RECOGNITION OF CODING REGIONS IN GENOME ALIGNMENT

Astakhova T.V.[*1], *Petrova S.V.*[1], *Tsitovich I.I.*[2], *Roytberg M.A.*[*1]

[1] Institute of Mathematical Problems in Biology RAS, Puschino, Russia; [2] Institute of Information Transmission Problems RAS, Moscow, Russia
* Corresponding authors: e-mail:, Roytberg@impb.psn.ru

Keywords: *coding region, gene recognition, genome alignment, synonymous and non-synonymous substitution*

Summary

Motivation: Gene recognition is an old and important problem. Statistical and homology based methods work quite well, if one attempts to find long exons or full genes, but is unable to recognize relatively short coding fragments. Genome alignments and study of synonymous and non-synonymous substitutions provide opportunity to overcome this drawback. Our aim is to propose a criterion to distinguish short coding and non-coding fragments of genome alignment and to create an algorithm to locate alignment coding regions.

Results: We developed a method to locate aligned exons in a given alignment. First, we scan the alignment with a window of a fixed size (~40 bp) and assign the score H(P) to each window position P a score H(P). The value H(P) reflects, if numbers K_S of synonymous substitutions, K_N of non-synonymous substitutions and D of deleted symbols appear to be similar for coding regions. Second, we mark "exon-like" regions, ELRs, i.e. sequences of consecutive high-scoring windows. Presumably, each ELR contains one exon. Third, we highlight an exon within every ELR. All the steps have to be performed twice, for the direct and reverse complementary chains independently. Finally, we compare predictions for two chains to exclude possible predictions of "exon shadows" on complementary chain instead of real exons. Tests have shown that 93 % of marked ELRs have intersections with real exons and 93 % of aligned annotated exons intersect marked ELRs.The total length of marked ELRs is ~ 1.35 of the total length of annotated exons. About 75 % of the total length of predicted exons belong to annotated exons and more, than 70 % of the total length of correctly aligned annotated exons belong to predicted exons. The run-time of the algorithm is proportional to the length of a genome alignment.

Introduction

The existence of powerful genome alignment methods (Brudno *et al.*, 2003; Roytberg *et al.*, 2002) and availability of many complete genomes, including several eukaryotic, lead to revisions of classical problems of sequence analysis. Indeed, we can analyze pair-wise (or, if possible, multiple) sequence alignment instead of one genome sequence. In the case of the gene recognition problem, genome alignments allow to take advantage of two ideas. First, coding regions are, in general, more conserved than non-coding. Thus, one can attempt to recognize genes as a sequence of well-aligned genome fragments (Bafna, Huson, 2000; Batzoglou *et al.*, 2000; Novichkov *et al.*, 2001; Taher *et al.*, 2003). Such methods are effective for relatively distant species, but some genes can be unrecognizable because of low similarity between species. On the other hand, alignment of close genomes often gives many false positive exons, because of the existence of conserved non-coding regions (Shabalina, Kondrashov, 1999). Second, one can additionally pay attention to the difference between substitution patterns in the coding and non-coding regions, the former tend to be synonymous, i.e. to preserve a coded residue. The metods using alignment-based HMMs or pair HMMs (Meyer, Durbin, 2002; Pedersen, Hein, 2003), take into account the differences between various parts of a genome alignment implicitly, in course of HMM training. Despite the promising results, shown by the methods, we think that it is worth to learn explicitly, what benefit one can get from the differences in substitution patterns. The explicit usage of the differences is implemented in (Nekrutenko *et al.*, 2001), abilities of the approach were demonstrated in (Nekrutenko *et al.*, 2002). However, the goal of the paper (Nekrutenko *et al.*, 2001) is mainly to study the ability of the proposed criterion to recognize relatively long exons as a whole, authors make no attempt to recognize exon borders or short coding regions.

We propose a two-stage procedure combining prediction techniques of traditional identification of exons in DNA sequence and methods based on information about genome alignment. First, using investigation of substitution patterns, we perform an alignment filtration, i.e. locate "exon-like regions" (ELR) in the alignment. Then, the putative exon within ELR can be found using the classical statistical approach. Below we will demonstrate the advantages and drawbacks of the approach and will discuss possible ways for improving it.

Materials and Methods

General description of the approach. The algorithm works in four steps. The first three have to be performed independently for direct and reverse complementary complement chains. At the last step we compare the results obtained for two chains and prepare the final prediction. We start (the first step) with scanning the alignment with a window of a fixed size w and a given shift s. For each considered window, we make the decision if it is exon-like or not. Then (the second step), we reveal "exon-like" regions, ELRs. An ELR is a set of consecutive window positions (see details below). Any two ELRs marked on a chain do not intersect each other. Presumably, each ELR contains one exon. At this step, we work only with exon/nonexon marks of window positions, the marks were assigned on previous step. On the third step we reveal a putative exon for each ELR and assign a score to the exon. If ELR does not contain a pair assigned a aligned exons of high enough score, the ELR is to be rejected. Finally, we compare ELRs found on the direct and reverse chains. If two ELRs of different chains intersect, we keep only one, the ELR having an exon assigned a higher score.

