Actes des Journées Ouvertes Biologie, Informatique et Mathématiques, JOBIM 2007, Marseille, France (10-12 Juillet 2007), Brun C. and Didier G. (eds), pp.339-340

IMGT/LIGMotif: A tool to annotate V-, D- and J-GENE of immunoglobulin and T cell receptor of vertebrates

Jérôme Lane[1], Marie-Paule Lefranc[1,2] and Patrice Duroux[1]

[1] IMGT®, the international ImMunoGeneTics information system®, Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France
[2] Institut Universitaire de France, 103 Boulevard Saint Michel, 75005 Paris, France
{Jerome.LANE, Marie-Paule.LEFRANC, Patrice.DUROUX}@igh.cnrs.fr

Abstract: *The aim of this work is to investigate the bioinformatic problem of annotating immunoglobulin (IG) and T cell receptor (TR) loci in genomic DNA of vertebrates owing to the unusual structure of variable, diversity and joining types of genes symbolised by V-, D- and J-GENE, respectively. Indeed, conventional bioinformatic softwares which are based on standard structure of genes cannot identify them precisely. To solve this problem, a new tool named IMGT/LIGMotif specific to IG and TR has been developed for genomic annotation at IMGT®, the international ImMunoGeneTics information system®. At present time the software can predict these IG and TR genes for human and mouse. The processing of the annotation is based on the DESCRIPTION concept established in IMGT-ONTOLOGY, the first ontology in the domain of immunogenetics and immunoinformatics. The implemented algorithm combines a similarity search (BLAST) with a matching of V-, D- and J-GENE patterns.*

Keywords: Annotation, immunoinformatics, immunogenetics, IMGT, pattern matching, similarity search, immunoglobulin, T cell receptor.

1 Introduction

Major improvement in DNA sequencing has led to the availability of many new genome databases. Thus, the process of the annotation of genomic DNA must be enhanced to follow the evolution of the sequencing progress. This is a big challenge as they are many difficulties to overcome. The unusual structures of some types of genes like the variable, diversity and joining genes of immunoglobulins (IG) and T cell receptors (TR) symbolised by the labels V-, D- and J-GENE respectively, is one of them. Indeed, the coding sequences of the V-, D- and J-GENE (i.e. V-EXON, D- and J-REGION) are delimited in their 5' and/or 3' extremities by atypical structures like recombination signals (RS) and not only by conventional ones such as splicing sites, initiation and/or terminal translation signals. That is why conventional gene finding softwares are inefficient to predict these genes. Therefore, a Java annotation tool specific to IG and TR of vertebrates named IMGT/LIGMotif has been developed for genomic annotation at IMGT®, the international ImMunoGeneTics information system® [1]. A description of the principle and the algorithm of the program is given here.

2 Principle

The principle of the automatic annotation is based on the DESCRIPTION concept of IMGT-ONTOLOGY [2], the first ontology in the domain of immunogenetics and immunoinformatics. This concept corresponds to the definition of terms and rules which are necessary to describe the organization and components of IG and TR of vertebrates and to characterize their motifs. For example, 270 labels have been defined for the nucleotide sequences. The description takes root in the matching of patterns of V-, D- and J-GENE of IG and TR. The patterns are defined here as structures composed of motifs. A motif is defined as a short conserved, common and specific nucleotide sequence of IG and TR genes. Motifs can be either conventional as splicing sites and conserved amino acid codons or characteristic of the IG and TR like RS. In the patterns, motifs are separated from each other by intervals with a minimum and maximum nucleotides length.

3 Algorithm

The algorithm IMGT/LIGMotif has for objective the prediction of IG and TR genes and their annotation of a genomic sequence. It includes a matching of a set of patterns for V-, D- and J-GENE. A gene is predicted for a pattern match if the number of motifs found is sufficient. A predicted gene can have overlapping matches. In this case, matches are submitted to a selection based on the number of motifs found. Then, predicted genes are described using positions of motifs found. The pattern matching strategy is limited to the motifs. Thus, to improve the results, a similarity search is provided. A BLAST [3] is done against a set of long coding nucleotide IMGT reference sequences. The BLAST matches are filtered by given parameters like E-value and score. Finally, the BLAST and pattern matching permit to classify all predicted genes in three groups. The first group includes genes only predicted from the BLAST. The second group is composed of genes only predicted by the pattern matching. For these genes, only those composed of all the motifs of a pattern are kept. Finally, the third group includes genes that are both predicted from the BLAST and pattern matching.

4 Conclusion and perspectives

The combination of BLAST and pattern matching in IMGT/LIGMotif allows the prediction of IG and TR genes in genomic sequences and the description step of the annotation process by providing the delimitation of IMGT labels. The tool works on IG and TR loci of human and mouse. Preliminary work shows that the software works for IG and TR loci of other vertebrates species. Sets of patterns will be refined accordingly to the new data found on motifs.

Acknowledgements

We thank Gérard Mennessier for the first versions of IMGT/LIGMotif.

References

[1] M.-P. Lefranc IMGT, the international ImMunoGeneTics information system®: a standardized approach for immunogenetics and immunoinformatics, Nucl. Acids Res., 1(1):3.

[2] V. Giudicelli and M.-P. Lefranc, Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 15, 1047-1054, 1999.

[3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tool, J. Mol. Biol., 215(3):403-410, 1990.