Asian Journal of Information Technology 5(5) : 512-517, 2006

A New Approach to Spliced Alignment Gene Prediction Algorithm

Mohammad Zakir Hossain Sarker, Jubair Al Ansary and Md Shajjad Hossain Khan
Dewan Md. Rashed Iqbal
Department of CSE, East West University 43, Mohakhali, Dhaka-1212, Bangladesh

Abstract: Genomics is the field of study that seeks to understand the structure and function of all genes in an organism based on knowing the organism's entire Deoxyribonucleic Acid (DNA) sequence and extensive reliance on powerful computer technologies. The science of Bioinformatics, which is the bonding between molecular biology with computer science, needed genomic information to contribute in various disciplines. In recognition of that, many universities, government institutions and pharmaceutical firms have formed bioinformatics groups, consisting of molecular biologists and computer scientists. All of these bioinformatics groups are depended on laboratory experiments along with web base resources. These web recourses are time consuming and rigorous to access and use. This study has studied some of the existing gene prediction algorithms, which are used behind these web-based resources. Gene prediction algorithm predicts the probability of existing genes in the biological sequences such as protein sequences or DNA sequences. This study also describes a new approach to spliced alignment algorithm, one of the mostly used gene prediction algorithms. This new approach is more accurate and will overcome the complexities of the existing algorithm. And also will make the whole research procedure faster and easier.

Key words: Bioinformatics, gene prediction, genomics, DNA sequence, protein sequence

INTRODUCTION

According to Andreas D and Francis B.[1], today it is not possible to accurately determine more than about 1000 consecutive base pairs of a DNA sequence. This is a major limitation for large-scale sequencing projects. DNA fragment assembly is a technique that attempts to reconstruct the original DNA sequence from a large number of fragments, each several hundred base pairs long. The DNA fragment assembly is needed because current technology, such as gel electrophoresis, cannot directly and accurately sequence DNA molecules longer than 1000 bases. However, most genomes are much longer. For example, a human DNA is about 3.2 billion nucleotides in length and cannot be read at once (according to Brown, P.O. and Botstein, D.[2]).

Gelfand, M. S.[3] has described that computational gene prediction has been an active area of research for over 20 years. The algorithms that have been developed are traditionally categorized as either *ab initio* or alignment-based. Ab initio methods, such as GENESCAN and GENIE make perditions using only the DNA sequence to be annotated and a model of gene structure. Alignment-based methods such as PROCRUSTES and GENEWISE which attempts to align homologous proteins to genomic sequence. In this study we have studied 3 (three) popular and mostly used gene prediction algorithms. These are as follows:

- Gene prediction by GeneMark.hmm
- Gene Predictions via Spliced Alignment
- Genomic homology into gene prediction

After studying all of the above algorithms we have changed and modified the Spliced Alignment Algorithm and have given a new-look to it. Before going to more detail let's understand few of the basic terms.

WHAT IS GENE?

According to Jonathan E., Mihaela P. and Steven L[4] a single gene consists of a unique sequence of DNA that provides the complete instructions to make a functional product, called a protein. Genes instruct each cell type-such as skin, brain and liver-to make discrete sets of proteins. The nucleus contains long strands of DNA that encode this genetic information. A DNA chain is made up of four chemical bases: Adenine (A) and Guanine (G), which are called purines and Cytosine (C) and Thymine (T), referred to as pyrimidines (Fig. 1). Each base has a slightly different composition, or combination of oxygen, carbon, nitrogen and hydrogen. The chemical nature of the bases in double-stranded DNA creates a slight twisting force that gives DNA its characteristic gently

Corresponding Author: Mohammad Zakir Hossain Sarker, Department of CSE, East West University 43, Mohakhali, Dhaka-1212, Bangladesh

