Supplemental Data for PSI-BLAST Pseudocounts and the Minimum Description Length Principle

Stephen F. Altschul[1], E. Michael Gertz, Richa Agarwala,
Alejandro A. Schäffer and Yi-Kuo Yu

National Center for Biotechnology Information,
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894

[1]To whom correspondence should be addressed. Email: altschul@ncbi.nlm.nih.gov

Supplemental Data A: The Model Description Length for Large n

An expression for the description length of the model, given in reference (30) of the main text, is

$$DL_n(\mathcal{M}) = \frac{k}{2}\log_2\frac{n}{2\pi} + \log_2 \int\limits_{\Theta} \sqrt{|I(\theta)|}d\theta.$$

The Fisher information of a multinomial model is $I(\theta) = 1/\prod_{i=1}^{k+1}\theta_i$, where θ_i represents one of the $k+1$ components of the vector θ. The range of integration Θ is the set of all probability vectors. Observing that $\theta_{k+1} = 1 - \sum_{i=1}^{k}\theta_k$, we may write the integral as

$$\int\limits_{\Theta} \sqrt{|I(\theta)|}d\theta = \int\limits_{\sum_{i=1}^{k}\theta_i\leq 1;\theta_i\geq 0} \cdots \int \sqrt{\frac{1}{(1-\sum_{i=1}^{k}\theta_i)\prod_i \theta_i}} \prod_{i=1}^{k} d\theta_i \ .$$

With the substitution $x_i \equiv \sqrt{\theta_i}$, the region of integration Θ is transformed to the intersection of the unit sphere S^k with the set for which $x_i \geq 0$ for all i. Moreover, the integrand is transformed to

$$\frac{2^k \prod_i dx_i}{\sqrt{1-\sum_{i=1}^{k} x_i^2}},$$

which is symmetric under the transformation $x_i \mapsto -x_i$ for any i. Thus the integral can be transformed to an integral over S^k.

$$\int\limits_{\Theta} \sqrt{|I(\theta)|}d\theta = \int\limits_{S^k} \frac{\prod_i dx_i}{\sqrt{1-\sum_{i=1}^{k} x_i^2}} = \Omega_k \int\limits_{0}^{1} \frac{r^{k-1}dr}{\sqrt{1-r^2}},$$

where $r^2 \equiv \sum_i x_i^2$, and $\Omega_k = 2\pi^{k/2}/\Gamma(\frac{k}{2})$ is the surface area of S^k. But

$$\begin{aligned}\Omega_k \int\limits_{0}^{1} \frac{r^{k-1}dr}{\sqrt{1-r^2}} &= \Omega_k \int\limits_{0}^{\pi/2} (\sin\phi)^{k-1}\, d\phi \\ &= \frac{\Omega_k}{2} 2^{k-1}\frac{\Gamma(k/2)\Gamma(k/2)}{\Gamma(k)} = \pi^{k/2}2^{k-1}\frac{\Gamma(k/2)}{\Gamma(k)} \ ,\end{aligned}$$

and one may use the gamma function multiplication formula $\Gamma(z)\Gamma(z+\frac{1}{2}) = \sqrt{\pi}\Gamma(2z)2^{1-2z}$ to conclude

$$\int\limits_{\Theta} \sqrt{I(\theta)}\, d\theta \ = \frac{\pi^{(k+1)/2}}{\Gamma(\frac{k+1}{2})} = \frac{\Omega_{k+1}}{2} \ .$$

