Statistical investigations into immune system diversity in mice and man

Yuval Elhanati[1], Zachary Sethna[1], Curtis G. Callan, Jr.[1], Thierry Mora[2] and Aleksandra M. Walczak[3]

Short Abstract — The ability of the adaptive immune system to respond to arbitrary pathogens stems from the broad diversity of immune cell surface receptors. This diversity originates in a stochastic DNA editing process. We can infer the biological details of this process from sequence data using statistical algorithms based on the transfer matrix method from statistical mechanics. Using samples from mice of different ages, we track the time evolution of mice immune system, from embryo to adult. Comparing the models we learn from humans and mice, we also find surprising differences between species in term of immune diversity.

Keywords — immune receptor diversity, TCR, statistical inference, mice

I. INTRODUCTION

THE immune system in mammals and other vertebrates is able to recognize and respond to specific pathogenic threats via membrane receptors on T cells. When a T cell receptor (TCR) successfully binds to a pathogenic molecule (antigen), an immune response is initiated. Thus, the effectiveness of the system against a wide range of pathogens depends on a large diversity of different membrane receptors.

This diversity is generated by random combinations of genomic elements – V/D/J genes, with random insertions and deletions between them. Not all recombination scenarios are equally likely, and the same receptor sequence may be obtained in several different ways [1].

Systematic quantification of the probability distributions of the possible recombinations, based on available sequence data, is indispensable for the study of the immune system diversity, in different animals and under varying conditions

II. METHODS

Inferring the distributions from sampled receptor sequences is a computationally hard problem, naïvely requiring enumerating every possible recombination scenario for every sequence [2]. However, this problem can be shown to be analogous to the calculation of observables in a statistical mechanics system as averages over an ensemble. Breaking down the random generation model into a chain of states, the partition function can be calculated efficiently using the transfer matrix method.

We developed and implemented such a computational method, based on the forward-backward algorithm, that efficiently infers a generation model for the rearrangement process. The inferred gene selection, deletion and insertion profiles characterize the generation machinery. Furthermore, the model can be used to study the entire distribution, generate synthetic sequences from it and calculate the probability of generation of any receptor sequence. This is a powerful diagnostic tool, applicable to any reasonable sample.

III. RESULTS

We previously inferred a TCR generation model for healthy humans [2]. Now, we learned generation models from TCR sequences taken from the blood and thymus of mice of different ages, quantifying the changes in the recombination process that occur from embryo to young adult [4]. We find a rapid increase with age in the number of random insertions and a dramatic increase in diversity. Since the blood accumulates cells over time, blood repertoires are mixtures of different statistical recombination processes and we unravel the mixture statistics to obtain a picture of the time evolution of the early immune system. Also, comparing with the previous human results, we find even the adult mice repertoire much less diverse despite the similar pathogenic challenge.

IV. CONCLUSIONS

We can quantify and infer the generation process of TCRs using efficient statistical methods. We applied this procedure to samples taken from mice in different ages, tracking the maturation process of the immune system and finding surprising differences in diversity from humans.

REFERENCES

[1] Murphy KP, Travers P, Walport M, Janeway C (2008) *Janeway's Immunobiology.* Garland, New York.
[2] Murugan A, et al. (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. PNAS 109: 16161-16166.
[3] Elhanati et al. (2016) RepgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data. Bioinformatics 32 (13): 1943-1951.
[4] Sethna et al. (2017) Insights into immune system development and function from mouse T-cell repertoires. PNAS 114: 2253-2258.

[1]Joseph Henry Laboratories, Princeton University, Princeton, NJ
[2]Laboratoire de Physique Statistique, CNRS and ENS, Paris, France
[3]Laboratoire de Physique Théorique, CNRS and ENS. Paris, France