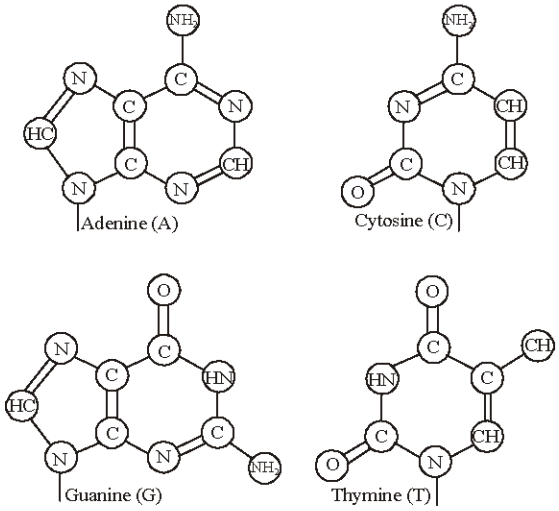

Fig. 1: The four DNA bases

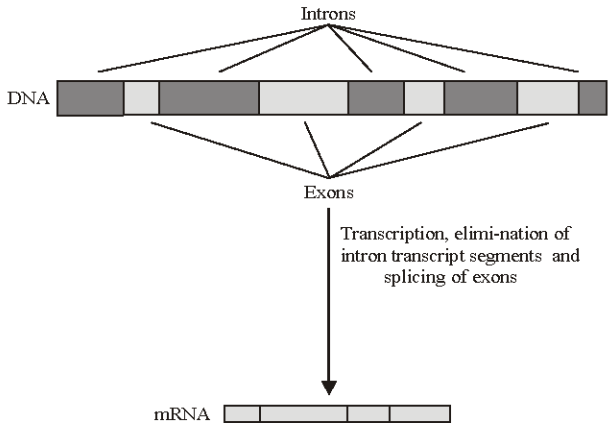

Fig. 2: Exons and Introns

coiled structure, known as the double helix. A-T and G-C base pairs are said to be complementary. One strand of DNA can act as a template to direct the synthesis of a complementary strand.

Andrade M. A. and Sander C.[5] say that only a small percentage of the 3 billion bases in the human genome becomes an expressed gene-product. Approximately 1 percent of our genome that is expressed, 40% is alternatively spliced to produce multiple proteins from a single gene.

THE CORE GENE SEQUENCE: INTRONS AND EXONS

It is described in Huang, X., Adams, M.D., Zhou, H. and Kerlavage, A.R.[6] that genes make up about 1% of the total DNA in our genome. In the human genome, the coding portions of a gene, called exons, are interrupted by dominant sequences, called introns. Both exons and introns are "transcribed" (copied) into messenger Ribonucleic Acid (mRNA), but before it is transported to the ribosome, the primary mRNA transcript is edited. This editing process removes the introns, joins the exons together and adds unique features to each end of the transcript to make a "mature" mRNA. Fig. 2 shows the Exons and Introns.

GENE PREDICTION USING COMPUTERS

According to Claverie J., Notredame C.[7], not all of the human genome sequence is annotated and not all of the known sequence has been assigned a particular position in the genome. When the complete mRNA sequence for a gene is known, computer programs are used to align the mRNA sequence with the appropriate region of the genomic DNA sequence. This provides a reliable indication of the beginning and end of the coding region for that gene. In the absence of a complete mRNA sequence, the boundaries can be estimated by ever-improving, but still inexact, gene prediction software.

The problem is the lack of a single sequence pattern that indicates the beginning or end of a eukaryotic gene. Fortunately, the middle of a gene, referred to as the core gene sequence, has enough consistent features to allow more reliable predictions.

Open reading frames, stretches of DNA, usually greater than 100 bases not interrupted by a stop codon such as TAA, TAG or TGA; start codons such as ATG (here: Adenine (A), Guanine (G) nd Thymine (T)); specific sequences found at splice junctions, a location in the DNA sequence where RNA removes the non-coding areas to form a continuous gene transcript for translation into a protein; and gene regulatory sequences. This process is dependent on computer programs that search for these patterns in various sequence databases and then make predictions about the existence of a gene[8].

VARIOUS GENE PREDICTION ALGORITHMS

We have studied and understood few gene prediction algorithms and then modified the Spice Alignment Algorithm to get better result.

Gene prediction by genemark.hmm: A general pattern recognition algorithm should be able to compute the probability that a particular functional sequence A underlies a given sequence S, P(A\S) = P(a1, a2, ..., aL\b1, b2, ..., bL) (Gelfand, M.S., Mironov, A.A. and Pevzner, P.[9]).

