RECOGNITION OF CODING REGIONS IN GENOME ALIGNMENT

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

[1] *Institute of Mathematical Problems in Biology, Russian Academy of Sciences, ul. Institutskaia, 4, Pushchino, Russia, 142290, e-mail: Roytberg@impb.psn.ru;* [2] *Institute of Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per., 19, Moscow, Russia, 127994*

[*] *Corresponding author*

Abstract: Gene recognition is an old and important problem. Statistical and homology-based methods work relatively well, if one tries to find long exons or full genes but are unable to recognize relatively short coding fragments. Genome alignments and study of synonymous and non-synonymous substitutions give a chance 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 aligned coding regions. We have 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 a score to each window position. The score reflects if numbers K_S of synonymous substitutions, K_N of non-synonymous substitutions, and D of deleted symbols look like those for coding regions. Second, we mark the 'qualified exon-like' regions, QELRs, i.e., sequences of consecutive high-scoring windows. Presumably, each QELR contains one exon. Third, we point out an exon within every QELR. All the steps have to be performed twice, for the direct and reverse complement chains independently. Finally, we compare the predictions for two chains to exclude any possible predictions of 'exon shadows' on complementary chain instead of real exons. Tests have shown that ~ 93 % of the marked QELRs have intersections with real exons and ~ 93 % of the aligned annotated exons intersect the marked QELRs. The total length of marked QELRs is ~ 1.30 of the total length of annotated exons. About 85 % of the total length of predicted exons belongs to annotated exons. The runtime of the algorithm is proportional to the length of a genome alignment.

Key words: coding region; gene recognition; genome alignment; synonymous and non-synonymous substitution

1. INTRODUCTION

Existence of powerful genome alignment methods (Roytberg et al., 2002; Brudno et al., 2003;) and availability of many complete genomes, including several eukaryotic ones, lead to new formulations of classic problems of sequence analysis. Indeed, we can analyze pairwise (or, if possible, multiple) sequence alignment instead of one genome sequence. In case of gene recognition, the problem of genome alignments allows one to exploit two ideas. First, coding regions are, in general, more conservative than the non-coding ones. Thus, one can try to recognize genes as sequence of well-aligned genome fragments (Bafna and Huson, 2000; Batzoglou et al., 2000; Novichkov et al., 2001; Taher et al., 2003).

Such methods are efficient for relatively distant species, but some genes can be unrecognizable because of a low interspecies similarity. From the other hand, alignment of close genomes often gives many false positive exons because of existence of conservative non-coding regions (Shabalina and Kondrashov, 1999). Second, one can additionally pay attention to the difference between substitution patterns in coding and non-coding regions; the former tend to be synonymous, i.e., preserve a coded residue.

The methods using alignment-based HMMs or pair HMMs (Meyer and Durbin, 2002; Pedersen and 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 these methods, we think that it is worth learning explicitly what benefit one can get from the differences in substitution patterns.

The explicit usage of the differences is implemented by Nekrutenko et al. (2001); the abilities of this approach were demonstrated by Nekrutenko et al. (2002). However, the goal of the paper by Nekrutenko et al. (2001) was mainly to study the ability of the proposed criterion to recognize relatively long exons as a whole; authors did not try to recognize the 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 with classic statistical approach. Below, we will demonstrate advantages and drawbacks of the approach and will discuss possible ways to improve it.

2. METHODS AND ALGORITHMS

General description of the approach. The algorithm works in four steps. Three first steps have to be performed independently for the direct and reverse complement chains. At the last step, we compare the results obtained for two chains and prepare the final prediction. We start (first step) with scanning of the alignment with a window of a fixed size w and a given shift s. For each considered window, we make a decision if it is exon-like or not. Then (second step), we reveal the '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. During this step, we work only with exon/non-exon marks of window positions, the marks were assigned at previous step. At the third step, we reveal a putative exon for each ELR and ascribe the exon with a score. If ELR does not contain a pair of aligned exons of high enough score, the ELR is to be rejected. Finally, we compare ELRs found on the direct and inverse chains. If two ELRs from different chains intersect each other, we keep only one of them, the ELR having an exon of higher score.

