A Bayesian semi-parametric approach for the analysis of T-cell receptor diversity

N. Sepúlveda[a,b], *C. D. Paulino*[b,c], *M. Guindani*[d], *Peter Mueller*[e]

[a] London School of Hygiene and Tropical Medicine, London, United Kingdom
[b] Center of Statistics and Applications of University of Lisbon, Lisbon, Portugal
[c] Instituto Superior Técnico, Lisbon, Portugal
[d] University of Texas M. D. Anderson Cancer Center, Houston, USA
[e]University of Texas at Austin, USA

e-mails: nuno.sepulveda@lshtm.ac.uk
dpaulino@math.ist.utl.pt
mguindani@mdanderson.org
pmueller@math.utexas.edu

Abstract: In immunology, the T-cell receptor (TCR) diversity is instrumental to understand how the immune system correctly discriminates harmful microorganisms from body components. As in species diversity studies, this quantity is only accessible by means of estimation based on a sample of TCR sequences. Similar to species-abundance distribution in Ecology, the so-called sequence-abundance distribution is the core of the whole analysis. The primary aim is to estimate the total number of unseen TCR types by fitting simple parametric models to that distribution [1]. However, theoretical studies on T-cell physiology have recently shown that the sequence-abundance distribution results from a complex and intricate mixture of different TCR populations (reviewed in ref. [2]). Thus, although fitting well to available data, current parametric solutions for TCR estimation seem unlikely to provide reliable estimates. Here we propose a flexible Bayesian semi-parametric model where TCR sequences are sampled according to a Poisson distribution with a given rate that, in turn, varies with another but yet unknown probability distribution [3]. We assume a Dirichlet process for this second level distribution using an appropriate Gamma distribution as the centre measure. We illustrate the proposed model with previously published data on type I diabetes, a disease caused by erroneous immune responses to pancreatic cells. We also simulate data from realistic immunological settings to demonstrate the superiority of our method in relation to current parametric counterparts.

Keywords: Poisson mixture models; Bayesian parametric and semiparametric analysis.

This work has been supported by FCT - Projects PEst-OE/MAT/UI 0006/2014 and PTDC/MAT/118335/2010.

Referências

[1] Sepúlveda, N., Paulino, C. D., Carneiro, J. (2010). Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J. Immunol. Methods* 353:124–137.

[2] Sepúlveda, N. (2009). *How is the T-cell repertoire shaped?* PhD thesis, University of Oporto, Portugal.

[3] Guindani, M., Sepúlveda, N., Paulino, C. D., Mueller, P. (2014). A Bayesian semiparametric approach for the differential analysis of sequence counts data. *Appl. Statist.* 63:385–404.