Introduction to data analysis for natural and social sciences

Marco Casari

June 27, 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 (RepSeq) data. 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. In particular mRNA data is used, but with suitable parametrisation of the noise distribution also gDNA data can be analysed.

Once the parameters are learned, Bayesian inference can be used to obtain the posterior probability of expansion and predict the dynamics of individual clonotypes. Possible applications are the description of response to natural stimuli or strong perturbations (e.g. vaccination, acute infection), tracking clonotypes in different tissues, the description of evolution of chronic infections (e.g. HIV).

Some characteristics distinguish the model from standard approaches of differential expression analysis. The possibility to work in a Bayesian framework is one, others are the manipulation of finite cell counts given an activated clonotype, instead of working with average gene expression data, and the use of counts distribution as prior for the evaluation of likelihood of expansion.

Finally, validation of the model is performed on data gathered from a longitudinal experiment encompassing vaccination against yellow fever, which poses as acute infection in humans.

2.1 Modeling repertoire variation

The model consists in three main parts, identified by: noise model, clone size distribution and dynamical model.

The noise model for an individual clonotype is a conditional probability $P(n|f)$ of cell counts n given the true frequency f. This definition is necessary because frequency f is not known and RepSeq experiments supply a TCR read count (i.e. a proxy for cell count) which depends on the frequency. Three probability distributions have been considered for the noise model: Poisson distribution, negative binomial and a mixed negative binomial-Poisson model. Poisson distribution is set with mean $\bar{n} = fN_{\text{read}}$ where N_{read} is the total number of counts, but it is not effective in reproducing observed noise because of its smaller variance. The negative binomial distribution has mean $\bar{n}$ and variance $\bar{n} + a\bar{n}^{\gamma}$, where the two parameters a and γ are used to learn the noise variance. The third

[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.

model considers explicitly the cell count m of the clonotype in the sample used to infer the general model. It is built by summing over $m \in \{1, \dots, M\}$ the product of the negative binomial $P(m|f)$ with mean fM and the Poisson distribution $P(n|m)$ with mean $\frac{m}{M} N_{\text{read}}$, where M is the total number of cells in the sample, which is an additional parameter of the general model. This distribution is preferred because it has the best performance on RepSeq data.

The clone size distribution $\rho(f)$ of a given clonotype is a power law with exponent ν. To avoid divergence, a minimum frequency $f_{\min}$ is set and its value corresponds to the frequency of lymphocytes if each one were an individual clonotype. There is one constraint on clonotype frequencies, namely they must sum up to 1, and it is satisfied when number of clonotypes N is large and $\langle f \rangle = \frac{1}{N}$. The joint distribution of frequencies is the product of $\rho(f)$ over the entire repertoire and it is set non-zero when the constraint holds. Normally the constraint denies immediate factorisation of the joint distribution, except when it is verified, for instance during the sampling procedure when it is enforced.

The dynamical model $G(f', t'|f, t)$ describes the probability of expansion or contraction of a clonotype from frequency f at time t to frequency f' at a successive time t'. For the purposes of the article, t is before vaccination and t' is after vaccination. To explicit $G(f', t'|f, t)$, it is assumed that a fraction α of each clonotype reacts to a stimulus and its expansion or contraction does not depend on its size. To quantify the response, a log fold change $s = \ln\left(\frac{f'}{f}\right)$ is defined and the infinitesimal variation in frequency of the dynamical model is expressed as infinitesimal variation of function

$$\rho_s(s) = (1 - \alpha)\delta(s - s_0) + \alpha\rho_{\text{exp}}(s - s_0) \quad ,$$

where $s_0 < 0$ is a measure of contraction extended to the entire clonotype to compensate for the expansion and preserve the constraint on frequencies, while ρ_{exp} modulates the expansion of activated clonotypes. The form of ρ_{exp} is chosen according to the case under analysis.

Parameters are evaluated in two consecutive steps. First, replicate RepSeq experiments are performed at time t and parameters of distributions $P(n|f)$ and $\rho(f)$ are inferred simultaneously. Second, repertoires at different times t and t' are compared to evaluate the parameters of distribution $G(f', t'|f, t)$.

2.2 Inferring the noise profile from replicate experiments