JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 8, 2011

Pp. 925–939
DOI: 10.1089/cmb.2011.0039

Research Articles

# The Complexity of the Dirichlet Model for Multiple Alignment Data

YI-KUO YU and STEPHEN F. ALTSCHUL

## ABSTRACT

**A model is a set of possible theories for describing a set of data. When the data are used to select a maximum-likelihood theory, an important question is how many effectively independent theories the model contains; the log of this number is called the model's complexity. The Dirichlet model is the set of all Dirichlet distributions, which are probability densities over the space of multinomials. A Dirichlet distribution may be used to describe multiple-alignment data, consisting of *n* columns of letters, with *c* letters in each column. We here derive, in the limit of large *n* and *c*, a closed-form expression for the complexity of the Dirichlet model applied to such data. For small *c*, we derive as well a minor correction to this formula, which is easily calculated by Monte Carlo simulation. Although our results are confined to the Dirichlet model, they may cast light as well on the complexity of Dirichlet mixture models, which have been applied fruitfully to the study of protein multiple sequence alignments.**

**Key words:** alignment, computational molecular biology, dynamic programming, multiple alignment, sequence analysis.

## 1. INTRODUCTION

When attempting to describe a set of data, one frequently may choose among many alternative theories. In general, the more free parameters a theory has available, the better it will be able to describe the data. However, a theory with too many free parameters will tend to "overfit" the data—modeling its noise rather that its regularities—and this can lead to poor predictions on new data. One approach to avoiding this problem is the Minimum Description Length (MDL) principle (Grünwald, 2007). In short, consider a theory to be a probability distribution over the space of all possible data and a model to be a parametrized set of theories. The MDL principle then implies that one should seek to minimize the description length of the data given a model, plus the description length of the model. Formally, the description length of the data given a model is the negative log probability of the data as implied by the maximum-likelihood theory from the model. The description length or complexity of a model is the log of the number of effectively independent theories it contains. A central element of MDL theory is the formal definition of model complexity, and its calculation for specific models. This article studies the complexity of the Dirichlet model applied to multiple alignment data.

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

Multiple alignments of protein or DNA data can be formally rather complex, for example, allowing sequences to contain gap characters, or imposing additional structure, such as sequence weights, or a phylogenetic tree relating the sequences. In this article, we consider only a simplified version of multiple alignment data, consisting of $n$ columns, with each column containing $c$ letters chosen from an alphabet of size $L$; no "gap" characters are allowed. A simple, default way to model such data is by means of a multinomial distribution, which assigns each letter type a probability. However, when the sequences are related and properly aligned, the various letters tend to appear with differing frequencies from one column to another, and the multinomial model can take no account of this.

An elegant approach to describing multiple alignment data for protein sequences is the Dirichlet mixture model (Brown et al., 1993; Sjölander et al., 1996). This model postulates that the data in a given column can be described by a multinomial distribution, but that this multinomial may vary from column to column. It postulates further that the multinomials themselves are drawn randomly and independently, following a Dirichlet mixture distribution, which is a simple linear combination of several Dirichlet distributions. We have not been able to analyze the complexity of Dirichlet mixture models, and so here study the simpler case of single Dirichlet models, which often are rich enough to describe well multiple DNA sequence alignment data. Elsewhere (Ye et al., 2011), we generalize our results by heuristic argument to the Dirichlet mixture models appropriate for protein sequences.

Below, we first review some basics of MDL theory and of Dirichlet models. Second, we derive a closed-form analytic formula for the complexity of the Dirichlet model applied to multiple alignment data in the limit of large $n$ and $c$. Third, we derive an expression for the complexity of the Dirichlet model for arbitrary $c$ in the form of a definite $L$-dimensional integral. Finally, we describe how this integral may be evaluated accurately and efficiently using Monte Carlo simulation, even when $L$ is large.

## 2. MATHEMATICAL BASICS

### 2.1. The minimum description length principle

The minimum description length principle (Grünwald, 2007) has a substantial body of theory, but we summarize here only those elements we will need. A specific theory is taken to be equivalent to a probability distribution over the space of all possible sets of data. We will assume that alternative theories may be parametrized by $\theta$, which lies within the $k$-dimensional space $\Theta$, and a model $\mathfrak{M}$ is the set of all theories in $\Theta$. The complexity of $\mathfrak{M}$, represented by COMP($\mathfrak{M}$), can be thought of as the log of the effective number of independent theories $\mathfrak{M}$ contains, a notion that can be formalized as described in Grünwald (2007). COMP($\mathfrak{M}$) depends both on $\mathfrak{M}$ and on the quantity of data it is used to describe, and may be obtained by integrating over $\Theta$ a measure of the density of independent theories. In brief, if the data consist of $n$ independent observations, then given certain reasonable assumptions, it can be shown that for large $n$

$$\mathrm{COMP}(\mathfrak{M}, n) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\Theta}\sqrt{|\mathbf{J}_\theta|}d\theta + o(1), \tag{1}$$

where $|\mathbf{J}_\theta|$ is the determinant of the Fisher information matrix for $\mathfrak{M}$ (Grünwald, 2007). Intuitively, the greater the value of this determinant, and the greater $n$, the greater the density of independent theories. In this article, we derive $|\mathbf{J}_\theta|$ for the Dirichlet model, and seek to evaluate the integral in eq. (1). It is traditional to use logs to the base 2 in eq. (1), and therefore to express model complexity in bits.

### 2.2. The Dirichlet model

A finite alphabet $\mathcal{A}$ consists of $L$ letters, which can be conveniently identified with the first $L$ natural numbers. A multinomial distribution over $\mathcal{A}$ is defined by a vector $\vec{p}$ of $L$ positive probabilities that sum to 1. Because of this linear constraint, the space $\Omega_L$ of all multinomial distributions over $\mathcal{A}$ is $(L-1)$-dimensional. A Dirichlet distribution is a probability density over $\Omega_L$. It is parametrized by an $L$-dimensional vector $\vec{\alpha}$ of positive real numbers, and has probability density defined as

$$f(\vec{x}) \equiv Z\prod_{j=1}^{L} x_j^{\alpha_j - 1}, \tag{2}$$

where $Z \equiv \Gamma(\sum_j \alpha_j)/\prod_{j=1}^{L} \Gamma(\alpha_j)$ is a scalar chosen so that integrating $f(\vec{x})$ over $\Omega_L$ yields 1. It is convenient to define $\alpha \equiv \sum_j \alpha_j$, and to define $q_j \equiv \alpha_j/\alpha$. Then $(\vec{q}, \alpha)$ is an alternative parametrization of the Dirichlet distribution, which still has only $L$ free parameters because the $q_j$ must sum to 1. The set of all Dirichlet distributions over $\Omega_L$ forms the Dirichlet model $\mathcal{D}_L$.

A multiple alignment consists of $n$ columns of letters from $\mathcal{A}$, with each column containing $c$ letters. Then, to say that the alignment is described by a particular Dirichlet distribution is shorthand for saying that the Dirichlet distribution generates a particular multinomial $\vec{p}$ for each column, and the letters in that column are then generated by $\vec{p}$.

## 3. THE COMPLEXITY OF THE DIRICHLET MODEL FOR ALIGNMENTS WITH LARGE *n* AND *c*

### *3.1. The Fisher information matrix*

Our goal is to calculate the integral in eq. (1), where $\mathbf{J}_\theta$ is the Fisher information matrix for the Dirichlet model. When $L$ is large, evaluating the determinant of this $L \times L$ matrix is a potentially challenging problem, both analytically and computationally. However, as we will see, the matrix has the special form

$$\mathrm{M_{j,j'}} = \delta_{\mathrm{j,j'}}\ \mathrm{D_j} - \mathrm{D}. \tag{3}$$

In Appendix A, we prove the useful lemma that the determinant of a matrix of this form can be written as the product

$$\det \mathbf{M} = \mathrm{F} \prod_{\mathrm{j}=1}^{\mathrm{L}} \mathrm{D_j}, \tag{4}$$

where

$$F \equiv 1 - D \sum_{j=1}^{L} D_j^{-1}. \tag{5}$$

This will greatly simplify the computation of the integral in question.

