Diversity and overlap analysis in TCR populations

Michał Seweryn and Grzegorz A Rempała
mseweryn@cph.osu.edu, grempala@cph.osu.edu

Problem

Maintaining a proper diversity of T-cell receptor populations (TCR's) is crucial for the immune system's ability to recognize a vast variety of foreign antigens and to avoid autoagression. Due to large diversity of TCR's and the sampling error (of high-throughput sequencing) the standard diversity and overlap measures of the contingency table analysis are insufficient. Applying information theory here we have developed some new ones specifically for TCR data analysis.

Basic Concepts

Each population of TCR's corresponds to a vector of counts $\boldsymbol{c}_i = \left(c_{1,i}, \ldots, c_{m_i,i}\right), i = 1, \ldots, s$. Let $X = (X_1, \ldots, X_{m_i}), \sum_{k=1}^{m_i} X_k = n$ be a sample of size n. We define 'sample coverage' $C := \sum_{l=1}^{m_i} p_l I_l$, where $I_l = 1$ if $X_l > 0$ and $I_l = 0$ otherwise and it's Good and Turing estimator $\hat{C} = \frac{s_n(1)}{n}$, where $s_n(1)$ is the number of singletons.

Let $\mathcal{F}_c$ be the *fingerprint* or *diversity* of the population $\boldsymbol{c}$ – a vector given by $\mathcal{F}_c = (s(1), \ldots, s(\max_i c_i))$, $s(k) = card\{i : c_i = k\}$. A nonnegative, real function of the fingerprint is called a measure (index) of diversity.

Set $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots \boldsymbol{c}_n$ to be populations and let $supp(\boldsymbol{c}_i)$ denote the support of $\boldsymbol{c}_i$. The overlap between vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ is then $\mathcal{O}_n = \bigcap_{k=1}^{n} supp(\boldsymbol{c}_k)$. Any nonnegative, real function G, such that $G\left(\mathcal{O}_n \boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\right)$, is an overlap measure (index).

Let $\boldsymbol{c}$ - population and D - monotone diversity measure, $\mathcal{F}_{I(m)}$ the fingerprint of a uniform population with m different receptors. We define ENS (effective number of species) as the smallest solution of the equation $D(\mathcal{F}_{I(y)}) = D(\mathcal{F}_c)$.

Previous work – diversity and overlap indices

Renyi's (and Shannon's) entropy $H_\alpha(\mathcal{F}_c) = \frac{1}{1-\alpha} \log\left(\sum_k s(k) \left(\frac{k}{n}\right)^\alpha\right), \alpha \geq 0,$ and $H_1(\mathcal{F}_c)$

Simpson's index $ISI := \exp(H_2(\mathcal{F}_c))$, Chao-Shen's index $H_1(\mathcal{F}_c) = -\sum_k s(k) \frac{k}{n(1-(1-k/n)^n)}$

For two populations $\boldsymbol{c}, \boldsymbol{c}'$ define Jaccard's index $J(\boldsymbol{c}, \boldsymbol{c}') = \frac{\sum_i \min(c_i, c_i')}{\sum_i (c_i + c_i') - \sum_i \min(c_i, c_i')}$

For $(\boldsymbol{p}_1, \boldsymbol{p}_2)$ – pair of normalized populations: Morisita-Horn's index and Renyi's divergence

$$MH(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{2\sum_i p_{i,1} p_{i,2}}{\sum_i p_{i,1}^2 + \sum_i p_{i,2}^2}, \quad F_\alpha(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{1}{\alpha - 1} \log\left(\sum_i \frac{p_{i,1}^\alpha}{p_{i,2}^{\alpha-1}}\right), \alpha \geq 0$$