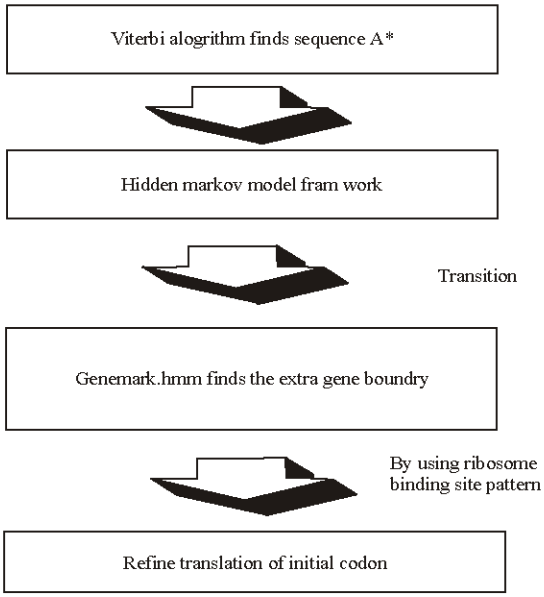

Fig. 3: Major processes of genemark. HMM

As per Loytynoja A, Milinkovitch M.C.[10] the core GeneMark.hmm procedure computes the P(A\S) value and eventually, defines the functional sequence A* having the largest value P(A*\S) among all possible A. The functional sequence A*, the output of the algorithm, describes the most likely annotation of the DNA sequence S.

To further improve the prediction of the translation start position the model of the Ribosome-Binding Site (RBS) was derived. This model was used to refine translation initiation codon prediction at the post-processing step.

Since only in a few cases is the precise position of the translation initiation codon known from an experiment. However, the database annotation of the initiation codon represents the expert decision summarizing much indirect evidence and is thought to be close to the real one. Also there were genes missed by GeneMark.hmm, mainly due to overlaps that were recovered by GeneMark. However, the GeneMark.hmm program made several new predictions and some of them were confirmed by similarity search. Figure 3 shows the Major Processes of GeneMark.hmm.

Gene predictions via spliced alignment: Gelfand, M. S., Mironov, A. A. and Pevzner, P.[9] describes that this algorithm explores all possible exon assemblies in polynomial time and finds the multiexon structure with the best fit to a related protein. This is the main feature of the algorithm distinguishing it from other programs. Figure 4 shows the Major Processes in Spliced Alignment Algorithm.

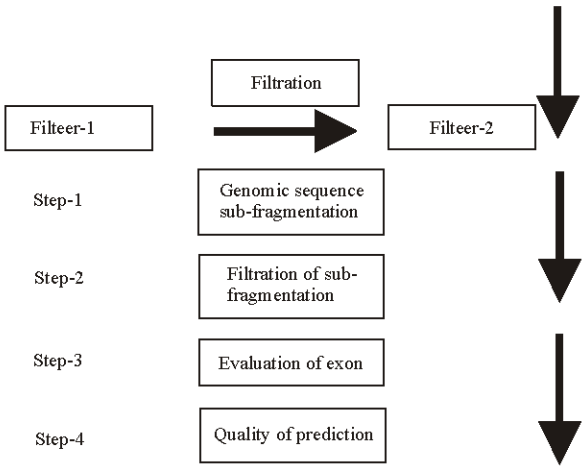

Fig. 4: Major processes in spliced alignment algorithm

Filtration: Initially all *initial exons* bounded by a start codon, ATG and a candidate donor site. GT, *internal exons* bounded by an acceptor site AG and a donor site. GT, *terminal exons* bounded by an acceptor site AG and a stop codon. TGA, TAA, or .TAG are considered (and denote the left and right boundaries of a coding region, respectively).

Filtration consists of two weak filters removing clearly abnormal exons and a final filter of adjustable strictness. The first filter removes exons with weak splicing sites as estimated by positional nucleotide weight matrices. The threshold is set very low and only two actual acceptor sites are filtered out at this step.

At the second step the genomic sequence is divided into sub fragments of length 10 kb with 2.5 kb of overlap. Further filtration is performed independently in each sub fragment and the candidate exons are evaluated by a scoring function taking into account strength of the splicing sites and the coding potential.

Overall, the two preliminary filters decrease the number of candidate exons approximately 15-fold, while losing 12 actual exons in the entire sample. One thousand highest scoring exons are retained in each sub fragment. This filter loses 10 actual exons: 1 initial, 6 internal and 3 terminals. It should be noted that the statistical properties of these exons are so unusual that any conventional gene recognition algorithm will likely lose them. At the same time, preliminary filtering sharply decreases the number of candidate exons, making the final filter more robust.