To derive an expression for $\mathcal{D}_L$'s Fisher information matrix, first consider a particular column of observations $\vec{y}$, in which the counts for the various letters are given by $\vec{c}$, so that $\sum_{j=1}^{L} c_j = c$. As described in Sjölander et al. (1996) and Altschul et al. (2010), given the Dirichlet distribution with parameters $\vec{\alpha}$, the probability of observing the data in this column is given by

$$P(\vec{y}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + c)} \prod_{j=1}^{L} \frac{\Gamma(\alpha_j + c_j)}{\Gamma(\alpha_j)}. \tag{6}$$

Note that $P(\vec{y})$ represents the probability for the specific column $\vec{y}$, not for all columns that yield letter count vector $\vec{c}$.

Taking logs of both sides of eq. (6), we have

$$\ln P(\vec{y}) = \lg(\alpha) - \lg(\alpha + c) + \sum_{j=1}^{L} \big[\lg(\alpha_j + c_j) - \lg(\alpha_j)\big], \tag{7}$$

where $\lg(x)$ represents $\ln \Gamma(x)$.

The $(j, j')$th entry of the Fisher information matrix is given by

$$\mathbf{J}_{j,j'}(\vec{\alpha}) = E\left[ - \frac{\partial^2 \ln P(\vec{y})}{\partial \alpha_j \partial \alpha_{j'}} \right] = \sum_{\vec{y}} P(\vec{y}) \frac{\partial^2 [-\ln P(\vec{y})]}{\partial \alpha_j \partial \alpha_{j'}}. \tag{8}$$

An alternative form for eq. (6), useful for calculating expectation values, is described in Appendix B.

From eq. (7), we may express the quantity averaged in eq. (8) as

$$\frac{\partial^2[-\ln P(\vec{y})]}{\partial\alpha_j\partial\alpha_{j'}} = \delta_{j,j'}\left[\psi'(\alpha_j) - \psi'(\alpha_j + c_j)\right] - [\psi'(\alpha) - \psi'(\alpha + c)], \tag{9}$$

where the trigamma function $\psi'(x)$ is the first derivative of the digamma function $\psi(x) \equiv d\lg(x)/dx$. Since $c$ is held fixed in our analysis, the only term in eq. (9) that has a nontrivial average is $\delta_{j,j'}\psi'(\alpha_j + c_j)$. Therefore

$$\begin{aligned}\mathbf{J}_{j,j'}(\vec{\alpha}) &= \delta_{j,j'}\left\{\psi'(\alpha_j) - E[\psi'(\alpha_j + c_j)]\right\} - [\psi'(\alpha) - \psi'(\alpha + c)] \\ &= \delta_{j,j'}D_j - D,\end{aligned} \tag{10}$$

where

$$D_j = \psi'(\alpha_j) - E[\psi'(\alpha_j + c_j)]; \tag{11}$$

$$D = \psi'(\alpha) - \psi'(\alpha + c). \tag{12}$$

Our Fisher information matrix therefore takes the form of eq. (3), so we can calculate $|\mathbf{J}(\vec{\alpha})|$ using eqs. (4) and (5).

### 3.2. The asymptotic behavior of $|\boldsymbol{J}(\alpha\vec{\mathbf{q}})|$

We will consider later the exact calculation of $D_j$ and $D$ of eqs. (11) and (12). In the case that $\alpha_j$ and $\alpha$ are large, however, we can derive closed-form asymptotic expressions for these quantities, and for the $F$ that they imply.

To start, we note that for large $x$, $\psi'(x) = 1/x + 1/(2x^2) + \mathcal{O}(x^{-3})$. Therefore, eq. (12) implies that

$$D = \frac{1}{\alpha} - \frac{1}{\alpha + c} + \mathcal{O}(\alpha^{-3}) = \frac{c}{\alpha(\alpha + c)} + \mathcal{O}(\alpha^{-3}). \tag{13}$$

Furthermore, as we show in Appendix C, for large $\alpha_j$, $E[\psi'(\alpha_j + c_j)] = \psi'(\alpha_j + E[c_j]) + \mathcal{O}(\alpha_j^{-3})$, and $E[c_j] = c\alpha_j/\alpha$. Writing $\alpha_j$ as $\alpha q_j$, these imply that

$$D_j = \frac{1}{\alpha q_j} - \frac{1}{\alpha q_j + c q_j} + \mathcal{O}(\alpha^{-3}) = \frac{c}{q_j\alpha(\alpha + c)} + \mathcal{O}(\alpha^{-3}). \tag{14}$$

Unfortunately, because $D$ and $D_j$ appear only as a ratio in eq. (5), using eqs. (13) and (14) in an attempt to approximate $F$ yields $1 - D\sum_{j=1}^{L} {D_j}^{-1} \approx 1 - \sum_{j=1}^{L} q_j = 0$. If one wishes to express $F$ accurately to the order $\alpha^{-2}$, one needs to obtain $D_j$ and $D$ to the order $\alpha^{-4}$. We show how to do this in Appendix C, and that the resulting asymptotic formula for $F$ takes the form

$$F = \frac{(L-1)(c-1)}{2\alpha(\alpha + c)} + \mathcal{O}(\alpha^{-3}). \tag{15}$$

Now, using eq. (4) and the asymptotic eqs. (14) and (15) for the $D_j$ and $F$, we are able to calculate the determinant we require in the limit of large $\alpha$:

$$|\mathbf{J}(\alpha\vec{q})| \approx \frac{(L-1)(c-1)c^L}{2\alpha^{L+1}(\alpha + c)^{L+1}\prod_{j=1}^{L} q_j}. \tag{16}$$

It will be useful to derive an expression for $|\mathbf{J}(\alpha\vec{q})|$ in the small $\alpha$ limit as well. In short, analysis of the exact expressions (26) and (27) derived below shows that as $\alpha \to 0$, $D = 1/\alpha^2 + \mathcal{O}(1)$ and $D_j = 1/(\alpha^2 q_j) + \left(\sum_{k=1}^{c-1} 1/k\right)/(\alpha q_j) + \mathcal{O}(1)$, implying $F = \alpha\sum_{k=1}^{c-1} 1/k + \mathcal{O}(\alpha^2)$. This implies that for small $\alpha$,

$$|\mathbf{J}(\alpha\vec{q})| \approx \frac{\sum_{k=1}^{c-1} 1/k}{\alpha^{2L-1}\prod_{j=1}^{L} q_j}. \tag{17}$$

