Published online 16 December 2008 *Nucleic Acids Research, 2009, Vol. 37, No. 3* ***815–824***
doi:10.1093/nar/gkn981

PSI-BLAST pseudocounts and the minimum description length principle

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

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

Received October 15, 2008; Revised November 18, 2008; Accepted November 20, 2008

ABSTRACT

Position specific score matrices (PSSMs) are derived from multiple sequence alignments to aid in the recognition of distant protein sequence relationships. The PSI-BLAST protein database search program derives the column scores of its PSSMs with the aid of pseudocounts, added to the observed amino acid counts in a multiple alignment column. In the absence of theory, the number of pseudocounts used has been a completely empirical parameter. This article argues that the minimum description length principle can motivate the choice of this parameter. Specifically, for realistic alignments, the principle supports the practice of using a number of pseudocounts essentially independent of alignment size. However, it also implies that more highly conserved columns should use fewer pseudocounts, increasing the inter-column contrast of the implied PSSMs. A new method for calculating pseudocounts that significantly improves PSI-BLAST's retrieval accuracy is now employed by default.

INTRODUCTION

The scores of local protein sequence alignments are generally constructed as the sum of 'substitution scores' for aligning pairs of amino acids, and 'gap scores' for aligning runs of amino acids in one sequence with null characters inserted into the other (1). Given certain generally applicable conditions, the substitution scores s_{ij} for aligning amino acids i and j can be written in the log-odds form $s_{ij} = \left[\ln(q_{ij}/p_i p_j)\right]/\lambda$. Here, the p_i are 'background frequencies' with which the amino acids typically occur in proteins, the q_{ij} are 'target frequencies' characterizing aligned amino acid pairs in the 'true alignments' sought, and λ is a scale constant (2,3). All reasonable substitution matrices are implicitly of log-odds form, and the most widely used are explicitly so constructed (4–6). This log-odds formalism carries over naturally to protein profiles or position specific score matrices (PSSMs), where the scores for aligning amino acids to a specific column generally are constructed by the formula $s_i = [\ln(q_i/p_i)]/\lambda$, where q_i is the predicted probability that a properly aligned homologous protein has amino acid i in that column. A central problem in the construction of PSSMs is therefore how to estimate the q_i from a multiple alignment.

Several issues arise in converting a vector of observed counts $\mathbf{c}$ into a vector of predicted probabilities $\mathbf{q}$. The first is that the sequences in a multiple alignment are rarely independent, but are rather related to one another by a complex phylogenetic tree. If each amino acid in the alignment is given equal weight, there is a danger that a large number of closely related sequences will outvote a smaller number of more diverse sequences, thus squandering the available information. To address this problem, a large number of weighting schemes have been proposed (7–16), which assign lower weights to data from closely related sequences. These weights cannot properly be derived from the data in a single alignment column, and therefore generally are constructed considering larger alignment regions. We employ below the method used by PSI-BLAST (17,18), which is a modification of that described by Henikoff and Henikoff (11).

Once weights have been applied to the sequences in a multiple alignment, the raw amino acid count vector $\mathbf{c}$ is converted into an 'observed frequency' vector $\mathbf{f}$. If one assumes, as discussed below, that the observed data are equivalent to n, not necessarily integral, independent observations, then the weighted count vector is $n\mathbf{f}$. Unless otherwise specified, it will be assumed below that when we speak of observed frequencies or counts, we mean $\mathbf{f}$ and $n\mathbf{f}$.

It would be possible simply to adopt $\mathbf{f}$ as the predicted probabilities $\mathbf{q}$, and this is indeed the maximum

*To whom correspondence should be addressed. Tel: +1 301 435 7803; Fax: +1 301 480 2288; Email: altschul@ncbi.nlm.nih.gov likelihood estimate. However, an obvious shortcoming to this approach is that, due to small sample sizes, it is likely the observed frequencies of several amino acids will be 0. The scores assigned to these amino acids will then be $-\infty$, a reasonable result only if one truly believes it is impossible for these amino acids to appear in the column in question. One way around this problem is a Bayesian approach, in which a prior probability distribution is specified over the space of amino acid distributions. Mathematically, it is convenient to specify such a prior as a Dirichlet distribution. If the mean prior probability for amino acid i is chosen to be p_i, then the expected posterior probability for amino acid i will be proportional to $nf_i + mp_i$. This approach is equivalent to adding m 'pseudocounts' to the n effective observations, with the pseudocounts distributed proportionately to the p_i. The number of pseudocounts depends upon how peaked is the Dirichlet prior chosen.

