Translated codons (Trons) useful for direct matching of a genomic DNA sequence and a protein sequence or profile

Osamu Gotoh [1]
gotoh@saitama-cc.go.jp
[1] Department of Biochemistry, Saitama Cancer Center Research Institute
818 Komuro, Ina-machi, Saitama 362 Japan

1 Introduction

Locating protein coding exons on a genomic DNA sequence is the initial and difficult step of predicting functions of the genes embedded in that part of the genome. Recent rapid development in this area has attained an accuracy of nearly 80 % for correctly predicting coding exons (CDSs) in mammalian genes [1]. However, for homology modeling of a protein structure, for example, we need perfectly accurate prediction of CDSs. This final bit of improvement could be obtained by directly matching the DNA sequence with a known protein sequence(s) of a homologous gene(s) [2] [4]. A new convention of encoding a DNA sequence into a series of 23-letter alphabets was devised for better performance of this type of analyses. With this convention, we developed a DNA vs. protein sequence alignment algorithm which allows for frame shift errors and long gaps corresponding to introns, in the standard framework of dynamic programming paradigm. The precision of prediction was further improved by consideration of coding potentials and splicing signals. More than 60 *Caenorhabditis elegans* cytochrome P450 (*CYP*) genes were examined by this method. The quality of the predictions was assessed by multiple alignment of conceptually translated protein sequences, and conservation patterns of intron insertions sites. The results indicated that the annotated CDS information in public databases contains considerable amount of miss-assignments.

2 Methods

A remarkable feature of the universal genetic code is that the second nucleotide in a codon is most influential for the specificity. In fact, all codons for an amino acid have a unique nucleotide at the second position, except for Ser (UCN and AGY) and termination (UAR and UGA) codons. Thus only 23 letters are necessary and sufficient to unambiguously encode both the original nucleotide sequence and conceptually translated amino acid sequences in the three frames. We propose to call each translated codon 'tron', and express them by the standard one-letter amino-acid codes, except for 'J', 'O', and 'U' that represent translated AGY, UAR, and UGA codons, respectively. A usual 20 × 20 amino-acid similarity matrix is easily expanded to 23 × 20 tron vs. amino- acid similarity matrix.

A special form of gap-penalty function was used. An insertion or deletion (indel) of k nucleotides was penalized in the usual way by an affine function if k is a multiple of 3, but otherwise an additional penalty was given to allow but disfavor potential frame shifts. An insertion of nucleotides longer than a given threshold was regarded as an intron, and a constant penalty was imposed independently of the length. A bonus was given such a long insertion if its ends conform to canonical exon-intron boundary signals and the reading frame continues. In spite of the rather complicated form of a gap penalty function, the computation time is proportional to the product of the lengths of the sequences under comparison as in the case of alignment of two sequences of the same kind [3].

3 Results

More than 60 putative genes coding for cytochrome P450 (CYP) enzymes were found in the *C. elegans* genomic sequence database maintained at the Sanger Center. Nearly a half of them were also available in GenBank with annotated CDSs. The exon-intron organizations of all of the genes were predicted by the method mentioned above. The genes were classified into three paralogue groups based on similarity in the translated amino acid sequences. No intron-insertion site is conserved across the groups. The gene organizations are also very divergent within each group, although most of the insertion sites are shared by more than two genes. Multiple alignments of translated amino acid sequences according to the GenBank annotations contained many indels which were absent in the corresponding alignment obtained from our prediction of CDSs. This observation indicates that the CDSs suggested in the public databases are affected by considerable amount of miss-assignments.

Acknowledgements

This work was supported in part by a Grant-in-Aid (No. 09272225) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Burge, C. and Karlin, S., "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol.*, 268:78-94, 1997.

[2] Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. "Gene recognition via spliced sequence alignment," *Proc. Natl. Acad. Sci. USA*, 93:9061-9066, 1996.

[3] Gotoh, O., "Optimal sequence alignment allowing for long gaps," *Bull. Math. Biol*, 52:359-373, 1990.

[4] Huang, X. and Zhang, J., "Methods for comparing a DNA sequence with a protein sequence," *Comput. Applic. Biosci.*, 12:497-506, 1996.