Mixture Models for Analyzing Antigen Receptor Data

Joshua Greene[1], Leszek Ignatowicz[2], and Grzegorz Rempala[1]

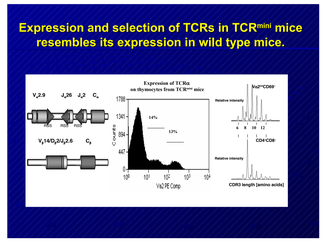
Department of Biostatistics[1], Center for Biotechnology and Molecular Medicine[2], Georgia Health Sciences University, Augusta, GA 30912, USA

1. Introduction

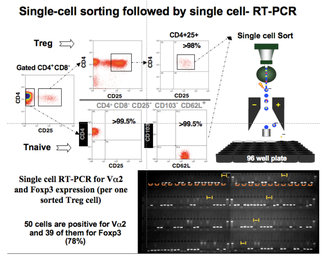
Currently one of the standard ways of analyzing a vertebrate immune system is to sequence and compare the counts of specific antigen receptor clones derived from various tissues under different experimental or clinical conditions. In particular, one can compare the immune systems of elderly subjects with the immune systems of subjects from other age groups. The statistical challenges involved in these analyses are difficult and are not readily addressed by standard contingency table analyses due to the serious under-sampling of the receptor populations. In order to address these challenges hierarchical clustering algorithms for analyzing diversity among TCR populations will be derived by using the multivariate Poisson-lognormal model. Further, richness estimates will be obtained by using the multivariate Poisson-lognormal model, which will be compared with richness estimates obtained from a complete hierarchical Bayesian richness analysis.

2. Data Collection Procedures

Since we currently do not have the ability to collect TCR data from humans, our statistical analyses will be conducted on TCR data collected from the transgenic mouse animal model. The pictures below describe the construction of the animal model and the data collection procedure used to obtain the TCR data respectively.



TCR^{mini} preliminary data. The leftmost panel demonstrates how TCR^{mini} mice are constructed while the other panels demonstrate that the TCR expression in the thymus in TCR^{mini} mice closely follows the TCR expression profile observed in wild type mice (i.e. only most mature thymocytes express high levels of TCR). Left panel shows spectratyping of CDR3 before (upper) and after (lower) thymic selection.



Method of data collection for antigen receptors. In order to examine the distribution of different antigen receptors (TCR) on CD4+T cells, our collaborators use the single cell sorting followed by single cell RT-PCR. In this figure, two populations of CD4+ T cells were discriminated via flow cytometry based on different expression of CD4, CD25 and CD62L surface markers. The monoclonal antibodies conjugated with different fluorochromes specific for these molecules were used to stain and then to sort individual CD4+ T cells into 96 well plate. Two populations of CD4+ T cells were isolated that have different function in the immune system (effector or regulatory). Next mRNA from each cell was used to produce cDNA, which then was used to amplify by RT-PCR the CDR3 fragments (one that encodes variable region of TCRa chain). Finally CDR3α PCR product is sequenced and obtained sequences are aligned and analyzed [3][1].

3. Methods

The dissimilarity measures for hierarchical clustering given here were also used by Rempala et al.[2] and both the multivariate model and richness estimates given here are straightforward extensions of the bivariate model and richness estimates considered by Rempala et al.[2]:

Multivariate Poisson-lognormal (MPLN) model

$$p(k_1,\ldots,k_m,\mu,\Sigma) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}[\Pi_{i=1}^{m}\, g_{k_i}(\mu_i,\sigma_i,u_i)]\times \phi(u_1\ldots u_m;\rho)du_1\ldots du_m \quad k_i\geq 0 \tag{1}$$

where μ and Σ denote the mean vector and variance-covariance matrix for the log abundances of the TCR species respectively; $\phi(u_1\ldots u_m;\rho)$ denotes the normal multivariate density with correlations $\rho_{i,j}(i\neq j)$, zero means, and unit variances; and $g_{k_i}(\mu_i,\sigma_i,u_i) = \frac{\exp(u_i\sigma_i k_i+\mu_i k_i+e^{-(u_i\sigma_i+\mu_i)})}{k_i!}$, $k_i\geq 0$ is the re-parameterized Poisson distribution.

Dissimilarity measures for hierarchical clustering

- $\mathcal{Q}$-induced dissimilarity measure:

$$\mathcal{D}(p_\theta^{(1)},p_\theta^{(2)}) = 1-\frac{2\mathcal{Q}(p_\theta^{(1)},p_\theta^{(2)})}{\mathcal{Q}(p_\theta^{(1)},p_\theta^{(1)})+\mathcal{Q}(p_\theta^{(2)},p_\theta^{(2)})} \tag{2}$$

- Morisita-Horn dissimilarity index $\mathcal{D}_{MH}$: Take in (2) $\mathcal{Q}=\mathcal{Q}_{MH}$ where

$$\mathcal{Q}_{MH}(p_\theta^{(1)},p_\theta^{(2)}) = \Sigma_{k,l\geq 1}\, klp_\theta(k,l) \tag{3}$$

- Correlation-based dissimilarity index $\mathcal{D}_\rho$: Take in (2) $\mathcal{Q}=\mathcal{Q}_\rho$ where

$$\mathcal{Q}_\rho(p_\theta^{(1)},p_\theta^{(2)}) = |\Sigma_{k,l\geq 0}\, \tilde{k}\tilde{l}p_\theta(k,l)| \tag{4}$$

