Sequence analysis

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

Giulio Isacchini1,2, Carlos Olivares1, Armita Nourmohammad2,3,4, Aleksandra M. Walczak1,T,∗ and Thierry Mora1,T,∗

1Laboratoire de physique de l’Ecole normale superieure (PSL University), CNRS, Sorbonne Universite, Universite de Paris, 75005 Paris, France, 2Max Planck Institute for Dynamics and Self-organization, 37077 Gottingen, Germany, 3Department of Physics, University of Washington, Seattle, WA 98195, USA and 4Fred 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.

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 pathogens 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

V C
(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 \( P_{\text{gen}} \), the probability in the periphery \( P_{\text{post}} \), 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.

### 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