### 3.3. The complexity of $\mathcal{D}_{\mathrm{L}}$ for large n and c

To derive a formula for COMP $(\mathcal{D}_L, n, c)$ in the limit of large $n$ and $c$, we first consider the integral of eq. (1). Note that this integral involves all the parameters $\alpha_j$ arbitrarily close to 0. However, as we argue below,

as $c$ grows the relative contribution to the integral from small $\alpha_j$ vanishes, so in the limit of large $c$ it is valid to use the asymptotic eq. (16) throughout the range of integration. In this limit, we can for simplicity replace $c-1$ by $c$ in the numerator of eq. (16). This allows us to write

$$\begin{aligned} I &= \int_0^\infty \cdots \int_0^\infty \sqrt{|\mathbf{J}(\vec{\alpha})|}\, d\vec{\alpha} \\ &= \int_{\Omega_L} d\vec{q} \int_0^\infty \sqrt{|\mathbf{J}(\alpha\vec{q})|}\ \alpha^{L-1} d\alpha \\ &\approx \int_{\Omega_L} d\vec{q} \int_0^\infty \sqrt{\frac{L-1}{2} \frac{c^{L+1}}{\alpha^{L+1}(\alpha+c)^{L+1}\prod_j q_j}}\ \alpha^{L-1} d\alpha \\ &= \sqrt{\frac{L-1}{2}}\, c^{\frac{L+1}{2}} \left[\int_{\Omega_L} \prod_j q_j^{-1/2} d\vec{q}\right] \left[\int_0^\infty \sqrt{\frac{\alpha^{L-3}}{(\alpha+c)^{L+1}}} d\alpha\right] \\ &\equiv \sqrt{\frac{L-1}{2}} c^{\frac{L+1}{2}} [I_1][I_2], \end{aligned} \tag{18}$$

where $I_1$ represents the integral over $\vec{q}$ and $I_2$ the integral over $\alpha$. As described in the supplementary material to Altschul et al. (2009),

$$I_1 = \frac{\pi^{L/2}}{\Gamma(L/2)}, \tag{19}$$

so we turn our attention to $I_2$. Letting $\beta \equiv \alpha/c$ and then $\tan^2\phi \equiv \beta$, we can write this integral as

$$\begin{aligned} I_2 &= \frac{1}{c}\int_0^\infty \sqrt{\frac{\beta^{L-3}}{(\beta+1)^{L+1}}} d\beta = \frac{1}{c}\int_0^{\frac{\pi}{2}} \frac{\tan^{L-3}\phi}{\sec^{L+1}\phi} 2\tan\phi\ \sec^2\phi\ d\phi \\ &= \frac{2}{c}\int_0^{\frac{\pi}{2}} \sin^{L-2}\phi \cos\ \phi\ d\phi = \frac{2}{c(L-1)}. \end{aligned} \tag{20}$$

Combining eqs. (18), (19), and (20) yields

$$I \approx \sqrt{\frac{2}{L-1}}\, c^{\frac{L-1}{2}} \frac{\pi^{L/2}}{\Gamma(L/2)}. \tag{21}$$

Finally, substituting eq. (21) for the integral in eq. (1) yields

$$\mathrm{COMP}(\mathcal{D}_L, n, c) = \frac{L}{2}\log n + \frac{L-1}{2}\log c + A_L + o(1), \tag{22}$$

where $A_L$ is an alphabet-size dependent constant given by

$$A_L = -\log\Gamma(L/2) - \frac{1}{2}\log(L-1) - \frac{L-1}{2}. \tag{23}$$

Specifically, for protein sequences $A_{20} = -30.093$ bits, and for DNA sequences $A_4 = -2.292$ bits.

Eq. (22) is valid only for large $n$ and $c$. However, as we show in the next section, it requires only a minor adjustment for small $c$. Provocatively, when $L \gg 1$, one can use Stirling's approximation to rewrite eqs. (22) and (23) as

$$\mathrm{COMP}(\mathcal{D}_L, n, c) \approx \frac{1}{2}\log\left[\frac{(enc/L)^L}{2\pi c}\right]. \tag{24}$$

Note here that $nc/L$ is the number of observations per model parameter.

We omit a formal proof that it is valid to calculate the integral as we have done, in the limit of large $c$, by using the asymptotic formula (16) for $|\mathbf{J}(\alpha\vec{q})|$ over the complete domain of integration. However, in outline, the required reasoning proceeds as follows. First, after transforming into the $(\vec{q}, \alpha)$ coordinate system, the

integral is split into large-$\alpha$ and small-$\alpha$ domains. Using eq. (17), an asymptotic analysis of the integral over the small-$\alpha$ domain shows that it grows as $\sqrt{\log c}$, whereas the integral over the large-$\alpha$ domain grows as $c^{(L-1)/2}$, as described above. Therefore, in the limit of large $c$, the small-$\alpha$ domain may be ignored when one takes the log of the sum of these two integrals. Furthermore, because $\alpha = c\tan^2\phi$ after the substitutions above, the lower limit of the domain of the large-$\alpha$ integral, expressed as $\phi$, shrinks to zero as $c$ grows. Similar arguments may be applied to $\vec{q}$ near the boundaries of $\Omega_L$.

## 4. THE COMPLEXITY OF THE DIRICHLET MODEL FOR ARBITRARY *c*

### *4.1. Exact formulas for* $D_j$ *and* D

The derivation of eq. (22) assumes $n$ and $c$ are large. Whereas in practice $n \gg 100$ for typical multiple alignment data sets to which eq. (22) might be applied, $c$ can be quite small (Ye et al., 2011). Therefore, we here take the alternative approach of deriving a calculable expression for the Dirichlet model's $|\mathbf{J}_\theta|$, and of estimating the definite integral of eq. (1) by Monte Carlo simulation. Given eqs. (4), (5), and (10), we need only derive an expression for $D_j = E[\psi'(\alpha_j) - \psi'(\alpha_j + c_j)]$.

By definition of the trigamma function, and using eq. (38) from Appendix B,

$$\begin{aligned} D_j &= E\left[\sum_{k=0}^{c_j-1}\frac{1}{(\alpha_j+k)^2}\right] \\ &= \frac{\Gamma(\alpha)/\Gamma(\alpha_j)}{\Gamma(\alpha-\alpha_j)}\sum_{c_j=0}^{c}\frac{c!}{c_j!(c-c_j)!}\times\int_0^1 x^{c_j+\alpha_j-1}\left[\sum_{k=0}^{c_j-1}\frac{1}{(\alpha_j+k)^2}\right](1-x)^{c-c_j+\alpha-\alpha_j-1}dx. \end{aligned} \tag{25}$$

In eq. (25), the term $1/(\alpha_j+k)^2$ is present only in $c_j = k+1$ to $c_j = c$. This allows us to write

$$\frac{1}{(\alpha_j+k)^2}\sum_{c_j=k+1}^{c}\frac{c!}{c_j!(c-c_j)!}x^{c_j}(1-x)^{c-c_j} = \frac{1}{(\alpha_j+k)^2}\left[1-\sum_{h=0}^{k}\frac{c!}{k!(c-k)!}x^h(1-x)^{c-h}\right].$$

Given that

$$\int_0^1 x^{u-1}(1-x)^{v-1}dp = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)},$$

we may then rewrite $D_j$ as

$$\begin{aligned} D_j &= \frac{\Gamma(\alpha)/\Gamma(\alpha_j)}{\Gamma(\alpha-\alpha_j)}\sum_{k=0}^{c-1}\frac{1}{(\alpha_j+k)^2}\left[\frac{\Gamma(\alpha-\alpha_j)}{\Gamma(\alpha)/\Gamma(\alpha_j)} - \sum_{h=0}^{k}\frac{c!}{h!(c-h)!}\frac{\Gamma(\alpha+c-\alpha_j-h)}{\Gamma(\alpha+c)/\Gamma(\alpha_j+h)}\right] \\ &= \sum_{k=0}^{c-1}\frac{1}{(\alpha_j+k)^2}\left[1-\sum_{h=0}^{k}\frac{c!}{h!(c-h)!}\frac{\Gamma(\alpha)\Gamma(\alpha_j+h)\Gamma(\alpha+c-\alpha_j-h)}{\Gamma(\alpha_j)\Gamma(\alpha-\alpha_j)\Gamma(\alpha+c)}\right] \\ &= \sum_{k=0}^{c-1}\frac{1}{(\alpha_j+k)^2}\left[1-G\sum_{h=0}^{k}\frac{\Gamma(\alpha_j+h)\Gamma(\alpha+c-\alpha_j-h)}{h!(c-h)!}\right], \end{aligned} \tag{26}$$

where $G = c!\Gamma(\alpha)/[\Gamma(\alpha_j)\Gamma(\alpha-\alpha_j)\Gamma(\alpha+c)]$. Note that the inner sum increases by a single term for each increase in $k$, so that $D_j$ can be computed in time proportional to $c$. The expression for $D_j$ in eq. (26) is exact for arbitrary values of $\alpha_j > 0$ and arbitrary integers $c > 0$. The exact series expression for $D$ for arbitrary $c$ and $\alpha$ can also be written as

$$D = \sum_{k=0}^{c-1}\frac{1}{(\alpha+k)^2}. \tag{27}$$

Eqs. (26) and (27) allow us to calculate $F$ using eq. (5), and therefore to calculate the integrand in eq. (1) at arbitrary $\vec{\alpha}$.