Analysis of window position. Let w be the size of the window. For a window at the position P, i.e. for the fragment of alignment from position P to position P+w-1, the program calculates its score H(P). The score characterizes the presence of stabilizing selection at the protein level. The basic characteristics of the window at a given position of alignment are: 1) FMatch – fraction of match alignment positions, i.e. superposition of identical nucleotides; 2) Probability $\mathrm{Pr}(K_{T,} K_S, D)$ to obtain K_S or more synonymous substitutions, if K_T random independent substitutions were performed and D codons are deleted. We calculate $\mathrm{Pr}(K_{T,} K_S, D)$ for three possible frames. The score H(P) is a negative binary logarithm of the minimum of the three probabilities. The window position P is "exonic", if both FMatch(P) and the score H(P) exceed a threshold.

Exon-like regions. A region is a set of consecutive windows, i.e. windows at positions P, P+s, P=2s, …, where s is a given shift. A region starts at the beginning of the 1st window and ends at the end of the last window. An exon-like region (ELR) is a region meeting the following conditions: 1) The 1st window of the region contains a putative acceptor site or START codon; the last window of the region contains a putative donor site or STOP codon (see below); 2) A region is "exone-dense", i.e. the difference between the number of non-exonic and exonic windows within a consecutive part of a region cannot exceed a threshold InnerCut; 3) The number of non-exonic windows at the beginning and at the end of a region cannot exceed a threshold EdgeCut; 4) a region is not a part of another fragment meeting conditions 1– 3.

A putative acceptor (donor) site is aligned, i.e. present in both sequences, dinucleotide "AC" ("GT"), having Berg-von Hippel score (Berg, von Hippel, 1987) exceeding given cut-off. Putative START- and STOP-codons also have to be presented in both sequences and to be aligned.

Putative exons. Putative exon is a part of an exon-like region, starting with an acceptor or START and ending with a donor or STOP. We assign a statistical score S(E) and alignment score A(E) to every putative exon E. The score S(E) is the sum of scores of exons, given on every genomic sequence, calculated by the method described in (Gelfand *et al.*, 1996). The value S(E) depends on scores of splicing sites, codon potential and exon length. Alignment score A(E) reflects the difference between the ratios K_S/K_N for the exon calculated for the considered chain and the reverse chain. The score G(R) of an ELR is the sum of the maximum values of S(E) and A(E) for putative exons belonging to the region. If the value G(R) is below a cut-off (currently, 2.25), the region R will be rejected. Otherwise, the exon corresponding to the maximum score S(E) is considered as a predicted exon for the region R.

Genome alignment and gene annotation. The program implementation of the method was tested on the alignment of syntenic regions of the 6th *Homo sapiens* chromosome and the 17th *Mus musculus* chromosome of ~700000 nucleotides length. The alignment was obtained by the OWEN program (Ogurtsov *et al.*, 2002). There are 56 genes annotated on the human sequence and 58 genes annotated on the mouse sequence. Alternative splicing variants are given for 17 human genes, and for only one mouse gene. Mouse genes contain 567 annotated exons, 479 of these are aligned correctly with the corresponding human exons. Incorrect alignment of the other genes can be mostly explained by the inconsistency of exon the annotation in the human and mouse genome. The total length of all the annotated mouse exons is 100,869, the average exon length is 178.

Testing parameters. We have used the following parameters values (see above): 1)Window size w = 40; window offset s = 10 bp; 2) FMatch cut-off for "exonic" window FMatchMin = 0.65; H(P) cut-off for "exonic" window H_Min = 2.25; 3) ELR cut-offs InnerCut = 6; EdgeCut = 6; 4) minimum score of an acceptor splicing sites ACC_Score = -17; minimum score of a donor splicing site DON_Score $\geq$ -7; 5) the minimum length of an inner exon is 40 bp; the minimum length of start or exon is 15 bp.

Results and Discussion

Testing results. The algorithm produces two types of objects (see Materials and Methods): exon-like regions (ELR), and putative exons. Below we give all the results related to the mouse chromosome. The results for the human chromosome are very similar.

We have revealed 628 ELRs, the total length of a revealed ELR is 13,694 (136% of the total length of annotated exons), the average ELR length is 218 (122 % of the average exon length). 586 of these (93 %) have intersection with an annotated exon. Only 35 of the 479 correctly aligned exons (7 %) do not intersect ELRs.

