Scope: Apply mixture model a better mixture model methodology to strain data.

Current, Application-focused, paper: [**SplitStrains**] Last-Century, method-focused, paper (by our very own MTW!): [**CasRobWel2004**]

1 Recap of Split Strains

1.1 Data

[**SplitStrains**] uses 5 datasets to test their method(s). All of the datasets appear to be artificial or generated in some way.

1.2 Method

Input data is aligned using the BWA-MEM tool. I still have to figure out exactly what this is, what it does, (and why???). They define a feature vector, $x^i := (p^i_A, p^i_C, p^i_G, p^i_T, d^i)$, where:

- $p^i_l \in [0, 100]$ is "the percentage of base [l] at position i",
- d^i is "the number of aligned reads at position i",
- and x^i satisfies $\sum_{l\in L} p^{(i)}_l = 100$

And more generally,

- d_i is "the number of reads that map to position i",
- $k_i := d_i \max_{l\in L} p^i_l$ i.e. the maximum value in x^i

$P(D|H_0) = \prod_i ($

2 More General Thoughts

Let: B the set of letters and, say, $k := |S|$. $\bar{S} \in S^{\times|S|} = S^{\times k}$ is some defined arrangement of $\bar{S}$ such that $i \neq j \implies L_i \neq L_j$. (This may also be understood as a bijection between S and $1 : k$) $e_\bullet : S \to \mathbb{K}^k$ given by $l \mapsto \sum_{i=1}^k 1_{l=\bar{L}_i} e_i = e_{\{j|L_j=l\}}$ Let's call $V_L = sp\{e_l \mid l \in L\}$. Now, let $e_\bullet : L^{\times r} \to V_L^{\otimes r}$be given by $w \mapsto e_{w^1} \otimes \cdots \otimes e_{w^r}$ $e_\bullet : \bigcup_{r\in\mathbb{N}} L^{\times r} \to \bigoplus_{r\in\mathbb{N}} V_L^{\otimes r}$ be given by $w \mapsto e_{w^1} \otimes \cdots \otimes e_{w^\#}$ $e_\bullet : P(\bigcup_{r\in\mathbb{N}} L^{\times r}) \to \bigoplus_{r\in\mathbb{N}} V_L^{\otimes r}$ be given by $W \mapsto \sum_{w\in W} \frac{1}{|W|} e_w$.

$\mathcal{X} = (x_i \subset \bigcup_{r\in\mathbb{N}} B^{\times r})_{i=1}^n$ is the sequence of observations. $\mathbb{F}_n = e_{\mathcal{X}}$ $\mathcal{Y} = [y_i \subset \bigcup_{r\in\mathbb{N}} B^{\times r}]_{i=1}^n$ is an array of observations. $\mathcal{Y} = G\mathcal{X}$

It is routine to verify that for every $\Omega(L) =$

$e_\bullet : \bigcup_{r\in\mathbb{N}} L^{\times r} \to$ given by $w \mapsto e_{w^1} \otimes \cdots \otimes e_{w^\#}$ $\mathbb{F}_m := \frac{1}{m} \sum_{i=1}^m \frac{1}{|x_i|} e_{x_i}$