Stirling's approximation $\Gamma(x+1) \approx \sqrt{2\pi x}\ x^x e^{-x}$ yields

$$\log_2\Gamma\left(\frac{k+1}{2}\right) \approx \frac{1}{2} + \frac{1}{2}\log_2(\pi) + \frac{k}{2}\log_2\left(\frac{k-1}{2}\right) - \frac{k}{2}\log_2(e) + \frac{1}{2}\log_2(e) \ ,$$

so that

$$\log_2 \int_\Theta \sqrt{I(\theta)}d\theta \approx \frac{k+1}{2}\log_2(\pi) - \frac{1}{2} - \frac{1}{2}\log_2(\pi) - \frac{k}{2}\log_2\left(\frac{k-1}{2}\right) + \frac{k}{2}\log_2(e) - \frac{1}{2}\log_2(e)$$

$$\approx \frac{k}{2}\log_2\left(\pi e \frac{2}{k-1}\right) - \frac{1}{2} - \frac{1}{2}\log_2(e)$$

$$\approx \frac{k}{2}\log_2\left(\pi e \frac{2}{k}\right) - \frac{k}{2}\log_2\left(1 - \frac{1}{k}\right) - \frac{1}{2} - \frac{1}{2}\log_2(e)\ .$$

Note that $\log_2(1-x) = \ln(1-x)/\ln(2) \approx -x/\ln(2) = -x\log_2(e)$, giving

$$\log_2 \int_\Theta \sqrt{I(\theta)}d\theta \approx \frac{k}{2}\log_2\left(\pi e \frac{2}{k}\right) - \frac{1}{2}\ .$$

Adding this to $\frac{k}{2}\log_2\frac{n}{2\pi}$, we obtain

$$DL_n(\mathcal{M}) \approx \frac{k}{2}\log_2\frac{ne}{k} - \frac{1}{2} \text{ bits}\ .$$

Supplemental Data B: Assumptions Used in the Text

In this section, we provide a more formal description of our use of the MDL principle to optimize PSI-BLAST pseudocounts. We describe the approximations that are used in our calculations, and the assumptions that we make to justify them, by relating the MDL principle to the Bayesian approach to modeling a column in a multiple alignment.

The Bayesian approach begins with prior probability distribution $w(\theta)$ over all amino acid frequency vectors. The probability for observing a set x^n of amino acids is then given by

$$P(x^n) = \int_\Theta P(x^n \mid \theta)\, w(\theta)\, d\theta, \qquad \mathbf{1}$$

where Θ is the set of all frequency vectors, and θ ranges over Θ. The minimum description length principle has connections to the Bayesian approach. It has been shown that if $P(x^n \mid \theta)$ is a multinomial, and $w(\theta)$ is proportional to the square root of the Fisher information, then for large n the minimum combined description length of the model and the data approaches $-\log_2 P(x^n)$ bits; see citation (30) of the main text for an overview of these results. The prior distribution proportional to the square root of the Fisher information is known as Jeffrey's prior.

Jeffrey's prior is an uninformative prior. Its use indicates that one has no good prior information about the distribution of θ. To the contrary, the BLOSUM matrices were developed by observing the amino acid substitution frequencies in alignments at a certain evolutionary distance; in the case of BLOSUM-62, the frequencies among those sequences that align with less

than 62% identity. By using the information inherent in BLOSUM-62 to calculate data dependent pseudocounts, one may reduce the minimum description length of columns of biologically accurate multiple alignments.

The use of pseudocounts comes at the cost of adding another parameter α to the model. The number of bits used to encode α is thus part of the description length of the model. Traditionally, PSI-BLAST has used the small integer m, rather than α, to specify the number of pseudocounts, and we have not observed a need to specify the number of pseudocounts with high precision. Thus we make the following simplifying assumption.

Assumption 1 *In our calculations, the contribution of the parameter α to the description length of the model may be safely ignored.*

Information theory implies that if one encodes a column of amino acids with observed frequency $\mathbf{f}$ using a code optimized for a column with frequency $\mathbf{q}$, then the change in the description length of the data approaches $nD(\mathbf{f}\,||\,\mathbf{q})$ for large n. For large n, however, the MDL principle suggests that the model may be described in fewer bits than would be required to encode $\mathbf{q}$ exactly, and therefore there is an error implied in using $nD(\mathbf{f}\,||\,\mathbf{q})$ as the change in the description length of the data. It is shown at the end of this section that this error does not increase as n grows if the following two assumptions hold.