An interesting case arises when $c = 1$. Here one obtains the exact results $D = 1/\alpha^2$, and $D_j = 1/(q_j\alpha^2)$. Consequently, $F = 1 - D\sum_{j=1}^{L} D_j^{-1} = 0$. This supports the correctness of our asymptotic results above for the factor $F$, which also yield zero upon substituting $c = 1$. Intuitively, alignments that are only one letter deep lack sufficient information to distinguish among Dirichlet distributions within $\mathcal{D}_L$, underdetermining $\vec{\alpha}$ and therefore yielding zero independent distributions. Such individual sequences, in contrast, are sufficient to distinguish among alternative multinomial distributions.

### 4.2. *Monte Carlo evaluation of the definite integral*

Monte Carlo evaluation of the definite integral requires only a way to sample points randomly from the integration domain, and a way to calculate the integrand at each point, as described above. Practical problems with accuracy can arise, however, if the integrand varies greatly over its domain, or even diverges. Especially in high dimensions, a "flat" integrand is desirable, and it is of course trivial to find the definite integral when the integrand is constant.

Using Monte Carlo techniques to evaluate the definite integral of eq. (1) presents two main challenges: first, that the domain of integration is infinite in all $L$ dimensions; second, that the integrand can vary greatly. Both problems can be mitigated by carefully chosen changes of variable. As we describe these changes below, we omit detailed derivation of the Jacobian determinants that accompany them.

Eqs. (26), (27), (4), and (5) express the integrand as an easily calculated function of $\vec{\alpha}$. We begin by replacing the variables of integration $\vec{\alpha}$ with the alternative parameters $(\vec{q}, \alpha)$, where $\alpha_j \equiv \alpha q_j$. This confines the $L-1$ free parameters of $\vec{q}$ to the finite domain $\Omega_L$, but leaves $(0, \infty)$ as the domain of the remaining parameter $\alpha$.

Deferring our discussion of $\alpha$, we consider first the variables $\vec{q}$ in greater detail. As implied by an examination of the integral $I_1$ in section 3.3, the integrand diverges near the boundaries of $\Omega_L$, even though the definite integral remains finite. This divergence, however, increases the variance of Monte Carlo estimates, because they are affected disproportionately by $\vec{q}$ sampled near the boundaries of $\Omega_L$. Fortunately, we can eliminate this problem by sampling $\vec{r}$ uniformly from the unit $(L-1)$-sphere (Marsaglia, 1972), and letting $q_j \equiv r_j^2$; note that $\vec{q}$ is thereby automatically confined to $\Omega_L$. The domain of integration remains finite, but after multiplication by the Jacobian determinant, the integrand becomes essentially flat as a function of $\vec{r}$.

Turning to the final variable $\alpha$, there are many ways to render its domain finite. However, guided by an analysis of the integral $I_2$ in section 3.3, we choose to replace $\alpha$ by $\gamma$ using the equation

$$\alpha \equiv Lc\frac{\gamma^2}{1-\gamma^2}. \tag{28}$$

This has two desired effects. First, it transforms the domain of integration to the finite (0, 1). Second, it renders the integrand flat both as $\gamma$ approaches 0, and as $\gamma$ approaches 1, with a smooth transition between these two regimes near the center of $\gamma$'s domain.

With these changes of variable, accurate Monte Carlo estimation of the definite integral in question becomes tractable, even for large $L$. We have implemented a program to perform this estimation, and applied it to a variety of $L$ and $c$. Generalizing eq. (22) to small $c$ by writing

$$\mathrm{COMP}(\mathcal{D}_L, n, c) = \frac{L}{2}\log n + \frac{L-1}{2}\log c + \Delta_{L,c} + A_L + o(1), \tag{29}$$

our calculated values of $\Delta_{L,c}$ are given in Table 1. Using $10^7$ sample points for each definite integral, the standard errors for our estimates of $\Delta_{L,c}$ are all $\leq 0.0004$ bits. The differing qualitative behavior of $\Delta_{L,c}$ for $L = 2$ and $L \geq 3$ may be understood, in the light of eq. (20), as arising from the differing qualitative behavior near $\phi = 0$ of $\sin^{L-2}\phi$ for these $L$.

As can be seen by examining Table 1, and as our theory from the previous section asserts, for each $L$ the correction term $\Delta_{L,c}$ approaches 0 as $c$ gets large. Remarkably, the asymptotic eq. (22) is accurate to within 1 bit in all instances, and to within 0.4 bits except when $L = 2$ and $c \leq 3$. Thus, in many practical applications, $\Delta_{L,c}$ may be completely ignored, or approximated by a small constant dependent on $L$ and typical values of $c$.

TABLE 1. VALUES OF $\Delta_{L,c}$, IN BITS

| | c | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| L | 2 | 3 | 4 | 6 | 10 | 20 | 40 | 100 |
| 2 | −0.725 | −0.480 | −0.369 | −0.259 | −0.165 | −0.084 | −0.037 | −0.003 |
| 3 | −0.055 | 0.071 | 0.111 | 0.136 | 0.144 | 0.136 | 0.118 | 0.090 |
| 4 | 0.198 | 0.235 | 0.229 | 0.210 | 0.183 | 0.148 | 0.117 | 0.082 |
| 5 | 0.310 | 0.285 | 0.254 | 0.215 | 0.179 | 0.141 | 0.109 | 0.074 |
| 6 | 0.360 | 0.293 | 0.250 | 0.207 | 0.170 | 0.134 | 0.103 | 0.070 |
| 7 | 0.378 | 0.287 | 0.241 | 0.198 | 0.164 | 0.129 | 0.099 | 0.067 |
| 8 | 0.381 | 0.277 | 0.231 | 0.191 | 0.159 | 0.126 | 0.096 | 0.065 |
| 9 | 0.375 | 0.266 | 0.222 | 0.186 | 0.156 | 0.124 | 0.094 | 0.063 |
| 10 | 0.367 | 0.256 | 0.216 | 0.182 | 0.154 | 0.121 | 0.092 | 0.062 |
| 12 | 0.346 | 0.243 | 0.208 | 0.178 | 0.151 | 0.119 | 0.090 | 0.061 |
| 15 | 0.320 | 0.231 | 0.201 | 0.174 | 0.148 | 0.117 | 0.088 | 0.059 |
| 20 | 0.294 | 0.222 | 0.196 | 0.171 | 0.145 | 0.113 | 0.086 | 0.057 |

All values are calculated by Monte Carlo simulation, as described in the text, using $10^7$ random points, and have a standard error $\leq 0.0004$ bits. For protein sequences $L = 20$ and additional values of $\Delta_{20,c}$ are provided in Ye et al. (2011).

