Bioinformatics, 36(16), 2020, 4510–4512
doi: 10.1093/bioinformatics/btaa574
Advance Access Publication Date: 17 June 2020
Applications Note

Sequence analysis

SOS: online probability estimation and generation of T-and B-cell receptors

Giulio Isacchini[1,2], Carlos Olivares[1], Armita Nourmohammad[2,3,4], Aleksandra M. Walczak[1,†,*] and Thierry Mora[1,†,*]

[1]Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France, [2]Max Planck Institute for Dynamics and Self-organization, 37077 Göttingen, Germany, [3]Department of Physics, University of Washington, Seattle, WA 98195, USA and [4]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[†]The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
*To whom correspondence should be addressed.
Associate Editor: Peter Robinson

Received on April 3, 2020; revised on May 18, 2020; editorial decision on June 9, 2020; accepted on June 10, 2020

Abstract

Summary: Recent advances in modelling VDJ recombination and subsequent selection of T- and B-cell receptors provide useful tools to analyse and compare immune repertoires across time, individuals and tissues. A suite of tools—IGoR, OLGA and SONIA—have been publicly released to the community that allow for the inference of generative and selection models from high-throughput sequencing data. However, using these tools requires some scripting or command-line skills and familiarity with complex datasets. As a result, the application of the above models has not been available to a broad audience. In this application note, we fill this gap by presenting Simple OLGA & SONIA (SOS), a web-based interface where users with no coding skills can compute the generation and post-selection probabilities of their sequences, as well as generate batches of synthetic sequences. The application also functions on mobile phones.

Availability and implementation: SOS is freely available to use at sites.google.com/view/statbiophysens/sos with source code at github.com/statbiophys/sos.

Contact: aleksandra.walczak@phys.ens.fr or thierry.mora@phys.ens.fr

1 Introduction

The adaptive immune system recognizes pathogens through the generation of a highly diverse repertoire of T- and B-cell receptors (TCR and BCR) which have the potential to recognize even unknown pathogen and initiate an immune response. To produce this diversity, it exploits a highly stochastic process named V(D)J recombination. In addition, to block possible auto-reactive receptors, a selection process is mounted in the thymus for T cells, and a similar process of central tolerance is implemented for B cells. Probabilistic models of TCR and BCR have been proposed (Elhanati *et al.*, 2014; Murugan *et al.*, 2012; Ralph and Matsen, 2016) based on immune repertoire sequencing data (Bradley and Thomas 2019; Georgiou *et al.*, 2014; Heather *et al.*, 2017; Minervina *et al.*, 2019). Software has been developed to infer the probability of generation of any BCR or TCR (IGoR; Marcou *et al.*, 2018), and to evaluate this probability for both nucleotide and amino-acid sequences (OLGA; Sethna *et al.*, 2019). Another tool (SONIA; Sethna *et al.*, 2020) was released to infer the selective pressures acting on the receptors and used to predict the probability of naive sequences in the periphery (Isacchini *et al.*, 2020). To make these tools available to a broader audience, we provide a new web tool which allows for the analysis of single TCR and BCR sequences.

2 Features

As explained in the introductory 'About' tab, the web tool evaluates the generation and post-selection probability of single naive TCRs and BCRs in different species based on the specific sequence the user inputs manually. The engine is based on two pieces of python software, OLGA and SONIA and shipped with pre-trained models of recombination and selection for the following loci: human alpha and beta chains or TCR (TRA and TRB), human heavy and light chain of unmutated BCR (IGH, IGK and IGL) and mouse TRB.

After choosing the species and receptor chain in the 'Evaluate' tab, the user inputs a Complementary Determining Region 3

