Model Sampling

According to Misha, the samples have roughly the same number of cells. Since the model involves expansion and contraction in the second condition, some normalization in the second condition is needed such that it produces roughly the same total number of cells as those in the first condition, consistent with the observed data. One approach (the one taken below) is to normalize at the level of clone frequencies.

Direct Sampling

The frequencies of the first condition, f_i, are sampled from $\rho(f)$ until they sum to 1 (i.e. until before they surpass 1, with a final frequency added that takes the sum exactly to 1). An equal number of log-fold changes, s_i, are sampled from $P(s)$. The normalized frequencies of the second condition are then $f'_i = f_i e^{s_i} / \sum_j f_j e^{s_j}$. Counts from the two conditions are then sampled from $P(n|f)$ and $P(n'|f')$, respectively. As a final step, unseen clones, i.e. those with $(n, n') = (0, 0)$, are discarded.

Effective Sampling

For an efficient implementation, the procedure should avoid sampling the numerous clones that produce $(n, n') = (0, 0)$, since these are discarded. Such a procedure follows.

First, the (f, s)-plane is partitioned into two regions, $D = \{(f, s)|f < f_0, fe^s < f_0\}$ and its complement, $\bar{D}$, with f_0 chosen such that clones sampled from $\bar{D}$ are often *seen*, i.e. $n + n' > 0$ (a minority of clones sampled in $\bar{D}$ will nevertheless give $n + n' = 0$; these are discarded). In contrast, frequencies sampled from D will be small, so that most will be unseen, and we must condition on the clone being seen when sampling from this regime. Moreover, their average is unaffected by the long-tailed behaviour of the distribution in the large-frequency regime and thus is well-approximated by the corresponding ensemble average. We use this latter fact when computing the renormalization of the frequencies of the second condition.

We compute the mass in D as $P_D = \int_f \mathrm{d}f \rho(f) \sum_s P(s) \mathbb{1}((f, s) \in D)$.

We sample (f, s) in $\bar{D}$ until the sum of the first condition's frequencies, $\sum_i f_i$ added to the expected sum in D, $P_D N_{cl} \langle f \rangle_{P(f|D)}$, equals 1,

$$1 = \sum_{i=1}^{N_{\bar{D}}} f_i + P_D N_{cl} \langle f \rangle_{P(f|D)} \, , \tag{1}$$

where N_{cl} is the total number of clones in the repertoire. The number sampled from $\bar{D}$, $N_{\bar{D}}$, is determined from that expression self-consistently by substituting $N_{cl} = N_{\bar{D}}/(1 - P_D)$ obtained from $N_{\bar{D}} + P_D N_{cl} \equiv N_{cl}$. The normalization for the second condition's frequencies is then

$$Z = \sum_{i=1}^{N_{\bar{D}}} f_i e^{s_i} + P_D N_{cl} \langle f e^s \rangle_{P(f,s|D)} \tag{2}$$

such that the second condition's frequencies are $f'_i = f_i e^{s_i}/Z$. Molecule counts are then sampled from $P(n|f)$ and $P(n'|f')$.

We then sample from D conditioned on the clone being seen, i.e. having produced a finite number of molecules in either of the two conditions. We thus sample $N_D = P(n + n' > 0|D) P_D N_{cl}$ clones from $P(f, s|D, n + n' > 0)$. To avoid having to sample over the joint distribution of n and n', we condition on the 3 regions of finite counts in both conditions, $(n, 0)$, $(0, n')$, and (n, n'), in which n and n' can be sampled independently. Note the presence of the normalization factor, Z, in

$$P(n + n' > 0|D) = \int_f \mathrm{d}f \rho(f) \sum_s P(s)(1 - P(n = 0|f) P(n' = 0|f' = fe^s/Z)) \, . \tag{3}$$

and

$$P(f, s|D, n + n' > 0) = \frac{\rho(f) P(s)(1 - P(n = 0|f) P(n' = 0|f' = fe^s/Z))}{P(n + n' > 0|D) P_D} \, . \tag{4}$$

We then concatenate the N_D sampled counts from D and the $N_{\bar{D}}$ sampled counts (with $(n, n') = (0, 0)$ realizations discarded) from $\bar{D}$ to obtain the full data set.