### 4.3. Perspective on the complexity of the Dirichlet model

It is worth trying to gain a conceptual perspective on eq. (29). First, as should be intuitively clear, the greater the number of columns $n$ in one's data, or the number of observations $c$ per column, the easier it is to distinguish among alternative Dirichlet distributions, and thus the greater the complexity of the model. However, why the differing dependencies on $n$ and $c$? Consider the alternative parametrization of a Dirichlet distribution $(\vec{q}, \alpha)$ described above. The location parameters $\vec{q}$ can be viewed as specifying a multinomial distribution with $L-1$ free parameters. If one were to describe the data using only a multinomial model, there would be $nc$ observed letters, and the complexity of the model would grow as $\frac{L-1}{2}\log nc$. Using instead a Dirichlet model introduces the additional parameter $\alpha$. Increasing $n$ corresponds to sampling more multinomial distributions, and thus can be seen as relevant to estimating $\alpha$, the concentration of these distributions about their mean. That $\alpha$ is a single parameter means the additional model complexity it introduces should grow as $\frac{1}{2}\log n$. In contrast, increasing $c$ corresponds only to taking more observations from each multinomial sampled. While a larger $c$ helps to constrain the location parameters of a Dirichlet distribution, it is in essence of no help in constraining the distribution's concentration.

In practice, multiple alignment data frequently consist of $n$ columns in which $c$ varies by column, but we do not propose to analyze with rigor a generalization of the problem to this case. However, the perspective described above suggests that, for such data, it is appropriate to extend eq. (29) simply by using $\bar{c}$, the mean number of observations per column, in place of $c$.

## 5. CONCLUSION

We have derived an analytic formula, eq. (22), for the complexity $\mathrm{COMP}(\mathcal{D}_L, n, c)$ of the Dirichlet model $\mathcal{D}_L$ applied to multiple alignment data with large $n$ and $c$. To calculate this complexity for small $c$, we have derived an easily evaluated expression for the determinant required by the relevant definite integral. Using this expression, we have applied Monte Carlo estimation to find $\mathrm{COMP}(\mathcal{D}_L, n, c)$ for arbitrary $c$, as given by eq. (29).

The Dirichlet models studied here, although tractable, are too simple to describe accurately multiple alignment data from protein sequences. In a companion article (Ye et al., 2011), we study Dirichlet mixtures (Brown et al., 1993; Sjölander et al., 1996; Altschul et al., 2010), which are better suited to proteins but too complex to analyze rigorously. There, we extend by informal arguments the results of this article to the more general and more broadly applicable case of Dirichlet mixtures.

## APPENDIX

### A. The determinant of a matrix with special form

To prove the lemma described in eqs. (3–5), we first establish a general equality in linear algebra:

$$\det\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix} = \det\mathbf{A}\cdot\det\left[\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right] = \det\mathbf{D}\cdot\det\left[\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right], \tag{30}$$

where $\mathbf{A}$ and $\mathbf{D}$ are *invertible* square matrices, while $\mathbf{B}$ and $\mathbf{C}$ are rectangular matrices. The equality (30) follows from the decompositions

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{C} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{A}^{-1}\mathbf{B}\\ \mathbf{0} & \mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{B}\\ \mathbf{0} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0}\\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I}\end{bmatrix},$$

and the fact that $\det[\mathbf{XY}] = \det\mathbf{X}\cdot\det\mathbf{Y}$ when $\mathbf{X}$ and $\mathbf{Y}$ are square matrices of the same dimension.

To proceed, we rewrite the matrix $\mathbf{M}$ of eq. (3) as the product

$$\mathbf{M} = \mathbf{d}\cdot\left[\mathbf{I}-\mathbf{b}\mathbf{b}^{T}\right]\cdot\mathbf{d}, \tag{31}$$

where $\mathbf{d}$ is a diagonal matrix with elements $d_{j,j'} = \delta_{j,j'}\sqrt{D_j}$, and $\mathbf{b}$ is a vector whose components are $b_j = \sqrt{D/D_j}$. Eq. (31) implies that

$$\det\mathbf{M} = \det\mathbf{d}\cdot\det\left[\mathbf{I}-\mathbf{b}\mathbf{b}^{T}\right]\cdot\det\mathbf{d}. \tag{32}$$

Using eq. (30), we observe that

$$\det\left[\mathbf{I}-\mathbf{b}\mathbf{b}^{T}\right] = \left(1-\mathbf{b}^{T}\mathbf{b}\right) \tag{33}$$

by considering the determinant of the matrix

$$\begin{bmatrix}1 & \mathbf{b}^{T}\\ \mathbf{b} & \mathbf{I}\end{bmatrix}.$$

Since matrix $\mathbf{d}$ consists of diagonal elements only, we may write the determinant of $\mathbf{M}$ as

$$\det\mathbf{M} = \left(\prod_{j=1}^{L}\sqrt{D_j}\right)^{2}\left[1-\sum_{j=1}^{L}\left(\sqrt{\frac{D}{D_j}}\right)^{2}\right] = F\prod_{j=1}^{L}D_j,$$

as stated in eqs. (4) and (5).

### B. The expectation of a function of $c_j$

To evaluate the asymptotic behavior of $D_j$ and $F$, it is necessary to calculate expectation values of functions that depend only on the count vector $\vec{c}$. For this purpose, it is convenient to elaborate eq. (6) as follows:

$$\begin{aligned}P(\vec{y}) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha+c)}\prod_{j=1}^{L}\frac{\Gamma(c_j+\alpha_j)}{\Gamma(\alpha_j)}\\ &= \frac{\Gamma(\alpha)}{\prod_{j=1}^{L}\Gamma(\alpha_j)}\int_{\Omega_L}\prod_{j=1}^{L}x_j^{c_j+\alpha_j-1}d\vec{x}.\end{aligned} \tag{34}$$

Focusing now on $\vec{c}$, eq. (34) implies that we may write the expectation value of a function $K(\vec{c})$ as

$$\begin{aligned}E[K(\vec{c})] &= \sum_{\vec{y}}P(\vec{y})K(\vec{c})\\ &\frac{\Gamma(\alpha)}{\prod_{j=1}^{L}\Gamma(\alpha_j)}\sum_{\substack{c_1,c_2,\ldots,c_L\geq 0\\ \sum_j c_j = c}}\frac{c!}{\prod_{j=1}^{L}c_j!}\int_{\Omega_L}\prod_{j=1}^{L}x_j^{c_j+\alpha_j-1}K(\vec{c})d\vec{x}.\end{aligned} \tag{35}$$