Coverage corrected diversity indices

Now we aim to estimate the diversity of a population of T-cells given sample $\mathcal{X}$ of size n. We define a family of 'sample based' diversity measures which in a natural way overemphisize rare species in the case of undersampling error, we also use a Horvitz-Thompson type correction for bias. Let $\hat{C}$ be the Good-Turing estimator of the sample coverage, we define

$$\hat{H}_{\alpha\hat{C}}(\mathcal{X}) = \frac{\log\left(\sum_k s_n(k) \left(\frac{k}{n}\right)^{\alpha\hat{C}}\right)}{1 - \alpha\hat{C}}, \quad \hat{H}^{(n)}_{\alpha\hat{C}}(\mathcal{X}) = \frac{\log\left(\sum_k \frac{s_n(k) k^{\alpha\hat{C}}}{n^{\alpha\hat{C}}(1-(1-k/n)^n)}\right)}{1 - \alpha\hat{C}}$$

Set $\hat{\boldsymbol{p}}$ to be the MLE of the normalized population vector $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}} = \hat{C}\hat{\boldsymbol{p}}$. Let $0 < \alpha < \infty$ and assume that $H_\alpha(\boldsymbol{p}) < \infty$. If $\alpha < 1$ or if $\alpha > 1$ and $\sum_k p_k \log^r(1/p_k) < \infty$ for some $r > 0$ then

$$H^{(n)}_\alpha(\tilde{\boldsymbol{p}}) \xrightarrow{a.s.} H_\alpha(\boldsymbol{p}) \quad \text{and} \quad H^{(n)}_{\hat{C}\alpha}(\tilde{\boldsymbol{p}}) \xrightarrow{a.s.} H_\alpha(\boldsymbol{p}). \quad \text{If } \alpha = 1, \text{ then } H^{(n)}_1(\tilde{\boldsymbol{p}}) \xrightarrow{a.s.} H_1(\boldsymbol{p}),$$

and on the set $\{\hat{C} < 1 \text{ i.o.}\}$, $H^{(n)}_{\hat{C}}(\tilde{\boldsymbol{p}}) - \frac{\log R^{(n)}_1(\tilde{\boldsymbol{p}})}{1 - \hat{C}} \to H_1(\boldsymbol{p})$, where $R^{(n)}_1(\tilde{\boldsymbol{p}}) := \sum \frac{\tilde{p}_i}{1 - (1 - \tilde{p}_i)^n}$.

New overlap indices

We consider a slightly more general form of the Morisita-Horn index, which allows it to put more weight on rare (resp. abundant) receptors. For $(\boldsymbol{p}_1, \boldsymbol{p}_2)$ a pair of normalized populations and $\alpha, \beta \in (0, \infty)$ the power-geometric (or PG) index of overlap is given by

$$PG_{\alpha,\beta}(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{\sum p_{i1}^\alpha p_{i2}^\beta}{\sum p_{i1}^{2\alpha} + \sum p_{i2}^{2\beta}}.$$

In analogy with the adjustment of diversity indices, and in the notation as above, we may consider $PG^{(n)}_{\hat{C}_1\alpha, \hat{C}_2\beta}(\tilde{\boldsymbol{p}}_1, \tilde{\boldsymbol{p}}_2)$ as the sample-coverage and Horvitz-Thompson adjusted PG index. Assume that $\sum p_{i1}^\alpha < \infty$ and $\sum p_{i2}^\beta < \infty$, as well as $\sum p_{i1} \log^{r_1} \frac{1}{p_{i1}} < \infty$ for some $r_1 > 0$, if $\alpha > 1$ and $\sum p_{i2} \log^{r_2} \frac{1}{p_{i2}} < \infty$ for some $r_2 > 0$, if $\beta > 1$. Then

$$PG^{(n)}_{\hat{C}_1\alpha, \hat{C}_2\beta}(\tilde{\boldsymbol{p}}_1, \tilde{\boldsymbol{p}}_2) \xrightarrow{a.s.} PG_{\alpha,\beta}(\boldsymbol{p}_1, \boldsymbol{p}_2).$$