A problem with the Dirichlet prior approach is that it ignores information about amino acid relationships that is present even in standard substitution matrices. For example, the BLOSUM-62 matrix (6) implies that the observation of a single leucine should increase the predicted probabilities for the observation of other hydrophobic amino acids, but a Bayesian approach using a Dirichlet prior such as discussed above will decrease the predicted probabilities for all non-observed amino acids. Accordingly, two main approaches have been proposed to balance prior knowledge of amino acid relationships and observed data. The first is the Dirichlet mixture method (19,20). This elegant formalism assumes a prior on the space of amino acid distributions that consists of a mixture of a number of Dirichlet distributions. These distributions can be seen as representing typical biases found frequently in protein positions. For example, one Dirichlet prior may favor aromatic residues, another charged residues, etc. A Dirichlet mixture prior can capture information about amino acid relationships and allow the observation of one amino acid to increase the predicted probability of another. The second approach is the data-dependent pseudocount method (17,21), which is employed by PSI-BLAST. As discussed in greater detail below, this approach predicts target frequencies by adding pseudocounts to observed counts, but lets the pseudocounts depend upon the observed data. In the limit of a large number of observations, the frequencies predicted by both the Dirichlet mixture and data-dependent pseudocount methods approach the observed frequencies. One advantage of the data-dependent pseudocount method is that the scores it implies can easily be engineered to reduce to any specified substitution matrix in the case of a column with $n = 1$. Used in the same algorithmic context, the two methods have roughly equivalent success in the recognition of subtle biological relationships (22).

One question that arises in applying the data-dependent pseudocount method is how many pseudocounts to employ, and whether this number should depend in any way upon the data. In the absence of theory, this question has to date been treated empirically. In this article, we study the question of pseudocount number through the lens of the minimum description length (MDL) principle (23), which we review briefly below. The MDL principle is of utility in choosing among models of varying complexity, usually corresponding to their number of degrees of freedom. We argue that varying the number of pseudocounts can be understood as varying model complexity, so that the MDL principle should apply. We find that this principle suggests that the number of pseudocounts used should depend upon the data observed. An implementation of the implied procedure improves PSI-BLAST database search accuracy to a modest but statistically significant degree. This success suggests that the minimum description length principle may provide a fruitful perspective for considering other aspects of protein profile construction.

MATERIALS AND METHODS

Evaluating search accuracy

In this article, we evaluate the search accuracy of a baseline version of PSI-BLAST (blastpgp release 2.2.17) and several variants. The evaluation is based on a 'gold standard' for determining whether two sequences are related. We employ the ASTRAL 40 subset (24) of release 1.71 of the Structural Classification of Proteins (SCOP) database (25,26), excluding sequences from superfamilies with only one member. We divide these sequences into a training query set (odd numbered sequences when listed lexicographically) and a test query set (even numbered sequences). We use the training set to optimize parameter settings, but evaluate the resulting search programs using the test set.

PSI-BLAST is most effective at constructing PSSMs when it compares queries to a large sequence database, for which the true and false sequences relationships are in general unknown. Accordingly, our protocol has two phases. In the first phase, we use PSI-BLAST to compare each query to a frozen version (available from the authors upon request) of the non-redundant (nr) protein sequence database maintained by the National Center for Biotechnology Information (27) for four rounds, or less if convergence is reached earlier, saving the PSSM used in the last round as a 'checkpoint'. In the second phase, we use PSI-BLAST to compare the checkpointed PSSM to the SCOP database. For a given query sequence, we consider any sequences returned that were classified by SCOP as from the same superfamily to be true positives and any sequences classified as from different folds to be false positives. Sequences that are from the same fold but from different superfamilies are treated as neither true nor false positives because it is difficult to determine whether such sequences are in fact homologous. The training set contains 3609 queries having 111 809 true positives, and the test set 3609 queries with 109 133 true positives.