If $K$ depends on only one of the $c_j$, we may integrate away all the $x_{j'\neq j}$ to simplify eq. (35). This amounts to writing

$$\sum_{\substack{c_1,c_2,\ldots,c_L\geq 0\\ \sum_j c_j = c}} = \sum_{c_j=0}^{c} \sum_{\substack{\{c_{j'\neq j}\geq 0\}\\ \sum_{j'} c_{j'} = c - c_j}}, \tag{36}$$

and

$$\frac{c!}{\prod_{j=1}^{L} c_j!} = \frac{c!}{c_j!(c-c_j)!}\frac{(c-c_j)!}{\prod_{j'\neq j} c_{j'}!}, \tag{37}$$

so that

$$\begin{aligned}
E[K(c_j)] &= \sum_{\vec{y}} P(\vec{y})K(c_j)\\
&= \frac{\Gamma(\alpha)}{\prod_{j=1}^{L}\Gamma(\alpha_j)}\sum_{c_j=0}^{c}\frac{c!K(c_j)}{c_j!(c-c_j)!}\int_{\Omega_L} x_j^{c_j+\alpha_j-1}(1-x_j)^{c-c_j}\prod_{j'\neq j} x_{j'}^{\alpha_{j'}-1}d\vec{x}\\
&= \frac{\Gamma(\alpha)}{\prod_{j=1}^{L}\Gamma(\alpha_j)}\sum_{c_j=0}^{c}\frac{c!K(c_j)}{c_j!(c-c_j)!}\int_0^1 x_j^{c_j+\alpha_j-1}(1-x_j)^{c-c_j+\sum_{j'\neq j}\alpha_{j'}}dx_j\\
&\quad\times\int_0^1\delta\left((1-x_j)\left[\sum_{j'\neq j}\tilde{x}_{j'}-1\right]\right)\prod_{j'\neq j}\left[\tilde{x}_{j'}^{\alpha_{j'}-1}d\tilde{x}_{j'}\right]\\
&= \frac{\Gamma(\alpha)}{\prod_{j=1}^{L}\Gamma(\alpha_j)}\sum_{c_j=0}^{c}\frac{c!K(c_j)}{c_j!(c-c_j)!}\int_0^1 x_j^{c_j+\alpha_j-1}(1-x_j)^{c-c_j+\alpha-\alpha_j}dx_j\\
&\quad\times(1-x_j)^{-1}\int_{\Omega_{L-1}}\prod_{j'\neq j}\left[\tilde{x}_{j'}^{\alpha_{j'}-1}d\tilde{x}_{j'}\right]\\
&= \frac{\Gamma(\alpha)/\Gamma(\alpha_j)}{\Gamma(\alpha-\alpha_j)}\sum_{c_j=0}^{c}\frac{c!K(c_j)}{c_j!(c-c_j)!}\int_0^1 x^{c_j+\alpha_j-1}(1-x)^{c+\alpha-c_j-\alpha_j-1}dx.
\end{aligned} \tag{38}$$

### C. *The asymptotic behavior of* D, $D_j$, *and* F

As mentioned in the main text, to obtain $F$ to the order $\alpha^{-2}$, we need to calculate $D_j$ and $D$ to the order $\alpha^{-4}$. Here we provide the details needed for this task.

One way to express the trigamma function is

$$\psi'(x) = \sum_{k=0}^{\infty}\frac{1}{(x+k)^2},$$

which in the large $x$ limit yields the asymptotic form

$$\psi'(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} + \mathcal{O}(x^{-5}). \tag{39}$$

For notational convenience, the $n$th derivative of the digamma function will be written as $\psi^{(n)}(x) \equiv d^n\psi(x)/dx^n$. That is, $\psi'(x)$ and $\psi^{(1)}(x)$ represent the same function $d\psi(x)/dx$. To seek the asymptotic behavior of $D_j$, we introduce $k_j \equiv E[c_j]$, and expand the function $\psi^{(1)}(\alpha_j + c_j)$ around $\alpha_j + k_j$ prior to taking the average. In other words, we write

$$\psi^{(1)}(\alpha_j + c_j) = \psi^{(1)}(\alpha_j + k_j) + \sum_{\ell=1}^{\infty}\frac{(c_j-k_j)^\ell}{\ell!}\psi^{(\ell+1)}(\alpha_j + k_j), \tag{40}$$

which leads to

$$E\left[\psi^{(1)}(\alpha_j + c_j)\right] = \psi^{(1)}(\alpha_j + k_j) + \sum_{\ell=2}^{\infty} \frac{E\left[(c_j - k_j)^{\ell}\right]}{\ell!} \psi^{(\ell+1)}(\alpha_j + k_j). \tag{41}$$

The fact that $\psi^{(\ell+1)}(\alpha_j + k_j)$ is of order $\alpha^{-\ell-1}$ makes eq. (41) an asymptotic expansion, provided that $E[(c_j - k_j)^{\ell}]$ becomes independent of $\alpha$ in the large-$\alpha$ limit. We shall show this later by evaluating the leading behavior of $E[(c_j - k_j)^{\ell}]$. From eq. (41), it is evident that we need to compute up to at least the third central moment to obtain $D_j$ to the order $\alpha^{-4}$.

Introducing the new variable $v \equiv c_j - 1$, we may now use eq. (38) of Appendix B to calculate $k_j$:

$$\begin{aligned} k_j \equiv \sum_{\vec{y}} P(\vec{y}) c_j &= \frac{\Gamma(\alpha)/\Gamma(\alpha_j)}{\Gamma(\alpha - \alpha_j)} c \sum_{v=0}^{c-1} C_v^{c-1} \int_0^1 x^{v+\alpha_j} (1-x)^{c-1-v+\alpha-\alpha_j-1} dx \\ &= \frac{\Gamma(\alpha)/\Gamma(\alpha_j)}{\Gamma(\alpha - \alpha_j)} c \int_0^1 x^{\alpha_j} (1-x)^{\alpha-\alpha_j-1} dx \\ &= c \frac{\Gamma(\alpha)}{\Gamma(\alpha_j)\Gamma(\alpha - \alpha_j)} \frac{\Gamma(\alpha_j + 1)\Gamma(\alpha - \alpha_j)}{\Gamma(\alpha + 1)} = c \frac{\alpha_j}{\alpha}, \end{aligned} \tag{42}$$

as mentioned in the main text.

To compute the second central moment, we write

$$E\left[(c_j - k_j)^2\right] = E[c_j(c_j - 1)] + E[c_j] - k_j^2 = E[c_j(c_j - 1)] + k_j - k_j^2, \tag{43}$$

and

$$\begin{aligned} E[c_j(c_j - 1)] &= \frac{\Gamma(\alpha)/\Gamma(\alpha_j)}{\Gamma(\alpha - \alpha_j)} c(c-1) \sum_{v=0}^{c-2} C_v^{c-2} \int_0^1 x^{v+\alpha_j+1} (1-x)^{c-2-v+\alpha-\alpha_j-1} dx \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha_j)\Gamma(\alpha - \alpha_j)} c(c-1) \int_0^1 x^{\alpha_j+1} (1-x)^{\alpha-\alpha_j-1} dx \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha_j)\Gamma(\alpha - \alpha_j)} c(c-1) \frac{\Gamma(\alpha_j + 2)\Gamma(\alpha - \alpha_j)}{\Gamma(\alpha + 2)} \\ &= c(c-1) \frac{\alpha_j(\alpha_j + 1)}{\alpha(\alpha + 1)}. \end{aligned} \tag{44}$$

Therefore,

$$E[(c_j - k_j)^2] = \frac{c\alpha_j}{\alpha} \left[ \frac{(c-1)(\alpha_j + 1)}{\alpha + 1} + 1 - \frac{c\alpha_j}{\alpha} \right] = c q_j (1 - q_j) \frac{\alpha + c}{\alpha + 1}. \tag{45}$$

The calculation of the $\ell$th central moment of $c_j$ employs the same idea. We first express

$$\begin{aligned} c_j^r &= \sum_{m=0}^{r-1} g_{m;r} c_j (c_j - 1) \cdots (c_j - m) \\ &= \sum_{m=0}^{r-1} g_{m;r} \frac{c_j!}{(c_j - m - 1)!}, \end{aligned} \tag{46}$$

where the coefficients $g_{m;r}$ depend on $r$ and can be obtained in the following way. By setting $c_j = 1$ in eq. (46), we zero all terms except $g_{0;r}$ on the right hand side, and therefore see that $g_{0;r} = 1$. With $g_{0;r}$ known, we may then set $c_j = 2$; the only nonzero terms on the right hand side of eq. (46) are then $2g_{0;r} + 2!g_{1;r}$. With the left hand side now equal to $2^r$, and $g_{0;r} = 1$, we easily solve for $g_{1;r} = 2^{r-1} - 1$. We may then set $c_j = 3, 4, \ldots$ etc. to solve respectively for $g_{2;r}, g_{3;r}, \ldots$ etc.

