Otto gene prediction tool[50]. I would use the RefSeq if my tool is used only for human gene prediction. Otherwise I would go with the available databases, or if possible form a database of my own based on the redistricted organisms for my tool.

For Otto, the cutoff for annotation of a gene when comparing to RefSeq is that the genomic sequence has to match at least 50% of its length to the RefSeq. The sequence identity must be greater than 92%. For my tool, the threshold will be heuristically determined once the tool is run on benchmarks of known genes.

For this phase the **SeqAlignFPGA tool**[45] which I am currently speeding up under my SURA would be used. It uses a prefix trie model and backward search algorithm to match the query sequence with the reference gnome. The computational time is linear with the length of the query sequence. It is **more accurate and several to tens of times faster than BLAST**.

4.2. Alternative splice site prediction

The second phase involves alternative splice site prediction. Here I plan to **use a customized variation of MaxEntScan.** The tool is based on maximum entropy principle. Although, it is the most accurate and sensitive tool till date, it lacks specificity as compared to other good tools[48]. Unlike the usual method for training it with the generalized data, I **plan to make a model based on the results of the first approach.** This means that only families of DNA which have a significant match with the query sequence (above the threshold) shall be entertained for the so called "prior data". This would increase the low specificity of the tool and further increase the sensitivity.

4.3. Ab-initio gene prediction based on standard HMM model

Now the second phase would device the gene into exons and introns. The third phase uses standard HMM for ab initio gene prediction. It is well known from literature that using different content sensors and thus different models for coding and non-coding regions is always a good way of proceeding with gene prediction [51]. In case of neural networks based approach and specifically **standard** HMMs, this is even more critical since high sequence similarity is needed. The training set generated from the results of phase 1 will thus be further enhanced for construction of two exclusive standard HMMs. Moreover, the reason for my choice of standard HMM instead of generalized HMM is that it is an integrated model. Most of the content sensors would be covered here.

Researches in the field of machine learning have developed a Fast **Two "Level HMM Decoding Algorithm for Vocabulary Handwriting Recognition (FTLDA)**[49]**.** This algorithm breaks up the computation of words into two levels: state (or word) level and character (or letter) level. An analogous decoding algorithm for standard HMM for gene prediction is also possible by using patterns (hexamers, dinucleotides) for the first level and the individual nucleotide bases for the second level. **This would speedup the decoding process by as much as 15 times for 240,000 nts, the size of a large human gene.**

Finally, similar to Otto, **the result will be validated by comparison with EST, protein, and genomic sequence databases** using SeqAlignFPGA (or BLAST, since it is widely used) and also using **phylogenetic programs like Paup or Phylip (after doing the SeqAlignFPGA)** to identify homology and paralogy. For example, in the case of protein comparison, the sequence will be translated and matched against protein database. Similarly the gnomic sequence will be matched with EST. The reduction in computational time in the above phases will also enable this stage to complete faster.

5. RESULTS

This report:

(i) Successfully reviewed the main mathematical models and computational algorithmic approaches to predicting genes in eukaryotic genomes and underlines their intrinsic advantages and limitations.

(ii) Presented a method for a gene prediction using a three phase approach based on combination of comparison based (sequence alignment/database search) and ab-inito (standard HMM) methods.

(iii) Speeded up the tool by using custom modifications (i) using SeqAlignFPGA for database search instead of BLAST; and (ii) using modified FTLDA for HMM decoding algorithm. Since the critical stage is the FTLDA algorithm predicted speedup is ~15 times for a large human gene as compared to other combined approaches. The time for the extra second phase will be amortised due to high speed-up obtained in the first phase.

This approach is **likely to identify many types of functional and structural genes, transposons, retrotransposons, CG Islands, SSRs, interspersed repeats, tandem repeats, segment duplications, ALU repeats (in GG rich regions), homologs/ paralogs, small sections such as SNP as well as large isochors.** Non-coding exons can be predicted during the second phase of alternative splicing. Overlapping genes can also be predicted (multiple ORFs would have common series of nucleotide) in phase 1 and 2.

It might be hard to identify pseudogenes, promoter and terminators based on the standard HMM used (according to the past results, performance is likely to degrade). Depending upon the results of the test benchmarks**, a possible way to identify the end regions is through incorporating tools like PROMOSER, TransTerm, FindTerm etc. before the third phase.** This way genes in which (i) CpG Islands (present close to promoters, which is 10 to 35 upstream of the start codon) are likely to be methylated; and/or (i)having promoters which are GC or CpG rich instead of AT will be classified as pseudogenes and these results would be incorporated into the standard HMM model. Similarly other factors can also be incorporated.

Comparison with EST, protein, and genomic sequence databases using SeqAlignFPGA/ BLAST and using phylogenetic programs, of the predicted gene shall be done and the average specificity and sensitivity will be reported. Although, this approach is time consuming, I feel this is the **best way measure the success of the tool**.

Experiments to execute include initial benchmark sequences (of human, drosophila and mouse) which are already known and present in database are required to be run as

query sequences with this tool to optimise both the default configuration(E value, filters etc.) and the hidden parameters (various thresholds, markov model order) involved.