Moreover, we consider a different approach based on $(m \times n)$ contingency table $\boldsymbol{C} = [c_{ij}]$ with columns representing n different population and rows representing m receptors. Let $\boldsymbol{P} = [p_{ij}] := [\frac{c_{ij}}{\sum_{kl} c_{kl}}]$ be the normalized matrix with columns $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$, $\boldsymbol{p}_{i\circ} = \sum_j p_{ij}$, $\boldsymbol{p}_{\circ j} = \sum_i p_{ij}$ and $\boldsymbol{P}_\circ = (p_{\circ 1}, \ldots, p_{\circ n})$, $\boldsymbol{P}^\circ = (p_{1\circ}, \ldots, p_{m\circ})$, as well as $\boldsymbol{Q} = \boldsymbol{P}_\circ \bigotimes \boldsymbol{P}^\circ := [p_{i\circ}\, p_{\circ j}]$. Define

$$I_\alpha(\boldsymbol{C}) = 1 - F_\alpha(\boldsymbol{P}, \boldsymbol{Q}) / H_{2-\alpha}(\boldsymbol{P}_\circ) \text{ and } Q_\alpha(\boldsymbol{C}) = 1 - I_\alpha(\boldsymbol{C}).$$

Note that for $\alpha \in (0, 2)$ we have $0 \leq Q_\alpha(\boldsymbol{C}) \leq 1$, $Q_\alpha(\boldsymbol{C}) = 0$ iff $\boldsymbol{p}_1 = \boldsymbol{p}_2 = \cdots = \boldsymbol{p}_n$ and if the vectors $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n$ form an orthogonal system, then $Q_\alpha(\boldsymbol{C}) = 1$.

Let $\hat{\boldsymbol{P}}$ be the empirical MLE of $\boldsymbol{P}$, then we also have that $I_\alpha(\hat{\boldsymbol{P}}) \xrightarrow{a.s.} I_\alpha(\boldsymbol{P})$.