Inference

The sampling procedure normalizes $N_{cl}\langle f\rangle$ and $N_{cl}\langle fe^s\rangle$. From sampled dataset, the first condition arises from expressing the average over n and n' and pulling out of the sum the unseen contribution, $n + n' = 0$,

$$\begin{aligned}\langle f\rangle =& P(0,0)\langle f\rangle_{P(f|0,0)} + \sum_{n+n'>0} P(n,n')\langle f\rangle_{P(f,s|n,n')}\\ \langle f\rangle \approx& P(0,0)\langle f\rangle_{P(f|0,0)} + \sum_{(n,n')\in\mathcal{D}} \frac{\#(n,n')}{N_{cl}}\langle f\rangle_{P(f,s|n,n')}\\ \langle f\rangle =& P(0,0)\langle f\rangle_{P(f|0,0)} + \frac{1}{N_{cl}}\sum_{i=1}^{N}\langle f\rangle_{P(f,s|n_i,n_i')} \ ,\end{aligned}$$

where in the second line we have introduced the total number of clones in the repertoire to estimate the probability $P(n, n'))$ from the number of occurences of the pair in the dataset, $\mathcal{D}$. In the last line we resum over clones. The normalization condition is that the above expression should equal $1/N_{cl}$, the constraint thus reduces to

$$1 = N_{cl}\langle f\rangle = P(0,0)N_{cl}\langle f\rangle_{P(f,s|0,0)} + \sum_{i=1}^{N}\langle f\rangle_{P(f,s|n_i,n_i')} \ ,$$

We estimate N_{cl} using the identity, $N_{cl} \equiv N/(1 - P(0, 0))$ (which contains information about the normalization of the second frequency).

Does $\langle f\rangle$ above equal $\int_{f_{min}}^{1} f\rho(f)\mathrm{d}f$ and does it equal $1/N_{cl}$, where $N_{cl} \equiv N/(1 - P(0, 0))$?

The second normalization obeys the same reasoning with the constraint $\langle fe^s\rangle = 1/N_{cl}$,

$$1 = \sum_{i=1}^{N}\langle fe^s\rangle_{P(f,s|n_i,n_i')} + P(0,0)N_{cl}\langle fe^s\rangle_{P(f,s|0,0)} \ .$$

I. FULL REPERTOIRE FORMULATION

For a sample dataset with N_{samp} clones the generative model of the full repertoire is

$$P(\vec{f},\vec{s},n,n') = \prod_{i=1}^{N_{samp}} P(n|f_i)P(n'|f_i e^{s_i}/Z)\rho(f_i)\rho(s_i)\rho(s_{N_{cl}}) \prod_{j=N_{samp}+1}^{N_{cl}-1} \rho(f_j)\rho(s_j) \ , \tag{5}$$

where $\vec{f} = (f_1,\ldots,f_{N_{cl}})$, $\vec{s} = (s_1,\ldots,s_{N_{cl}})$, with the normalization constraint $f_{N_{cl}} = 1-\sum_{i=1}^{N_{cl}-1} f_i$, and $Z = Z(\vec{f},\vec{s}) = \sum_{i=1}^{N_{cl}} f_i e^{s_i}$ and the total number of clones, N_{cl}, must be determined self-consistently from $N_{cl} = N_{samp}/(1-P(0,0))$ where the marginal likelihood is

$$P(n,n') = \int \mathrm{d}\vec{f}_1^{N_{samp}}\mathrm{d}\vec{s}_1^{N_{samp}} \prod_{i=1}^{N_{samp}} P(n|f_i)\rho(f_i)\rho(s_i) \int \mathrm{d}\vec{f}_{N_{samp}+1}^{N_{cl}-1}\mathrm{d}\vec{s}_{N_{samp}+1}^{N_{cl}} P(n'|f_i e^{s_i}/Z)\rho(s_{N_{cl}}) \prod_{j=N_{samp}+1}^{N_{cl}-1} \rho(f_j)\rho(s_j)$$

with notation $\vec{f}_i^{\,j} = (f_i,\ldots,f_j)$ and same for s. $\vec{f}_{N_{samp}+1}^{N_{cl}}$ and $\vec{s}_{N_{samp}+1}^{N_{cl}}$ appear only in Z.