Assumption 2 *The frequencies $\mathbf{f}$ used in the computation of α are bounded away from zero.*

Assumption 3 *The pseudocount parameter, α converges to zero at a rate no slower that a constant times $1/\sqrt{n}$.*

Assumption 2 holds because PSI-BLAST applies a fixed number of pseudocounts, thus bounding each f_i away from zero, before optimizing α. Assumption 3 holds if the number of pseudocounts m is considered to be constant for large n, as is done in PSI-BLAST, or if m is allowed to grow as $n^{1/3}$, as is suggested by equation **9** of the main text.

A vector $\mathbf{q}$ of target frequencies is encoded using a deterministic rule that associates each possible value of $\mathbf{q}$ with a vector $\mathbf{g}$ that can be represented exactly. Our calculations assume a rule that does not depend on α. As α grows, fewer of the vectors $\mathbf{g}$ that may be represented exactly are associated with a $\mathbf{q}$ in the image of Θ under $M'(\alpha)$, which we denote Θ_α, than are associated with a $\mathbf{q}$ in Θ. Thus, the description length of the model decreases with increasing α.

Mathematically, for a given α, the model is encoded by the prior probability

$$w_\alpha(\theta) \propto \begin{cases} \sqrt{I(\theta)} & \text{if } \theta \in \Theta_\alpha; \text{ and} \\ 0 & \text{otherwise,} \end{cases}$$

where $I(\theta)$ is the Fisher information. We assign zero probability to those θ outside of Θ_α to indicate that, according to the model, those probabilities are not credible. The combined description length of the data and model is

$$-\log_2 \int_{\Theta_\alpha} P(x^n \mid \theta)\, w_\alpha(\theta)\, d\theta. \qquad \mathbf{2}$$

The description length of the model is the difference of **2** and the description length of the data, $nH(\mathbf{f}) + nD(\mathbf{f} \,||\, \mathbf{q})$. By definition of $w_\alpha(\theta)$, the integral in **2** is

$$\frac{1}{\int_{\Theta_\alpha} \sqrt{I(\theta)}\, d\theta} \int_{\Theta_\alpha} P(x^n \mid \theta) \sqrt{I(\theta)}\, d\theta. \qquad \mathbf{3}$$

If most of the mass of $P(x^n \mid \theta)\sqrt{I(\theta)}$ lies in Θ_α, the integral **3** may be approximated by

$$\frac{1}{\int_{\Theta_\alpha} \sqrt{I(\theta)}\, d\theta} \int_{\Theta} P(x^n \mid \theta) \sqrt{I(\theta)}\, d\theta, \qquad \mathbf{4}$$

were the limits of integration of $P(x^n \mid \theta)\sqrt{I(\theta)}$ have changed from Θ_α to Θ. But then we may follow the derivation in reference (30) of the main text to find that the model length is approximately

$$\frac{k}{2} \log_2 \frac{n}{2\pi} + \int_{\Theta_\alpha} \sqrt{I(\theta)}\, d\theta$$

Thus, we make the following assumption

Assumption 4 *The approximate change in the description length of the model for a fixed n, as the pseudocount parameter α increases from β to γ, is*

$$-\log_2 \int_{\Theta_\gamma} \sqrt{I(\theta)}\, d\theta + \log_2 \int_{\Theta_\beta} \sqrt{I(\theta)}\, d\theta.$$

The Error in the Description Length of the Data

When we apply pseudocounts with pseudocount parameter α, we wish to model the data using the target frequencies $\mathbf{q}$, related to the observed frequencies $\mathbf{f}$, by the formula

$$\mathbf{q} = [\alpha M + (1 - \alpha) I]\, \mathbf{f}. \qquad \mathbf{5}$$

MDL theory suggests that the description length of the model should grow asymptotically as $\frac{19}{2} \log_2(n)$ bits, and this is less than the $19 \log_2(n)$ bits that are required to describe exactly

the observed counts of amino acids in n independent observations. Thus, when describing the model, one may describe $\mathbf{q}$ only approximately.