6. DISCUSSION AND FUTURE WORK

This tool is predicted to work well under all circumstances except for end regions for which an alternate is provided. This tool is predicted to be more accurate than existing tools and due to alternative measures used, also computationally less intensive than other combined approaches. Key for the success of this tool for better prediction is separation of introns and exons sites before using standard HMM models.

Although, I tried my best but no tools are free from defects. Limitations remain because of limitations of the database (only a fraction of the genes have been identified) (phase 1) and the limitation of the ab-initio methods which propagate in the HMM (phase 3).

Our approach will help design better gene prediction tools based on the three phases involved. Our methodology for speedup will also result in development of faster and reliable tools. Future work includes incorporating 3D-DNA structure prediction using approach analogous to protein structure prediction using online multiplayer games [46] and other methods.

To conclude, from this term paper assignment, I learned that metabolism involved in the transcription and translation of gene into protein is complex and thus, gene prediction is indeed a complex process. There are many factors involved and no matter how far we go, there will be problems that will still be waiting for us to solve. This term paper increased my appreciation for all the researchers, both computer scientist and biologist working in this domain which if looked superficially, is a next-to-random process. Therefore the proposed tool, like other tools is likely to have drawbacks.

References:

[1] “Gene Prediction”. Online. [Available] http://www.wikipedia.org
[2] Wang, Chen and Li, "A brief review of computational gene prediction methods", Genomic Proteomics, Vol.2, No.4, pp. 216-221, 2004
[3] Rabindra Ku.Jena, Musbah M.Aqel, Pankaj Srivastava, and Prabhat K.Mahanti, "Soft Computing Methodologies in Bioinformatics", European Journal of Scientific Research, Vol.26, No.2, pp.189-203, 2009
[4] D. Sundar, “BEL418, Bioinformatics” Spring, 2012. Indian Institute of Technology Delhi, India
[5] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee and Byoung- Tak Zhang, "miTarget: microRNA target gene prediction using a support vector machine", BMC Bioinformatics, Vol.7, No.411, pp.1-14, 2006
[6] Marijke J. van Baren and Michael R. Brent, "Iterative gene prediction and pseudogene removal improves genome annotation", Genome Research, Vol.16, pp.678-685, 2006
[7] Biju Issac and Gajendra Pal Singh Raghava, "EGPred: Prediction of Eukaryotic Genes Using Ab Initio Methods after combining with sequence similarity approaches", Genome Research, Vol.14, pp.1756-1766, 2004
[8] Katharina J Hoff, Maike Tech, Thomas Lingner, Rolf Daniel, Burkhard Morgenstern and Peter Meinicke, "Gene prediction in metagenomic fragments: A large scale machine learning approach", BMC Bioinformatics, Vol. 9, No.217, pp.1-14, April 2008.
[9] Mai S. Mabrouk, Nahed H. Solouma, Abou-Bakr M. Youssef and Yasser M. Kadah, "Eukaryotic Gene Prediction by an Investigation of Nonlinear Dynamical Modeling Techniques on EIIP Coded Sequences", International Journa Kleffe,J., Hermann,K., Vahrson,W., Wittig,B. and Brendel,V. (1996)
Logitlinear models for the prediction of splice sites in plant pre-mRNA
sequences. Nucleic Acids Res., 24, 4709±4718. l of Biological and Life Sciences, Vol. 3, No.4, pp. 225-230, 2007
[10] Shaun Mahony, Panayiotis V. Benos, Terry J.Smith and Aaron Golden, Self-organizing neural networks to support the discovery of DNA-binding motifs", Neural Networks, Vol.19, pp.950-962, 2006
[11]Fickett,J.W. (1996) Finding genes by computer: the state of the art. Trends Genet., 12, 316-320.
[12]Yeh,R.-F., Lim,L.P. and Burge,C.B. (2001) Computational inference of homologous gene structures in the human genome. Genome Res., 11 , 803-816.
[13] Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. Bioinformatics, 17, S140-S148.
[14]Pedersen,A.G. and Nielsen,H. (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. In Gaasterland,T., Karp,P., Karplus,K., Ouzounis,C., Sander,C. and Valencia,A. (eds), The Fifth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp. 226-233
[15] “Gene Finding”. Online. [Available] “http://ismb01.cbs.dtu.dk/ GeneFinding.html#A304”
[16] A.Salamov and V.Solovyev, Online. [Available] http://genomic.sanger. ac.uk/
[17]Mathé, Catherine and Sagot, Marie‑France and Schiex, Thomas and Rouzé, Pierre, "Current methods of gene prediction, their strengths and weaknesses", Nucleic Acids Research, 19, pp. 4103-4117.
[18] Tolstrup,N., Rouze,P. and Brunak,S. (1997) A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res., 25, 3159-3163.
[19] Kleffe,J., Hermann,K., Vahrson,W., Wittig,B. and Brendel,V. (1996) Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. Nucleic Acids Res., 24, 4709-4718.
[20] Jiang,J. and Jacob,H.J. (1998) EbEST: an automated tool using expressed sequence tags to delineate gene structure. Genome Res., 8, 268-275.
[21] Birney,E. and Durbin,R. (1997) Dynamite: a ¯exible code generating language for dynamic programming methods used in sequence comparison. Proc. Int. Conf. Intell. Syst. Mol. Biol., 5, 56 64.
[22]Rogozin,I.B., Milanesi,L. and Kolchanov,N.A. (1996) Gene structure prediction using information on homologous protein sequence. Comput. Appl. Biosci., 12, 161-170.
[23]Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) Gene recognition via spliced sequence alignment. Proc. Natl Acad. Sci. USA, 93, 9061-9066.
[24]Novichkov,P.S., Gelfand,M.S. and Mironov,A.A. (2001) Gene recognition in eukaryotic DNA by comparison of genomic sequences. Bioinformatics, 17, 1011-1018.
[25]Batzoglou,S., Pachter,L., Mesirov,J., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res., 10, 950-958.
[26]Wiehe,T., Gebauer-Jung,S., Mitchell-Olds,T. and Guigo,R. (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. Genome Res., 11, 1574-1583.
[27]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 Res., 8, 967-974
[28]Pachter,L., Alexandersson,M. and Cawley,S. (2002) Applications of generalized pair hidden Markov models to alignment and gene finding problems. J. Comput. Biol., 9, 389-399. Pennisi,E. (1999) Keeping genome databases clean and up to date.
Science, 286, 447-450

