OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

The Construction and Use of Log-Odds Substitution Scores for Multiple Sequence Alignment

Stephen F. Altschul[1]*, John C. Wootton[1], Elena Zaslavsky[2], Yi-Kuo Yu[1]

1 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Center for Translational Systems Biology and Department of Neurology, Mount Sinai School of Medicine, New York, New York, United States of America

Abstract

Most pairwise and multiple sequence alignment programs seek alignments with optimal scores. Central to defining such scores is selecting a set of substitution scores for aligned amino acids or nucleotides. For local pairwise alignment, substitution scores are implicitly of log-odds form. We now extend the log-odds formalism to multiple alignments, using Bayesian methods to construct "BILD" ("Bayesian Integral Log-odds") substitution scores from prior distributions describing columns of related letters. This approach has been used previously only to define scores for aligning individual sequences to sequence profiles, but it has much broader applicability. We describe how to calculate BILD scores efficiently, and illustrate their uses in Gibbs sampling optimization procedures, gapped alignment, and the construction of hidden Markov model profiles. BILD scores enable automated selection of optimal motif and domain model widths, and can inform the decision of whether to include a sequence in a multiple alignment, and the selection of insertion and deletion locations. Other applications include the classification of related sequences into subfamilies, and the definition of profile-profile alignment scores. Although a fully realized multiple alignment program must rely upon more than substitution scores, many existing multiple alignment programs can be modified to employ BILD scores. We illustrate how simple BILD score based strategies can enhance the recognition of DNA binding domains, including the Api-AP2 domain in *Toxoplasma gondii* and *Plasmodium falciparum*.

Citation: Altschul SF, Wootton JC, Zaslavsky E, Yu Y-K (2010) The Construction and Use of Log-Odds Substitution Scores for Multiple Sequence Alignment. PLoS Comput Biol 6(7): e1000852. doi:10.1371/journal.pcbi.1000852

Editor: Adam Siepel, Cornell University, United States of America

Received October 30, 2009; **Accepted** June 3, 2010; **Published** July 15, 2010

Funding: This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (SFA, JCW, Y-KY), and NIH NIAID contract HHSN266200500021C (EZ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

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

Introduction

Protein and DNA sequence alignment is a fundamental tool of computational molecular biology. It is used for functional prediction, genome annotation, the discovery of functional elements and motifs, homology-based structure prediction and modeling, phylogenetic reconstruction, and in numerous other applications. The effectiveness of alignment programs depends crucially upon the scoring systems they employ to evaluate possible alignments. For pairwise alignments, scores typically are defined as the sum of "substitution scores" for aligning pairs of letters (amino acids or nucleotides), and "gap scores" for aligning letters in one sequence with null characters between letters in the other. Substitution and gap scores may be generalized to multiple alignments, i.e. those involving three or more sequences.

Most useful local pairwise alignment algorithms allow gaps and explicitly assign them scores [1–4]. However, many local multiple alignment algorithms do not allow gaps, or allow them only implicitly as spacers between distinct ungapped alignment blocks. Indeed the alignments recorded in some protein family databases are explicitly constructed with ungapped alignment blocks separated by variable length spacers [5], and it has been argued that this formalism corresponds well to the observed relationships imposed by protein structure [6]. Short ungapped blocks are also used in the DNA context, to represent, for example, transcription factor binding sites.

Many pairwise substitution scores have been developed for protein [7–20] and DNA [21,22] sequence comparison, and a statistical theory for substitution scores has been developed for local alignments without gaps [23,24]. It is not trivial to generalize pairwise scoring systems to multiple alignments, and the following four principal approaches have been proposed to this long-standing problem: A) *Tree scores*. An evolutionary tree can be defined relating the sequences in question, with each sequence residing at one leaf of the tree. By reconstructing letters at the internal nodes of the tree, the score for an aligned column of letters is defined as the sum of pairwise substitution scores for all edges of the tree [25,26]. B) *Star scores*. As a special case of tree-scores, a single "consensus" letter can be defined for an alignment column. The column score is defined as the sum of pairwise scores for the consensus letter to each letter in the column. The tree in question reduces to a star, with the consensus at the central node. C) *Sum-of-the-Pairs or SP scores*. A column score can be constructed as the sum of substitution scores for all pairs of letters in the column [27,28]. D) *Entropy scores*. Scores can be based on the entropy of the letter frequencies observed in a column [29]; these scores have become particularly popular for DNA alignments. All these approaches are open to refinement, for example by weighting the pairwise scores of the sequences involved.

All reasonable substitution scores for pairwise local alignment are implicitly log-odds scores [23,30], which compare the

Author Summary

Multiple sequence alignment is a fundamental tool of biological research, widely used to identify important regions of DNA or protein molecules, to infer their biological functions, to reconstruct ancestries, and in numerous other applications. The effectiveness and accuracy of sequence comparison programs depends crucially upon the quality of the scoring systems they use to measure sequence similarity. To compare pairs of DNA or protein sequences, the best strategy for constructing similarity measures has long been understood, but there has been a lack of consensus about how to measure similarity among multiple (i.e. more than two) sequences. In this paper, we describe a natural generalization to multiple alignment of the accepted measure of pairwise similarity. A large variety of methods that are used to compare and analyze DNA or protein molecules, or to model protein domain families, could be rendered more sensitive and precise by adopting this similarity measure. We illustrate how our measure can enhance the recognition of important DNA binding domains.

probabilities of aligning two letters under models of relatedness and non-relatedness, and the most popular are explicitly so constructed [7,8,14]. We argue that multiple alignment column scores should be similarly constructed, based upon explicit target frequency predictions for columns from accurate alignments of related sequences. For this purpose, we propose, the method with the strongest theoretical foundation relies upon the specification of a Bayesian prior, over the space of multinomial distributions for describing alignment columns representing true biological relationships [31,32]. We call column scores based on such a formalism "Bayesian Integral Log-odds" or BILD scores. Although these scores are implicit in earlier work, their full generality and utility has not been recognized. They may be calculated efficiently, and may be generalized to allow for the differential weighting of sequences in a multiple alignment. We also consider an alternative approach that allows log-odds column scores to be derived from any pairwise substitution matrix.

Given their form, multiple alignment log-odds scores can be used directly to define the proper extent of multiple alignment blocks, and to derive natural scores for profile-profile comparison. We show that they also arise from the perspective of the Minimum Description Length Principle [33], which allows them to be combined naturally with other information theoretic measures. Other direct applications are specifying when a sequence should be included in a multiple alignment at all, and when an alignment of many related sequences is better split into several alignments each involving fewer sequences.

Efficient methods for calculating BILD scores allow them to be incorporated into Gibbs sampling algorithms for ungapped local multiple alignment. Most practical protein applications, however, require provisions for gaps. We describe two methods for extending an ungapped local multiple alignment produced by the Gibbs sampling strategy to a gapped alignment, the first using asymmetric affine gap costs, and the second hidden Markov models. In the latter, column BILD scores inform the construction of position-specific gap costs, and yield gapped alignments in greater conformity with considerations of protein structure. We illustrate the applications of the programs by using them to uncover previously undescribed Api-AP2 domains of *Toxoplasma gondii* and *Plasmodium falciparum*.

Multiple sequence alignment comprises a diverse set of problems and approaches. Many sophisticated statistical inference techniques have been applied to the multiple alignment problem and to the related problem of phylogenetic reconstruction, e.g. [34–37]. It is not our purpose here to develop a new multiple alignment program. Rather, we seek only to argue that the "substitution scores" for multiple alignment columns which lie at the core of most multiple alignment methods can in many cases be improved. Although many statistical alignment methods are Bayesian-based, the BILD scores directly implied by Bayesian reasoning have been heretofore unrecognized.

Methods

Multiple Alignment Log-Odd Scores

Log-odds pairwise substitution scores can be written $s_{i,j}=\log\left(q_{i,j}/p_ip_j\right)$. Here, $q_{i,j}$ is the frequency with which residues i and j correspond in accurate alignments of related sequences, and p_i is the background probability with which residue i occurs. The base of the logarithm is arbitrary, and merely defines a scale for the scoring system. We henceforth assume that unless the natural logarithm is specified, all logarithms are base 2, and the resulting scores are therefore in the units of bits [30]. Note that no target frequencies $q_{i,j}$ are uniquely optimal for pairwise sequence alignment, because different $q_{i,j}$ are appropriate for comparing sequences diverged by different amounts of evolution [7,8,13,30]. This perception gives rise to families of substitution matrices, such as the PAM [7,8] and BLOSUM [14] series for protein comparison.

To generalize log-odds scores to multiple alignments, we first develop some notation. We consider the alphabet A from which the letters in our sequences are drawn to consist of L elements, which for convenience we represent by the numbers 1 through L. An ungapped column from a multiple alignment of M sequences is a vector $\vec{x}$, each of whose components x_1 through x_M takes on a value in A. In essence, the log-odds approach compares two theories, one in which all the letters aligned are related or homologous, and the other in which none are. Each theory implies a probability for observing any given set of data. For the alignment column $\vec{x}$, we define $Q(\vec{x})$ as the probability of observing the data under the assumption of relatedness, and $P(\vec{x})$ under the assumption of non-relatedness. Then the log-odds score for this column is defined as

$$S(\vec{x})=\log\frac{Q(\vec{x})}{P(\vec{x})}. \tag{1}$$

Assuming background probabilities p_1 through p_L for the various letters, $P(\vec{x})$ is given simply by

$$P(\vec{x})=\prod_{k=1}^{M}p_{x_k}. \tag{2}$$

We will consider one primary strategy for deriving $Q(\vec{x})$. As with pairwise scores, all sets of multiple alignment column scores with negative expected value are implicitly log-odds scores [23,30]. However, unless their values for $Q(\vec{x})$ are explicitly constructed in a sensible way, log-odds scores are unlikely to perform well in the applications suggested below.

For alignments of more than two sequences, there are of course other possibilities than for all or none of the sequences to be related. However, as we will describe below, scores of the form of equation (1) can be applied to the comparison of sequences where

only a subset are related, by adding indicator variables to include or exclude sequences.

Log-odds scores S for alignment columns immediately suggest substitution scores R for aligning two different columns of letters. Specifically, letting $\overrightarrow{xy}$ be the concatenation of the vectors $\vec{x}$ and $\vec{y}$, define

$$R(\vec{x},\vec{y}) = S(\overrightarrow{xy}) - S(\vec{x}) - S(\vec{y}) = \log \frac{Q(\overrightarrow{xy})}{Q(\vec{x})Q(\vec{y})}. \quad (3)$$

These column-column alignment scores may be used consistently in progressive alignment algorithms, which proceed by aligning the most closely related sequences first [38,39], although as will be discussed below problems may arise in the definition of gap scores. They may also be used for profile-profile alignment, a topic of considerable recent interest [40–48].

BILD Scores

For multiple alignments, perhaps the best approach to defining and calculating $Q(\vec{x})$ is a Bayesian one [31,32]. (An alternative approach, based on pairwise scoring matrices, is described in Text S1.) Assume that the letters in a specific column from an accurate alignment of related sequences are generated independently, but with probabilities q_1 through q_L that in general differ from the background probabilities. Assume further that it is possible to assign a prior probability distribution Θ_0 to the multinomial distributions $\vec{q}$ associated with columns of related letters. This prior Θ_0 can be derived from a detailed study of related protein or DNA sequences.

Although the data $\vec{x}$ associated with a specific column generally have no temporal or other privileged order, assume for convenience that they are observed sequentially, in the order x_1 to x_M. Then we may apply Bayes' theorem to transform the prior distribution Θ_0 to a posterior Θ_1, after the observation of x_1. More generally, each subsequent observation x_k can be seen to transform the prior Θ_{k-1} into a posterior distribution Θ_k. We may then use the chain rule to write

$$Q(\vec{x}) = \mathrm{Prob}(\vec{x}|\Theta_0) = \prod_{k=1}^{M} \mathrm{Prob}(x_k|\Theta_{k-1}). \quad (4)$$

The individual terms in this product may be calculated by integrating over all possible multinomial distributions $\vec{q}$:

$$\mathrm{Prob}(x_k|\Theta_{k-1}) = \int q_{x_k} \Theta_{k-1}(\vec{q})\, d\vec{q}. \quad (5)$$

Finally, combining equations (1), (2) and (4) yields

$$S(\vec{x}) = \sum_{k=1}^{M} \log \frac{\mathrm{Prob}(x_k|\Theta_{k-1})}{p_{x_k}}. \quad (6)$$

We call scores defined in this way Bayesian Integral Log-odds or BILD scores. They can be understood simply as the sum of log-odds scores for the individual letters observed in a column, with the "target frequency" for each letter x_k calculated based upon the prior distribution Θ_0, and the "previously observed" letters x_1 through x_{k-1}. Even though, by this formula, the log-odds score for a letter varies with its position in the column, the total column score is nevertheless invariant under permutation of the column's letters.