Putative exons, highlighted within ELRs, show the following results. 408 correctly alignment annotated exons (or 85 %) intersect corresponding putative exons. Usually, the intersection between putative and annotated exon cover almost all the annotated exon (87 % on average), 215 correctly aligned exons (50 %) are recognized correctly. At least one exon border is recognized correctly for 406 exons (84 %). The common part of all the correctly aligned exons and corresponding putative exons constitute 74 % of the total length of correctly aligned annotated exons. Approximately the same part of the total length of putative exons belongs to annotated exons.

Discussion. The algorithm addresses two problems. First, it approximately locats the area cohere it is reasonable to search for exons (generation of exon-like regions). Second, highlights out putative exons within the ELRs. The problems are relatively independent, i.e. we can use an arbitrary gene recognition algorithm to solve the second problem, when the first is already solved.

Our main efforts were targeted at the first problem, and the algorithm efficiently solve it. Taking into account its linear run-time, the algorithm can serve as a useful filtration tool for any exon-recognition algorithm, working with genome alignments. We envigage ways of improving the method. In particular, we plan to gather statistics of the possible values of pairs (K_N, K_S) in the coding and non-coding regions, and define the scoring function H(P) using the maximum likelihood principle.

Putative exons show a weaker correlation with annotated exons than the ELRs. We plan to improve significantly this part of the algorithm. For example, we plan to generate for a given ELR several putative exons having different frames and the link them to predict the whole gene. Another possible development of the project is to realign genomes in the vicinity of putative exon borders. General genome alignment algorithms often misalign conservated positions of splicing sites.

Aknowledgements

Authors are thankful to A.Kondrashov, A.Ogurtsov, S.Shabalina and S.Sunyaev for helpful discussions. The work has been supported by the Russian Found for Basic Research (project nos. 03-04-49469, 02-07-90412) and by grants from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003) and NWO.

References

Bafna V., Huson D.H. The conserved exon method for gene finding // Proc. Int. Conf. Intell. Syst. Mol. Biol. 2000. V. 8. P. 3–12.

Batzoglou S., Pachter L., Mesirov J.P., Berger B., Lander E.S. Human and mouse gene structure: comparative analysis and application to exon prediction // Genome Res. 2000. V. 10(7). P. 950–8.

Berg O.G., von Hippel P.H. Selection of DNA binding sites by regulatory proteins.Statistical-mechanical theory and application to operators and promoters // J. Mol. Biol. 1987. V. 193. P. 723–750.

Brudno M., Malde S., Poliakov A., Chuong B. Do, Couronne O., Dubchak I., Batzoglou S. Global alignment: finding rearrangements during alignment // Bioinformatics. 2003. V. 19. P. 54–62.

Gelfand M.S., Podolsky L.I., Astakhova T.V., Roytberg M.A. Recognition of genes in human DNA sequences // J. Mol. Biol. 1996. V. 3, N 2. P. 223–234.

Meyer I.M., Durbin R. Comparative ab initio prediction of gene structures using pair HMMs // Bioinformatics. 2002. V. 18(11). P. 1546–1547.

Nekrutenko A., Chung Wen-Yu., Li Wen-H. An evolutionary approach reveals a high protein-coding capacity of the human genome // Trends in Genetics. 2003. V. 19, N 6. P. 306–310.

Nekrutenko A., Makova K., Wen-Hsiung Li. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study // Genome Res. 2001. V. 12. P. 198–202.

Novichkov P.S., Gelfand M.S., Mironov A.A. Gene recognition in eukaryotic DNA by comparison of genomic sequences // Bioinformatics. 2001. V. 17(11). P. 1011–8.

Ogurtsov A.Y., Roytberg M.A., Shabalina S.A, Kondrashov A.S. OWEN: aligning long collinear regions of genomes // Bioinformatics. 2002. V. 18. P. 1703–1704.

Pedersen J.S., Hein J. Gene finding with a hidden Markov model of genome structure and evolution // Bioinformatics. 2003.V. 19(2). P. 219–27.

Roytberg M.A., Ogurtsov A.Y., Shabalina S.A., Kondrashov A.S. A hierarchical approach to aligning collinear regions of genomes // Bioinformatics. 2002. V. 18. P. 1673–1680.

Shabalina S.A., Kondrashov A.S. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes // Genet. Res. 1999. V. 74. P. 13–22.

Taher L., Rinner O., Garg S., Sczyrba A., Brudno M., Batzoglou S., Morgenstern B.A. AGenDA: homology-based gene prediction // Bioinformatics. 2003. V. 19(12). P. 1575–7.