EMBnet.journal
Bioinformatics in Action

Elhassani et al. (2021) EMBnet.journal 26(Suppl A), e957
http://dx.doi.org/10.14806/ej.26.A.957

Predicting L-PART1 exon using deep learning

Merouane Elazami Elhassani[1,2]✉, Loic Maisonnasse[1], Antoine Olgiati[1], Jérôme Rey[1], Veronique Giudicelli[1], Patrice Duroux[1], Sofia Kossida[1]

[1]IMGT®, The International ImMunoGeneTics Information System®, Institute of Human Genetics (IGH), Scientific Research National Center (CNRS), University of Montpellier, Montpellier, France

[2]ATOS Montpellier, immeuble Archimède, Montpellier, France

Competing interests: MEE none; LM none; AO none; JR none; VG none; PD none; SK none

Identifying and annotating Immunoglobulins (IG), T cell receptors (TR) genes of jawed vertebrate species effectively and precisely is still an arduous task, essentially due to the continuous avalanche of large genomic sequences produced by NGS technologies. In this study, to predict the L-PART1 exon (the first exon of IG and TR variable V-GENE), different deep neural network-based models (DNN, CNN, RNN, CNN-RNN) were trained in a supervised manner, which automatically learns features from annotated IG and TR genes. Those models will then be used to predict the L-PART1 exon within newly given sequences. Correct detection of this component would dramatically increase the chances to find the subsequent components constituting the V-GENE of the IG and TR. Sequence data were extracted from IMGT/LIGM-DB, the IMGT®1 nucleotides database: around 2756 L-PART1 of 354 species tagged as positive samples, negative samples were generated with different noise strategies. The final training datasets were formed by shuffling a portion of positive and negative samples. The majority of the trained models showed promising results. However, further studies are needed to investigate the possibility of predicting and locating all the components within the V-GENE of IG and TR.

Acknowledgements

We thank all members of the IMGT® team for their expertise and constant motivation. IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), 5th PCRDT Quality of Life and Management of Living Resources (QLG2 2000 01287), and 6th PCRDT Information Science and Technology (ImmunoGrid, FP6 IST 028069) programmes of the European Union (EU). IMGT® received financial support from the GIS IBiSA, the Agence Nationale de la Recherche (ANR) Labex MabImprove (ANR 10 LABX 53 01), the Région Occitanie Languedoc Roussillon (Grand Plateau Technique pour la Recherche (GPTR), BioCampus Montpellier. IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), the University of Montpellier, and the French Infrastructure Institut Français de Bioinformatique (IFB) ANR 11 INBS 0013. IMGT® is a registered trademark of CNRS. IMGT® is member of the International Medical Informatics Association (IMIA) and a member of the Global Alliance for Genomics and Health (GA4GH). This work was granted access to the High Performance Computing (HPC) resources of Meso@LR and of Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation 036029 (2010 2021) made by GENCI (Grand Equipement National de Calcul Intensif).

References

1. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G et al. (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res., **43**(Database issue):D413-22. http://dx.doi.org/10.1093/nar/gku1056