Note that it is possible to express

$$E\left[c_j(c_j-1)\cdots(c_j-m)\right] = \frac{\Gamma(\alpha)c(c-1)\cdots(c-m)}{\Gamma(\alpha_j)\Gamma(\alpha-\alpha_j)}\int_0^1 x^{\alpha_j+m}(1-x)^{\alpha-\alpha_j-1}dx$$
$$= \frac{\Gamma(\alpha)c(c-1)\cdots(c-m)}{\Gamma(\alpha_j)\Gamma(\alpha-\alpha_j)}\frac{\Gamma(\alpha_j+m+1)\Gamma(\alpha-\alpha_j)}{\Gamma(\alpha+m+1)}$$
$$= c(c-1)\cdots(c-m)\frac{\alpha_j(\alpha_j+1)\cdots(\alpha_j+m)}{\alpha(\alpha+1)\cdots(\alpha+m)}. \tag{47}$$

This means that for $\ell \geq 2$ we may write

$$E\left[(c_j-k_j)^\ell\right] = (-1)^{\ell-1}(\ell-1)k_j^\ell + \sum_{r=2}^{\ell} C_r^\ell E\left[c_j^r\right](-k_j)^{\ell-r}$$
$$= (-1)^{\ell-1}(\ell-1)k_j^\ell + \sum_{r=2}^{\ell} C_r^\ell (-k_j)^{\ell-r}$$
$$\times\left\{k_j + \sum_{m=1}^{r-1} g_{m;r}\frac{c!}{(c-m-1)!}\frac{\alpha_j(\alpha_j+1)\cdots(\alpha_j+m)}{\alpha(\alpha+1)\cdots(\alpha+m)}\right\}. \tag{48}$$

As an example, we may use this approach to obtain the third central moment

$$E\left[(c_j-k_j)^3\right] = c\ q_j(1-q_j)(1-2q_j)\frac{(\alpha+c)(\alpha+2c)}{(\alpha+1)(\alpha+2)}. \tag{49}$$

Because $c$ is a fixed positive integer, in the large-$\alpha$ limit we may write the leading term in eq. (48) as

$$E\left[(c_j-k_j)^{2\ell}\right] \approx (2\ell-1)!!\ \frac{c^\ell\alpha^{2\ell-2}\ (\alpha+c)}{(\alpha+1)(\alpha+2)\cdots(\alpha+2\ell-1)}\ q_j^\ell(1-q_j)^\ell \tag{50}$$
$$\underset{\alpha\to\infty}{\longrightarrow} (2\ell-1)!!\ c^\ell\ q_j^\ell(1-q_j)^\ell, \tag{51}$$
$$E\left[(c_j-k_j)^{2\ell+1}\right] \approx \frac{\ell}{3}\ (2\ell+1)!!\ \frac{c^\ell\ \alpha^{2\ell-1}\ (\alpha+c)\ (1-2q_j)}{(\alpha+1)(\alpha+2)\cdots(\alpha+2\ell)}q_j^\ell(1-q_j)^\ell \tag{52}$$
$$\underset{\alpha\to\infty}{\longrightarrow} \frac{\ell}{3}(2\ell+1)!!\ c^\ell\ (1-2q_j)\ q_j^\ell(1-q_j)^\ell. \tag{53}$$

In the limit of large $\alpha_j$ (or $k_j$), we note that

$$\frac{\psi^{(1+\ell)}(\alpha_j+k_j)}{\ell!} = \frac{(-1)^\ell}{q_j^{\ell+1}(\alpha+c)^{\ell+1}}\left[1+\frac{\ell+1}{2q_j(\alpha+c)}+\mathcal{O}\left(\frac{1}{(\alpha_j+k_j)^2}\right)\right]. \tag{54}$$

Therefore, the absence of $\alpha$ dependence and the presence of the factorials in eqs. (51) and (53) means that the expansion in eq. (41) is asymptotic in $\alpha$. That is, the larger $\alpha$, the more terms in the expansion one may retain to improve accuracy before the series becomes divergent.

With the aim of obtaining $D$ and the $D_j$ to the order $\alpha^{-4}$, we now continue the investigation of their large-$\alpha$ behavior. Using eq. (39), we write the asymptotic expression for $D$ as

$$D = \psi^{(1)}(\alpha) - \psi^{(1)}(\alpha+c)$$
$$= \left(\frac{1}{\alpha}+\frac{1}{2\alpha^2}+\frac{1}{6\alpha^3}+\mathcal{O}(\alpha^{-5})\right) - \left(\frac{1}{\alpha+c}+\frac{1}{2(\alpha+c)^2}+\frac{1}{6(\alpha+c)^3}+\mathcal{O}(\alpha^{-5})\right)$$
$$= \frac{c}{\alpha(\alpha+c)}\left[1+\frac{1}{2}\left(\frac{1}{\alpha}+\frac{1}{\alpha+c}\right)+\frac{1}{6}\left(\frac{1}{\alpha^2}+\frac{1}{\alpha(\alpha+c)}+\frac{1}{(\alpha+c)^2}\right)\right]+\mathcal{O}(\alpha^{-5}). \tag{55}$$

Next, noting that $k_j = cq_j$, we find that

$$\begin{aligned}
D_j &= \psi^{(1)}(\alpha_j) - E[\psi^{(1)}(\alpha_j + c_j)] \\
&= \frac{c}{\alpha_j(\alpha+c)}\left[1 + \frac{1}{2q_j}\left(\frac{1}{\alpha}+\frac{1}{\alpha+c}\right) + \frac{1}{6q_j^2}\left(\frac{1}{\alpha^2}+\frac{1}{\alpha(\alpha+c)}+\frac{1}{(\alpha+c)^2}\right)\right] \\
&\quad - \frac{E[(c_j-k_j)^2]}{2!}\psi^{(3)}(\alpha_j+k_j) - \frac{E[(c_j-k_j)^3]}{3!}\psi^{(4)}(\alpha_j+k_j) + \mathcal{O}(\alpha^{-5}) \\
&= \frac{c}{\alpha_j(\alpha+c)}\left[1 + \frac{1}{2q_j}\left(\frac{1}{\alpha}+\frac{1}{\alpha+c}\right) + \frac{1}{6q_j^2}\left(\frac{1}{\alpha^2}+\frac{1}{\alpha(\alpha+c)}+\frac{1}{(\alpha+c)^2}\right)\right] \\
&\quad - \left[cq_j(1-q_j)\frac{\alpha+c}{\alpha+1}\right]\left[\frac{1}{q_j^3(\alpha+c)^3} + \frac{3}{2q_j^4(\alpha+c)^4}\right] \\
&\quad - \left[cq_j(1-q_j)(1-2q_j)\frac{(\alpha+c)(\alpha+2c)}{(\alpha+1)(\alpha+2)}\right]\frac{(-1)}{q_j^4(\alpha+c)^4} + \mathcal{O}(\alpha^{-5}) \\
&= \frac{c}{\alpha_j(\alpha+c)}\left\{1 + \frac{1}{2q_j}\left(\frac{1}{\alpha}+\frac{1}{\alpha+c}\right) + \frac{1}{6q_j^2}\left(\frac{1}{\alpha^2}+\frac{1}{\alpha(\alpha+c)}+\frac{1}{(\alpha+c)^2}\right)\right. \\
&\quad \left. - \frac{\alpha(1-q_j)}{q_j(\alpha+c)(\alpha+1)} - \frac{3}{2}\frac{\alpha(1-q_j)}{q_j^2(\alpha+c)^2(\alpha+1)} + \frac{\alpha(1-q_j)(1-2q_j)(\alpha+2c)}{q_j^2(\alpha+c)^2(\alpha+1)(\alpha+2)}\right\} + \mathcal{O}(\alpha^{-5}) \\
&= \frac{c}{\alpha_j(\alpha+c)}\left\{1 + \frac{\alpha}{(\alpha+c)(\alpha+1)} + \frac{2\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)} + \frac{1}{2q_j}\left[\frac{1}{\alpha}+\frac{1}{\alpha+c}\right.\right. \\
&\quad \left. - \frac{2\alpha}{(\alpha+c)(\alpha+1)} + \frac{3\alpha}{(\alpha+c)^2(\alpha+1)} - \frac{6\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)}\right] + \frac{1}{6q_j^2}\left[\frac{1}{\alpha^2}+\frac{1}{\alpha(\alpha+c)}\right. \\
&\quad \left.\left. + \frac{1}{(\alpha+c)^2} - \frac{9\alpha}{(\alpha+c)^2(\alpha+1)} + \frac{6\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)}\right]\right\} + \mathcal{O}(\alpha^{-5}).
\end{aligned} \tag{56}$$