Denote the score of a chain Γ by $|\Gamma|$ an exon score is now defined as either Partition Function Rescoring which is according to Gelfand, M.S., Podolsky, L.I., Astakhova, T.V. and Roytberg, M.A.[11],

Where c is some fixed constant.

$$P(e) = \sum_{\Gamma \ni e} e^{c|\Gamma|}$$

Or Best Chain Rescoring,

$$B(e) = \max_{\Gamma \ni e} |\Gamma|$$

The ranking is performed independently for initial, internal and terminal exons. The proportion of these three classes of exons in the filter output is 1:3:1, respectively.

Thus the filtering is controlled by two switches (the maximal number of exons in chains E = 1, 2, or 3 and the use of P or B scores) and the *filtration stringency* parameter F. This parameter determines the number of exons dependent on the genomic sequence length. Following preliminary analysis, three values of this parameter have been considered: 1 exon per 14 nucleotides (F =14, weak filtration), 1 exon per 33 nucleotides (F= 33, moderate filtration).

And 1 exon per 100 nucleotides (F =100 strong filtration). Note that if E =1 all filters coincide. Single-exon genes were considered separately in an analogous manner. The minimum length of such genes was set to 180 nucleotides; one candidate exon was retained per 200 bp of the genomic sequence. The quality of prediction was assessed using the correlation coefficient between the predicted and the actual genes,

$$C = \frac{T_P . T_N - F_P . F_N}{\sqrt{(T_P + F_P).(T_N + F_N).(T_P + F_N).(T_N + F_P)}}$$

Where T_P and T_N are the numbers of correctly predicted coding (true positive) and non coding (true negative) nucleotides, respectively, FN is the number of missed coding (false negative) nucleotides and FP is the number of non coding nucleotides predicted to be coding (false positive) as per Mount D.W.[12].

Since the targets have been chosen by the BLAST database search (via *Entrez*), many targets have only local similarities with the analyzed genes and thus produce artifacts when the (global) spliced alignment is performed.

Genomic homology into gene prediction: Wiley S.R.[13] has described that, TWINSCAN is a new gene-structure prediction system that directly extends the probability model of GENSCAN allowing it to exploit homology between two related genomes. Separate probability models are used for conservation in exons, introns, splice sites and UTR. TWINSCAN is specifically designed for the analysis of high-throughput genomic sequences containing an unknown number of genes.

TWINSCAN is based on GENSCAN ++, C ++ re implementation of GENSCAN. In order to train and test our algorithm, we needed annotated genomic sequences and their homologs. The sequences should be large and contain a mixture of complete and incomplete gene, single and multi exon genes and genes on g both strands.

The new algorithm performance is often bench marked with using dataset and methods of Burset Guigo. This data set consists of 570 short genomic sequences (average length 5074 bp) completing one complete multiexon gene with out alternatively spliced form.

A common scenario is the analysis of High-Throughput Genomic (HTG) sequences, which are generally 100-200 Kb in length and contain an unknown number of genes. Probability models the GENSCAN model is based on an explicit state duration Hidden Markov.

Model (HMM). Each state of the HMM corresponds to one of the seven categories with which all nucleotides are ultimately labeled- promoter, 5' UTR, exon, intron, etc.

TWINSCAN consist of the new, joint probability model on DNA sequence and conservation sequence, together which the same optimization algorithm used by GENSCAN[14].

PROPOSED SPLICED ALIGNMENT ALGORITHM FOR GENE PREDICTIONS

The Spliced Alignment Filtration consists of two weak filters removing clearly abnormal exons and a final filter of adjustable strictness. The first filter removes exons with weak splicing sites as estimated by positional nucleotide weight matrices. The threshold is set very low and only two actual acceptor sites are filtered out at this step. In this *filtration stringency* parameter F This parameter determines the number of exons dependent on the genomic sequence length. Following preliminary analysis, three values of this parameter have been onsidered: 1 exon per 14 nucleotides (F =14, weak filtration), 1 exon per 33 nucleotides (F= 33, moderate filtration) and 1 exon per 100 nucleotides (F =100 strong filtration).

Instead of using biological probes, randomly select short probes (e.g. 12 bps) from each fragment. Uses exact pattern matching in determining the relative positions (i.e. probes occurrences) of the input fragments, including their reverse complements.

