Gene prediction by multiple syntenic alignment

Said S. Adi* **and Carlos E. Ferreira**

Institute of Mathematics and Statistics (IME)
University of São Paulo (USP)
Rua do Matão 1010 – Cidade Universitária
05508-900 – São Paulo (SP) – Brazil

Summary

Given the increasing number of available genomic sequences, one now faces the task of identifying their functional parts, like the protein coding regions. The gene prediction problem can be addressed in several ways. One of the most promising methods makes use of similarity information between the genomic DNA and previously annotated sequences (proteins, cDNAs and ESTs). Recently, given the huge amount of newly sequenced genomes, new similarity-based methods are being successfully applied in the task of gene prediction. The so-called **comparative-based** methods lie in the similarities shared by regions of two evolutionary related genomic sequences. Despite the number of different gene prediction approaches in the literature, this problem remains challenging. In this paper we present a new comparative-based approach to the gene prediction problem. It is based on a syntenic alignment of three or more genomic sequences. With syntenic alignment we mean an alignment that is constructed taking into account the fact that the involved sequences include conserved regions intervened by unconserved ones. We have implemented the proposed algorithm in a computer program and confirm the validity of the approach on a benchmark including triples of human, mouse and rat genomic sequences.

1 Introduction

The gene prediction problem can be defined as the task of finding the genes encoded in a genomic sequence of interest. In other words, given a DNA sequence, we would like to correctly pinpoint the positions of the exons that compose one or all of its genes. Like the search for promoters, CpG islands and others functional genomic regions, the search for genes has an undeniable practical importance.

The genes of most eukaryotic organisms are neither continuous nor contiguous. They are separated by long stretches of intergenic DNA and their coding fragments, named *exons*, are interrupted by non-coding ones, the *introns*. Besides the exons and introns, the eukaryotic genes include a number of others elements, like $5'$-UTR, $3'$-UTR and splicing sites. A typical multi-exon eukaryotic gene has the structure shown in Figure 1.

Gene prediction methods can be roughly classified into two main categories, called *ab initio* or intrinsic methods and *similarity-based* or extrinsic methods (see [17] for a extensive review on this topic). The first ones ([9],[25], [28]) rely on statistical information that alone, or in conjunction with some signals previously identified in the target sequence, allow the identification of its coding, non-coding and intergenic regions. The most recent intrinsic methods make use

*To whom correspondence should be addressed.

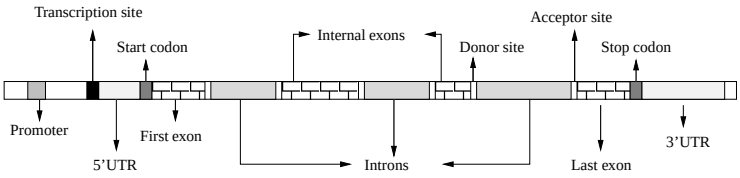

Figure 1: Simplified structure of a multi-exon eukaryotic gene

of Hidden Markov Models (HMMs) ([8], [14], [15], [16]) in order to combine both the signal and statistical information concerning the genes being searched for. The similarity-based methods ([5], [10], [12], [19]) rely on the homology between the genomic sequence and a fully annotated transcript sequence, like cDNAs, ESTs or proteins, in order to accomplish their task. Recently, with the huge amount of newly sequenced genomes, new similarity-based methods are being successfully applied in the task of gene prediction. In some way different from traditional extrinsic methods, the so-called *comparative-based* methods ([6], [21], [22], [23], [29]), pioneered by Batzoglou *et al.* with ROSETTA[4], rely upon the similarities between regions of two unannotated and evolutionary related genomic sequences in order to find the genes encoded in each of them. The main assumption of these methods is that the functional parts of an eukaryotic genomic sequence, the exons, tend to be more conserved than the non-functional ones, the introns. Despite the enormous progress made to date (see Brent and Guigó [7] for the latest survey on this topic), the gene identification problem remains open and continues to be an interesting subject of research.