BILD scores have some conceptual connections to star- and entropy-based multiple alignment scoring systems. The simplest generalization of star scores imposes a prior probability distribution on the consensus letter, but still assumes a probabilistic pairwise substitution model. As we describe in Text S1, this yields a class of log-odds scores we call MELD scores. BILD scores arise, in contrast, by thinking of the "consensus" not as an ancestral letter, but rather as a generative probabilistic model, and by integrating over a prior distribution placed on this model.

Given observed and background letter distributions $\vec{q}$ and $\vec{p}$, entropy scores have been defined variously, and conceptually distinctly, as: i) $\sum p_j \log 1/p_j - \sum q_j \log 1/q_j$, the entropy difference between $\vec{p}$ and $\vec{q}$; ii) $\log L - \sum q_j \log 1/q_j$, the entropy difference between a uniform distribution on L letters and $\vec{q}$; and iii) $\sum q_j \log q_j/p_j$, the relative entropy of $\vec{q}$ and $\vec{p}$. Definitions i) and ii) differ only by a constant. One may refine any of these definitions by taking $\vec{q}$ to be a posterior letter distribution, derived from a prior and a set of observations. Both BILD and entropy-based scores can be viewed as the sum of scores derived from the probabilities for individual observations. The central distinction is that BILD scores estimate the probability for a given such observation using only "earlier" ones, whereas entropy scores estimate this probability using the complete collection of observations.

Dirichlet Distributions

Although the definition of BILD scores is valid for any prior distribution Θ_0 one wishes to specify, it is in general impractical to calculate the Θ_k, or the integral in equation (5), except when Θ_0 takes the form of a Dirichlet distribution [49], or a mixture of a finite number of Dirichlet distributions [31,32]. In this case, as described below, all the Θ_k are also Dirichlet distributions, or Dirichlet mixtures, and $\mathrm{Prob}(x_k|\Theta_{k-1})$ is easily calculated. Therefore, for mathematical as opposed to biological reasons, we always assume that BILD scores are defined using a Dirichlet or Dirichlet mixture prior. The family of Dirichlet mixtures, however, is rich enough that it can capture well much relevant prior knowledge concerning relationships among the various amino acids or nucleotides.

We review here the essentials of Dirichlet distributions. A multinomial distribution on L letters is specified by an L-dimensional vector $\vec{q}$, within the simplex defined by $0 \le q_j \le 1$, and $\sum_{j=1}^{L} q_j = 1$. The requirement that the q_j sum to 1 renders the space of multinomials $L-1$ dimensional. A Dirichlet distribution, defined over this space, is parametrized by an L-dimensional vector $\vec{\alpha}$ with all α_j positive. We shall sometimes refer to such a distribution by its parameters $\vec{\alpha}$, and we define α^* as the sum of the α_j. The Dirichlet distribution $\vec{\alpha}$ is given by the probability density function

$$\rho(\vec{q}) = \mathcal{Z} \prod_{j=1}^{L} q_j^{\alpha_j - 1}, \quad (7)$$

where the normalizing scalar $\mathcal{Z} = \Gamma(\alpha^*) \Big/ \Pi_{j=1}^{L} \Gamma(\alpha_j)$ ensures that integrating ρ over its domain yields 1. Here $\Gamma(x) \equiv \int_0^\infty t^{x-1} e^{-t} dt$, is the Gamma function, and $\Gamma(n) = (n-1)!$ for positive integral n. The uniform density is a special case that arises when all the α_j are 1.

Dirichlet distributions have two convenient properties. First, the expected frequency of letter a implied by $\vec{\alpha}$ is α_a/α^*. Second, the posterior distribution yielded by Bayes' theorem, after the observation of the letter a, is a Dirichlet distribution $\overrightarrow{\alpha'}$ with $\alpha'_a = \alpha_a + 1$, but with all other parameters equal to those of $\vec{\alpha}$.

To illustrate how to calculate BILD scores using these properties, consider the case of DNA comparison (with the numbers 1 through 4 identified respectively with the nucleotides A, C, G and T), with uniform background probabilities $p_j = 0.25$, and a Dirichlet prior Θ_0 given by the parameter vector (1,1,1,1). By equation (4), the target frequency Q associated with the alignment column "AATC" is given by $\text{Prob}[\text{A}|\Theta_0]\cdot\text{Prob}[\text{A}|\Theta_1]\cdot\text{Prob}[\text{T}|\Theta_2]\cdot\text{Prob}[\text{C}|\Theta_3] = \text{Prob}[\text{A}|(1,1,1,1)]\cdot\text{Prob}[\text{A}|(2,1,1,1)]\cdot\text{Prob}[\text{T}|(3,1,1,1)]\cdot\text{Prob}[\text{C}|(3,1,1,2)] = \frac{1}{4}\cdot\frac{2}{5}\cdot\frac{1}{6}\cdot\frac{1}{7} = \frac{1}{420}$. Thus the score for the column is $-\log 420 + \log 256 = -0.714$ bits. In contrast, for the column "AAAC", $Q = \frac{1}{4}\cdot\frac{2}{5}\cdot\frac{3}{6}\cdot\frac{1}{7}$, and the score for this column is $-\log 140 + \log 256 = 0.871$ bits.

The essence of a Dirichlet distribution is perhaps best understood through the alternative parametrization $(\vec{\beta}; \beta^*)$, where $\beta_j = \alpha_j/\alpha^*$, and $\beta^* = \alpha^*$. Because the β_j must sum to 1, there are still only L independent parameters. The vector $\vec{\beta}$ describes the center of mass of the distribution, while β^* indicates how concentrated the distribution is about this point. Large values of β^* correspond to distributions with most of their mass near $\vec{\beta}$, whereas values of β^* near 0 correspond to distributions with most of their mass near the boundaries of the simplex. It is frequently sensible, although not necessary, to choose a prior Θ_0 whose $\vec{\beta}$ is identical to the background frequencies $\vec{p}$. In this case, $\text{Prob}(x_1|\Theta_0) = p_{x_1}$, and the first summand in equation (6) is always 0. In other words, no letter in a column, considered in isolation, carries any information as to whether the column represents a true biological relationship.

Dirichlet Mixtures

Single Dirichlet distributions frequently are adequate for capturing prior knowledge concerning "true" alignment columns of related DNA sequences, but this is not the case for proteins. Most simply, distinct regions of multinomial space, representing different collections of amino acids, should have high prior probabilities. In order to address the deficiency of single Dirichlet distributions, Brown *et al*. [31] proposed the use of Dirichlet mixture priors. A Dirichlet mixture is simply the weighted sum of C distinct Dirichlet distributions. It is specified by C positive "mixture parameters" m_1 through m_C that sum to 1, and a set of L standard Dirichlet parameters, $\alpha_{i,1}$ through $\alpha_{i,L}$, for each of the C component Dirichlet distributions. (It will be useful later to define α_i^* as $\sum_{j=1}^{L}\alpha_{i,j}$.) In all, because of the restriction on the sum of the m_i, a Dirichlet mixture has $C(L+1)-1$ independent parameters. The Dirichlet components of a mixture generally are thought of as describing various types of positions (e.g. hydrophobic, charged, aromatic) typically found in proteins.

Bayes' theorem implies that, given a C-component Dirichlet mixture as a prior, the posterior distribution after the observation of a single letter is also a C-component Dirichlet mixture [31,32]. Brown *et al*. [31] proposed Dirichlet mixture priors in the context of deriving "substitution" scores for aligning amino acids to columns from a multiple protein sequence alignment. This restricted context can be understood as comprehending a single summand from equation (6). BILD scores extend Brown *et al*.'s sequence-profile alignment scores to comprehensive scores for multiple alignment columns.

Generalizing the development above, we describe here how to calculate the probability of a particular observation a given a Dirichlet mixture prior Θ_{k-1}, and how to calculate the posterior Θ_k resulting from this observation. First, given a Dirichlet mixture, with parameters m_i and $\alpha_{i,j}$, the probability of observing letter a is given simply by

$$\text{Prob}(a) = \sum_{i=1}^{C} m_i \frac{\alpha_{i,a}}{\alpha_i^*}, \tag{8}$$

which follows directly from the definition of Dirichlet mixtures, and the result for single Dirichlet distributions. Second, given the observation of letter a, and a Dirichlet mixture prior parametrized as above, the parameters m_i' and $\alpha_{i,j}'$ of the posterior distribution may be calculated as follows:

$$\begin{aligned}
&\text{(i) For } i \text{ from 1 to } C, \text{ define } \tilde{m}_i := m_i \frac{\alpha_{i,a}}{\alpha_i^*};\\
&\text{(ii) For } i \text{ from 1 to } C, \text{ define } m_i' := \frac{\tilde{m}_i}{\sum_{i=1}^{C}\tilde{m}_i};\\
&\text{(iii) For } i \text{ from 1 to } C \text{ and } j \text{ from 1 to } L, \text{ define}\\
&\quad \alpha_{i,j}' := \alpha_{i,j} + 1 \text{ if } j = a,\\
&\quad \text{and } \alpha_{i,j}' := \alpha_{i,j} \text{ otherwise.}
\end{aligned} \tag{9}$$

In short, first multiply the mixture parameters m_i by the Bayesian factors $\alpha_{i,a}/\alpha_i^*$ and normalize, and then add 1 to each $\alpha_{i,a}$. Mathematics establishing the validity of this procedure appears in [32]. Their development is more complex than we require here, because we modify the Dirichlet mixture parameters only one observation at a time. We note that given the m_i' and $\alpha_{i,j}'$, it is simple to invert procedure (9) to determine the m_i and $\alpha_{i,j}$. This is useful for applications such as the Gibbs sampling algorithm discussed below.

Many multiple alignment problems involve subsets of sequences that are much more closely related to one another than to the other sequences being considered, and this may yield suboptimal results, because a large number of closely related sequences can "outvote" a few more divergent sequences. One remedy has been to assign each sequence a numerical weight, with closely related sequences down-weighted [50–61]. Also, subsumed in such weights may be the recognition that the total number of effective observations represented by an alignment column may be smaller than the number of sequences it comprehends [4,62,63]. Thus, for certain applications it may be desirable to generalize BILD scores to weighted sequences. To do so, we need to define the concept of the probability of a "fractional observation" of a letter, and describe as well how a posterior distribution is calculated after such a fractional observation. Arguments supporting how this may be done can be extracted from the mathematical development in [32]. Both equation (8) and the first step of procedure (9) involve multiplication by the factors $\alpha_{i,a}/\alpha_i^*$. For the fraction Δ of an observation of letter a, these factors must be replaced by the alternative factors

$$f_i(\Delta) = \frac{\Gamma(\alpha_{i,a}+\Delta)}{\Gamma(\alpha_{i,a})}\frac{\Gamma(\alpha_i^*)}{\Gamma(\alpha_i^*+\Delta)}. \tag{10}$$

Also, in the last step of procedure (9), the quantity Δ rather than 1 must be added to each $\alpha_{i,a}$. The factors $f_i(\Delta)$ are identical to the original factors when $\Delta = 1$, and all $f_i(\Delta)$ approach 1 as Δ approaches 0, as some reflection shows they must.

Finally, note that equation (10) may be applied to $\Delta > 1$ as well as $\Delta \leq 1$, and may be useful even when all observations are unitary. Thus, by aggregating observations, the BILD score for a

Table 1. Dirichlet mixture priors for protein sequence comparison.

Name of prior	Name on UCSC website	Number of components	$\mathcal{D}_2$ (bits)	Equivalent PAM matrix	$\hat{\mathcal{D}}$ (bits)	Equivalent PAM matrix
Θ_0^A	uprior	9	1.44	80	2.85	20
Θ_0^B	byst	9	0.91	130	2.34	35
Θ_0^C	recode3	20	0.61	175	1.88	55
Θ_0^D	recode4	20	0.37	245	1.63	70
Θ_0^E	fournier	20	0.18	360	0.92	125

$\mathcal{D}_2$ is the relative entropy [30] of the pairwise substitution matrix implied by the Dirichlet mixture prior. $\hat{\mathcal{D}}$ is the mean relative entropy of the multinomial distribution (Text S2).

doi:10.1371/journal.pcbi.1000852.t001

column containing L' unique letters may be calculated with L' summands, rather than the M summands of equation (6). For a single Dirichlet prior, $Q(\vec{x})$ reduces to the simple formula

$$Q(\vec{x}) = \frac{\Gamma(\alpha^*)}{\Gamma(\alpha^* + c^*)} \prod_{j=1}^{L} \frac{\Gamma(\alpha_j + c_j)}{\Gamma(\alpha_j)}, \qquad (11)$$

where c_j is the count of letter j, and c^* is the total count of all residues. Only the numerator inside the product varies from column to column within an alignment, yielding further efficiency for calculation.