When summing the terms associated with $q_j^{-2}$ inside the second pair of square brackets, we find that the final contribution becomes of order $\alpha^{-5}$, and thus may be dropped from the analysis. Also, the terms associated with $q_j^{-1}$, when added, can be rearranged as

$$\begin{aligned}
&\frac{1}{\alpha} + \frac{1}{\alpha+c} - \frac{2\alpha}{(\alpha+c)(\alpha+1)} + \frac{3\alpha}{(\alpha+c)^2(\alpha+1)} - \frac{6\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)} \\
&\qquad = \frac{c}{\alpha(\alpha+c)} - \frac{1}{(\alpha+c)(\alpha+1)} + \mathcal{O}(\alpha^{-3}).
\end{aligned} \tag{57}$$

Consequently, we have

$$\begin{aligned}
D_j = \frac{c}{\alpha_j(\alpha+c)}&\left\{1 + \frac{\alpha}{(\alpha+c)(\alpha+1)} + \frac{2\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)}\right. \\
&\left. + \frac{1}{2q_j}\left[\frac{c}{\alpha(\alpha+c)} - \frac{1}{(\alpha+c)(\alpha+1)}\right]\right\} + \mathcal{O}(\alpha^{-5}).
\end{aligned} \tag{58}$$

Eqs. (55) and (58) allow us to express $F$ to the order $\alpha^{-2}$. We first calculate

$$\frac{D}{D_j} = \frac{\alpha_j}{\alpha}\left[1 + \frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\alpha+c}\right) + \frac{1}{6}\left(\frac{1}{\alpha^2} + \frac{1}{\alpha(\alpha+c)} + \frac{1}{(\alpha+c)^2}\right)\right]$$
$$\times\left\{1 - \frac{\alpha}{(\alpha+c)(\alpha+1)} - \frac{2\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)}\right.$$
$$\left. - \frac{1}{2q_j}\left[\frac{c}{\alpha(\alpha+c)} - \frac{1}{(\alpha+c)(\alpha+1)}\right] + \frac{\alpha^2}{(\alpha+c)^2(\alpha+1)^2}\right\} + \mathcal{O}(\alpha^{-3})$$
$$= q_j\left\{1 + \frac{1}{2}\left[\frac{1}{\alpha} + \frac{1}{\alpha+c} - \frac{2\alpha}{(\alpha+c)(\alpha+1)}\right]\left(1 - \frac{\alpha}{(\alpha+c)(\alpha+1)}\right)\right.$$
$$+ \frac{1}{6}\left[\frac{1}{\alpha^2} + \frac{1}{\alpha(\alpha+c)} + \frac{1}{(\alpha+c)^2}\right] - \frac{2\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)}$$
$$\left. - \frac{1}{2q_j}\left[\frac{c}{\alpha(\alpha+c)} - \frac{1}{(\alpha+c)(\alpha+1)}\right]\right\} + \mathcal{O}(\alpha^{-3})$$
$$= q_j\left\{1 + \frac{1-q_j^{-1}}{2}\left[\frac{c}{\alpha(\alpha+c)} - \frac{1}{(\alpha+c)(\alpha+1)}\right] + \frac{3}{2}\frac{1}{(\alpha+c)(\alpha+1)}\right.$$
$$\left. + \frac{1}{6}\left[\frac{1}{\alpha^2} + \frac{1}{\alpha(\alpha+c)} + \frac{1}{(\alpha+c)^2}\right] - \frac{2\alpha(\alpha+2c)}{(\alpha+c)^2(\alpha+1)(\alpha+2)}\right\} + \mathcal{O}(\alpha^{-3}) \quad (59)$$
$$= q_j\left\{1 + \frac{1-q_j^{-1}}{2}\left[\frac{c}{\alpha(\alpha+c)} - \frac{1}{(\alpha+c)(\alpha+1)}\right]\right\} + \mathcal{O}(\alpha^{-3}), \quad (60)$$

where the final expression comes from the fact that the last three terms in eq. (59) sum to order $\alpha^{-3}$. We may now compute $F$ to order $\alpha^{-2}$:

$$F = 1 - D\sum_{j=1}^{L} D_j^{-1} = \left(\frac{c}{2\alpha(\alpha+c)} - \frac{1}{2(\alpha+c)(\alpha+1)}\right)\sum_{j=1}^{L}(1-q_j) + \mathcal{O}(\alpha^{-3})$$
$$= (L-1)\left(\frac{c}{2\alpha(\alpha+c)} - \frac{1}{2(\alpha+c)(\alpha+1)}\right) + \mathcal{O}(\alpha^{-3})$$
$$= \frac{(L-1)(c-1)}{2\alpha(\alpha+c)} + \mathcal{O}(\alpha^{-3}),$$

the result shown in eq. (15) of the main text.

## ACKNOWLEDGMENTS

We thank Drs. John Spouge and Xugang Ye for helpful conversations. This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

## DISCLOSURE STATEMENT

No competing financial interests exist.

## REFERENCES

Altschul, S.F., Gertz, E.M., Agarwala, R., et al. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.* 37, 815–824.

Altschul, S.F., Wootton, J.C., Zaslavsky, E., et al. 2010. The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comp. Biol.* 6, e1000852.

Brown, M., Hughey, R., Krogh, A., et al. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families, 47–55. *In* Hunter, L., Searls, D., and Shavlik, J., eds. *Proc. First Int. Conf. on Intelligent System for Mol. Biol.* AAAI Press, Menlo Park, CA.

Grünwald, P.D. 2007. *The Minimum Description Length Principle.* MIT Press, Cambridge, MA.

Marsaglia, G. 1972. Choosing a point from the surface of a sphere. *Ann. Math. Stat.* 43, 645–646.

Sjölander, K., Karplus, K., Brown, M., et al. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* 12, 327–345.

Ye, X., Yu, Y.-K., and Altschul, S.F. 2011. On the inference of Dirichlet mixture priors for protein sequence comparison. *J. Comput. Biol.* 18 (this issue).

Address correspondence to:
*Dr. Stephen F. Altschul*
*National Center for Biotechnology Information*
*National Library of Medicine*
*National Institutes of Health*
*Bethesda, MD 20894*

*E-mail:* altschul@ncbi.nlm.nih.gov