On generative models of T-cell receptor sequences

Giulio Isacchini,[1,2,*] Zachary Sethna,[3,*] Yuval Elhanati,[3] Armita Nourmohammad,[1,4] Aleksandra M. Walczak,[2] and Thierry Mora[2]

[1]*Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany*

[2]*Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, and Université de Paris, 75005 Paris, France*

[3]*Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA*

[4]*Department of Physics, University of Washington, 3910 15th Avenue Northeast, Seattle, WA 98195, USA*

In this comment on *Davidsen et al.,* "Deep generative models for T cell receptor protein sequences", *eLife* 2019;8:e46935, we compare the performance of the variational auto-encoder presented in that article to a previously proposed approach for which a software implementation, SONIA, has been recently released. We find that SONIA performs as well as the variational auto-encoder, at a lower computational cost.

Davidsen et al. [1] describe an elegant approach for learning the distribution of T-cell receptor beta sequences (TCR), based on a Variational Auto-Encoder (VAE). The method makes it possible to generate new sequences with the same statistics as real repertoires, and to evaluate the frequency of individual sequences. Its main strength is that it does not take any information about the origin of these sequences through VDJ recombination and thymic and peripheral selection. Yet it manages to extract statistical regularities imprinted by these processes.

An alternative approach [2–4] has been to learn interpretable, knowledge-guided models of VDJ recombination and selection. First, a recombination model for the probability of generation of a sequence σ, denoted by $P_{\rm gen}(\sigma)$, is learned from failed nonproductive rearrangements which are free of selection biases [2, 4]. This model describes in detail the probabilities of V, D, and J usages, and of deletion and insertion profiles. Second, a model of selection is learned on top of the generation probability $P_{\rm gen}$ to describe the distribution of productive sequences, $P_{\rm post}(\sigma) = Q(\sigma)P_{\rm gen}(\sigma)$, where $Q(\sigma) = (1/Z)\exp[\sum_i \theta_i x_i(\sigma)]$ is a selection factor calculated through a combination of sequence features $x_i(\sigma)$: sequence length, identity of amino acids at each position, and the choice of V and J segments [3].

Davidsen et al. [1] compared their VAE approach to a reduced version of this selection model (not examined in [3]), which they call `OLGA.Q`. Because no software implementation was provided with the original article, only VJ choice and sequence length were included in $Q(\sigma)$, leading to mediocre performance relative to the VAE. Here, using SONIA, a new publicly available software package of selection inference, we fill this gap to offer a more complete comparison. Fig. 1 repeats the analyses of Fig. 2, 3, and 4 of the commented article, but shows the results of the full selection model of [3] ('SONIA'). In constrast to `OLGA.Q`, SONIA performs at least as well as the VAE in predicting individual sequence frequencies (Fig. 1a). The VAE captures some of the rules of recombination and selection encoded by SONIA, but not as much as it did with `OLGA.Q` (Fig. 1b). In return, SONIA seems

*These authors contributed equally.

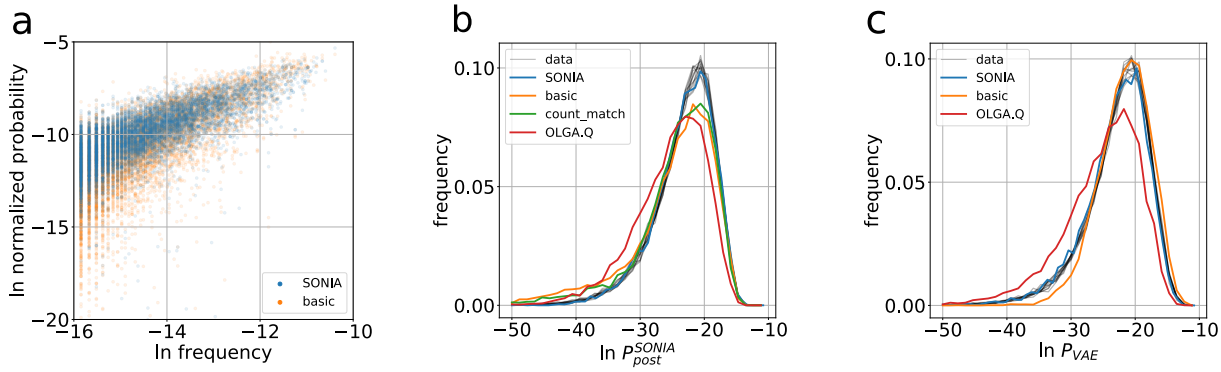

FIG. 1: **A model of generation and selection describes repertoire statistics well**. **a.** Comparison of SONIA versus VAE (`basic` version) prediction for TCR frequencies using the entire 666 strong cohort from Ref. [5], responding to the lower left plot of Fig. 2 of the commented paper. A universal generation model P_{gen} was trained from unique out-of-frame nucleotide sequences across donors. Then the selection model Q was trained using $2 \cdot 10^5$ unique productive amino acid sequences that were absent from the test set. Pearson correlation coefficients are $r^2 = 0.464$ for SONIA, and $r^2 = 0.447$ for the VAE. **b.** Distribution of $P_{\text{post}}^{\text{SONIA}}$ of TCRs from 11 individuals from Ref. [6], as well as sequences generated by SONIA, the `basic` and `count_match` versions of the VAE, and `OLGA.Q`, responding to Fig. 3 of the commented paper. The SONIA model Q was trained on a set of 10^5 sequences, on top of the P_{gen} model trained for **(a)**. The match of distributions between SONIA and data shows that SONIA captures the sequence statistics well, while VAE sequences have a broader distribution. **c.** Distribution of P_{VAE} for the same sequences as in **(b)**, responding to Fig. 4 of the commented paper. SONIA-generated sequences show much smaller deviations from the data than `OLGA.Q`.

to capture most of the characteristics of the real sequences learned by the VAE (Fig. 1c).

In summary, both approaches, VAE and SONIA, perform equally well, with perhaps a slight advantage for the latter. SONIA is an order of magnitude faster than the VAE, which uses Monte-Carlo sampling to calculate predicted frequencies. The average computing time for $P_{\text{post}}^{\text{SONIA}}(\sigma)$ is 14 ms per sequence on a laptop computer and 3 ms on a 16-core computer, versus 0.18 s for the `basic` version of the VAE, and 0.87 s for its refined version `count_match` (no parallelization possible). SONIA took 30 minutes on a laptop computer (using an IGoR model trained on 10^5 sequences, which took 1 hour on a 64-core cluster node), versus 9.5 hours for VAE training.

These results suggest that, while knowledge-free approaches such as the VAE perform well, there is still value in preserving the structure implied by the VDJ recombination process as a baseline for learning complex distributions of immune repertoires. Extending the SONIA model considered here beyond a simple linear combination of features offers interesting directions for future improvement in repertoire modeling.

Code availability. All code for reproducing the figures of this comment can be found at `https://github.com/statbiophys/compare_selection_models_2019/`. The SONIA package upon which that code builds is available at `https://github.com/statbiophys/`

`SONIA/`.

---

[1] Davidsen K, et al. (2019) Deep generative models for T cell receptor protein sequences. *eLife* 8:e46935.

[2] Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* 109:16161–16166.

[3] Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci.* 111:9875–9880.

[4] Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* 9:561.

[5] Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* 49:659–665.

[6] De Neuter N, et al. (2019) Memory CD4+ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus. *Genes and Immunity* 20:255–260.