In this work we present a new comparative-based gene prediction method. It lies on the comparison of a number of genomic sequences in order to identify their genes. This comparison is made by means of a dynamic programming algorithm that aligns the input sequences taking into account the existence of intermittent similarities between them. The proposed approach has been implemented and validated on a data set including three triples of human, mouse and rat genomic sequences. To the best of our knowledge, this is one of the first gene prediction program that takes into account more than two sequences in order to find their protein coding regions.

2 Related Works

The first gene prediction program based on the comparison of two genomic sequences was proposed in 2000 by Batzoglou *et al.*. This program, called ROSETTA[4], takes as input two genomic sequences and gives as output the positions of their exons. This is done in two steps. First of all, the input sequences are aligned by means of a special alignment program called GLASS. The possible exons encoded on both sequences are then found by locating well-conserved regions within the resulting pairwise alignment. In the second step, the possible exons are assembled into a gene by searching for an optimal valid parse of these regions.

Besides ROSETTA, there are a number of another gene prediction tools (SGP-2[22], SGP-1[29], AGENDA[23], etc) that first compare the input sequences and, in a subsequent step, find the best chain of exons representing the genes being searched for. The main difference between these programs relies on the way the comparison of the two input sequences is done. Both SGP-1 and SGP-2, for example, make use of BLAST programs in order to compare the input

sequences. AGENDA, in its turn, finds the possible exons in each sequence by comparing them via DIALIGN[18], a program that assembly a pair-wise alignment from gap-free local segment alignments.

Different from the abovementioned programs, UTOPIA[6] and PROGEN[21] make their predictions in a single step. The main idea of these programs is to align the input sequences taking into account the possibility of starting an exon on one or both sequences. In a similar way, EXON_FINDER2[1] accomplishes its task by locating the exons during the alignment process. The main feature of this program is that similarities on both exons and introns are considered.

It is important to note that all the comparative-based programs developed so far take as input two genomic sequences in order to identify their genes. In this work we propose a new gene prediction program that compare more than two genomic sequences in searching for the genes encoded in each of them. To the best of our knowledge, this is one of the first gene prediction program that takes into account more than two sequences in order to accomplish its task. Recently, Siepel and Hausler [26] proposed a phylo-HMM based method that compares a number of sequences and predicts each of their conserved exons. The main restriction of this method is that a multiple alignment of the input sequences need to be available to ensure its correct functioning as well as the phylogeny of the species being compared.

3 Methods

Traditional alignment algorithms cannot be used directly in the task of comparing two genomic sequences sharing a number of similar regions intervened by regions with a low level of similarity. The Smith-Waterman algorithm[27], for example, is only suitable for the identification of a high-scoring similar region in two related sequences. On the other hand, the Needleman-Wunsch global alignment algorithm[20] will tend to align even unrelated regions of the sequences being compared. With these restrictions in mind, a new model to compare two sequences sharing intermittent similarities becomes necessary.

A way to align two sequences s and t sharing discontinuous regions of high similarity was suggested by Almeida *et al.* in [3]. In this work, the authors present a dynamic programming algorithm whose main idea is to heavily penalize mismatches and gaps inside similar regions of the two sequences and to penalize in a slightly way their occurrences inside regions with a low level of similarity. To this end, the score of a best alignment between prefixes of s and t is stored in two different sets of matrices: one for the conserved regions (S_e, I_e, D_e) and another for the regions where differences are more probably to occur (S_i, I_i, D_i). This type of pairwise alignment, where regions with distinct levels of similarity are taken into account, can be referred as *syntenic alignment*.