Therefore, for target frequencies $\mathbf{q}$, the description length of the data is approximately

$$nH(\mathbf{f}) + nD(\mathbf{f}\,||\,\mathbf{g}), \qquad \mathbf{6}$$

where $\mathbf{g}$ is one of the roughly $n^{19/2}$ frequency vectors that may be used to describe the model. The expression **6** is an approximation in the sense that the description length of any particular encoding of the data is only approximated by a calculation using the relative entropy. The expression is exact in that $D(\mathbf{f}\,||\,\mathbf{g})$ is the correct relative entropy.

The density in Θ of the probability vectors $\mathbf{g}$ that may be represented exactly is a discretization of Jeffrey's prior. One may show that $\mathbf{g}$ may be chosen so that

$$\max_i |q_i - g_i| = \mathcal{O}(1/\sqrt{n}).$$

The notation $\mathcal{O}(1/\sqrt{n})$ indicates a function that converges to zero no more slowly than a constant times $1/\sqrt{n}$, as n tends to infinity.

Because one must use $\mathbf{g}$ rather than $\mathbf{q}$ to encode the data, there is an error in the estimate $nD(\mathbf{f}\,||\,\mathbf{q})$ of the description length of the data. Specifically,

$$\begin{aligned} nD(\mathbf{f}\,||\,\mathbf{g}) - nD(\mathbf{f}\,||\,\mathbf{q}) &= n\sum_{i=1}^{20} f_i \log_2(f_i/g_i) - n\sum_{i=1}^{20} f_i \log_2(f_i/q_i) \\ &= n\sum_{i=1}^{20} f_i \log_2(g_i/q_i) \\ &= \frac{n}{\log_2(e)} \sum_{i=1}^{20} \frac{f_i}{q_i}(g_i - q_i) + n\mathcal{O}(1/n), \end{aligned}$$

where we have used a first-order series expansion of $\log_2(x)$ about q_i.

For any nonzero f_i,

$$\begin{aligned} \frac{f_i}{q_i} &= \frac{f_i}{\sum_{j=1}^{20} \left((1-\alpha)\delta_{ij} + \alpha m_{ij}\right) f_j} \\ &= 1 - \frac{\alpha}{(1-\alpha) f_i} \sum_{j=1}^{20} m_{ij} f_j + \mathcal{O}\left(\frac{\alpha^2}{(1-\alpha)^2 f_i^2}\right). \end{aligned}$$

Because a small number of pseudocounts are applied to the observed frequencies before the optimal pseudocounts are calculated, f_i is bounded away from zero. Therefore, for small α,

$$\frac{f_i}{q_i} = 1 + \mathcal{O}(\alpha).$$

We therefore find that

$$nD(\mathbf{f}\,||\,\mathbf{g}) - nD(\mathbf{f}\,||\,\mathbf{q}) = n\mathcal{O}(\alpha) \times \mathcal{O}(1/\sqrt{n}) + n\mathcal{O}(1/n).$$

So if the error in using $nD(\mathbf{f}\,||\,\mathbf{q})$ as the change in the description length of the data is not to grow with increasing n, then α must converge to zero at a rate of at least $\mathcal{O}(1/\sqrt{n})$.

Supplemental Data C: $D(\mathbf{f}\,||\,\mathbf{q})$ is nondecreasing in α

Let $\mathbf{f}$ and $\mathbf{q}$ be related by the formula **5**, and let

$$r(\alpha) = D(\mathbf{f}\,||\,\mathbf{q}) = \sum_{i=1}^{20} f_i \log(f_i/q_i).$$

