image from https://stats.stackexchange.com/questions/439716/questions-about-pmf-of-multinomial-distribution-with-indicator-variables

Multinomial

Parameters	$n > 0$ number of trials (integer) $p_1, \ldots, p_k$ event probabilities ($\Sigma p_i = 1$)
Support	$x_i \in \{0, \ldots, n\}, \quad i \in \{1, \ldots, k\}$ $\Sigma x_i = n$
pmf	$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$

Maximum Likelihood estimate of p_i is x_i / n

Seq 1: ELVESLIVER
Seq 2: ELVESLIVED
Seq 3: ELMESLIMED
Seq 4: FLEESLIVES
Seq 5: FLEASLIVES

ELVISLIVES ? Consider isoleucine at Position #4...

Dirichlet priors and variation of preferred residues among sites

In general, amino acid residues in proteins will be a particular one of 20 possible types.

Before we see any sequence data for a specific group of proteins that we want to study, we have little information regarding what residues will be found at a particular site...

... (exception: we can reasonably guess that the first site in a protein will be occupied by methionine) ...

... but we do know that some amino acid types tend to be more common in general than others.

How can we take into account the fact that certain residues are favored at certain sites and also that we have **a priori** knowledge about certain residues in proteins tending to be more common than others?

One relatively successful approach is to model the **a priori** distribution of residue probabilities at a site with a Dirichlet distribution.

Assume we study a particular site (alignment column) in a protein group of interest.

Let π_i be the probability of observing amino acid type i at this site when we randomly choose a sequence (*notice the new notation*).

We know $\pi_1 + \pi_2 + \ldots + \pi_{20} = 1$ but we do not know what the π_i values actually are.

A Dirichlet distribution can summarize our *a priori* knowledge.

The probability density $f()$ of a Dirichlet distribution is denoted Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_{20})$ if:

$$f(\pi_1, \ldots, \pi_{20} | \alpha_1, \ldots, \alpha_{20}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{20})} \pi_1^{\alpha_1 - 1} \cdots \pi_{20}^{\alpha_{20} - 1}$$

where

$$\alpha_0 = \sum_{i=1}^{20} \alpha_i.$$

When α_i is a positive integer, $\Gamma(\alpha_i) = (\alpha_i - 1)!$.

Facts regarding the Dirichlet:

$$E(\pi_i) = \frac{\alpha_i}{\alpha_0}$$

$$Var(\pi_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

For $i \neq j$,

$$cov(\pi_i, \pi_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Notice that the covariances can never be positive.

Before we see any data, our best guess about π_i is $E(\pi_i) = \alpha_i/\alpha_0$.

The α_i are sometimes referred to as "pseudocounts" because they represent our prior knowledge as if we had already seen counts of α_i for each residue type i.

What is the best guess about π_i after we make n_0 independent observations of residues at a particular site and observe that n_i of these are amino acid type i?

The probability density for $\pi_1, \ldots, \pi_{20}$ after we observed some data is called the *a posteriori* distribution.

In this case, algebra shows that the *a posteriori* distribution for $\pi_1, \ldots, \pi_{20}$ would be Dirichlet$(\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_{20} + n_{20})$.

In other words, our a posteriori expectation of π_i is $(\alpha_i + n_i)/(\alpha_0 + n_0)$.

A refinement of HMMs

Proteins have different kinds of sites. For example, some sites are on the surface and exposed to H_2O and other sites are on the interior and not exposed to H_2O.

Sites on surface tend to be occupied by hydrophilic amino acids . Sites in interior tend to be occupied by hydrophobic ones.

To take the variety of types of sites into account, a mixture of Dirichlet prior distributions can be used.

Each component of the mixture represents a specific "type" of site (e.g. hydrophilic) and has an associated frequency within the mixture. Different mixture components have different preferred amino acid residues.

The different components in a mixture may be associate with different structural environments, for example.