The recurrences below can be used to compute the matrices S, I and D mentioned above. In what follows, k is a non-negative scalar used to penalize the beginning of a non-conserved region. With respect to h and g, they correspond to the penalty associated with opening and extending a gap, respectively. Finally, w is the score of a match/substitution involving two residues a and b of the sequences. The only restriction imposed on these parameters is that, given the low expected level of similarity of the non-conserved regions, the value of h_i and g_i need to be smaller than that associated with h_e and g_e. Due to this same fact, the values associated with $w_i(a, b)$ need to be smaller than that of $w_e(a, b)$ when $a \neq b$.

$$S_e[i][j] = w_e + \max \left\{ \begin{array}{l} S_e[i-1][j-1], D_e[i-1][j-1], I_e[i-1][j-1], \\ S_i[i-1][j-1], D_i[i-1][j-1], I_i[i-1][j-1]. \end{array} \right.$$

$$I_e[i][j] = \max \left\{ \begin{array}{l} S_e[i][j-1] - (h_e + g_e), D_e[i][j-1] - (h_e + g_e), I_e[i][j-1] - g_e, \\ S_i[i][j-1] - (h_e + g_e), D_i[i][j-1] - (h_e + g_e), I_i[i][j-1] - g_e. \end{array} \right.$$

$$D_e[i][j] = \max \left\{ \begin{array}{l} S_e[i-1][j] - (h_e + g_e), D_e[i-1][j] - g_e, I_e[i-1][j] - (h_e + g_e), \\ S_i[i-1][j] - (h_e + g_e), D_i[i-1][j] - g_e, I_i[i-1][j] - (h_e + g_e). \end{array} \right.$$

$$S_i[i][j] = w_i + \max \left\{ \begin{array}{l} S_i[i-1][j-1], D_i[i-1][j-1], I_i[i-1][j-1], \\ S_e[i-1][j-1] - k, D_e[i-1][j-1] - k, I_e[i-1][j-1] - k. \end{array} \right.$$

$$I_i[i][j] = \max \left\{ \begin{array}{l} S_i[i][j-1] - (h_i + g_i), D_i[i][j-1] - (h_i + g_i), I_i[i][j-1] - g_i, \\ S_e[i][j-1] - (k + h_i + g_i), D_e[i][j-1] - (k + h_i + g_i), \\ I_e[i][j-1] - (k + g_i). \end{array} \right.$$

$$D_i[i][j] = \max \left\{ \begin{array}{l} S_i[i-1][j] - (h_i + g_i), D_i[i-1][j] - g_i, I_i[i-1][j] - (h_i + g_i), \\ S_e[i-1][j] - (k + h_i + g_i), D_e[i-1][j] - (k + g_i), \\ I_e[i-1][j] - (k + h_i + g_i). \end{array} \right.$$

Like any pairwise alignment algorithm, the one based on the above recurrences can be used as the core of any progressive multiple alignment algorithm, like the center star approximation developed by Gusfield[13]. In short, given k input sequences, the center star algorithm first makes a pairwise alignment of all these sequences and takes as center that one most similar with the another $k-1$ sequences. Chosen the center sequence c, all the pairwise alignments between c and the another input sequences are grouped into a multiple alignment following the *once a gap always a gap* approach. In this method, a column consisting only of spaces is inserted into the multiple alignment being constructed whenever the pairwise alignment chosen to be grouped includes a new space in c.

Considering the possibility of constructing a multiple alignment of k sequences by means of $k-1$ pairwise syntenic alignments, we have developed a new gene prediction method lying on genomic sequences comparison. Different from traditional comparative-based methods, the one proposed here makes use of more than two sequences in order to identify their coding regions. Multiple comparison can enable us to overcome the main drawback presented by the existing comparative-based methods. Since their predictions rely in comparing only two sequences, these methods suffer from the high number of false positives when two close genomes are used as input.

The first step of our approach is to construct a multiple syntenic alignment by using the center star approximation algorithm in conjunction with the pairwise syntenic alignment algorithm.