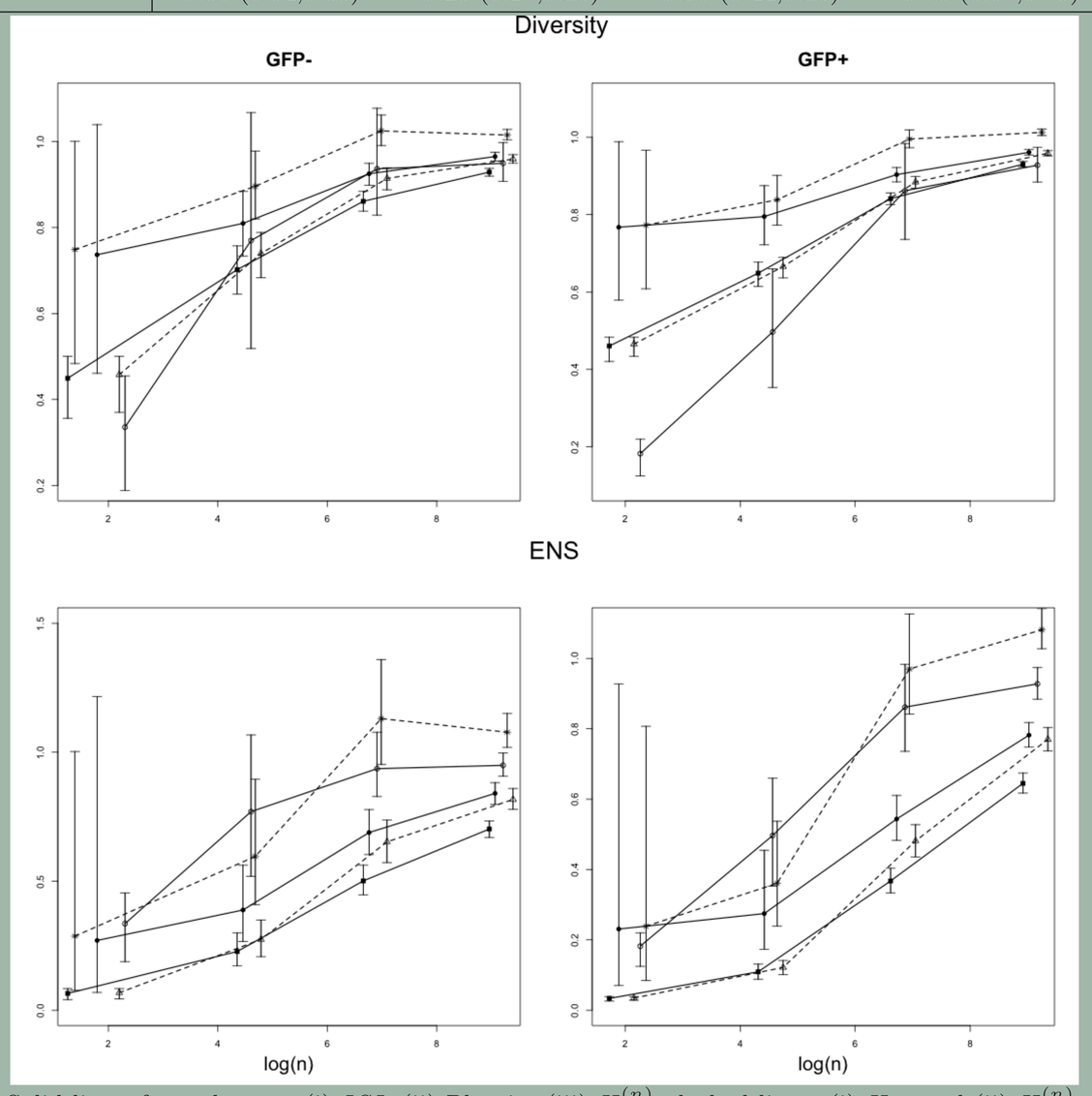
Acknowledgements

This research was partially supported by US NIH grant R01CA-152158 (GAR, MS) and US NSF grant DMS-1106485 (GAR). The supports are gratefully acknowledged.

Performance of new diversity indices

We analyze two TCR datasets obtained from high-throughput sequencing experiments conducted in the molecular immunology lab of Prof Leszek Ignatowicz. One dataset consists of the so called "regulatory" T-cells (GFP^+) the second one of the so-called "naive" T-cells. Diversity and ENS (based on 500 repetitions) is reported relatively to the values in the complete set.