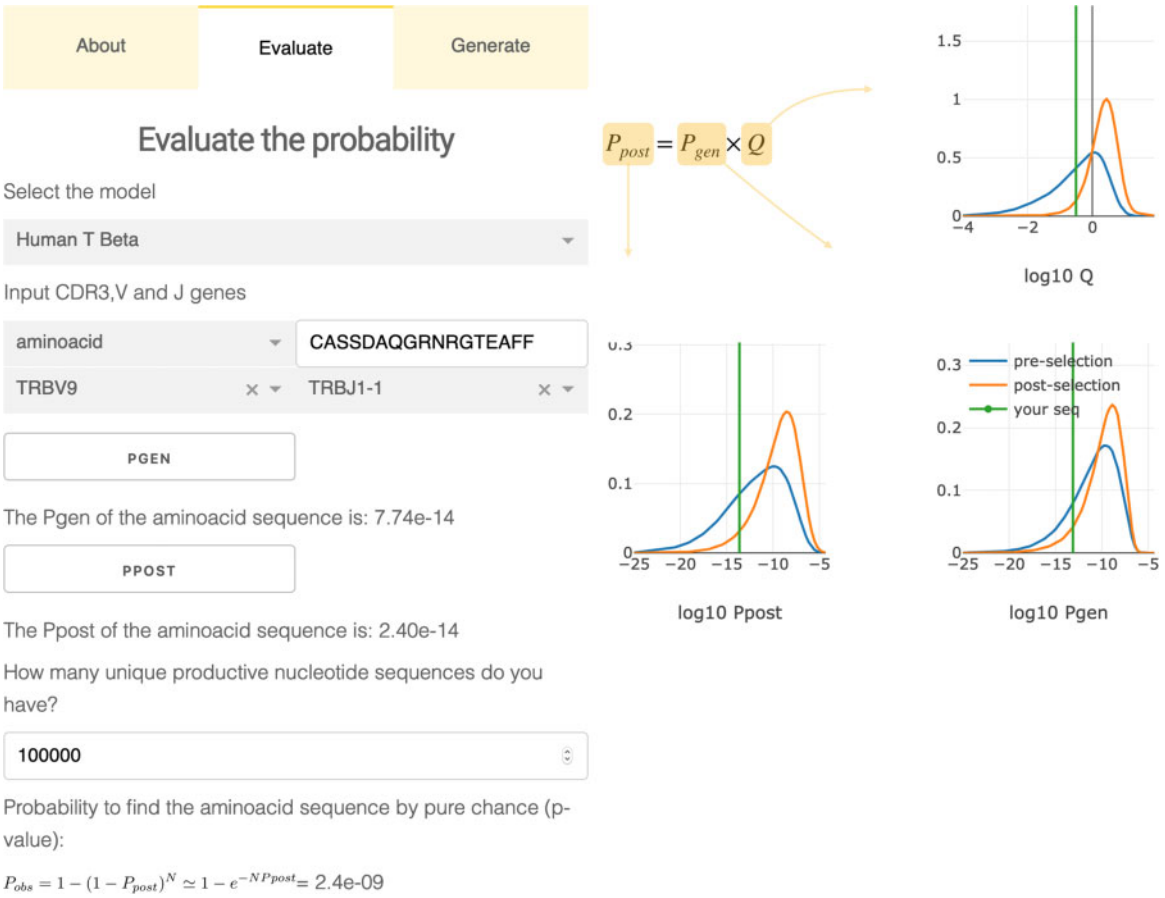

Fig. 1. SOS web interface. The user inputs a CDR3 sequence (amino acid or nucleotides) and V and J segments. The programme outputs the generation probability P_{gen}, the probability in the periphery P_{post} and evaluates a *P*-value corresponding to the probability of finding that sequence by chance in a repertoire of size *N* (input by user). An additional tab allows for the generation of synthetic repertoires

(CDR3), either as a nucleotide or an amino acid sequence, and optionally V and J germline genes from dropdown lists. The server outputs the generation probability (P_{gen}, conditioned on sequence productivity), and the post-selection probability (P_{post}), as shown in Figure 1 (left). When V and J are not specified, the programme sums over all possibilities for these segments to calculate the total probability of the CDR3.

To help interpret the result and assess how the sequence of interest compares to others, P_{gen}, P_{post} and the selection factor $Q = P_{post}/P_{gen}$ are plotted as green vertical lines on histograms of random sequences (Fig. 1, right). These random sequences are drawn either from the pre-selection distribution P_{gen} (blue line) obtained by generation with an IGoR-trained recombination model or from the post-selection distribution P_{post} (orange line) obtained by weighting the same sequences by their selection factor Q, which was shown to describe the data well (Elhanati *et al.*, 2014). That feature only works when V and J and specified. The tool also provides an estimation of the probability to observe the sequence in a generic repertoire. The user inputs the size N of the sequenced repertoire (unique productive nucleotide sequences), and the tool outputs the probability of observing the sequence within a repertoire of that size, given by $1 - (1 - P_{post})^N$.

Using the 'Generate' tab, the user can synthetize a specified number of receptor sequences from P_{gen} or P_{post}, after choosing the species and chain type from dropdown lists. The file with the generated sequences, composed of the CDR3 sequence (nucleotide and amino-acid translation), V and J segments, is available for download as a CSV file. The user may fix the seed of the random number generator for reproducibility.

3 Discussion

The interface can be used by investigators to evaluate how surprised one should be to find a given sequence in one or multiple repertoires. It could help distinguish receptors with a specific function from chance detections. The tool can also be used to evaluate the potential of certain receptors (in particular antibodies, albeit in their unmutated version) for vaccination or therapeutic purposes. The web interface is also available on mobile phones without the plotting options.

Funding

AN, AMW and GI were supported in part by the DFG grant (SFB1310) for Predictability in Evolution. AN and GI were supported in part by the MPRG funding through the Max Planck Society. CO, GI, TM and AMW were supported in part by grant ERC POC no. 824735. GI, TM and AMW were supported in part by grant ERC CoG n. 724208.

Conflict of Interest: The authors have no conflicts of interest

References

Bradley,P. and Thomas,P.G. (2019) Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Annu. Rev. Immunol.*, **37**, 547–570.

Elhanati,Y. *et al.* (2014) Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci. USA*, **111**, 9875–9880.

Georgiou,G. *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.*, **32**, 158–168.

Heather,J.M. *et al.* (2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief. Bioinf.*, **19**, 554–565.

Isacchini,G. *et al.* (2020) On generative models of T-cell receptor sequences. *Phys. Rev. E.*, 101, 062414.

Marcou,Q. *et al.* (2018) High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, **9**, 561.

Minervina,A. *et al.* (2019) T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. Transpl. Int., 32: 1111–1123.

Murugan,A. *et al.* (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. USA*, **109**, 16161–16166.

Ralph,D.K. and Matsen,F.A. (2016) Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.*, **12**, e1004409.

Sethna,Z. *et al.* (2019) OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, **35**, 2974–2981.

Sethna,Z. *et al.* Y. (2020). Population variability in the generation and thymic selection of T-cell repertoires, 1–17. *arXiv:2001.02843*.