The Choice of Priors

Only the research team that first proposed Dirichlet mixtures for protein sequence comparison has derived, from analyses of large protein alignment collections, sets of Dirichlet mixture prior parameters [31,32]. Twelve such sets, involving various numbers of Dirichlet components, can currently be found at http://compbio.soe.ucsc.edu/dirichlets/index.html. We list five of these in Table 1, which we refer to as Θ_0^A through Θ_0^E.

Proteins diverged by different degrees of evolutionary change are best studied using pairwise substitution matrices with different relative entropies [30], and the analogous claim should hold for Dirichlet mixture priors. A Dirichlet mixture prior implies a background amino acid frequency distribution $\vec{p}$, as well as a symmetric pairwise substitution matrix, by means of the formula $s_{i,j} = \log\left[Q(i,j)/p_i p_j\right]$. The relative entropies $\mathcal{D}_2$ of the substitution matrices implicit in the priors Θ_0^A through Θ_0^E range from 1.44 bits, roughly equivalent to that of the PAM-80 matrix [7,8], which is appropriate for fairly close evolutionary relationships, to 0.18 bits, roughly equivalent to that of the PAM-360 matrix, which is appropriate only for extremely distant relationships (Table 1).

As well as $\mathcal{D}_2$, one may calculate the mean relative entropy $\hat{\mathcal{D}}$ of the multinomial distributions $\vec{q}$ described by a Dirichlet mixture prior to the background frequencies $\vec{p}$ (see Text S2). For Θ_0^A to Θ_0^E, $\hat{\mathcal{D}}$ ranges from 2.85 to 0.92 bits (Table 1). That $\hat{\mathcal{D}}$ has a much greater value than $\mathcal{D}_2$ indicates that on average much more information is available per position from an accurate multiple alignment of many related sequences than from a single sequence. We note that, in lieu of using different priors, the effective relative entropy of a particular Dirichlet mixture may be tuned by scaling the weights of the sequences to which it is applied [43].

Standard pairwise substitution matrices are constructed from sets of proteins with certain background amino acid frequencies $\vec{p}$, and are non-optimal for the comparison of proteins with compositions that differ greatly from $\vec{p}$ [64]. Similarly, a Dirichlet mixture prior has an implicit background amino acid composition $\vec{p}$, and should not be optimal when applied to proteins with compositions that differ greatly from $\vec{p}$. It is possible to adjust standard matrices for use with non-standard compositions [64,65], and we will discuss elsewhere an analogous strategy that can be applied to adjust Dirichlet mixture priors.

Single Dirichlet priors may be appropriate for DNA sequence comparison. The uniform density, arising when all $\alpha_j = 1$ ($\alpha^* = 4$), has frequently been advocated in the absence of prior knowledge, and "Jeffreys' prior" [66], which is uninformative in a deeper sense, corresponds to all $\alpha_j = 0.5$ ($\alpha^* = 2$) [33]. When specific prior knowledge concerning an application domain is available, however, there is generally not a strong argument for using uninformative priors. For related DNA sequences, the columns of accurate alignments are sometimes dominated by one or two nucleotides, suggesting that all α_j should be smaller than 1. Furthermore, it usually makes sense for the α_j to be proportional to the background frequencies p_j. If this is stipulated, the specification of a Dirichlet prior reduces to the specification of α^*. Assuming a uniform nucleotide composition, the values of $\mathcal{D}_2$ and $\hat{\mathcal{D}}$ implied by α^* from 0.5 to 4.0 are given in Table 2. An empirical study of transcription factor binding sites [67] concludes that, at least for the analysis of such sites, α^* should be 1 or lower.

Local Alignment Width and Local Multiple Alignment

A direct application of multiple alignment log-odds scores is to determining local alignment width. As formulated by Smith and Waterman [1], an optimal local alignment is one that maximizes an alignment score but is of arbitrary width. Such scores should fall on the log side of the "log-linear phase transition" [68], which

Table 2. Relative entropies for DNA sequence comparison.

α^*	$\mathcal{D}_2$ (bits)	$\hat{\mathcal{D}}$ (bits)
0.5	0.792	1.387
1.0	0.451	1.062
1.5	0.294	0.860
2.0	0.208	0.721
2.5	0.155	0.621
3.0	0.120	0.545
3.5	0.096	0.485
4.0	0.078	0.437

See footnote to Table 1.

doi:10.1371/journal.pcbi.1000852.t002

implies that for ungapped local alignments, substitution scores must be of log-odds form [23,30].

Equation (1) explicitly generalizes pairwise log-odds scores to the multiple alignment case. They are positive for some alignment columns, negative for others, and must have negative expected value. Therefore it is appropriate to define an optimal ungapped multiple alignment as one with maximal aggregate log-odds score. This immediately allows one to define the proper width or extent of an ungapped multiple DNA or protein alignment, without resorting to the *ad hoc* principles frequently required for other scoring systems [69]. Although the Smith-Waterman algorithm can be applied to optimize log-odds-scored local multiple alignments, it is too slow for most purposes. Nevertheless, once relative offsets have been fixed for a set of sequences, it is trivial to determine an optimal ungapped local multiple alignment along the single implied diagonal.

The ungapped local multiple alignment problem may be formulated as seeking segments of common width W within multiple DNA or protein sequences that, when aligned, optimize a defined objective function. We take this function here to be the aggregate log-odds score for the aligned columns. One way to approach this optimization is by means of a Gibbs sampling strategy, as described by Lawrence *et al*. [69]. Log-odds scores can be used to adjust W dynamically, by applying the Smith-Waterman algorithm to the diagonal implied by a provisional alignment, without the need for an arbitrary parameter or an *ad hoc* optimization. They may also be used to determine dynamically whether or not a sequence should participate in the multiple alignment at all, for which purpose it is useful first to consider log-odds scores from the perspective of the Minimum Description Length Principle.

Log-Odds Scores and the Minimum Description Length Principle

The Minimum Description Length (MDL) Principle provides a criterion for choosing among alternative theories for describing a set of data [33,49]. To simplify greatly, it suggests that given a set of alternative theories T_i to describe a set of data D, that theory should be chosen which minimizes DL_i, defined as the sum of $L(T_i)$, the description length of the theory, and $L(D|T_i)$, the description length of the data given the theory. By convention, description lengths are measured in bits.

From information theory [70], the information associated with an event of probability p is $-\log_2 p$ bits. Focusing on actual encoding schemes for probabilistic events can unduly complicate MDL analyses. Accordingly, we here follow the approach of section 3.2.2 of [33], in which description lengths are allowed to be non-integral, and are identified with negative log probabilities. Thus, if the data can be described probabilistically, $L(D|T_i) = -\log[Prob(D|T_i)]$. The length of the theory $L(T_i)$ is defined as the number of bits needed to specify the free parameters of T_i, i.e. those that are fitted to the data [33].

For local multiple alignment, the theory T_0 that the input sequences are unrelated has only the background probabilities $\vec{p}$ as parameters, whose description length we will call L_p. The data D is comprised of M sequences, with lengths N_1 through N_M, and consisting of the letters $y_{i,j}$. Then $DL_0 = L_p + L(D|T_0) = L_p - \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log p_{y_{i,j}}$. The theory T_1 states that segments of width W beginning at positions s_i within the various sequences are related, and that the probability of the data $\vec{x}_c$ within each column c of the implied alignment is $Q(\vec{x}_c)$; the probability of the rest of the data may be described with the background frequencies $\vec{p}$. The free parameters are $\vec{p}$, the vector of starting positions $\vec{s}$, and W. Each s_i may take on one of $N_i - W + 1$ values, so its description length is approximately $\log N_i$, if W is not too large compared to N_i. Thus, we have $L(T_1) = L_p + L_W + \sum_{i=1}^{M} \log N_i$, where L_W is the description length of W. (If all feasible widths are taken to be equally likely, L_W is just $\log[\min_i N_i]$. Other encodings have L_W grow slowly with W [33,49].) It is apparent that $L(D|T_1) = -\sum_{c=1}^{W} \log Q(\vec{x}_c) - \sum \log p_{y_{i,j}}$, where the latter sum is taken only over those letters not participating in the local multiple alignment. Everything simplifies when we consider the difference in the total description lengths of the two theories:

$$DL_0 - DL_1 = S - \sum_{i=1}^{M} \log N_i - L_W, \tag{12}$$

where $S = \sum_{c=1}^{W} \log \frac{Q(\vec{x}_c)}{P(\vec{x}_c)}$ is simply the log-odds score for the implied alignment. In other words, T_1 is preferred whenever S exceeds $\sum_{i=1}^{M} \log N_i + L_W$. As described in Text S3, this prescription is related to the statistical theory for ungapped local alignments [23].

To allow one or more sequences to be excluded from the multiple alignment, we consider not 2, but 2^M theories, distinguished by M binary indices I_i, which take on the value 1 to indicate that sequence i participates in the alignment, and 0 otherwise. These theories need not be *a priori* equally likely; if necessary, for i from 1 to M we can specify prior probabilities π_i that sequence i contains a segment related to segments in the other sequences. Let us consider the difference in the description lengths of two theories, T_0' and T_1', that differ only in their index I_i. Theory T_1' incurs the cost $-\log \pi_i$ for the prior probability that $I_i = 1$, and also requires describing the location of the related segment, which costs $\log N_i$ bits. In contrast, theory T_0' incurs only the cost $-\log(1-\pi_i)$, so T_1' costs $\delta_i = \log N_i + \log \frac{1-\pi_i}{\pi_i}$ more bits to describe than T_0'. Thus, for T_1' to be preferred, the log-odds score of the multiple alignment must increase by at least δ_i when the segment from the ith sequence is added. If π_i is close to 1, δ_i can be negative, and is $-\infty$ if $\pi_i = 1$. In short, the greater the prior probability that a given sequence contains a relevant segment, the lower the score of such a segment need be for inclusion in the alignment.

The change in the log-odds score with the addition of a segment from the ith sequence depends upon which other sequences, and which of their segments, participate in the alignment. Consequently, the values of the indicator variables I_i must be part of the larger optimization, and their selection can be readily incorporated into a Gibbs sampling algorithm. The MDL Principle can also be extended to the case where a single sequence may contain more than one copy of a pattern, and, as previously described [62,71,72] and discussed in Text S4, to the clustering of multiple alignments into subfamilies.

Gap Scores

Although our central concern is to define a new type of multiple alignment substitution score, many important applications require the construction of gapped multiple alignments, and these generally entail scores for insertions and deletions. Multiple alignment gap scores should be defined in a manner consistent with the substitution scores used [73], so we will consider what type gap scores might fruitfully be paired with BILD scores.

Just as the log-odds perspective places pairwise substitution scores in a probabilistic framework [7,8,23,30], so pairwise gap scores can be viewed as specifying probabilities for insertions and deletions within biologically accurate alignments [74–82]. For pairwise alignments, "affine" gap scores, of the form $-(a+bk)$ for a gap of length k [83–85], are those most commonly used [3,4], although more complex gap scores have frequently been proposed [86–89]. When there is an essential asymmetry between the sequences being aligned, differing scores may be assigned to gaps within the two sequences. Furthermore, when substitution and gap scores are properly integrated and both expressed in the units of bits, the two parameters of affine gap scores can be understood to specify jointly the average frequencies and lengths of gaps in the alignments sought [82]. If gaps are to be introduced into the BILD score formalism, an immediate problem is which, if any, letters from individual sequences should be understood as insertions with respect to the "canonical" pattern. In other words, it appears a canonical width for the multiple alignment must somehow be chosen, with respect to which gaps arising in the alignment of individual sequences can be assessed.

Profile-Sequence Alignment

For simplicity, suppose we have a "canonical" multiple alignment A, i.e. one with a specified number of columns, to which we wish to align a single sequence S, to produce a new multiple alignment A'. It is reasonable to define the alignment score of A' as the pre-existing alignment score for A plus the incremental pairwise score for aligning A and S. This pairwise alignment involves substitutions (letters from S aligned to columns from A), insertions (runs of letters from S that are not aligned to any columns from A), and deletions (runs of columns from A that are not aligned to any letters from S). BILD scores for the columns of A' arise naturally when one defines the substitution scores for aligning A to S as incremental BILD scores. It remains then only to define gap scores for insertions and deletions in the alignment of A and S.