Assume $f_i > 0$ for all i. The case in which some $f_i = 0$ can be handled by omitting those terms from the sum, as they do not contribute to $r(\alpha)$. We may write $r'(\alpha)$, the derivative of $r(\alpha)$ with respect to α, as the sum

$$r'(\alpha) = -\sum_{i=1}^{20} f_i \frac{\sum_{j=1}^{20}(M_{ij} - \delta_{ij})f_j}{\sum_{j=1}^{20}(\alpha M_{ij} + (1-\alpha)\delta_{ij})f_j} = -\sum_{i=1}^{20} \frac{w_i}{1 + \alpha(w_i/f_i)},$$

where $w_i = \sum_{j=1}^{20}(M_{ij} - \delta_{ij})f_j$. Note that

$$r'(0) = -\sum_{i=1}^{20} w_i = \sum_{i=1}^{20}\sum_{j=1}^{20} M_{ij}f_j - \sum_{i=1}^{20} f_i = 0.$$

The second derivative of $r(\alpha)$ with respect to α,

$$r''(\alpha) = \sum_{i=1}^{20} \frac{w_i^2/f_i}{(1 + \alpha(w_i/f_i))^2},$$

is nonnegative for all $\alpha > 0$. Therefore, $r(\alpha)$ is strictly increasing for α in 0 to 1, unless $w_i = 0$ for all i, in which case $r(\alpha)$ is identically zero. However, if all w_i are zero, then $\mathbf{f} = M\mathbf{f}$ and the frequencies $\mathbf{f}$ are exactly equal to background frequencies implicit in M.

Supplemental Data D: Small α limit

We wish to show that the approximate decrease in the model description length, given by the difference

$$-\log_2 \int_{\Theta_\alpha} \sqrt{I(\theta)}\, d\theta + \log_2 \int_{\Theta_0} \sqrt{I(\theta)}\, d\theta, \qquad \textbf{7}$$

behaves as a constant times $\sqrt{\alpha}$ when α is small. Equivalently, we wish to show that the derivative of **7** with respect to $\gamma \equiv \sqrt{\alpha}$ is finite and positive at $\gamma = 0$. By the rules of differentiation of the logarithm, it suffices to show that

$$\frac{d}{d\gamma} \int\limits_{\Theta_{\alpha=\gamma^2}} \sqrt{I(\theta)}\, d\theta \Bigg|_{\gamma=0}$$

is finite and negative.

A point θ in Θ_α has 19 degrees of freedom, which may be represented by the 19 components θ_i for $i = 1, \ldots 19$. To simplify notation, we write $\theta_{20} = 1 - \sum_{i=1}^{19} \theta_i$, but note that $d\theta = \prod_{i=1}^{19} d\theta_i$.

For each θ in Θ_0, let

$$y_i = \sum_{j=1}^{20} [(1-\alpha)\delta_{ij} + \alpha m_{ij}]\theta_j, \qquad \textbf{8}$$

for $i = 1, \ldots, 19$. These 19 equations define a differentiable path from each point θ in Θ_0 to a point $\mathbf{y}$ in Θ_α. As with θ, we simplify notation by writing $y_{20} = 1 - \sum_{i=1}^{19} y_i$. Because the columns of M must sum to one, it follows that

$$y_{20} = \sum_{j=1}^{20} [(1-\alpha)\delta_{20,j} + \alpha m_{20,j}]\theta_j.$$

By the fundamental theorem of calculus

$$\frac{d}{d\gamma} \int\limits_{\Theta_{\alpha=\gamma^2}} \sqrt{I(\theta)}\, d\theta \Bigg|_{\gamma=0} = \oint\limits_{\partial\Theta_0} \sqrt{I(\theta)}\, \frac{d\mathbf{y}}{d\gamma} \cdot \hat{n}\, dS,$$

where $\hat{n}$ indicates the outward facing normal to the surface $\partial\Theta_0$ at any value of θ. The surface $\partial\Theta_0$ may be divided into 20 pieces, which we denote $\mathcal{T}_k = \{\theta \mid \theta_k = 0 \text{ and } \theta \in \Theta_0\}$ for $k = 1, \ldots, 20$.