Given the multiple syntenic alignment, we consider the conserved regions in each sequence as the core of possible exons. These regions are thus extended and trimmed until a splicing site be found. Here, the *log-likelihood ratio* of each potential splicing site (`AG/GT` and `ATG/TAG,TAA, TGA`) is used to discriminate between false and true signals. Given an input genomic sequence, the log-likelihood ratio of each possible splicing site is calculated by using a set of conditional probability matrices described by Salzberg in [24]. Given a probability matrix M, the probability $P(s)$ of a random sequence s of length l be a true splicing site can be calculated as $P(s) = \sum_{i=1}^{l} M[s[i]][i]$, where $s[i]$ is the symbol at position i of s. The results of the previous step are a cluster of exons to each conserved region of the input sequences. To each of these exons, we associate a score calculated by the sum of the scores of each pairwise sub-alignment involving its sequence. The next step is to assembly these exons in order to construct the genes encoded in each sequence.

The exons found in the last step are assembled by searching for a maximum path in a weighted directed acyclic graph representing all the possible chains of exons. In this graph, each vertex is associated with a possible exon and there is an arc to each pair of compatible exons. An exon e_i is said compatible with an exon e_{i+1} if e_i finishes before e_{i+1}. Beside this, if e_i is a start or internal exons, e_{i+1} needs to be an internal or end exon. The weight of each arc equals the score of the exon represented by its rightmost vertex.

An example of such a graph for a set of exons $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$, where $e_1 = \{15, 76, f, 256\}$, $e_2 = \{25, 87, f, 45\}$, $e_3 = \{13, 87, f, 156\}$, $e_4 = \{231, 245, i, 85\}$, $e_5 = \{356, 412, i, 90\}$, $e_6 = \{238, 300, i, 56\}$ and $e_7 = \{459, 578, l, 152\}$ is shown in Figure 2. Here, the first and second fields of each quadruple are, respectively, the left and right coordinates of the exon. The type of exon (f = first, i = internal and l = last) is represented by the third field. The last field is the exon related score.

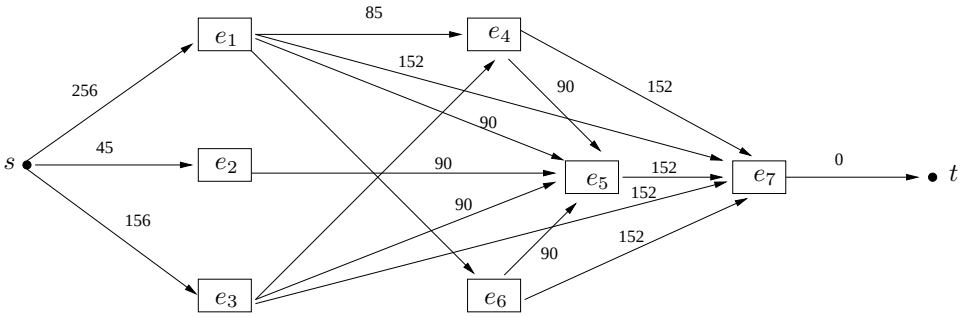

Figure 2: An example of a graph representing all the possible chains of the exons in E.

4 Implementation

The mentioned ideas were implemented in a new comparative-based gene prediction program called EXON_FINDER3. It takes as input a number k of sequences $s_1, \ldots, s_k$ in FASTA format and returns the exons encoded in each sequence with the associated multiple syntenic alignment.

EXON_FINDER3 has been implemented in ANSI C and can be downloaded at the following URL: http://www.ime.usp.br/∼said/programs/. Its main data structures are the set of dynamic

programming matrices and the directed acyclic graph used in the assembly step. The dynamic programming matrices are filled in row by row following the recurrences presented in Section 3. By making a traceback in the computed matrices, we construct the pairwise alignments that reveal the conserved regions shared by a pair of sequences. Given the syntenic alignment for all the possible pairs of input sequences, they are grouped into a multiple syntenic alignment A following the ideas presented by Francis *et al.* in [11]. Taking s_1 as the center sequence, let $A_{1,j}$ be the optimal pairwise syntenic alignment of s_1 and s_j. Define c_o and c_m to be, respectively, the longest sequence of spaces inserted before the first and last symbol of s_1 in all $k-1$ pairwise alignments involving this sequence. Similarly, let c_i be the longest sequence of spaces inserted between the i-th and the $i+1$-th symbols of s_1. Given these values, set A to contain initially a single row $S = c_0 \bullet s_1[1] \bullet c_1 \bullet s_1[2] \bullet \ldots s_1[n] \bullet c_k$. For each s_j ($2 \leq j \leq k$), add s_j to A by inserting columns of spaces in $A_{1,j}$ until the sequence s_1 of this alignment becomes identical to S.