There is an essential asymmetry in gap scores for aligning A to S, relevant in many biological applications. For proteins, the columns of A represent canonical positions, present in most sequences of a protein family, and it should accordingly be very costly to delete any of these columns. In contrast, individual proteins often contain long loops not present in the great majority of related sequences [90,91], so even long insertions should not be very costly. Uniform but asymmetric affine insertion and deletion scores can capture this simple idea, and we have implemented them in one program described in the Results section below. These scores can be derived from the average frequencies and lengths [82] of insertions and deletions with respect to canonical protein family multiple alignments.

Just as incremental BILD substitution scores change as more sequences are added to a multiple alignment, so it is possible to let insertion and deletion scores change as well, and vary by position. In the context of Hidden Markov Models [76–81], many methods for doing this have been described. Below, we implement one simple procedure that depends only upon the BILD scores of multiple alignment columns, and not upon the relatively sparse gaps observed in any particular alignment.

Progressive Multiple Alignment and Profile-Profile Alignment

Formula (3) permits BILD substitution scores to be used for progressive multiple alignment. However, as described above, gaps scores pose a particular problem, because to define insertions and deletions one needs to construct a canonical alignment, and this is difficult for a small number of sequences. For example, when just two proteins are aligned, it is quite possible that gaps in both sequences would ultimately be seen as insertions with respect to a model describing the whole protein family, but there is no obvious way to determine this in advance. (The problem does not arise when substitution and gap scores are defined using the sum-of-pairs or SP formalism [27,28], for which no canonical alignment is necessary [73].) Accordingly, the approach we take below is eschew gaps at first, and thereby construct a canonical multiple alignment whose columns represent positions present in the majority of sequences. Only then do we realign individual sequences to this model, allowing gaps.

There has been considerable recent interest in aligning profiles that describe different protein families [40–48]. If BILD substitution scores, defined by equation (3), are to be used for this purpose, it would seem that we face the same problem for gaps that we do for progressive multiple alignment. Specifically, an insertion with respect to one profile is seen as a deletion with respect to the other, so how may one determine which, if either, perspective to adopt in a model describing both? However, so long as this goal is only to compare pairs of profiles, and not to proceed further, this problem may be elided. It is consistent to define pairwise gap costs for the alignment of two profiles, just as one would for the alignment of two sequences, without reference to a canonical alignment, and the substitution scores of equation (3) can be used sensibly with such gap costs. The gap costs chosen may depend upon the profiles being aligned, and may therefore be asymmetric and position specific. We leave for elsewhere the comparative evaluation of profile-profile alignment using substitution scores defined by equation (3), and those defined in other ways [40–48].

Results

Substitution scores for multiple alignment columns form only one element of successful multiple alignment programs. Depending upon their specific purposes, such programs may also employ gap scores, sequence weights, heuristic optimization algorithms, low-complexity filters, discontiguous patterns, provisions for no or multiple copies of a pattern within a sequence, the search for multiple distinct patterns, statistical assessments, etc. It is not our purpose here to develop a fully realized program to outperform existing state-of-the-art programs that involve multiple alignment. Rather, we seek only to argue that the use of explicitly constructed log-odds substitution scores can in many cases add values to these methods.

The programs we consider below have been constructed for evaluation purposes, to isolate the contribution of log-odds scores as much as possible. These programs are parsimonious in their complexity and use of free parameters, and employ various ideas that have appeared frequently elsewhere, and for which no novelty is claimed.

A. Ungapped Multiple Local Alignment Using Gibbs Sampling

BILD scores find perhaps their purest application in the ungapped local alignment problem described above, so it is worth studying them in this restricted context. The Gibbs sampling approach to finding optimal local multiple alignments was introduced by Lawrence et al. [69], and this algorithm can easily be modified to employ BILD scores. Potential advantages are improved sensitivity and the automatic definition of domain boundaries. Evaluation ideally requires a set of proteins with ungapped domains whose correct alignment is structurally

validated, but such sets are unfortunately very rare. Nevertheless, the collection of ungapped helix-turn-helix (HTH) domains in [69] provides a limited test set for analyzing BILD scores in the absence of gaps. As we describe in Text S5, with Tables S1 and S2, BILD scores achieve success on two fronts. First, they have greater average sensitivity than the entropy-based scores proposed by Lawrence et al. [69], in yielding accurate alignment from fewer sequences; second, they recognize with good precision the extent of the structurally-defined domains, and therefore do not require a prior specification of alignment width.

B. Extension to Gapped Local Alignment

Local multiple alignment programs generally must allow for gaps, either implicitly or explicitly. However, even for aligning gapped domains, the search for ungapped local alignments can be a fruitful first step. BILD scores can play an important role at this stage in defining the common core of a protein family, and can be adapted in subsequent stages to score gapped multiple alignments. As a proof of principle, we here develop a relatively simple algorithm, Program 1, that uses BILD scores as part of a gapped multiple alignment strategy. We describe this program's architecture and motivation below, and use a standard artificial test set to evaluate its ability to recognize the boundaries of local motifs, and to properly construct gapped local alignments. We then describe in section C how Program 1 may be refined through the consideration of features of protein structure, and illustrate the application of our methods to the delineation of a protein domain family.

Program 1 architecture. Input: A set of putatively related protein sequences potentially containing zero, one, or multiple instances of a common pattern. The sequences are in a standard unaligned format such as fasta.

Goal: To find a gapped local multiple alignment that optimizes an objective function defined as the sum of column BILD scores, minus gap costs, minus costs for describing the start locations of patterns. The user may specify whether a single or multiple instances of the pattern in each sequence should be sought, as well as whether the pattern may be absent in some sequences.

Heuristic algorithm:

a) Execute the Gibbs sampling strategy (Text S5) to determine a preliminary pattern width, and a preliminary ungapped local alignment, allowing at most one instance of the pattern per sequence.

b) For each input sequence S, remove any and all segments of S from the multiple alignment, and construct a BILD-score based position-specific score matrix (PSSM) M_S from the remaining alignment. Allowing gaps with affine gap costs [83,85], optimally align the whole of M_S to a segment from S, using for this purpose a generalization of the semi-global alignment algorithm of Erickson and Sellers [92]. Consider all sequences S, whether or not they were identified as containing a pattern in the initial Gibbs sampling stage, or in subsequent gapped alignment iterations. Asymmetric gap costs for insertions and deletions may be specified. (Note that, as described in the hidden Markov model (HMM) literature [76–82], for bit scores to retain their meaning, a small penalty, equivalent to the log probability of *not* initiating a gap, must be assessed whenever a letter is aligned to a motif column. For our purposes, this penalty is best viewed as an additional "gap score", although it may be coded as a modification to the substitution scores. Also, when a letter is not aligned to a motif column, the number of observations and aggregate BILD score for that column do not change.) Retain the alignment if the score exceeds a calculated threshold. Multiple non-overlapping segments within S that align to M_S can be found using a greedy approach.

c) Collect all the aligned segments from step b) into a new, gapped multiple alignment, and return to step b). Iterate until the objective score function stops increasing.

d) Adjust the width of the original pattern to optimize the alignment score. Alternatively, this step may be inserted between iterations.

Program 1 motivation. The initial search for ungapped segments in the Gibbs sampling step can delineate a common core pattern width, and provisional amino acid frequencies for each column, shared by a set of sequences, even when most segments are at first partially misaligned. For sequences containing repeated or multiple distinct patterns, it may be useful to restrict the width of the pattern sought. The MDL Principle can be used to provisionally exclude some sequences from the alignment at this stage, which may then be included later. Adopting this core pattern generally minimizes the average number of gaps that subsequently need to be introduced when aligning to members of the family. Using Erickson-Sellers semi-global alignment conserves the pattern width W, recognizing the importance of complete domains, and thereby both reduces the noise from chance partial similarities and aids the discovery of long insertions. Gapped alignment avoids the imposition of a block structure that may not be universally appropriate. However, columns are not added to the evolving profile to represent insertions, which can be idiosyncratic in length and location. Deletions may be present, but these are generally short and in a small minority of the sequences. Thus, the W concatenated aligned columns are densely occupied by amino acid data and are highly informative. This compressed type of profile, like the similar representation of Neuwald and Liu [82], generally corresponds well to the core structural elements of a domain. The use of asymmetric gap costs (with greater penalties for deletions) captures the natural asymmetry implied in aligning a sequence to such a core model. Note that elsewhere Gibbs sampling has been extended directly to the construction of gapped alignments [93,94], whereas Program 1 takes the simpler approach of confining the Gibbs sampling stage to the discovery of a provisional ungapped pattern.

Program 1 performance. The evaluation of the performance of a multiple alignment program requires a collection of sequence sets for each of which the correct alignment is known [95,96]. Multiple alignment programs may focus on the construction of global alignments, or on the discovery of local patterns, and different collections are accordingly appropriate for their evaluation. Among those collections in common use, "ref1" from IRMBase [96], which we will call IRM-1, appears the most appropriate for our gapped local multiple alignment program. IRM-1 contains 60 sets of sequences, with the sequences in each set containing a single, possibly gapped, local motif, embedded within otherwise random sequence. The motifs were generated artificially using the Rose program for simulated evolution [97]. This construction, although not completely realistic, means, however, that the extent and correct alignment of the motifs within the various sequences are precisely known. The 60 sets are divided into three groups of 20, consisting respectively of sets of 4, 8 and 16 sequences.

First, we evaluated the ability of Program 1 to identify properly the left and right motif boundaries within the 60 IRM-1 sequence sets. The results, grouped by the number of sequences within the various IRM-1 sets, are shown in Table 3, with positive deviations referring to patterns identified by Program 1 that are too long.

Table 3. The recognition of motif boundaries.

Program 1									
	Deviation from true boundary								
IRM-1 subset	**<−3**	**−3**	**−2**	**−1**	**0**	**1**	**2**	**3**	**>3**
4		1	1	5	29	1	2	1	
8				1	31	6	1	1	
16	1		1		36	1			1
Total	1	1	2	6	96	8	3	2	1
DIALIGN-TX									
	Deviation from true boundary								
IRM-1 subset	**<−3**	**−3**	**−2**	**−1**	**0**	**1**	**2**	**3**	**>3**
4	2	2	1	6	23	4	2		
8				6	24	7	2	1	
16	1		3	3	11	10	9	3	
Total	3	2	4	15	58	21	13	4	

Counts were made of the deviations found by Program 1 and DIALIGN-TX of the left and right pattern boundaries (120 total) for the embedded motifs within the 60 IRM-1 sequence sets, divided into the sets involving 4, 8, and 16 sequences [96]. At all 120 boundaries of the reported patterns, both programs align in register at least 50% of the sequences. This consensus allows us to determine to what extent the programs report conserved regions that are too long or too short. Positive deviations in the table refer to patterns identified by the programs that are longer than the actual patterns. To make an equitable comparison of the two programs, several non-default options and procedures were employed, as follows: (1) Asymmetric affine gap costs were inappropriate for Program 1 because the Rose program [97] used in the construction of IRM-1 does not simulate the differential rates with which insertions and deletions occur within real protein motifs. Accordingly, we empirically assigned all gaps of length k a score of $-8.5-0.5k$ bits, which corresponds [82] to an average frequency of 0.67% for insertions (and similarly for deletions) beginning at each motif position, and an average insertion or deletion length of 3.4. (2) We ran DIALIGN-TX at its least sensitive setting, using the "-l2" option, to avoid the excessive extensions into randomly aligned flanking sequences that degrade the accuracy of motif boundary recognition with the more sensitive default setting. (3) For DIALIGN-TX, we defined the boundary of a conserved motif as the maximum left or right extent to which *all* of the set of sequences aligned in register were reported as conserved. An alternative criterion might be to take a majority vote on the left or right extent of the reported pattern, but this criterion often gave unreasonably long extensions with DIALIGN-TX, and so was not used. For Program 1 run with the 16-sequence input sets, two outliers were found (columns headed <-3 and >3). These are cases where roughly half the sequences in the set contained large insertions or deletions, leading Program 1 to misalign a substantial minority of sequences at one of the boundaries.

doi:10.1371/journal.pcbi.1000852.t003

Program 1 identifies 80% of the motif boundaries exactly, and 92% to within 1 residue. Furthermore, the accuracy of boundary detection clearly improves as the number of sequences considered increases.