From above two circumstances we decided the Spliced Alignment Algorithm can be enhanced through using a particular filter instead of three. And the way to chose the filter is the way structured pattern matching

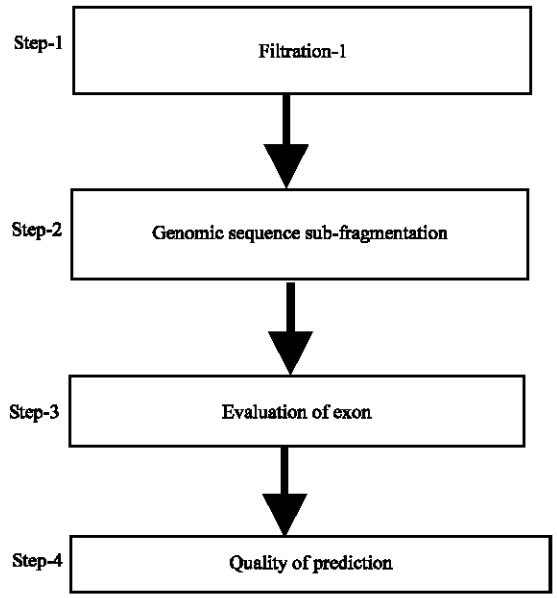

Fig. 5: Major processes for proposed spliced alignment algorithm

worked at the first phase. Where it said that the probe would be selected randomly. In our proposed algorithm we said the filter would be like a moderate one consisting of 32 nucleotides. And to overcome different length problem we would use structured alteration of this filter arrangement.

As in original algorithm all *initial exons* bounded by a start codon, ATG and a candidate donor site. GT, *internal exons* bounded by an acceptor site AG and a donor site. GT, *terminal exons* bounded by an acceptor site AG and a stop codon .TGA, TAA, or .TAG are considered. The proposed single filter will follow this basic structure. Additionally, to find the all-possible matches, it will be permutated it's sites for every sequence. And thus the algorithm will be closer to maximum accurate match. It has been found that for a single sequence, the filter will test 267 different variations. And those variations will represent all the probable gene forms. So the accuracy rate is expected to be more then 99%. Figure 5 shows the Major Processes for Proposed Spliced Alignment Algorithm.

METERIALS AND METHODS

This research is basically based on literature study. To understand the basics of Bio-Informatics initially we have studied few books like Andreas D. and Francis B.[1], Claverie J., Notredame C.[7], Mount D. W.[12] etc. and then gone through few proceedings and journal papers, articles from the Internet like Edwards D. and Batley J.[15]. Andrade M.A. and Sander C.[5], Narasimhan, G. [8,16-20]etc. After learning the basics of Bio-Informatics we have studied detail about Genomics since we planned to work on this topic. For doing so we have gone through various resources like Brown, P.O. and Botstein, D.[21], Cornish-Bowden A. and Cardenas M.L.[22], Cowley A.W.[23], Kim, S. and Segre, A.M.[24], Rudd S.[25], Stougaard J.[26], Ziv M. and Vienne D.[27-32] etc. Then we have rigorously gone through few Gene prediction algorithms i.e. Gene prediction by GeneMark.hmm, Gene prediction via Spliced Alignment and Genomic homology into gene prediction reading various resources like Gotoh, O.[33], Howe K.L., Chothia T. and Durbin R.[34], Huang, X., Adams, M.D., Zhou, H. and Kerlavage, A.R.[6,35,36] etc. After knowing the detail about these algorithms we have modified the Spliced Alignment algorithm which has been described in this study.

CONCLUSION

Computational gene prediction methods have yet to achieve perfect accuracy, even in the relatively simple prokaryotic genomes. Problems in gene prediction centre on the fact that many protein families remain uncharacterized. As a result, it seems that only approximately half of an organism's genes can be confidently predicted on the basis of homology to other known genes. In this study we have tried to understand the complexities of few gene prediction algorithms in Genomics. We have also tried to find out the pros and cons of those algorithms. Finally we have proposed a new approach to the Splice Alignment Algorithm considering the advantages and disadvantages of it. We hope the proposed algorithm will be able to overcome the complexities of the current algorithm and will ensure more accuracy. The future scope of this project is to implement and test of the proposed algorithm and eventually to serve the bioinformatics industry.