The multiple syntenic alignment A are searched for conserved regions in a second step. These regions are used as input to the procedure that constructs, for each input sequence s_i, a directed acyclic graph G_i. All these graphs are represented by a list of adjacency. The maximum path for each G_i is thus found by means of an algorithm based on a topological ordering of its vertex.

About the complexity of our approach, it takes $O(k^2m^2)$ (assuming that all the input sequences have the same length m) to construct the multiple alignment plus $O(\sum_{i=1}^{k} b_i^2)$ (where b_i are the number of possible exons in s_i) to find the maximum path in all G_i.

5 Results

In order to evaluate our approach, we have tested our program on a benchmark including three triples of single gene sequences from human, mouse and rat. These sequences were compiled from the HOMOLOGENE database. This system allows to find homologous genes among several sequenced eukaryotic genomes stored in GENBANK[2]. All the genes encoded in each sequence were evaluated experimentally and the sequences themselves have been used as a standard set to the evaluation of comparative-based gene prediction programs. Detailed information about these sequences are shown on Table 1.

The prediction made by our program are described on Table 2. It can be observed that half of the predicted exons were correctly identified in the first triple. Beside this, from a total of 34 predicted exons, only five do not overlap with an annotated exon. About the second triple, some false positives were predicted by our program. The same occurs in the last triple. Beside some false positives, the first exons of each sequence were missed by our program. This can be due to the small length of these exons. Despite these drawbacks, it is important to note that the results obtained are a bit better than that presented by another gene prediction program developed by us (EXON_FINDER2[1]) based on the comparison of two sequences only. The following coordinates were predicted by EXON_FINDER2 on a pair including the sequences HSU66875 and MMU34801: (66..123, 276..878, 976..1488) and (89..137, 208..805, 988..1500). The main differences between the predictions achieved by both programs become evident in Figure 3. This example is in accordance with the intuition that better results can be achieved when more than two sequences are compared in the search for their genes.

Triples:	Sequences	Length	Exons coordinates
1.	HSCKBG	4200	join(1148..1340, 1464..1618, 1691..1823, 2208..2379, 3050..3173, 3331..3520, 3600..3778)
	MUSCRKNB	4521	join(1711..1903, 2090..2244, 2322..2454, 2624..2795, 3381..3504, 3652..3841, 3925..4103)
	RATCKBR	4360	join(1276..1468, 1656..1810, 1888..2020, 2197..2368, 2941..3064, 3174..3363, 3443..3621)
2.	HSU66875	1569	join(802..874, 976..1112, 1281..1364)
	MMU34801	1910	join(729..801, 988..1124, 1202..1285)
	RNCOX6B	2569	join(1487..1762, 1951..2087, 2172..2366)
3.	HUMTHY1A	2806	join(27..63, 547..882, 1410..1522)
	MUSTHY1GC	3257	join(555..591, 1182..1520, 1907..2019)
	RNTHY1G	2863	join(31..67,735..1070,1473..1585)

Table 1: Triples used in testing our program.

Triples:	Predicted coordinates
1.	(1148..1340, 1464..1618, 1691..1754, 1783..1823, 2089..2379, 2957..3104, 3120..3219, 3331..3520, 3600..4081), (1581..1988, 2090..2109, 2120..2311, 2322..2454, 2624..2947, 3381..3504, 3527..3584, 3594..3861, 3925..4412) and (790..1030, 1203..1557, 1588..1675, 1677..1833, 1888..2030, 2085..2520, 2797..2917, 2941..3074, 3083..3123, 3140..3363, 3412..3992)
2.	(641..874, 931..1186, 1281..1312, 1325..1399), (625..801, 988..1031, 1036..1171, 1202..1286, 1294..1388, 1427..1500) and (1388..1762, 1871..1976, 1994..2107, 2172..2268, 2283..2535);
3.	(628..882, 1290..1424, 1451..1479, 1499..1624), (1263..1520, 1839..1941, 1948..2022, 2035..2171) and (630..1070, 1341..1453, 1473..1588, 1601..1656).