Most multiple alignment programs do not explicitly identify in their output conserved motifs as distinct from randomly aligned sequence. However, the output of program DIALIGN-TX [98], developed by the same research group that constructed IRM-1, displays the significantly conserved residues within each sequence in upper case letters, although these do not generally fall into completely consistent aligned columns. We have used this feature to compare the performance of DIALIGN-TX at identifying motif boundaries with that of Program 1 (Table 3, with details in the caption). DIALIGN-TX identifies 48% of motif boundaries exactly and 78% to within 1 residue, but its performance appears to degrade as the number of sequences considered increases. In summary, although existing multiple alignment programs such as DIALIGN-TX can do quite a good job at identifying the extent of common motifs embedded within random sequence, the use of BILD scores for this purpose can lead to noticeably improved precision.

The IRM database has been used previously to evaluate the performance of multiple alignment programs by computing how accurately they align the letters that are, by construction, "homologous" [96,99]. Given a set of sequences $\mathcal{S}$ from IRM, and the multiple alignment $\mathcal{M}$ produced by a particular program, the quality score for the program is defined to be the percentage, taken over all pairs of sequences within $\mathcal{S}$, of the homologous pairs of letters, within the annotated IRM-1 motif, that are aligned in $\mathcal{M}$ [99]. We used this measure to compare Program 1 to a variety of multiple alignment programs representative of distinct strategies: ClustalW [100]; PCMA [101]; MUSCLE [102,103]; ProbCons [104]; COBALT [99]; and DIALIGN-TX [98]. For each program, various quality score statistics for IRM-1 are presented in Table 4, along with aggregate program execution time. As can be seen, Program 1 performs better than or comparably to all the other multiple alignment programs, as measured by the various quality score statistics, and also runs substantially faster. Caution should be employed in interpreting Table 4, since Program 1 was explicitly designed for discovering single local patterns within otherwise unrelated sequences, while the other programs were primarily designed to construct global multiple alignments, and some use strategies or parameters that are not well adapted to local multiple alignment.

C. Protein Structure Considerations

As mentioned above, real protein domains are subject, on average, to much longer insertions than deletions, and this implies the utility of asymmetric affine gap costs for Program 1. The particular costs that are best will depend upon the statistical properties of gaps, and a possible refinement of Program 1 would be to adjust gap costs dynamically. From the analysis of a variety of protein families, we have found empirically that reasonable gap scores to use in conjunction with Θ_0^D Dirichlet mixture priors are $-8.5-k$ bits for a deletion of k motif positions (corresponding [82] to an initiation frequency per motif position of 0.28%, and a

Table 4. Multiple alignment accuracy.

	Quality Score Statistics				
Program	**Minimum**	**Mean**	**Median**	**% Perfect**	**Execution time (sec.)**
Program 1	60.7	95.0	99.8	48	18
DIALIGN-TX	37.6	94.2	98.4	38	95
PCMA 2.0	16.7	92.3	98.4	23	376
COBALT	45.6	95.1	98.0	22	303
ProbCons 1.10	16.7	82.8	92.2	27	506
MUSCLE 3.6	0.0	38.0	31.5	3	115
ClustalW 1.83	0.0	8.0	3.9	0	27

Quality score statistics were measured in the 60 sequence sets of the IRM-1 database [96]. "Percent perfect" refers to the proportion of the 60 datasets in which all homologous residues were correctly aligned. All programs were run with default parameters, except that Program 1 and DIALIGN-TX used the parameters detailed in Table 3. Because all programs other than Program 1 produce global multiple alignments as a matter of course, the quality score credits them for aligned residues independently of whether these residues are identified as lying within a conserved region. None of these programs explicitly identifies such regions, although DIALIGN-TX does so implicitly, as described in the caption to Table 3. Accordingly, in order not to artificially handicap Program 1 on this test, we calculated its quality scores by aligning, immediately adjacent to the conserved pattern it identifies within each sequence, and without gaps, all the residues deemed to lie beyond this pattern. In the small fraction of cases where the identified pattern stops short of the boundary of the embedded motif (see Table 3), this can produce a slightly better quality score than the pattern, considered in isolation, would yield. CPU execution times are for programs run on a Dual Pentium 4 Xeon 3.0 GHz CPU Linux computer with 64-bit architecture, and are averaged over three runs.
doi:10.1371/journal.pcbi.1000852.t004

mean length of 2.0), and $-9.25-0.25k$ bits for an insertion of length k into the motif (corresponding to a frequency of 0.87%, and a mean length of 6.3).

Protein structure implies more than an asymmetry between the frequency and length statistics of insertions and deletions. Reflecting the evolution of secondary structure elements and loops, certain motif positions are much less likely to be deleted than others and, similarly, insertions are much less likely to occur between certain pairs of motif positions than others. We describe below an extension of Program 1 to an HMM-based Program 2 that relies only upon column BILD scores to calculate position-specific gap score parameters. We then apply Programs 1 and 2 to the detection of Api-AP2 domains.

Program 2 motivation and architecture. Protein families or domains are often described by HMMs [76–81]. HMMs, in addition to specifying the probabilities for amino acids to occur in various profile positions, may specify distinct probabilities for insertions or deletions to occur in various locations. A more dynamic strategy for model construction than typical for HMMs may be based on the approach described above. As an example of a current strategy, the construction of a Pfam model [105–107] starts with a manually-curated gapped multiple alignment of selected members of the protein domain family, the "seed alignment", from which an HMM profile is built. The seed alignment and HMM are the static canonical entities that define a Pfam family. Then, as a separate procedure, sequence database search programs using this HMM are applied to identify and align additional family members. In contrast, our approach does not entail an initial manual alignment. We start with unaligned sequences, which may include a large proportion of flanking sequence and negative cases of proteins lacking the domain of interest. Moreover, any interesting new proteins discovered can readily be added to the input sequence set to compute a new model. This facilitates a flexible strategy of model updating as knowledge accumulates, although a static HMM could, of course, be retrieved at any stage, if desired.

When translated into the HMM formalism, specifying the asymmetric affine gap cost parameters of Program 1, two for insertions and two for deletions, is equivalent to specifying average frequencies and lengths for insertions and deletions [82], uniformly along the HMM. An HMM's free insertion and deletion parameters generally are optimized for the seed alignment provided. Given the sparsity of the seed data concerning the location of gaps, care must be taken to avoid overfitting [108–110]. In contrast, we here take the following approach to restricting gap locations based solely upon the BILD scores for columns in the core model, which are data-dense, combined with a few fixed parameters motivated by basic ideas concerning protein structure.

First, we observe that a high BILD score for a column C correlates with the column's importance, and indicates the column is unlikely to be deleted, consistent with a general tendency for conserved residues to occur within structural elements crucial for the folding energetics. Let $R(C)$ be the mean incremental BILD score for aligning a random residue to C. $R(C)$ is always negative, and large negative values of $R(C)$ correlate strongly with large positive BILD scores. We set the score (i.e. the log-probability) for extending a deletion through column C to an empirically chosen multiple F of $R(C)$. By default, F is 2.5. This has the desired effect of penalizing the deletion of columns with high BILD scores. An additional cost D_1 for the existence of a deletion (default: 8 bits) is left uniform throughout the HMM.

Second, we recognize that insertions are relatively unlikely to occur within regions of a protein that show a close clustering of more conserved positions. Let the normalized score $\bar{B}(C)$ be the BILD score for column C divided by the number of sequences it aligns. We simply disallow insertions anywhere between two columns C and C' separated by at most one intervening column, when both $\bar{B}(C)$ and $\bar{B}(C')$ exceed a set threshold T_I (default: 1 bit). Otherwise, the existence and extension costs, I_1 and I_2 for an insertion are left uniform throughout the HMM with default values of 9.25 and 0.25 bits, as for Program 1. This treatment is motivated by the typical unbroken patterns of local conserved clusters often observed in domain alignments, e.g. the alternating residues of a beta-strand face, or residue pairs within some turn geometries and cap structures. It may be fruitful to extend this insertion model, to conform with observed differences in the frequencies of long and short gaps [89,111], or to explicitly model the 3–4 spacing of conserved positions commonly seen in alpha-helices.

This simple approach to HMM parameter construction can of course be refined. Nevertheless, it captures central features of the location of insertions and deletions within proteins, without relying upon a preconstructed alignment, or on the relatively small sample of gaps present in a particular data set. Program 2 proceeds identically to Program 1, except that in place of the Erickson-Sellers algorithm with asymmetric affine gap costs, it uses the Viterbi algorithm to find an optimal path through the constructed HMM.

Application to Api-AP2 domains. To illustrate how our methods may be applied to typical problems, we consider the sequence-specific DNA recognition domains from the Api-AP2 transcription factor family of apicomplexan parasites. Multiple paralogous Api-AP2 domains in the translated proteomes of *Plasmodium* and *Cryptosporidium* parasites were initially discovered using PSI-BLAST searches by Balaji et al. [112], based on weak

similarity to the plant AP2 (APETALA2) transcription factors. These domains also have weak similarity to part of the HNH domain of homing endonucleases [113,114]. As the principal known sequence-specific DNA binding domains of Apicomplexa, Api-AP2 sequences represent a major lineage-specific gene expansion within the alveolate protists and are currently a topic of intense research [115,116]. They are believed to function in transcriptional activation crucial for parasite biology and development, and have potential as stage-specific anti-parasitic drug targets, due to the absence of AP2 homologs in the mammalian hosts. It is important to develop a system-level understanding of the Api-AP2 factors, and a prerequisite for this is to discover and annotate the entire complement of Api-AP2 proteins in each of these parasite genomes, possibly beyond the lists obtained from PSI-BLAST and the current HMM-based (Pfam model 00847) searches.

The Api-AP2 family presents a weakly-clustered pattern of amino acid conservation at variable spacing within the typically 50–60 amino acid domain. This profile is present one to approximately six times within otherwise extremely variable protein sequences of typically more than 1000 amino acids. The proteins show little significant homology outside the Api-AP2 regions: there are many low-complexity segments and occasional recognizable domains of other types, but the latter do not show any consistent relationship to the Api-AP2 regions. On the order of 20 to 80 Api-AP2 domains are encoded in each apicomplexan genome. Although relatively small, this domain has typical globular protein structure with a 3-strand beta-sheet packed against an alpha-helix, with several classes of beta-turn, and a longer loop. The more conserved positions occur mainly within the sheet, the helix and beta-turn structural elements. Consequently, multiple alignment profiles tend to show a loosely-patterned clustering of column scores, as is typical of globular domains.

Using PSI-BLAST searches of apicomplexan translated genome databases, we collected proteins containing at least one candidate Api-AP2 domain from *Toxoplasma gondii* (53 proteins) and *Plasmodium falciparum* (18 proteins; similar to the set identified by [112]). These sequences were used as input to develop the features of Programs 1 and 2 and to test their ability to construct protein domain profiles and discover additional domains. The results are outlined below, illustrated in Figures 1–3, and presented more fully in Tables S3 and S4 and their caption.

To explore how Programs 1 and 2 can tolerate negative cases, lacking the domain of interest, we spiked the Api-AP2 input sets with various proportions of (a) random sequences constructed by shuffling input sequences, or (b) real sequences lacking annotated conserved patterns, or (c) sequences that shared a conserved domain unrelated to Api-AP2. Spikes of types (a) and (b) comprising half of the total input sequences did not affect the final Api-AP2 models: the random and unrelated patterns in the spike sequences were all rejected (Figure 1) during or after the initial ungapped Gibbs sampling step, and this step runs faster if one specifies a prior expectation that a fraction of the input sequences do not contain the pattern of interest. If the Gibbs sampling stage of either Program 1 or 2 is run with a prior expectation that 10% of the input sequences do not contain an instance of the pattern, then all the random sequences and five of the Api-AP2 sequences are initially excluded, but subsequent gapped alignment steps recover segments from the initially rejected Api-AP2 sequences. The final result is the same whether or not sequences are excluded in the initial stage. With some spikes of type (c), the Gibbs sampling step converged on the competing domain instead of the Api-AP2 pattern. This suggests that input sequence domain parsing, e.g. by methods in [117,118], may sometimes be beneficial.

The amino acid frequencies observed within the core Api-AP2 model were strikingly similar for *Plasmodium* and *Toxoplasma* (Figure 2), consistent with an evolutionary expansion of this family from a single ancestral gene within the Alveolata, as proposed by [112]. Present day parasite lineages have evolved strikingly different codon and background amino acid content arising from genomic drift, e.g. in the very AT-rich *Plasmodium* and the more GC-enriched *Toxoplasma*. This contrast in background frequencies (Figure 2) demonstrates the value of log-odds scores for identifying a subtle pattern in very different sequence contexts.