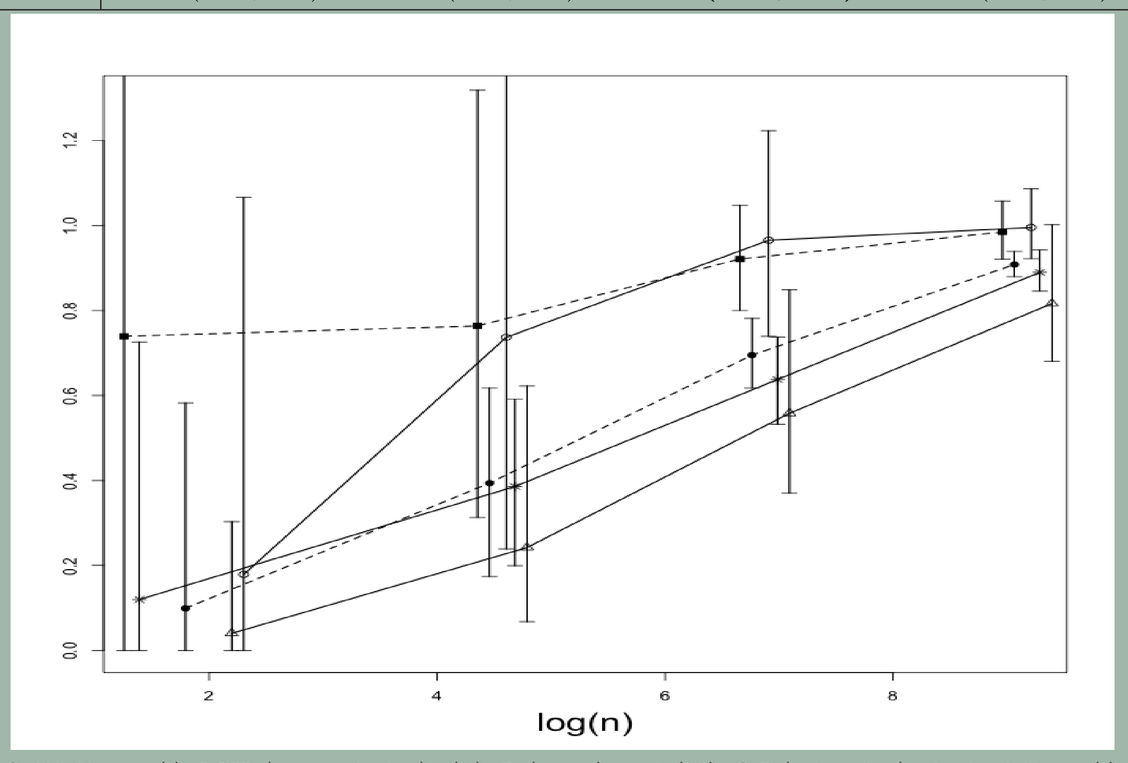
Stat/ENS	$n = 10^2$ $\hat{C} = 0.30$	$n = 10^3$ $\hat{C} = 0.62$	$n = 10^4$ $\hat{C} = 0.83$	$n = 10^5$ $\hat{C} = 0.94$
ISI	0.34 (0.19,0.45) **0.34 (0.19,0.45)**	0.77 (0.52,1.07) **0.77 (0.52,1.07)**	0.94 (0.83,1.08) **0.94 (0.83,1.08)**	0.95 (0.90,0.99) **0.95 (0.90,0.99)**
$H_{\hat{C}}$	0.46 (0.37,0.50) 0.07 (0.04,0.08)	0.74 (0.68,0.78) 0.27 (0.21,0.35)	0.92 (0.89,0.94) 0.65 (0.57,0.74)	0.96 (0.95,0.97) 0.83 (0.78,0.86)
$H^{(n)}_{\hat{C}}$	**0.75 (0.49,1.00)** 0.29 (0.077,1.00)	**0.90 (0.82,0.98)** 0.60 (0.41,0.90)	**1.02 (1.00,1.06)** 1.13 (0.95,1.35)	**1.01 (1.00,1.02)** 1.07 (1.01,1.15)
$H^{(n)}_1$	0.73 (0.46,1.04) 0.27 (0.06,1.22)	0.80 (0.73,0.89) 0.39 (0.26,0.56)	0.92 (0.90,0.95) 0.69 (0.60,0.78)	0.96 (0.95,0.97) 0.84 (0.79,0.88)
Plug-in H_1	0.45 (0.36,0.50) 0.06 (0.04,0.08)	0.70 (0.65,0.76) 0.23 (0.17,0.30)	0.86 (0.84,0.88) 0.50 (0.44,0.56)	0.93 (0.92,0.94) 0.70 (0.67,0.73)



Solid lines, from the top: (i) ISI, (ii) Plug-in, (iii) $H^{(n)}_1$; dashed lines: (i) $H_{\hat{C}}$, and (ii) $H^{(n)}_{\hat{C}}$.

Performance of new overlap indices

Stat	$n = 10^2$ $\hat{C}_1 = 0.25$ $\hat{C}_2 = 0.16$	$n = 10^3$ $\hat{C}_1 = 0.61$ $\hat{C}_2 = 0.40$	$n = 10^5$ $\hat{C}_1 = 0.83$ $\hat{C}_2 = 0.70$	$n = 10^6$ $\hat{C}_1 = 0.94$ $\hat{C}_2 = 0.91$
PG	**0.74 (0.00,4.2)**	**0.76 (0.31,1.31)**	0.92 (0.80,1.04)	**0.99 (0.92,1.05)**
I_1-ind	0.10 (0.00,0.59)	0.40 (0.18,0.62)	0.69 (0.62,0.78)	0.91 (0.88,0.95)
L	0.12 (0.00,0.73)	0.38 (0.20,0.59)	0.64 (0.53,0.74)	0.88 (0.84,0.94)
CJ	0.04 (0.00,0.30)	0.24 (0.06,0.62)	0.56 (0.37,0.85)	0.81 (0.68,1.01)
MH	0.17 (0.00,1.07)	0.74 (0.23,1.43)	**0.96 (0.73,1.22)**	0.99 (0.92,1.09)



Solid lines: (i) MH (open circles), (ii) L (stars) and (iii) CJ (triangles); dashed lines (i) $PG^{(n)}_{\hat{C}_1, \hat{C}_2}$ (squares), and I-index (filled circles).