Table -1. The values $Score_E(K_N, K_S)$

K_S / K_N	0	1	2	3	4	5
0	2.6	3.8	5.1	7.4	9.2	11.0
1	2.1	3.4	4.7	6.7	8.2	10.2
2	1.7	3.0	4.2	6.0	7.7	9.4
3	1.4	2.6	3.8	5.3	7.0	8.6
4	1.0	2.2	3.4	4.6	6.2	7.8
5	0.6	1.8	3.0	3.9	5.5	7.0
6	0.2	1.4	2.3	3.2	4.7	6.2
7	–0.2	1.0	1.8	2.7	4.0	5.4
8	–0.6	0.6	1.2	1.9	3.3	4.6
9	–1.0	0.2	0.7	1.1	2.6	3.8
10	–1.4	–0.2	0.3	0.8	1.8	3.0
11	–1.7	–0.6	–0.2	0.4	1.1	2.2
12	–2.1	–1.0	–0.5	–0.1	0.3	1.4
13	–2.5	–1.4	–0.9	–0.5	–0.1	0.6
14	–2.9	–1.8	–1.4	–1.0	–0.5	–0.2
15	–3.3	–2.5	–2.0	–1.5	–0.8	–1.0

Here, we set $Score_E(K_N, K_S) = 20$ if $K_S > 5$, and $Score_E(K_N, K_S) = -20$ if $K_S \leq 5$ and $K_N > 15$. The values are obtained from the statistics of windows of length 45.

Analysis of window position. Let w be a 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. We have tested two approaches to define the score $H(P)$, the 'theoretical' approach and the 'empiric' one. Within the *theoretical* approach, the basic characteristics of the window at a given position of alignment are (1) *FMatch*—the fraction of match alignment positions, i.e., superposition of identical nucleotides and (2) the probability $Pr(K_T, K_S, D)$ to obtain K_S or more synonymous substitutions if K_T random independent substitutions were performed and D codons are deleted. We calculate $Pr(K_T, K_S, D)$ for three possible frames. The score $H_T(P)$ is a negative binary logarithm of the minimum of the three probabilities. We say that the window position P is 'exonic', if both $FMatch(P)$ and the score $H_T(P)$ exceed the threshold.

Within the *empiric* approach, the score H_E is computed based on the statistics of the appearance of the windows with the given number of non-synonymous substitutions K_N and synonymous substitutions K_S in coding and non-coding regions. Table 1 shows the pre-computed *Scores* assigned to the different pairs (K_N, K_S); the values in Table 1 are obtained as log–likelihood ratios of the corresponding empiric frequencies (only windows without deletions were taken into account). Table 1 confirms significant difference between the two-dimensional distribution of (K_N, K_S) of the windows without deletions in coding and non-coding regions. If the window at the position P does not contain deletions ($D = 0$), then we get the value $H_E(P)$ from the pre-computed table (Table 1); the values in the table are obtained as log–likelihood ratios of the corresponding empiric frequencies. If $D > 0$ but the value *FMatch* exceeds a proper threshold, the value $H_E(P)$ is computed from the same table, but with recalculated values K_N and K_S.

Exon-like regions (ELRs). A region is a set of consecutive windows, i.e., the windows at positions P, $P + s$, $P + 2s$, ..., where s is a given shift. A region starts in the beginning of the first window and ends at the end of the last window. An exon-like region (ELR) is a region meeting the following conditions:

(1) the first 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 'exonic-dense', i.e., a difference between the numbers 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*; and
(4) a region is not a part of another fragment meeting conditions (1)–(3).

Putative exons and qualified ELRs (QELRs). Our algorithm first finds all ELRs in both chains and then reveals among them the 'qualified' ELRs (QELRs). The definition of QELR is based on the notion of a *putative exon.*

Putative exon is a part of an exon-like region starting with a putative acceptor site or START codon and ending with a putative donor site or a STOP codon. A putative acceptor (donor) site is aligned, i.e., present in both sequences, dinucleotide 'AC' ('GT'), its neighborhood has Berg–von Hippel score (Berg and Hippel, 1987) exceeding a given cut-off. Putative START- and STOP-codons also have to be present in both sequences and to be aligned. If the exon starts with a START-codon and/or ends with a STOP-codon, then it should not contain STOP-codons in the corresponding frame.

We assign each putative exon E with a statistical score $S(E)$ and an alignment score $A(E)$. The score $S(E)$ is calculated by the method described by Gelfand et al. (1996). The value $S(E)$ depends on the scores of splicing sites, codon potential, and exon length. Alignment score $A(E)$ reflects the difference between the ratios $K_S/\max(K_N, 1)$ for the exon calculated for the considered chain and the inverse chain.