The Api-AP2 pattern is present more than once in many of the input sequences as is often the case with eukaryotic multidomain proteins, potentially enabling these transcription factors to recognize combinations of DNA sites. The greedy algorithm included in Programs 1 and 2 allows such repeated domains to be identified. A total of 89 domains were found within the initial 53 *Toxoplasma* input sequences. Only 2 domains had borderline scores and may be candidates for classification as degenerate pseudo-domains. As shown in Table S3, repeats in the same protein can be very diverse in sequence. Programs 1 and 2 found several repeated domains additional to those reported in searches with Pfam model 00847, including some that differ from the canonical domain length by relatively long insertions in the central loop region (Figure 3, Table S3).

We conducted further searches of the *Toxoplasma* and *Plasmodium* databases based on the core Api-AP2 alignment obtained from Programs 1 and 2. These revealed new candidate proteins with Api-AP2 domains (Tables S3 and S4), not found with Pfam model 00847, some of which also show long loop insertions, but are otherwise strongly similar to the canonical Api-AP2 domain sequence (Figure 3). Including these new Api-AP2 cases, we have identified a total of 68 proteins (103 domains) for *Toxoplasma* and 29 proteins (50 domains) for *Plasmodium* (Tables S3 and S4).

It is not yet known if Api-AP2 domains with long insertions are active in DNA binding and transcriptional control, or whether any are inactive pseudo-domains, or are artifacts from errors in gene modeling. However, their occurrence illustrates that a relatively small minority of members of a domain family may contain long insertions, a general feature of protein evolution. Experimental studies confirm that many such long insertions, when artificially engineered into structural loops, have surprisingly low costs for the free energy of folding and little effect on the functional interactions of the proteins [90,91]. Thus, both the observed occurrence and the statistical thermodynamics of long insertions justify our treatment, described above, using asymmetric affine gap costs.

Discussion

We have described a natural generalization of log-odds substitution scores for pairwise alignments to substitution scores for multiple alignment columns. Multiple alignment log-odds scores probably are best derived using a Bayesian approach, yielding what we have called BILD scores. Log-odds scores imply scores for aligning multiple alignment columns to one another, or for aligning multiple alignment columns to single sequences, and it was in this latter context that the Bayesian approach was first formulated by Brown *et al*. [31]. In conjunction with the Minimum Description Length Principle, log-odds scores provide a means for determining the proper width or extent of a local multiple alignment, and for deciding whether a segment should be included in the alignment. They may also be used to cluster a set of related segments into subclasses; see Text S4 and [62,71,72].

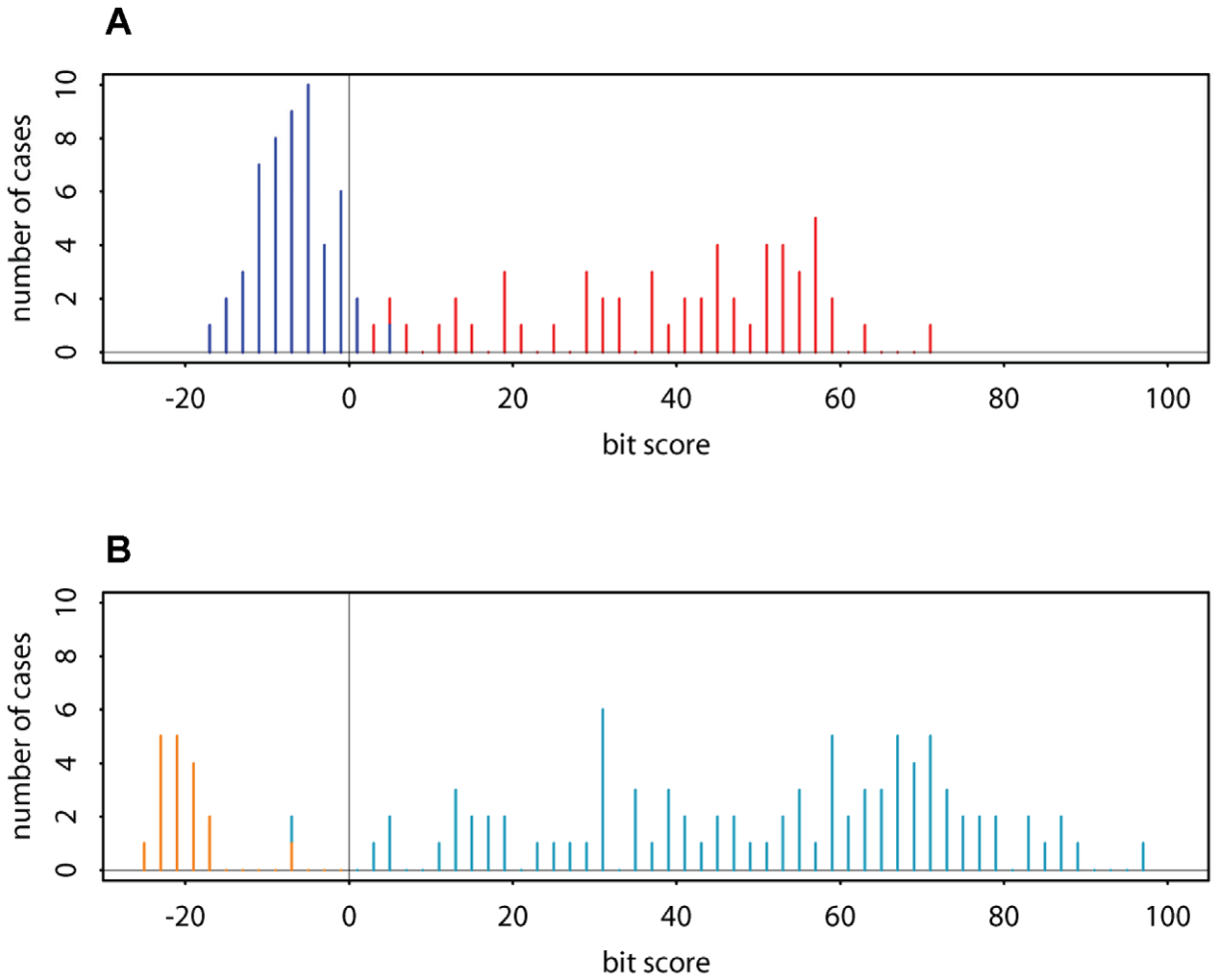

Figure 1. Distributions of bit scores from Api-AP2 domains and negative controls. The histograms in **A** and **B** represent data for both positive and negative cases reported by Program 1 at different intermediate stages of a run. The input file contained 107 amino acid sequences consisting of 54 *T. gondii* proteins with Api-AP2 domain candidates, and 53 random sequences obtained by shuffling the concatenated sequence of 53 of the 54 Api-AP2 proteins and cutting this shuffled string into the original lengths (method of [119]). The Dirichlet mixture prior Θ_0^D was specified. **A**: Results after the initial Gibbs sampling stage. The ungapped local alignment with optimal aggregate BILD score had width 53. For each sequence, we plot the incremental BILD score, resulting from the addition of a segment from that sequence to the alignment of all the other segments, minus the log of the effective length of that sequence. Scores from the real and random sequences are shown respectively in red and blue. If a prior probability for the existence of a domain in each sequence were specified, segments with scores below a calculated threshold would be rejected. Here, however, the Gibbs sampling step includes one ungapped segment from each of the 107 input sequences in the initial pattern it constructs. **B**: Results after the iterative gapped alignment stage. In each gapped alignment iteration of Program 1, the evolving length-53 pattern is aligned to each input sequence, perhaps multiple times, using a greedy application of the Erickson-Sellers algorithm. Incremental BILD scores are calculated from the current multiple alignment, excluding the sequence to which it is being realigned. Deletions of length k are assigned a score of $-8.5-k$ bits, and insertions of length k a score of $-9.25–0.25k$ bits. The cost for the existence of a pattern is based on assuming a mean of one instance per sequence, but with uniform probability at all positions of all sequences. In addition, the score for each aligned letter is adjusted slightly to reflect a small cost for not having a gap. At each iteration, the program reports segments with score ≥ -25 bits, but only segments with positive score are included in the next iteration. We show the data reported for the highest-scoring alignment; at this stage, at least one positively scoring segment derives from each of the 54 real sequences but only 2 segments (each with score less than -19 bits) derive from the 53 random sequences. 88 positive-scoring instances of the pattern are found, at least one from each of the real sequences, but none from the random sequences. In addition, 19 instances of the pattern with negative score are found, 2 of which derive from the random sequences. For an aligned segment, a log-odds bit score of 0 indicates an equal probability of being generated by the model implied by the other sequences, or at random by background amino acid frequencies. In **B**, the bars are colored according to the presence (cyan) or absence (brown) of strong sequence matches to the 3 beta-strands and the alpha-helix of the core Api-AP2 structure; the positions of these elements are shown in Figure 3. To qualify for a cyan bar, a sequence was required to contain either identities or high-structural-propensity substitutions that match the strongly conserved amino acids (with column BILD score ≥ 1.5 bits per residue) in the helix and at least 2 of the 3 beta-strands. The fairly clean separation, near 0 bits, of the cyan bars from the others indicates that a positive score is a good criterion for nominating a segment as an Api-AP2 candidate.
doi:10.1371/journal.pcbi.1000852.g001

One may compute rapidly the BILD score for a multiple alignment column, as well as the new score that results from the addition or subtraction of a single letter. This permits BILD scores to be used practically in Gibbs-sampling local multiple alignment programs. They can improve the performance of such programs, and remove the need for specifying the width of a pattern sought.

The proper description of protein domains in most cases requires a provision for gaps. We have implemented two relatively

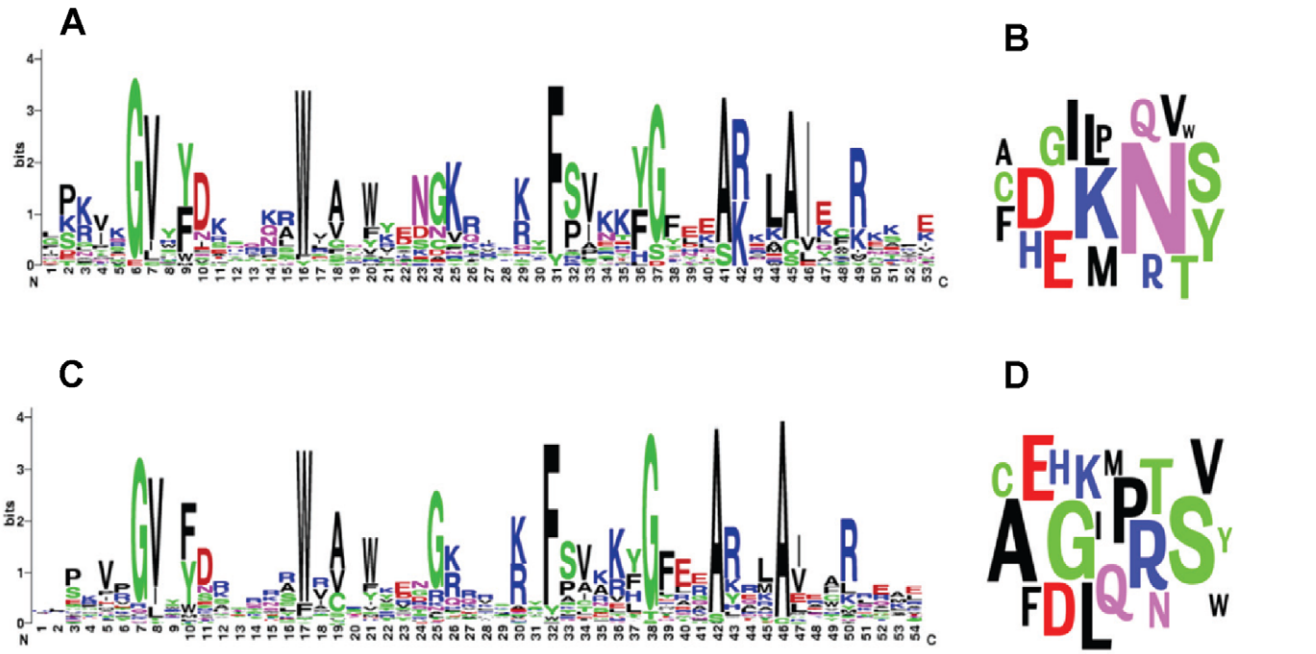

Figure 2. Near-identical Api-AP2 profiles from two parasites with very different background frequencies. For *P. falciparum* (**A**, **B**) and *T. gondii* (**C**, **D**), the logos [120] (http://weblogo.berkeley.edu/) represent the letters aligned in the columns of the core Api-AP2 patterns (**A**, **C**). In the letter clouds (http://www.wordle.net/advanced) (**B**, **D**), the area occupied by each letter indicates the background frequency of an amino acid in the input sequence set (compare Fig. 2.1 of [49]). Colors represent various amino acid classes. For both organisms, Programs 1 or 2, run with Dirichlet mixture priors Θ_0^B, Θ_0^C or Θ_0^D, converged on essentially the same 53- to 54-column core models that correspond to these logos. Api-AP2 models and logos almost identical to these were also obtained from other apicomplexan parasites *Cryptosporidium hominis*, *Babesia bovis*, *Theilleria parva*, and from the basal alveolate *Perkinsus marinus*, whereas the distantly related plant AP2 domains and HNH homing endonuclease/integrase domains gave distinct characteristic patterns similar in parts to Api-AP2 (data not shown). Thus, the core structural features of the Api-AP2 domain have been strongly conserved in long-diverged members of the Alveolata, following an ancestral gene expansion, whereas the background amino acid content of these organisms is strikingly different due to genome-wide drift.
doi:10.1371/journal.pcbi.1000852.g002

simple programs for extending a core ungapped pattern or profile to a gapped local multiple alignment. There are several key elements to our approach. First, the initial maximization of aggregate BILD scores using Gibbs sampling yields a core pattern and pattern length for further refinement. Second, the semi-global alignment of this pattern to the input sequences recognizes the importance of complete occurrences of the pattern. Third, the use of asymmetric affine gap costs (Program 1) recognizes that, with respect to the core pattern, long deletions generally are much more deleterious than long insertions. The placement of gaps can be refined using position-specific gap costs derived from column BILD scores (Program 2). Fourth, greedy alignment allows multiple instances of a pattern to be found within a single sequence. In conjunction with length-dependent gap costs, it discourages alignments spanning more than one instance of a pattern, but can still uncover long insertions. Fifth, iteration permits the core model to be refined, improving the discrimination of true relationships from chance similarities. This strategy,