[29]Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. Genome Res., 11, 1952-1957. [30]Rogozin,I.B., D'Angelo,D. and Milanesi,L. (1999) Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. Gene, 226, 129-137.
[31]Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res., 11, 889-900.
[32]Guigo,R., Knudsen,S., Drake,N. and Smith,T. (1992) Prediction of gene structure. J. Mol. Biol., 226, 141-157
[33]Solovyev,V. and Salamov,A. (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In Gaasterland,T., Karp,P., Karplus,K., Ouzounis,C., Sander,C. And Valencia,A. (eds), The Fifth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp. 294-302
[34]Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res., 26, 1107-1115.
[35]Fields,C.A. and Soderlund,C.A. (1990) gm: a practical tool for automating DNA sequence analysis. Comput. Appl. Biosci., 6, 263-270
[36]Snyder,E.E. and Stormo,G.D. (1995) Identification of protein coding regions in genomic DNA. J. Mol. Biol., 248, 1-18
[37]Reese,M.G., Eeckman,F.H., Kulp,D. and Haussler,D. (1997) Improved splice site detection in Genie. In Istrail,S., Pevzner,P. and Waterman,M.
(eds), First Annual International Conference on Computational Molecular Biology (RECOMB). ACM Press, New York, NY, pp. 232-240.
[38]Dong,S. and Searls,D.B. (1994) Gene structure prediction by linguistic methods. Genomics, 23, 540-551
[39]Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268, 78-94.
[40] Milanesi,L., Kolchanov,N.A., Rogozin,I.B., Ischenko,I.V., Kel,A.E., Orlov,Y.L., Ponomarenko,M.P. and Vezzoni,P. (1993) GenView: a computing tool for protein-coding regions prediction in nucleotide sequences. In Lim,H.A., Fickett,J.W., Cantor,C.R. and Robbins,R.J. (eds), Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis. World Scientific Publishing, Singapore, pp. 573-588.
[41]Xu,Y., Mural,R.J. and Uberbaker,E.C. (1994) Constructing gene models from accurately predicted exons: an application of dynamic programming. Comput. Appl. Biosci., 10, 613-623
[42]Krogh,A. (1997) Two methods for improving performace of a HMM and their application for gene finding. In Gaasterland,T., Karp,P., Karplus,K., Ouzounis,C., Sander,C. and Valencia,A. (eds), The Fifth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp. 179-186.
[43] Henderson,J., Salzberg,S. and Fasman,K. (1997) Finding genes in human DNA with a hidden Markov model. J. Comput. Biol., 4, 127-141.
[44]Pennisi,E. (1999) Keeping genome databases clean and up to date. Science, 286, 447-450
[45]Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760.
[46]Popovic Z, Cooper S, Khatib F, Treuille A, Barbero J, et al. Predicting protein structures with a multiplayer online game. *Nature.* 2010;466:756–760
[47] Ahn L. v, Dabbish L. *Labeling images with a computer game. Proceedings of the SIGCHI Conference on Human Factors in computing Systems.* Vienna: ACM; 2004. pp. 319–326.
[48]NGRL Splice Site Tools Analysis. [Online]. Available:http://www.ngrl.org.uk/Manchester/sites/default/files/publications/Informatics/NGRL_Splice_Site_Tools_Analysis_2009.pdf
[49] Koerich, A.L.; Sabourin, R.; Suen, C.Y.; "Fast two-level HMM decoding algorithm for large vocabulary handwriting recognition," Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on , vol., no., pp. 232- 237, 26-29 Oct. 2004
[50]Finding genes by computational methods: An analysis of methods and programs. [Oniline]. Available: biochem218.stanford.edu/Projects%202002/Sen.pdf
[51]Do JH, Choi DK, Computational approaches to gene prediction, Journal of Microbiology 44(2):137–144, 2006.