We ascribe each exon-like region R with the score $G(R)$ that is a sum of the maximal values of $S(E)$ and $A(E)$ for putative exons belonging to the region. We say that the ELR R is a qualified ELR (QELR) if R meets the following conditions:

(1) the value $G(R)$ of the region exceeds the cut-off *ELRScoreCut* and

(2) the region does not intersect an ELR on the opposite chain or the intersecting ELR on the opposite chain has a lower value ELR_Score.

The result of the algorithm's work is lists of QELR for both chains. The exon E corresponding to the maximal score $S(E)$ among all putative exons within a QELR R is considered as a predicted exon for the region R.

Genome alignments. We used two sets of genome alignments. The first set is the alignment of syntenic regions of the *Homo sapience* chromosome 6 (GenBank ACCESSION NT_007592) and the *Mus musculus* chromosome 17 (GenBank ACCESSION NT_002588) of ~700 000 nucleotides long. The human sequence contains 55 annotated genes and the mouse sequence contains 58 annotated genes.

Alternative splicing variants are given for 17 human genes and for only 1 mouse gene. Mouse genes contain 567 annotated exons, 476 of them are aligned correctly with the corresponding human exons. Incorrect alignment of other genes mostly can be explained by inconsistency of exon annotation in human and mouse genome. The total length of all the annotated mouse exons is 93 162; the average length is 165. The alignment was obtained by OWEN program (Ogurtsov et al., 2002).

The second set is the set of 117 orthologous mouse and human genes from Batzoglou et al. (2000). The genes were also aligned with the OWEN

program. The mouse genes contain 476 exons; 397 of them are aligned correctly, while the other exons have incorrectly aligned ends. The total length of mouse genes is 105 450; the average length is 222.

The alignment of syntenic regions of the *Homo sapience* chromosome 6 and the *Mus musculus* chromosome 17 was used as the training data for the algorithm; the set of Batzoglou et al. (2000) was used as testing sets.

Finally, we have analyzed the four pairs of orthologous mouse and rat mRNA with atypical ratio of the numbers of synonymous and non-synonymous substitutions; the set was proposed by G. Bazykin.

Testing parameters. We used the following values of parameters (see above): (1)window size $w = 45$, window offset $s = 15$ bp; (2) *FMatch* cut-off for 'exonic' window $FMatchMin = 0.65$, $H(P)$ cut-off for 'exonic' window: $H_T\_Min = 1.2$ (for 'theoretical' score H_T), $H_E\_Min = 3.0$ (for 'empiric' score H_E); (3) ELR cut-offs $InnerCut = 6$, $EdgeCut = 6$; (4) minimal score of an acceptor splicing sites $ACC\_Score = -17$, minimal score of a donor splicing sites $DON\_Score >= -7$; and (5) the cutoff for the ELR score $G(R)$ is 2.5.

3. RESULTS AND DISCUSSION

3.1 Results

The algorithm produces two types of objects (see Materials and Methods): qualified exon-like regions (QELR) and putative exons. The results on QELR prediction are given in Table 2; the results on exon prediction in Table 3. All results are given for the mouse chromosome; the results for the human chromosome are very similar. Results for training and testing sets are in good agreement. For the testing set, we have not reported 40 QELR predicted on the inverse chain, because we have no information about exons on this chain (Batzoglou et al., 2000, also do not consider predictions on inverse chain). If we take into account these extra QELR, the percent of QELR (line '% Inters QELR') will fall to 87 %.

The goal of the presented algorithm is to locate the aligned exons, not to give their precise borders; we consider the latter problem as a separate task and now continue to work on it; the results to be reported later. For example, we will propose the method to process correctly the QELRs containing more than one exon; this is a common situation for genes with short introns.

To check the applicability of the method to genes with nonstandard relation between K_N and K_S, we have considered four pairs of orthologous mouse and rat mRNA (Table 4).