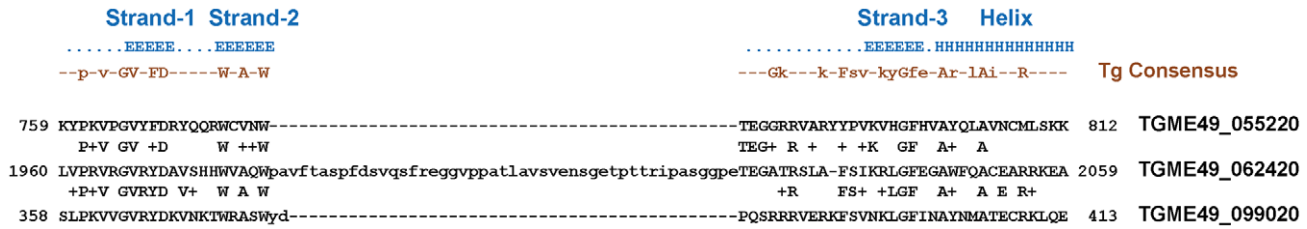

Figure 3. Large insertions in the central loop region of Api-AP2 domains. As a consequence of asymmetric gap costs, Programs 1 and 2 reported several positive Api-AP2 candidates which have long insertions but, in the other parts of the domain, show high-scoring matches to the canonical pattern. Here, the sequence of *T. gondii* protein TGME49_06420, which has a 45 amino acid insertion in the central loop region, is shown aligned with the two most-closely-matching domains of typical length. Program 2, run with Dirichlet mixture prior Θ_0^D and default parameters, assigned the insertion to the central loop location shown, which avoided the more conserved columns of the secondary structural elements indicated above the sequences. In contrast, Program 1 placed the same inserted residues in three separate locations, two of which would disrupt secondary structure. Moreover, with an established HMM search method [80] (http://hmmer.janelia.org/), only the right end alignment of this TGME49_06420 domain was found, but with a negative score well below the rejection threshold. Structural assignments E (beta-strand) and H (alpha-helix) are based on homologous experimental structures [121,122] (PDB codes 2gcc,3gcc,3igm).
doi:10.1371/journal.pcbi.1000852.g003

informed by considerations of protein structure, has proved a rapid and effective method for delineating protein families. Although our programs were developed only for research purposes, with the limited goal of testing the impact of BILD scores, their code is available upon request.

We have sought here primarily to describe the construction and potential uses of log-odds scores in the multiple alignment context. However, many avenues for further research, involving the development and benchmarking of complete multiple alignment programs, remain. To what extent can BILD scores improve the accuracy of profile-profile comparison programs? How does Erickson-Sellers semi-global alignment [92], with uniform asymmetric affine gap costs, compare to HMM [80,81] and other methods [6] in recognizing related sequence in database searches? We look forward to investigating some of these questions.

Supporting Information

Text S1 MELD Scores
Found at: doi:10.1371/journal.pcbi.1000852.s001 (0.05 MB PDF)

Text S2 The Mean Relative Entropy of Dirichlet Mixtures
Found at: doi:10.1371/journal.pcbi.1000852.s002 (0.05 MB PDF)

Text S3 The MDL Principle and Local Alignment Statistics
Found at: doi:10.1371/journal.pcbi.1000852.s003 (0.04 MB PDF)

Text S4 The MDL Principle and the Clustering of Multiple Alignments
Found at: doi:10.1371/journal.pcbi.1000852.s004 (0.05 MB PDF)

Text S5 Gibbs Sampling Algorithms and HTH Proteins
Found at: doi:10.1371/journal.pcbi.1000852.s005 (0.05 MB PDF)

Table S1 Helix-turn-helix proteins.
Found at: doi:10.1371/journal.pcbi.1000852.s006 (0.02 MB PDF)

Table S2 Number of sequences misaligned by Gibbs sampling programs. Sequence sets supplied to the BILD and Wadsworth samplers consist of the first *M* sequences listed in Table S1. For each sequence set, the BILD sampler determines an optimal motif width *W*. Both BILD and Wadsworth samplers optimize contiguous motifs of widths *W*, 17, 21 and 25. The number of sequences misaligned by the Wadsworth sampler are given in the table without parentheses; the number misaligned by the BILD sampler within parentheses.
Found at: doi:10.1371/journal.pcbi.1000852.s007 (0.02 MB PDF)

Table S3 Tables S3 and S4 show Api-AP2 domains and bit scores reported by Programs 1 and 2 for *Toxoplasma gondii* (Table S3) and *Plasmodium falciparum* (Table S4). Also shown are the bit scores obtained using HMMsearch database searches [Eddy SR (1998) *Bioinformatics* 14: 755–763] seeded with aligned Api-AP2 domains and with the current Pfam AP2 model number 00847 (http://pfam.sanger.ac.uk/family?entry = PF00847). Programs 1 and 2 were run with the Dirichlet mixture prior and default parameters described in the Results section and Figures 1 and 3. As input, we collected 68 amino acid sequences from *T. gondii* and 29 from *P. falciparum*, based on inspection of low-threshold PSI-BLAST and HMMsearch searches of the parasite genomic translation databases of ToxoDB [Gajria *et al.* (2008) *Nucleic Acids Res* 36: D553–556] and PlasmoDB [Aurrecoechea *et al.* (2009) *Nucleic Acids Res* 37: D539–543] (http://eupathdb.org/eupathdb/). These database searches were seeded with earlier alignments produced (as described in the Results section and Figure 1 legend) from more preliminary sets of 54 *T. gondii* and 18 *P. falciparum* sequences. We anticipated that the larger sets of input sequences might include some false positives; however, the final evolved models included at least one positive score from each of the 68 and 29 sequences, totaling 103 Api-AP2 domain candidates for *T. gondii* and 50 for *P. falciparum*. The corresponding core domain alignments assigned by Program 2 are shown, denoted respectively 'Tg-core' (with 53 columns in the evolved model plus 2 adjacent positively-scoring columns added from the left-flank) and 'Pf-core' (with 53 columns) respectively. These patterns exclude any insertions in individual sequences: the number of inserted residues is shown in a separate column. All of these domains have positive bit scores with Programs 1 and 2, except for the special case of domain 1.7, Table S3, which has been added manually to the *T. gondii* alignment. This domain is notable because of its occurrence within a multi-Api-AP2 protein and its strong match to the canonical 53-column pattern; however, it also has an unusually long insertion of 66 amino acids (assigned to the central loop by Program 2), the cost of which results in an overall negative score. The Tg-core (excluding domain 1.7) and the Pf-core alignments shown in Tables S3 and S4 were used as seed alignments for further analysis with HMMER version 2.3.2 (http://hmmer.janelia.org). HMMbuild and HMMcalibrate were used with default parameters to construct HMMs and calibrate their E-value distributions, and HMMsearch was used with a permissive E-value threshold of 100 to search the parasite genomic translation databases against these HMMs. These searches gave positive bit scores for the domains used for HMM construction (except domain 1.8, which was not reported, Table S3), as shown in the columns headed 'bits (HMMsearch, Tg-core seed)' and 'bits (HMMsearch, Pf-core seed)'. In some cases, HMMsearch alignments encompassed only part of the Api-AP2 pattern, either to the left or right of the central loop, denoted, respectively, 'LH only' and 'RH only' in comments columns. Note that all of the positively scoring sequences reported were present in the seed alignment, and no new Api-AP2 domain candidates were found in these HMMsearch database searches. HMMsearch scans of the same databases were also seeded by Pfam model 00847 (converted to a version 2.3.2 HMM with HMMbuild and HMMcalibrate as described above). The resulting bit scores are given in the columns headed 'bits (HMMsearch, Pfam00847 seed)'. The Pfam00847 model seed alignment contains both plant and apicomplexan AP2 domains, including some from *P. falciparum* but none from *T. gondii*. Consequently, the matches of Pfam00847 to the Api-AP2 domains are generally weaker than the matches obtained with the more specific models from Program 2 Api-AP2 core alignments, resulting in substantially lower bit scores. Several domain candidates (highlighted in color), 10 from *T. gondii* and 4 from *P. falciparum*, were not reported by the Pfam00847 HMMsearch above the E-value 100 threshold, and others (9 and 3 respectively) were given negative scores (and non-significant E-values). These low scores reflect misalignments (e.g. missed long insertions) in some cases. In other cases, limited deviations from the canonical conserved patterns occur, commonly in the first beta-strand. However, such deviant residues appear to be structurally compatible with the domain, with beta-strand-favoring propensities in most cases, suggesting that these examples may be authentic but non-canonical Api-AP2 domains. HMMER methodology is capable of identifying and aligning such domains if they are included in the seed alignment, as shown by the bit scores given by the Tg-core and Pf-core seeded searches. Indeed, the *T. gondii* domain 62.1 (TGME49_062420), which is the example with a 45/46 amino acid insertion shown in Figure 3, obtains a positive bit score with HMMsearch (and 46 inserted residues) when it is included in the Tg-core seed alignment (as in Table S3) but a negative score and only a partial alignment otherwise, as indicated in Figure 3 legend. In several cases, HMMsearch alignments

report different gap positions from Program 2, mostly with shorter insertions. In the case of domain 66.1 (Table S3) the alignment produced by HMMsearch appears to be more compatible with beta-strand propensities than the Program 2 alignment shown in Table S3, whereas in 6 other cases, the HMMsearch alignment appears more disruptive of secondary structure. This observation supports the potential benefit of incorporating secondary structure prediction into an HMM-based domain recognition strategy, as proposed by Won *et al.* [(2007) *BMC Bioinformatics* 8: 357]. Overall, the HMMsearch results shown in these Tables, compared with the Programs 1 and 2 output, show many more similarities than differences: the two approaches can achieve very similar results with appropriate inputs. Our examples also illustrate how the relatively simple BILD score based approaches, by reducing the strict dependence on seed alignments, might facilitate more automated processes for the discovery and reporting of protein domain families and more flexible updating strategies.

Found at: doi:10.1371/journal.pcbi.1000852.s008 (0.05 MB XLS)

Table S4 Api-AP2 domains and bit scores reported by Programs 1 and 2 for *Plasmodium falciparum*. For more details please see caption to Table S3.

Found at: doi:10.1371/journal.pcbi.1000852.s009 (0.03 MB XLS)

Acknowledgments

The authors thank Dr. Richa Agarwala for assistance in the benchmarking of Program 1 and other multiple alignment programs.

Author Contributions

Conceived and designed the experiments: SFA JCW EZ YKY. Performed the experiments: SFA JCW YKY. Analyzed the data: SFA JCW YKY. Wrote the paper: SFA JCW EZ YKY. Developed mathematical theory: SFA YKY.