Table 2: Exons predicted by our program

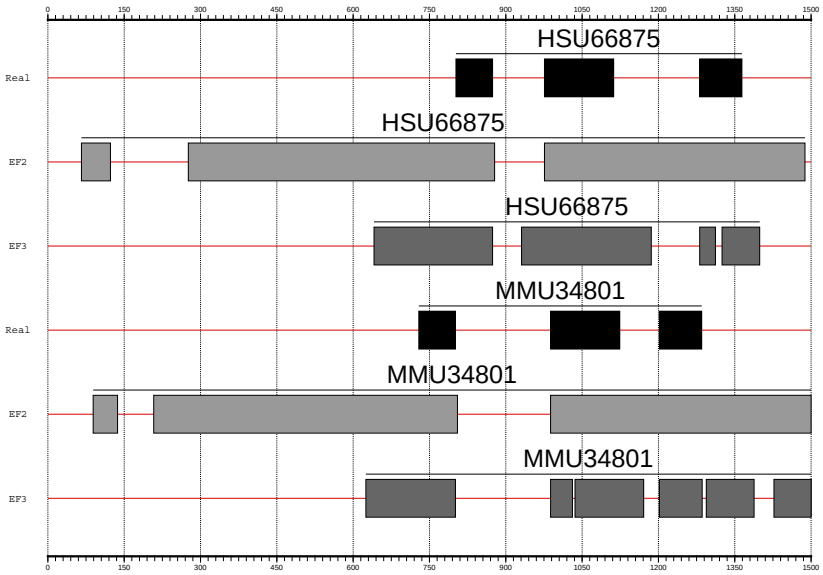

Figure 3: A graphical representation of the predictions made by EXON_FINDER2(EF2) and EXON_FINDER3(EF3) taking as input the sequences HSU66875 and MMU34801.

6 Discussion

Despite its practical importance and the number of methods developed to date, the gene identification remains an open and interesting problem. Given the increasing number of homologous sequences in the databases and the assumption that the exons tend to be more conserved than the introns inside a genome, comparative-based gene prediction programs starts to be extensively used in the task of gene identification. Since their predictions lies in comparing only two sequences, these methods suffer from the number of false positives when two close genomes are used as input.

In this work we presented a new program where more than two evolutionary related sequences can be compared in order to identify their genes. It is based on a multiple syntenic alignment of the input sequences. In other words, on an alignment that takes into account the fact that these sequences include a number of conserved regions, the exons, separated by unrelated ones, the introns and intergenic regions. To the construction of this alignment, the main idea is to heavily penalize mismatches and gaps inside the coding regions and to penalize in a slightly way its occurrences inside the non-coding regions of the sequences. This modified version of the Smith-Waterman algorithm are used as the core of the center star approximation algorithm. The resulting multiple alignment is post-processed and the possible exons are assembled by searching for a maximum path in a directed acyclic graph.

This approach was implemented and then tested on a benchmark including three triples of single gene sequences. The results obtained are very promising, despite some errors observed such as prediction of false positives and missing small exons. To the best of our knowledge, this is one of the first gene prediction program based on the comparison of several sequences in order to achieve its goals. Besides the possibility of predicting multiple genes in a sequence, the formulation of the exon assembly as a problem of searching for a maximum weight path in a DAG allows us to make use of well-studied algorithms to solve it and related variants.

7 Acknowledgement

We thank to Marie-France Sagot and Alair Pereira do Lago for their interesting and valuable comments during the development of the main ideas of this work. This work has been supported by FAPESP (00/06328), Pronex (107/97) and CNPq (300752/94-6).