We can use EM to learn this model's parameters: those of all the $\rho(s)$. We average the log-likelihood weighted by the posterior of the hidden variables conditioned on the data, and we have assumed that the parameters learned on same-day data apply to all clones, seen or unseen. Denoting parameters by λ,

$$Q(\lambda|\lambda') = \sum_{(n,n')_{obs}}^{N_{samp}} \int \mathrm{d}\vec{f}_1^{N_{cl}-1}\mathrm{d}\vec{s}_1^{N_{cl}} \rho(\vec{f},\vec{s}|n,n',\lambda') \log\left[P(\vec{f},\vec{s},n,n'|\lambda)\right] \ . \tag{6}$$

Maximizing Q with respect to λ is made easier since λ appears only in all the $\rho(s)$'s which appear as factors in $P(\vec{f},\vec{s},n,n'|\lambda)$ and so all but these factor's vanish. The resulting expressions require integration involving the posterior,

$$\rho(\vec{f},\vec{s}|n,n',\lambda') = \frac{P(\vec{f},\vec{s},n,n'|\lambda')}{P(n,n')} \ . \tag{7}$$

For a simple case of $\rho(s) = \frac{\lambda}{2}e^{-\lambda|s|}$, performing the derivative of Q and setting to zero gives,

$$\sum_{(n,n')_{obs}}^{N_{samp}} \int \mathrm{d}\vec{f}_1^{N_{cl}-1}\mathrm{d}\vec{s}_1^{N_{cl}} \rho(\vec{f},\vec{s}|n,n',\lambda') \sum_{i=1}^{N_{cl}} \frac{1-|s_i|\lambda}{\lambda} = 0 \ . \tag{8}$$

$$\frac{N_{samp}N_{cl}}{\lambda} = \sum_{(n,n')_{obs}}^{N_{samp}} \sum_{i=1}^{N_{cl}} \int \mathrm{d}\vec{f}_1^{N_{cl}-1}\mathrm{d}\vec{s}_1^{N_{cl}} |s_i| \rho(\vec{f},\vec{s}|n,n',\lambda') \ . \tag{9}$$

Writing out each of these vector integrals and identifying the jth integrand,

$$\sum_{(n,n')_{obs}}^{N_{samp}} \frac{1}{P(n,n')} \sum_{j=1}^{N_{cl}-1} \int \mathrm{d}\vec{f}_{1/j}^{N_{cl}-1}\mathrm{d}\vec{s}_{1/j}^{N_{cl}} \rho(s_{N_{cl}}) \left(\prod_{k=1/j}^{N_{cl}-1} \rho(f_k)\rho(s_k)\right) \tag{10}$$

$$\left[\Theta(N_{samp}-j)\left(\prod_{i=1/j}^{N_{samp}} P(n|f_i)\right) \int P(n|f_j)\rho(f_j)\mathrm{d}f_j \int \rho(s_j)|s_j| \left(P(n'|f_j e^{s_j}/Z) \prod_{i=1/j}^{N_{samp}} P(n'|f_i e^{s_i}/Z)\right) \mathrm{d}s_j + \right. \tag{11}$$

$$\left. \Theta(j-N_{samp}+1)\left(\prod_{i=1}^{N_{samp}} P(n|f_i)\right) \int \rho(f_j)\mathrm{d}f_j \int \rho(s_j)|s_j| \left(\prod_{i=1}^{N_{samp}} P(n'|f_i e^{s_i}/Z)\right) \mathrm{d}s_j\right] \tag{12}$$

where $/j$ denotes leaving out the jth index. Note that the integral expression in the first term includes a $P(n|f_j)$ factor, while those in the second term do not. Factorization of these integrals is prevented by Z, which contains all s and f variables. However, the many terms in the sum will decorrelate the resulting value so that the law of large numbers implies the sum should become nearly independent and converge to fixed value. In this case, the product in

the integral in both terms can be brought out of the integral over s_j and f_j.

