Introduction to data analysis for natural and social sciences

Marco Casari

June 26, 2023

1 Introduction

The present document constitutes the second part of the exam. A summary of article "Inferring the immune response from repertoire sequencing" is provided.[1] Results sections "Modeling repertoire variation" and "Inferring the noise profile from replicate experiments" are supplied with technical results.

2 Summary

The article introduces a probabilistic model able to estimate the noise and describe clonotypes expansion in longitudinal Repertoire Sequencing data (RepSeq). Next Generation Sequencing (NGS) provides large amounts of RepSeq data, which standard inference methods fails to describe accurately due to experimental noise and biological diversity. Many factors contribute to the noise: some are related to the experiment execution (e.g. sampling procedure, library preparation), others have biological origin (e.g. gene expression). These factors introduce variability in sequence counts which is not related to the natural variability of T-Cell Receptors (TCRs).

The model presented in the article is able to decouple the noise distribution from the clonotype counts distribution, learning parameters of both from RepSeq data. Additionally, once the parameters are learned, a Bayesian approach can be used to ...

2.1 Modeling repertoire variation

The model consists in three main parts, identified by the noise model, the clone size distribution $\rho(f)$ and the dynamical model $G(f', t'|f, t)$.

The noise model $P(n_i|f_i)$ is a conditional probability of cell counts n_i of the i-th clonotype given the true frequency f_i. This definition is necessary because frequencies f_i are not known, since RepSeq experiments supply a cell count for each clonotype and it is a noisy function of the frequency. Index $i = 1, \dots, N$ is a label to identify each clonotype, with N total number of clonotypes in the immune system. This value is not known either, the total number of counts N_{read} is used instead.[2]

Three functional forms has been selected for the noise model: Poisson distribution, negative binomial distribution and a two-step model. Poisson distribution is set with mean $\bar{n}_i = f_i N_{\text{read}}$ but it is not effective in reproducing observed noise because of its smaller variance. The negative binomial distribution has mean $\bar{n}_i$ and variance $\bar{n}_i + a\bar{n}_i^{\gamma}$

To build the model, three steps are involved, one for the characterisation of each model part.

, each one dedicated to is composed by three interacting parts, which are

2.2 Inferring the noise profile from replicate experiments

[1] Version 2 of the article is referenced. Full citation: Puelma Touzel M, Walczak AM, Mora T (2020) Inferring the immune response from repertoire sequencing. PLoS Comput Biol 16(4): e1007873. `https://doi.org/10.1371/journal.pcbi.1007873`.

[2] Clonotype index is omitted when the meaning of variables is clear.