References

[1] Adi, S.S. and Ferreira, C.E.: Gene prediction by syntenic alignment. *Lecture Notes in Computer Science*, 3594:246-250, 2005.

[2] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L.: GenBank. *Nucleic Acids Research*, 28(1):15-18, 2000;

[3] Almeida, N.F., Setubal, J.C. and Tompa, M.: On the use of don't care regions for protein sequence alignment. *IC Technical Reports 99-07* (1999).

[4] Batzoglou, S., Pachter, L., Mesirov, J., Berger, B. and Lander, E.S.: Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, 10:950-958, 2000.

[5] Birney, E. and Durbin, R.: Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 5:56-64, 1997.

[6] Blayo, P., Rouzé, P. and Sagot, M.-F.: Orphan gene finding - An exon assembly approach. *Theoretical Computer Science*, 290:1407-1431, 2003.

[7] Brent, M.R. and Guigó R.: Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.*, 14(3):264-272, 2004.

[8] Burge, C. and Karlin, S.: Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78-94, 1997.

[9] Fickett, J.W.: Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10:5303-5318, 1982.

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

[11] Chin, Francis Y.L., Ho, N.L., Lam, T.W., Wong, Prudence W.H. and Chan, M.Y.: Efficient Constrained Multiple Sequence Alignment with Performance Guarantee. *IEEE Computer Society Bioinformatics Conference (CSB'03)*, p. 337, 2003.

[12] Gelfand, M.S., Mironov, A.A. and Pevzner, P.A.: Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences*, 93(17):9061-9066, 1996.

[13] Gusfield, D.: Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bulletin of Mathematical Biology*, 55(1):141-154, 1993.

[14] Krogh, A., Mian, I.S. and Haussler, D.: A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research*, 22:4768-4778, 1994.

[15] Krogh, A.: Two methods for improving performance of a HMM and their application for gene finding. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (1997) AAAI Press, Menlo Park, CA, pp. 179-186.

[16] Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H.: A generalized Hidden Markov Model for the recognition of human genes in DNA. *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology* (1996) AAAI Press, Menlo Park, CA, pp. 134-142.

[17] Mathé, C., Sagot M.-F., Schiex, T. and Rouzé, P.: Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19):4103-4117, 2002.

[18] Morgenstern, B., Rinner, O., Abdeddaim, S., Haase, D., Mayer, K.F., Dress, A.W. and Mewes, H.-W.: Exon discovery by genomic sequence alignment. *Bioinformatics*, 18(6):777-787, 2002.

[19] Mott, R.: EST_Genome: a program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, 13:477-478, 1997.

[20] Needleman, S.B. and Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443-453, 1970.

[21] Novichkov, P.S., Gelfand, M.S. and Mironov, A.A.: Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, 17(11):1011-1018, 2001.

[22] Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W. and Guigó, R.: Comparative gene prediction in human and mouse. *Genome Research*, 13(1):108-117, 2003.

[23] Rinner, O. and Morgenstern, B.: AGenDA: Gene prediction by comparative sequence analysis. *In Silico Biology*, 2:195-205, 2002

[24] Salzberg, S.L.: A Method for Identifying Splice Sites and Translational Start Sites in Eukaryotic mRNA. *Computer Applications in the Biosciences*, 13(4):365-376, 1997.

[25] Shepherd, J.C.: Method to determine the reading frame of a protein from the purine / pyrimidine genome sequence and its possible evolutionary justification. *Proceedings of the National Academy of Sciences*, 78:1596-1600, 1981.

[26] Siepel, A. and Haussler, D.: Computational identification of evolutionarily conserved exons. *Proceedings of the Eighth Annual International Conference on Research in Computational Biology*, 177-186, 2004.

[27] Smith, T.F. and Waterman, M.S.: Identification of common molecular sub-sequences. *Journal of Molecular Biology*, 147(1):195-197, 1981.

[28] Staden, R. and McLachlan, A.D.: Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research*,10:141-156, 1982.

[29] Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. and Guigó, R.: SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.*,11(9):1574-1583, 2001.