$$\sum_{(n,n')_{obs}}^{N_{samp}} \frac{1}{P(n,n')} \sum_{j=1}^{N_{cl}-1} \int \mathrm{d}\vec{f}_{1/j}^{N_{cl}-1} \mathrm{d}\vec{s}_{1/j}^{N_{cl}} \rho(s_{N_{cl}}) \left(\prod_{k=1/j}^{N_{cl}-1} \rho(f_k)\rho(s_k) \right) \tag{13}$$

$$\left[\Theta(N_{samp}-j) \left(\prod_{i=1/j}^{N_{samp}} P(n|f_i)P(n'|f_i e^{s_i}/Z) \right) \int P(n|f_j)\rho(f_j)\mathrm{d}f_j \int \rho(s_j)|s_j| P(n'|f_j e^{s_j}/Z)\mathrm{d}s_j + \right. \tag{14}$$

$$\left. \Theta(j-N_{samp}+1) \left(\prod_{i=1}^{N_{samp}} P(n|f_i)P(n'|f_i e^{s_i}/Z) \right) \int \rho(s_j)|s_j|\mathrm{d}s_j \right] \tag{15}$$

In the first term, we see that this leaves an average of $|s|$ over the jth clone marginal, $P(n,n',f_j,s_j)$. In the second term, this leaves the average of $|s_j|$ over the prior, $\rho(s_j)$.

$$\sum_{(n,n')_{obs}}^{N_{samp}} \frac{1}{P(n,n')} \sum_{j=1}^{N_{cl}-1} \int \mathrm{d}\vec{f}_{1/j}^{N_{cl}-1} \mathrm{d}\vec{s}_{1/j}^{N_{cl}} \rho(s_{N_{cl}}) \left(\prod_{k=1/j}^{N_{cl}-1} \rho(f_k)\rho(s_k) \right) \tag{16}$$

$$\left[\Theta(N_{samp}-j) \left(\prod_{i=1/j}^{N_{samp}} P(n|f_i)P(n'|f_i e^{s_i}/Z) \right) \langle |s_j| \rangle_{P(n,n',f_j,s_j)} + \right. \tag{17}$$

$$\left. \Theta(j-N_{samp}+1) \left(\prod_{i=1}^{N_{samp}} P(n|f_i)P(n'|f_i e^{s_i}/Z) \right) \langle |s_j| \rangle_{\rho(s_j)} \right] \tag{18}$$

Note that j appears in the second term only as a label. Distributing the multi-integral into the two terms, we see that $\vec{f}_{N_{samp}+1}^{N_{cl}}$ and $\vec{s}_{N_{samp}+1}^{N_{cl}}$ integrate out so what remains is only integrated over the observed sample.

$$\sum_{(n,n')_{obs}}^{N_{samp}} \frac{1}{P(n,n')} \sum_{j=1}^{N_{cl}-1} [\tag{19}$$

$$\Theta(N_{samp}-j) \int \mathrm{d}\vec{f}_1^{N_{samp}} \mathrm{d}\vec{s}_1^{N_{samp}} \left(\prod_{i=1/j}^{N_{samp}} \rho(f_i)\rho(s_i) P(n|f_i)P(n'|f_i e^{s_i}/Z) \right) \langle |s_j| \rangle_{P(n,n',f_j,s_j)} + \tag{20}$$

$$\left. \Theta(j-N_{samp}+1) \int \mathrm{d}\vec{f}_1^{N_{samp}} \mathrm{d}\vec{s}_1^{N_{samp}} \left(\prod_{i=1}^{N_{samp}} \rho(f_i)\rho(s_i) P(n|f_i)P(n'|f_i e^{s_i}/Z) \right) \langle |s_j| \rangle_{\rho(s_j)} \right] \tag{21}$$