References

1. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197.
2. Sellers PH (1984) Pattern recognition in genetic sequences by mismatch density. Bull Math Biol 46: 501–514.
3. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444–2448.
4. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
5. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37: D205–D210.
6. Kann MG, Sheetlin SL, Park Y, Bryant SH, Spouge JL (2007) The identification of complete domains within protein sequences using accurate e-values for semi-global alignment. Nucleic Acids Res 35: 4678–4685.
7. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO, ed. Atlas of Protein Sequence and Structure. Washington, DC: Natl. Biomed. Res. Found., volume 5, suppl. 3. pp 345–352.
8. Schwartz RM, Dayhoff MO (1978) Matrices for detecting distant relationships. In: Dayhoff MO, ed. Atlas of Protein Sequence and Structure. Washington, DC: Natl. Biomed. Res. Found., volume 5, suppl. 3. pp 353–358.
9. Feng DF, Johnson MS, Doolittle RF (1985) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21: 112–125.
10. Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119: 205–218.
11. Rao JKM (1987) New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Peptide Protein Res 29: 276–281.
12. Risler JL, Delorme MO, Delacroix H, Henaut A (1988) Amino acid substitutions in structurally related proteins. J Mol Biol 204: 1019–1029.
13. Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256: 1443–1445.
14. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89: 10915–10919.
15. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL (1992) Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Prot Sci 1: 216–226.
16. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
17. Kann M, Qian B, Goldstein RA (2000) Optimization of a new score function for the detection of remote homologs. Proteins 41: 498–503.
18. Ng PC, Henikoff JG, Henikoff S (2000) PHAT: a transmembrane-specific substitution matrix. Bioinformatics 16: 760–766.
19. Müller T, Rahmann S, Rehmsmeier M (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics 17, Suppl. 1: S182–S189.
20. Goonesekere NC, Lee B (2008) Context-specific amino acid substitution matrices and their use in the detection of protein homologs. Proteins 71: 910–919.
21. States DJ, Gish W, Altschul SF (1991) Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. Methods 3: 66–70.
22. Chiaromonte F, Yap VB, Miller W (2002) Scoring pairwise genomic sequence alignments. In: Altman R, Dunker AK, Hunter L, Lauderdale K, Klein TE, eds. Proc. Pacific Symp. Biocomput. Mountain View, CA: World Scientific. pp 115–126.
23. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87: 2264–2268.
24. Dembo A, Karlin S, Zeitouni O (1994) Limit distribution of maximal nonaligned two-sequence segmental score. Ann Prob 22: 2022–2039.
25. Sankoff D (1975) Minimal mutation trees of sequences. SIAM J Appl Math 28: 35–42.
26. Sankoff D, Cedergren RJ (1983) Simultaneous comparison of three or more sequences related by a tree. In: Sankoff D, Kruskal JB, eds. Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, MA: Addison-Wesley. pp 253–263.
27. Murata M, Richardson JS, Sussman JL (1985) Simultaneous comparison of three protein sequences. Proc Natl Acad Sci USA 82: 3073–3077.
28. Bacon DJ, Anderson WF (1986) Multiple sequence alignment. J Mol Biol 191: 153–161.
29. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188: 415–431.
30. Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 219: 555–565.
31. Brown M, Hughey R, Krogh A, Mian IS, Sjölander K, et al. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In: Hunter L, Searls D, Shavlik J, eds. Proc. First Int. Conf. on Intelligent System for Mol. Biol. Menlo Park, CA: AAAI Press. pp 47–55.
32. Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, et al. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12: 327–345.
33. Grünwald PD (2007) The Minimum Description Length Principle. Cambridge, MA: MIT Press.
34. Xing EP, Karp RM (2004) MotifPrototyper: a Bayesian profile model for motif families. Proc Natl Acad Sci USA 101: 10523–10528.
35. Lunter G, Miklós I, Drummond A, Jensen JL, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6: 83.
36. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. PLoS Comput Biol 5: e1000392.
37. Satija R, Novák A, Miklós I, Lyngsø R, Hein J (2009) BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. BMC Evol Biol 9: 217.
38. Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351–360.
39. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
40. Pietrokovski S (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 24: 3836–3845.
41. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. strategies for structural predictions using sequence information. Protein Sci 9: 232–241.
42. Yona G, Levitt M (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 315: 1257–1275.
43. Edgar RC, Sjölander K (2003) SATCHMO: sequence alignment and tree construction using hidden markov models. Bioinformatics 19: 1404–1411.
44. Panchenko AR (2003) Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Res 31: 683–689.
45. Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326: 317–336.
46. Edgar RC, Sjölander K (2004) A comparison of scoring functions for protein sequence profile alignment. Bioinformatics 20: 1301–1308.

47. Wang G, Dunbrack RL, Jr. (2004) Scoring profile-to-profile sequence alignments. Protein Sci 13: 1612–1626.
48. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951–960.
49. MacKay DJC (2003) Information Theory, Inference, and Learning Algorithms. New York, NY: Cambridge University Press.
50. Altschul SF, Carroll RJ, Lipman DJ (1989) Weights for data related by a tree. J Mol Biol 207: 647–653.
51. Sibbald PR, Argos P (1990) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J Mol Biol 216: 813–818.
52. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9: 56–68.
53. Vingron M, Sibbald PR (1993) Weighting in sequence space: a comparison of methods in terms of generalized sequences. Proc Natl Acad Sci USA 90: 8777–8781.
54. Gerstein M, Sonnhammer ELL, Chothia C (1994) Volume changes in protein evolution. Appendix: A method to weight protein sequences to correct for unequal representation. J Mol Biol 236: 1067–1078.
55. Henikoff S, Henikoff JG (1994) Position-based sequence weights. J Mol Biol 243: 574–578.
56. Thompson JD, Higgins DG, Gibson TJ (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput Appl Biosci 10: 19–29.
57. Eddy SR, Mitchison G, Durbin R (1995) Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 2: 9–23.
58. Gotoh O (1995) A weighting system and algorithm for aligning many phylogenetically related sequences. Comput Appl Biosci 11: 543–551.
59. Krogh A, Mitchison G (1995) Maximum entropy weighting of aligned sequences of protein or DNA. In: Rawlings C, Clark D, Altman R, Hunter L, Lengauer T, et al. (1995) Proc. Third Int. Conf. on Intelligent System for Mol. Biol. Menlo Park, CA: AAAI Press. pp 215–221.
60. Bailey TL, Gribskov M (1996) The megaprior heuristic for discovering protein sequence patterns. In: States D, Agarwal P, Gaasterland T, Hunter L, Smith R, eds. Proc. Fourth Int. Conf. on Intelligent System for Mol. Biol. pp 15–24.
61. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, et al. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12: 387–394.
62. Brown DP, Krishnamurthy N, Sjölander K (2007) Automated protein subfamily identification and classification. PLoS Comput Biol 3: e160.
63. Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu YK (2009) PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Res 37: 815–824.
64. Yu YK, Wootton JC, Altschul SF (2003) The compositional adjustment of amino acid substitution matrices. Proc Natl Acad Sci USA 100: 15688–15693.
65. Yu YK, Altschul SF (2005) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics 21: 902–911.
66. Jeffreys H (1946) An invariant form of the prior probability in estimation problems. Proc Royal Soc London Series A 186: 453–461.
67. Nishida K, Frith MC, Nakai K (2009) Pseudocounts for transcription factor binding sites. Nucleic Acids Res 37: 939–944.
68. Vingron M, Waterman MS (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol 235: 1–12.
69. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262: 208–214.
70. Cover TM, Thomas JA (1991) Elements of Information Theory. New York, NY: Wiley.
71. Sjölander K (1998) Phylogenetic inference in protein superfamilies: analysis of SH2 domains. In: Glasgow J, Littlejohn T, Major F, Lathrop R, Sankoff D, et al. (1998) Proc. Sixth Int. Conf. on Intelligent System for Mol. Biol. Menlo Park, CA: AAAI Press. pp 165–174.
72. Brown DP (2008) Efficient functional clustering of protein sequences using the Dirichlet process. Bioinformatics 24: 1765–1771.
73. Altschul SF (1989) Gap costs for multiple sequence alignment. J Theor Biol 138: 297–309.
74. Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol 33: 114–124.
75. Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol 34: 3–16.
76. Tanaka H, Ishikawa M, Asai K, Konagaya A (1993) Hidden Markov models and iterative aligners: study of their equivalence and possibilities. In: Hunter L, Searls D, Shavlik J, eds. Proc. First Int. Conf. on Intelligent System for Mol. Biol. Menlo Park, CA: AAAI Press. pp 395–401.
77. Baldi P, Chauvin Y, Hunkapiller T, McClure MA (1994) Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci USA 91: 1059–1063.
78. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235: 1501–1531.
79. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Probabilistic models of proteins and nucleic acids. Cambridge, England: Cambridge University Press.
80. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
81. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14: 846–856.
82. Neuwald AF, Liu JS (2004) Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. BMC Bioinformatics 5: 157.
83. Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162: 705–708.
84. Fitch WM, Smith TF (1983) Optimal sequence alignments. Proc Natl Acad Sci USA 80: 1382–1386.
85. Altschul SF, Erickson BW (1986) Optimal sequence alignment using affine gap costs. Bull Math Biol 48: 603–616.
86. Waterman MS, Smith TF, Beyer WA (1976) Some biological sequence metrics. Adv Math 20: 367–387.
87. Miller W, Myers EW (1988) Sequence comparison with concave weighting functions. Bull Math Biol 50: 97–120.
88. Benner SA, Cohen MA, Gonnet GH (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol 229: 1065–1082.
89. Goonesekere NC, Lee B (2004) Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. Nucleic Acids Res 32: 2838–2843.
90. Ladurner AG, Fersht AR (1997) Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. J Mol Biol 273: 330–337.
91. Scalley-Kim M, Minard P, Baker D (2003) Low free energy cost of very long loop insertions in proteins. Protein Sci 12: 197–206.
92. Erickson BW, Sellers PH (1983) Recognition of patterns in genetic sequences. In: Sankoff D, Kruskal JB, eds. Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, MA: Addison-Wesley. pp 55–91.
93. Rocke E, Tompa M (1998) An algorithm for finding novel gapped motifs in dna sequences. In: RECOMB '98: Proceedings of the second annual international conference on Computational molecular biology. pp 228–233.
94. Wareham HT, Jiang T, Zhang X, Trendall CG (2000) Stochastic heuristic algorithms for target motif identification (extended abstract). Pac Symp Biocomput. pp 392–403.
95. Thompson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15: 87–88.
96. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics 6: 66.
97. Stoye J, Evers D, Meyer F (1998) Rose: generating sequence families. Bioinformatics 14: 157–163.
98. Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol 3: 6.
99. Papadopoulos JS, Agarwala R (2007) COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics 23: 1073–1079.
100. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
101. Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19: 427–428.
102. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
103. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
104. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.
105. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405–420.
106. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26: 320–322.
107. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281–288.
108. Yada T, Ishikawa M, Tanaka H, Asai K (1996) Extraction of hidden Markov model representations of signal patterns in DNA sequences. Pac Symp Biocomput. pp 686–696.
109. Won KJ, Prugel-Bennett A, Krogh A (2004) Training HMM structure with genetic algorithm for biological sequence analysis. Bioinformatics 20: 3613–3619.
110. Won KJ, Sandelin A, Marstrand TT, Krogh A (2008) Modeling promoter grammars with evolving hidden Markov models. Bioinformatics 24: 1669–1675.
111. Mott R (1999) Local sequence alignments with monotonic gap penalties. Bioinformatics 15: 455–462.
112. Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the

evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res 33: 3994–4006.
113. Magnani E, Sjölander K, Hake S (2004) From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. Plant Cell 16: 2265–2277.
114. Wuitschick JD, Lindstrom PR, Meyer AE, Karrer KM (2004) Homing endonucleases encoded by germ line-limited genes in Tetrahymena thermophila have APETELA2 DNA binding domains. Eukaryotic Cell 3: 685–694.
115. De Silva EK, Gehrke AR, Olszewski K, León I, Chahal JS, et al. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. Proc Natl Acad Sci USA 105: 8393–8398.
116. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, et al. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Mol Microbiol 71: 1402–1414.
117. Phuong TM, Do CB, Edgar RC, Batzoglou S (2006) Multiple alignment of protein sequences with repeats and rearrangements. Nucleic Acids Res 34: 5932–5942.
118. Raphael B, Zhi D, Tang H, Pevzner P (2004) A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res 14: 2336–2346.
119. Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18: 269–285.
120. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097–6100.
121. Allen MD, Yamasaki K, Ohme-Takagi M, Tateno M, Suzuki M (1998) A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. EMBO J 17: 5484–5496.
122. Lindner SE, De Silva EK, Keck JL, Llinás M (2010) Structural determinants of DNA binding by a P. falciparum ApiAP2 transcriptional regulator. J Mol Biol 395: 558–567.