- Overlap dissimilarity index $\mathcal{D}_{OV}$: Take in (2) $\mathcal{Q}=\mathcal{Q}_{OV}$ where

$$\mathcal{Q}_{OV}(p_\theta^{(1)},p_\theta^{(2)}) = \frac{\Sigma_{k,l\geq 0}\, p_\theta(k,l)}{2}\times \left(\frac{1}{\Sigma_{k>0}\, p_\theta^{(1)}(k)}+\frac{1}{\Sigma_{l>0}\, p_\theta^{(2)}(l)}\right) \tag{5}$$

- Mutual information dissimilarity index $\mathcal{D}_{MI}$: Take in (2) $\mathcal{Q}=\mathcal{Q}_{MI}$ where

$$\mathcal{Q}_{MI}(p_\theta^{(1)},p_\theta^{(2)}) = \Sigma_{k,l\geq 0}\, p_\theta(k,l)\log\left(\frac{p_\theta(k,l)}{p_\theta^{(1)}(k)p_\theta^{(2)}(l)}\right) \tag{6}$$

In the above p_θ is any bivariate probability distribution with corresponding marginal distributions $p_\theta^{(1)}$ and $p_\theta^{(2)}$, $\tilde{k}=(k-m_1)/s_1$, $\tilde{l}=(l-m_2)/s_2$, and m_i and s_i are, respectively, mean and standard deviation of $p_\theta^{(i)}$, $i=1,2$.

MPLN estimates of richness

$$\hat{M}_3 = D/(1-p_{\hat{\theta}}(0,\ldots,0)), \tag{7}$$

$$\hat{M}_4 = \Sigma_{k_i\geq 0,k_1+\ldots+k_m>0}\, f_{k_1,\ldots,k_m}/p_{\hat{\theta}}(k_1,\ldots,k_m) \tag{8}$$

where $D=\Sigma_{k_i\geq 0,k_1+\ldots+k_m>0}\, f_{k_1,\ldots,k_m}$.

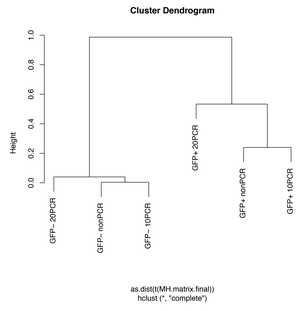
Complete hierarchical Bayesian estimate of richness

$$\hat{M}_{HB} = d_{HB}(k_1,\ldots,k_M) = \Sigma_f\,(\mathbb{1}(k_f>0)+\hat{\beta}_f\mathbb{1}(k_f=0)) \tag{9}$$

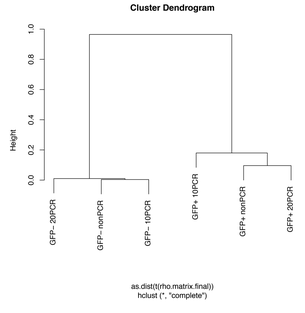
where k_f denotes the number of times that a particular TCR species f is present in a sample of TCR data and $\hat{\beta}_f$ denotes the estimated probability that a particular TCR species f is present in the TCR population.

4. Examples

The dataset used in these two examples were obtained from the lymph nodes of healthy TCR^{mini}Foxp3^{GFP+} and TCR^{mini}Foxp3^{GFP-} mice, where the sequences were obtained by using 0 cycles, 10 cycles, or 20 cycles of amplification followed by using 454 high throughput sequencing. The first picture illustrates the use of hierarchical clustering under the multivariate Poisson-lognormal model according to (3) using agglomerative clustering and complete linkage while the second picture illustrates the use of hierarchical clustering under the multivariate Poisson-lognormal model according to (4) using agglomerative clustering and complete linkage.



Hierarchical clustering dendrogram for the healthy TCR^{mini}Foxp3^{GFP+} and TCR^{mini}Foxp3^{GFP-} mice obtained under the multivariate Poisson-lognormal model according to (3) using agglomerative clustering and complete linkage.



Hierarchical clustering dendrogram for the healthy TCR^{mini}Foxp3^{GFP+} and TCR^{mini}Foxp3^{GFP-} mice obtained under the multivariate Poisson-lognormal model according to (4) using agglomerative clustering and complete linkage.

5. Conclusion

The hierarchical clustering dendrograms obtained in the examples by using the multivariate Poisson-lognormal model and the dissimilarity measures (3) and (4) show excellent consistency and give biologically plausible results. This is also in agreement with the performance of the clustering algorithms based on the bivariate Poisson-lognormal (BPLN) model used by Rempala et al.[2], which should not be surprising as the BPLN model is a special case of the MPLN model considered here.

6. References

[1] J.D. Freeman, R.L. Warren, J.R. Webb, B.H. Nelson, and R. A. Holt, Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing, Genome Res 19(10) (2009) 1817–1824.

[2] G. Rempala, M. Seweryn, and L. Ignatowicz, Model for comparative analysis of antigen receptor repertoires, Journal of Theoretical Biology 269(1) (2011) 1–15.

[3] R. L. Warren, B. H. Nelson, and R. A. Holt, Profiling model T-cell metagenomes with short reads, Sequence Analysis 25(4) (2009) 458-464.

Georgia Health Sciences University