Now, the integral in the first term is similar to the marginal, $P(n,n')$, but lacks the jth clone so we denote it $P_{/j}(n,n')$. The integral in the second term is the marginal (in this approximation where Z is fixed) and so cancels.

$$\sum_{(n,n')_{obs}}^{N_{samp}} \sum_{j=1}^{N_{cl}-1} \left[\Theta(N_{samp}-j) \frac{P_{/j}(n,n')}{P(n,n')} \langle |s_j| \rangle_{P(n,n',f_j,s_j)} + \Theta(j-N_{samp}+1) \langle |s_j| \rangle_{\rho(s_j)} \right] \tag{22}$$

We now can distribute the sum over j. Since the priors are all the same, we remove the j-dependence on the second term and the sum distributes over the second term simply as a factor equal to the number of clones. Switching the order for clarity,

$$\sum_{(n,n')_{obs}}^{N_{samp}} \left[(N_{cl}-N_{samp}) \langle |s| \rangle_{\rho(s)} + \sum_{j=1}^{N_{samp}} \frac{P_{/j}(n,n')}{P(n,n')} \langle |s_j| \rangle_{P(n,n',f_j,s_j)} \right] \tag{23}$$

Finally, we note that the fraction $\frac{P_{/j}(n,n')}{P(n,n')}$ leaves the jth marginal, $P_j(n,n')$ in the denominator, which combines with $P(n,n',f_j,s_j)$ to make the average over the jth posterior, $P(s_j,f_j|n,n')$,

$$\sum_{(n,n')_{obs}}^{N_{samp}} \left[(N_{cl}-N_{samp}) \langle |s| \rangle_{\rho(s)} + \sum_{j=1}^{N_{samp}} \langle |s_j| \rangle_{P(f_j,s_j|n,n')} \right] . \tag{24}$$

The same argument regarding the independence of j now applies to the second (previously first) term,

$$\sum_{(n,n')_{obs}}^{N_{samp}} \left[(N_{cl} - N_{samp})\langle|s|\rangle_{\rho(s)} + N_{samp}\langle|s|\rangle_{P(f,s|n,n')}\right] . \tag{25}$$

Putting this back into our EM solution equation,

$$\frac{N_{samp}N_{cl}}{\lambda} = \sum_{(n,n')_{obs}}^{N_{samp}} \left[(N_{cl} - N_{samp})\langle|s|\rangle_{\rho(s)} + N_{samp}\langle|s|\rangle_{P(f,s|n,n')}\right] . \tag{26}$$

$$\frac{1}{\lambda} = \frac{1}{N_{samp}} \sum_{(n,n')_{obs}}^{N_{samp}} \left[\left(1 - \frac{N_{samp}}{N_{cl}}\right)\langle|s|\rangle_{\rho(s)} + \frac{N_{samp}}{N_{cl}}\langle|s|\rangle_{P(f,s|n,n')}\right] \tag{27}$$

$$\frac{1}{\lambda} = \frac{1}{N_{samp}} \sum_{(n,n')_{obs}}^{N_{samp}} \left[P(n+n'=0)\langle|s|\rangle_{\rho(s)} + P(n+n'>0)\langle|s|\rangle_{P(f,s|n,n')}\right] , \tag{28}$$

where we have used the definition of $N_{cl} = N_{samp}/(1 - P(n+n'=0))$. The hidden and visible part of the repertoire contribute a weighted prior and posterior average respectively. The stable point of iterating this procedure occurs when the prior parameter, $\bar{s}$, equals the result, so that prior average equals posterior average.

$$\frac{1}{\lambda} = \frac{1}{N_{samp}} \sum_{(n,n')_{obs}}^{N_{samp}} \left[P(n+n'=0)\frac{1}{\lambda} + (1 - P(n+n'=0))\langle|s|\rangle_{P(f,s|n,n')}\right] \tag{29}$$

$$\frac{1}{\lambda} = \frac{1}{N_{samp}} \sum_{(n,n')_{obs}}^{N_{samp}} \langle|s|\rangle_{P(f,s|n,n')} . \tag{30}$$