REFERENCES

1. Andreas, D. and B. Francis, 2001. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Wiley Inter Science Press, 2nd Edition.
2. Bork, P. and R. Copley, 2001. Genome speaks, Nature,
3. Gelfand, M.S., 1996. Prediction of function in DNA sequence analysis, Proceedings of Fourth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, pp: 87-115.
4. Jonathan, E., P. Mihaela and L. Steven 1996. Computational Gene prediction using multiple Sources of Evidence, Proceedings of Fourth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, pp: 184-290.

5. Andrade, M.A. and C. Sander, 1997. Bioinformatics: From genome data to biological knowledge. Bio-Informatics, pp: 675-683.
6. Huang, X., M.D. Adams, H. Zhou and A.R. Kerlavage, 1997. A tool for analyzing and annotating genomic sequences, Genomics, pp: 37-45.
7. Claverie, J. and C. Notredame, 2004. Bioinformatics-A Beginner's Guide, WILEY-Dreametech India Pvt Ltd.
8. Concepts and Terms in Genetic Research-A Primer-An article available at http://www .niaaa.nih. gov/publications/arh26-3/165-171.htm last visited on December 29, 2005.
9. Gelfand, M.S., A.A. Mironov and P. Pevzner, 1996. Gene recognition via spliced sequence alignment, Proceedings of National Acad. Sci., USA, pp: 9061-9066.
10. Loytynoja, A. and M.C. Milinkovitch, 2003. A hidden Markov model for progressive multiple alignments, Bioinformatics, pp: 1505-1513.
11. Gelfand, M.S., L.I. Podolsky, T.V. Astakhova and M.A. Roytberg, 1996. Recognition of gene in human DNA sequences, J. Computational Biology, pp: 223-234.
12. Mount, D.W., 2004. Bioinformatics, Cold Spring Harbor Press.
13. Wiley, S.R., 1998. Genomics in the real world, Pharm. Des., pp: 417-422.
14. http://genes.mit.edu/ last visited on December 18, 2005.
15. Edwards, D. and J. Batley, 2004. Plant bioinformatics: From genome to phenome, Trends Biotechnol, pp: 232-237.
16. Narasimhan, G., 2005. What is Bioinformatics? Why Bioinformatics?,-Available at http://www.cs. fiu. edu/~giri/bioinf/SACM02.pdf.
17. For various databases - http://www.ncbi.nlm.nih.gov/ last visited on December 19, 2005.
18. http://www.biosino.org/hgp/Science-Pbo291(5507)1219.htm
19. http://www.oscargruss.com/CommercialBioinformatics.pdf.
20. http://www.ncbi.nlm.nih.gov/ last visited on December 27, 2005.
21. Brown, P.O. and D. Botstein, 1999. Exploring the new world of the genome with DNA micro-arrays, Nature Genetics, pp: 33-37.
22. Cornish-Bowden, A. and M.L. Cardenas, 2001. Functional genomics. Silent genes given voice, Nature, pp: 571-572.
23. Cowley, A.W., 1999. The emergence of physiological genomics, Nature, pp: 83-90.
24. Kim, S. and A.M. Segre, 1999. AMASS: A structured pattern matching approach to shotgun sequence assembly, J. Computational Biology, pp: 163-186.
25. Rudd, S., 2003. Expressed sequence tags: Alternative or complement to whole genome sequences? Trends Plant Sci., pp: 321-329.
26. Stougaard, J., 2001. Genetics and genomics of root symbiosis, Plant Biol., pp: 328-335.
27. Ziv, M. and D. Vienne, 2000. Proteomics: A link between genomics, genetics and physiology, Plant Mol. Biol., pp: 575-580.
28. http://www.cigenomics.bc.ca.
29. http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_1_1.shtml last visited on January 7, 2006.
30. http://marketing.appliedbiosystems.com
31. http://kidshealth.org/kid/talk/qa/what_is_gene.html
32. http://www.doegenomes.org/
33. Gotoh, O., 1997. Multiple sequence alignment: Algorithms and applications, Advanced Biophysics, pp: 159-206.
34. Howe, K.L., T. Chothia and R. Durbin, 2002. A generic framework for the integration of gene-prediction data by dynamic programming, Genome Res., pp: 1418-1427.
35. http://bip.weizmann.ac.il/index.html.
36. http://www.genomenewsnetwork.org/
. http://gdbwww.gdb.org/