Table -2. Qualified exon-like regions (QELR) predicted for the alignment of syntenic regions of the *Homo sapience* chromosome 6 and *Mus musculus* chromosome 17 (training set) and the set of 117 orthologous mouse and human genes from Batzoglou et al., 2000 (testing set)

	Training		Testing	
	Theor.	Empiric	Theor.	Empiric
N QELR	441	425	334	310
Tot L QELR	116591	116209	127827	144652
% Tot L Exon	125	125	121	137
Ave L QELR	264	273	339	414
% Ave L Exon	161	166	153	187
N Inters QELR	400	396	324	302
% Inters QELR	91	93	97	97
Covered Exon %	99	99	98	99
N Lost Exon	23	15	43	34
% Lost Exon	4	3	9	7

The data are given for both theoretical (columns 'Theor' and empiric versions of the scoring function $H(P)$. We use the following notation: 'N QELR', number of revealed QELRs; 'Tot L QELR', total length of revealed QELRs; '% Tot L Exon', ratio of the total length of revealed QELRs and the total length of annotated exons; 'Ave L QELR', average length of ELR; '% Ave L Exon', ratio of the average length of revealed QELRs and the average length of annotated exons; 'N Inters QELR', number of QELRs having intersection with an annotated exon; '% Inters QELR', percent of revealed QELRs having intersection with annotated exons; 'Covered Exon (%)', average part of the exon covered by intersecting QELR; 'N Lost Exon', number of 'lost exons', i.e., correctly aligned exons that do not intersect QELRs; and '% Lost Exon', percent of the lost exons among all correctly aligned exons.

Table -3. Correspondence between the predicted putative exons and annotated exons

	Training		Testing	
	Theor.	Empiric	Theor.	Empiric
% Lost Exon	27	31	30	36
Covered Exon (%)	87	90	88	88
Exactly Recogn (%)	41	39	43	39

'% Lost Exon', percent of correctly aligned annotated exons having no intersection with the predicted exons; 'Covered Exon (%)', average part of the correctly aligned annotated exon covered by intersecting predicted exon; and 'Exactly Recogn (%)', percent of the correctly aligned annotated exons that coincide with a predicted exon.

Table -4. Orthologous mouse and rat genes with nonstandard K_N and K_S ratio

Mouse GI	Rat GI	K_N	K_S	K_N/K_S
6678712	19705461	0.10	0.09	1.15
21312956	20806163	0.164	0.156	1.05
8394248	9507069	0.18	0.17	1.05
6753828	6978833	0.19	0.186	1.02

In all pairs, the program detected one QELR that contained the desired segment. In two cases (6 678 712 vs. 19 705 461 and 8 394 248 vs. 9 507 069), the coding region was predicted exactly. For the pair 6 753 828

vs. 6 978 833, the correct exon has rank 2 among all putative exons; the predicted exon has correct donor site and covers 77 % of the correct coding region. In the last case, the predicted QELR coincides with the correct coding region, but the predicted exon is significantly shorter.

3.2 Discussion

The algorithm addresses two problems. First, it approximately locates the area where it is reasonable to look for exons (generation of qualified exon-like regions, QELRs). Second, it points out the putative exons within QELRs. The problem is relatively independent, i.e., we can use arbitrary gene recognition algorithm to solve the second problem, when the first problem is already solved. We have studied whether the problems can be solved based on the difference of the substitution patterns in coding and non-coding regions.

Our main efforts were directed to the first problem, and the algorithm effectively solves it. Taking into account its linear runtime, the algorithm can serve as a useful filtration tool for any exon-recognition algorithm working with genome alignments. We have demonstrated that statistics of the possible values of pairs (K_N, K_S) in coding and non-coding regions can serve as the background to distinguish between the coding and non-coding fragments.

Putative exons show up worse correlation with the annotated exons than ELRs as well as the predictions made by the programs that use more sophisticated training technique (see Introduction). We plan to improve significantly this part of our algorithm. For example, we plan to generate for a given ELR several putative exons having different frames and 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 conservative positions of splicing sites.

AKNOWLEDGMENTS

Authors are thankful to A. Kondrashov, A. Ogurtsov, S. Shabalina, G. Bazykin, S. Sunyaev for helpful discussions. The work was supported by the Russian Foundation for Basic Research (grants Nos. 03-04-49369 and 02-07-90412); the Ministry for Industry, Science, and Technologies of the Russian Federation (grants Nos. 20/2002 and 5/2003); NIH (grant TW005899, co-PIs V. Tumanyan and M. Borodovsky), and NWO and ECO-NET grants. The work was partly performed during the visit of M.A. Roytberg to NCBI (January-March, 2004). We thank anonymous referees for their helpful comments.