Introduction to data analysis for natural and social sciences

Marco Casari

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 aspects.

2 Summary

The article introduces a probabilistic model able to estimate the noise and describes clonotypes expansion in longitudinal Repertoire Sequencing (RepSeq) data. Next Generation Sequencing (NGS) provides large amounts of RepSeq data, which standard inference methods fail 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 clonotypes count distribution, learning parameters of both from RepSeq data. In particular mRNA data is used, but with suitable parametrisations 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 as well as strong perturbations (e.g. vaccination, acute infection), tracking clonotypes in different tissues and the description of evolution of chronic infections (e.g. HIV).

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.

Some characteristics distinguish the model from standard approaches of differential expression analysis. One is the possibility to work in a Bayesian framework, 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 the likelihood of expansion.

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 are 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, as can be seen with experimental data in Figure 1. 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 model considers explicitly the cell count m of the

[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.

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 because 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 $s_0 < 0$ measures a 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 same time (same day in the article) 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

To infer the set of parameters θ_{null} of distributions $P(n|f)$ and $\rho(f)$ for a given clonotype, maximum likelihood estimation (MLE) is used on

$$P(n, n'|\theta_{\text{null}}) = \\ = \int_{f_{\min}}^{1} \mathrm{d}f \rho(f|\theta_{\text{null}}) P(n|f, \theta_{\text{null}}) P(n'|f, \theta_{\text{null}}) \quad ,$$

where n and n' are counts from two replicate experiments.

Since many clonotypes are missed by the experimental procedure (e.g. due to their small size) or inactive, only pairs (n, n') satisfying $n + n' > 0$ are observed. Their number is $N_{\text{obs}} \approx 10^5$, much smaller than the true value $N \approx 10^{10}$ for humans. Maximisation is then performed on the joint likelihood, obtained as product of $P(n, n'|\theta_{\text{null}}, n + n' > 0)$ over all observed pairs.

Experimental limitations affect also the normalisation condition, where the sum on clonotypes is now performed up to N_{obs} and terms for inactive clonotypes are evaluated separately. Estimated parameters do not change significantly between the new normalisation condition and the previous one applying on the entire repertoire.

Some of the inferred values are power law exponent $\nu \approx 2.1$ and minimum frequency $f_{\min} \approx 10^{-11}$, both compatible with observations in humans. The inferred parameters are consistent across donors and time.

The model is used to reproduce counts of replicate experiments. Results for the noise model of one experiment are collected in Figure 1: the Poisson distribution is under-dispersed with respect to the experimental data, while the mixed negative binomial-Poisson model shows the same variance. The mixed model reproduces the marginal distribution $P(n)$ (i.e. $P(n, n')$ marginalised over n') accurately with a power law behaviour, except for low counts due to experimental noise. Moreover, it fits more accurately the tails of $\rho(f)$, expressed as $P(n, n' = 0)$, than the negative binomial.

Three Hill diversities of the repertoire can be evaluated by using $\rho(f)$ and the estimate $N = \frac{N_{\text{obs}}}{1 - P(n=0)}$, where distribution $P(n)$ is $P(n|f)$ marginalised over f. Species richness is $N \approx 10^9$

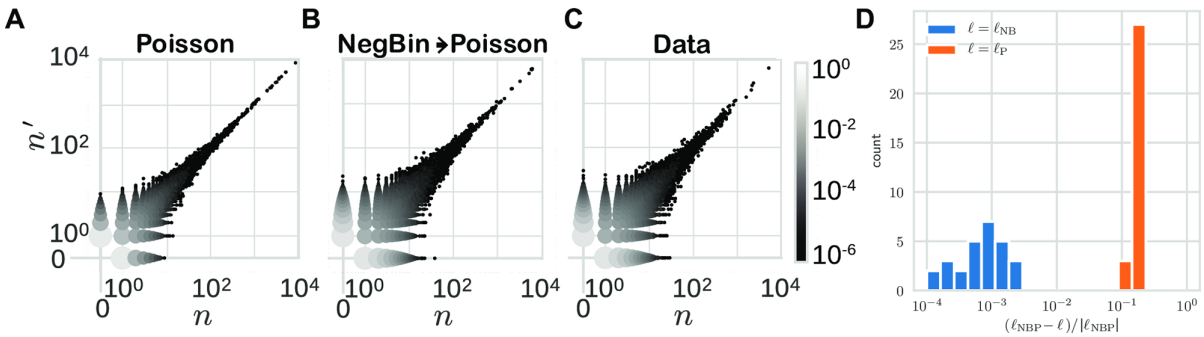

Figure 1: Probability distributions of count pairs obtained from inferred (A) Poisson distribution, (B) mixed negative binomial-Poisson model and (C) observed data of a replicate experiment. Figure (D) shows the comparison performed on the same dataset between log likelihoods of negative binomial (i.e. subscript "NB") and Poisson distribution (i.e. subscript "P") with respect to the negative binomial-Poisson model (i.e. subscript "NBP"). Cfr. Fig 2 in the article.

and is compatible with existing known boundaries of TCRs count.