Let us consider the case of $k = 1$, which is notationally simplest. The cases of $k = 2, \ldots, 19$ are similar. The case of $k = 20$ is treated separately below. On the surface $\mathcal{T}_1$,

$$\hat{n}\, dS = -\hat{e}_1 \prod_{j=2}^{19} d\theta_j,$$

where $\hat{e}_1$ is a unit vector along the first coordinate. Recalling that $\alpha = \gamma^2$, one may differentiate **8** to find

$$\frac{dy_i}{d\gamma} = 2\gamma \sum_{j=2}^{20} (m_{ij} - \delta_{ij})\theta_j.$$

Thus, near the surface of $\mathcal{T}_1$,

$$\frac{d\mathbf{y}}{d\gamma} \cdot \hat{n}\, dS = -2\gamma \sum_{j=2}^{20} m_{1j}\theta_j \prod_{j=2}^{19} d\theta_j.$$

If $\mathbf{y}$ is the image of a point θ on $\mathcal{T}_1$, then

$$\sqrt{I(\mathbf{y})} = y_1^{-1/2} \prod_{i=2}^{20} y_i^{-1/2} = \gamma^{-1} \left(\sum_{j=2}^{20} m_{1j}\theta_j \right)^{-1/2} \prod_{i=2}^{20} y_i^{-1/2}.$$

The surface integral over $\mathcal{T}_1$ is

$$\int_{\mathcal{T}_1} \lim_{\gamma \to 0} \sqrt{I(\mathbf{y})}\, \frac{d\mathbf{y}}{d\gamma} \cdot \hat{n}\, dS = -2 \int_{0 \le \theta_i \le 1} \cdots \int \left(\sum_{j=2}^{20} m_{1j}\theta_j \right)^{1/2} \prod_{i=2}^{20} \theta_i^{-1/2} \prod_{j=2}^{19} d\theta_j. \qquad \mathbf{9}$$

The sum $\sum_{j=2}^{20} m_{1j}\theta_j$ is strictly positive and bounded above by one. Thus the integrand in **9** is strictly negative, and the surface integral over $\mathcal{T}_1$ is bounded below by

$$-2 \int_{0 \le \theta_i \le 1} \cdots \int \frac{\prod_{i=2}^{19} \theta_i^{-1/2}}{\sqrt{1 - \sum_{i=2}^{19} \theta_k}} \prod_{i=2}^{19} d\theta_i.$$

This integral is recognizable as -2 times the integral of the square root of the Fisher information for 18 parameters, which is known to be positive and finite. Therefore, the surface integral over $\mathcal{T}_1$ is a finite, negative number. Similar arguments show that the surface integrals over $\mathcal{T}_k$ for $k = 2, \ldots, 19$ are finite negative numbers.

The surface $\mathcal{T}_{20}$ is the set of points for which $\sum_{j=1}^{19} \theta_j = 1$. On this surface, the values of any 18 variables determine the value of the remaining variable. Therefore, we may parameterize $\mathcal{T}_{20}$ in terms of θ_j for $j = 1, \ldots, 18$. With this parameterization

$$\hat{n}\, dS = \vec{e} \prod_{j=1}^{18} d\theta_j,$$

where $\vec{e}$ is the vector of length $\sqrt{19}$ with every component equal to 1. Therefore, on $\mathcal{T}_{20}$,

$$\frac{d\mathbf{y}}{d\gamma} \cdot \hat{n}\, dS = 2\gamma \sum_{i=1}^{19} \sum_{j=1}^{19} (m_{ij} - \delta_{ij})\theta_j \prod_{j=1}^{18} d\theta_j.$$

But

$$\sum_{i=1}^{19} \sum_{j=1}^{19} (m_{ij} - \delta_{ij})\theta_j = -1 + \sum_{i=1}^{19} \sum_{j=1}^{19} m_{ij}\theta_j = -\sum_{j=1}^{19} m_{20,j}\theta_j,$$

and so

$$\frac{d\mathbf{y}}{d\gamma} \cdot \hat{n}\, dS = -2\gamma \sum_{j=1}^{19} m_{20,j}\theta_j.$$

The rest of the argument that the surface integral over $\mathcal{T}_{20}$ is negative and finite is symmetric with the argument for $\mathcal{T}_1$.