Given a ranked list of search results classified as true or false positives, Receiver Operating Characteristic (ROC) analysis provides a useful tool for measuring search accuracy (28). In brief, suppose that as one descends through this list, the cumulative number of true positives is plotted against the cumulative number of false positives. If there are t possible true positives, then the ROC_n score is the

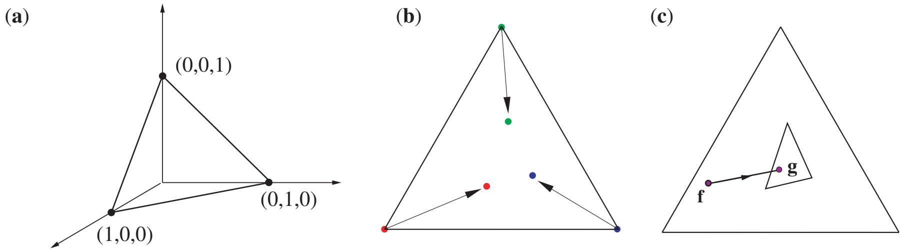

Figure 1. Frequency distribution space. (**a**) The frequencies for the case of three amino acids can be represented by points inside an equilateral triangle. (**b**) A substitution matrix maps the vertices of the simplex, each of which corresponds to the observation of a single amino acid, to target frequencies in the interior of the simplex. (**c**) The linear transformation implied by a substitution matrix may be applied to the whole of frequency distribution space, mapping any vector of observed frequencies **f** to a pseudocount frequency vector **g**.

area under this curve, up to n false positives, divided by $t \times n$. A ROC$_n$ score of 1.0 requires all true positives to be found before the first false positive and corresponds to perfect retrieval. In all our tests, we pool the search results from all queries, ranked by E-value, and evaluate search methods using the ROC$_{5000}$ score for the resulting list, which corresponds to about 1.4 errors per query. Other ROC$_n$ scores yield equivalent results, and we report both ROC$_{5000}$ and ROC$_{10\,000}$ scores below. Standard errors are calculated as described in Schäffer *et al.* (18).

Data-dependent pseudocounts

It is useful to develop a geometric understanding of the data-dependent pseudocount method. A set of target frequencies **q** is a vector with 20 elements. The elements of **q** are constrained to sum to 1, which reduces the degrees of freedom to 19. We represent the set of all possible **q** graphically as an equilateral triangle, by analogy to the case of only three amino acids (Figure 1a). In pairwise sequence comparison, each amino acid in the query sequence selects a particular row of the standard substitution matrix S to score amino acids with which it is aligned. A row of S in turn can be seen as corresponding to a vector of target frequencies (Figure 1b). An amino acid in the query can therefore be seen as a frequency vector **f** that happens to lie on a vertex of frequency space, and a score matrix can be seen as applying a linear transformation to **f** to obtain the frequency vector **g**, by the equation $\mathbf{g} = M\mathbf{f}$, where $M_{ij} = p_i \exp(\lambda S_{ij})$, and λ is the implicit scale of the substitution matrix (2,3). (Note that the substitution matrix S, the matrix of all target frequencies, and the linear transformation matrix M, while all related, are distinct.) This linear transformation can be applied to any observed frequency vector (Figure 1c), and one approach to estimating target frequencies **q** is simply to use this transformation, which yields the standard substitution matrix in the case of a single observation. It does not, however, have the desirable property of **q** approaching **f** for large numbers of observations.

The data-dependent pseudocount method estimates **q** with a linear combination of **f** and the **g** (Figure 2). Specifically,

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

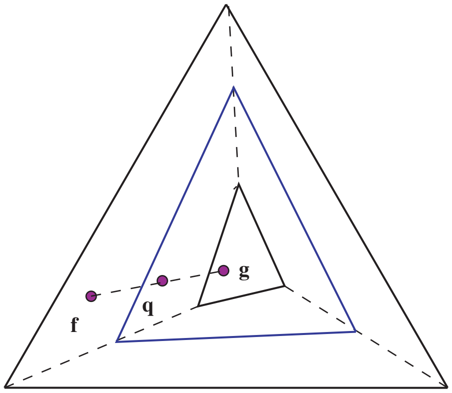

Figure 2. Linear transformations of frequency distribution space. A substitution matrix imposes a linear transformation M that maps each observed frequency vector **f** to a pseudocount vector **g**, and all of frequency distribution space to the smallest simplex shown. For values of α between 0 and 1, the use of pseudocounts imposes a linear transformation $M'(\alpha)$ that maps **f** to a point **q** on the line between **f** and **g**, and the frequency distribution space to the intermediate simplex shown.

where I is the identity matrix. The parameter α determines the relative weight given to the observed counts, which are proportional to **f**, and the data-dependent pseudocounts, which are proportional to **g**.

How should α depend on the effective number of counts n in a column? When n is 1, α should equal 1, so that **q** reduces to **g**. As n gets large, α should approach 0. Other than these limiting cases there are few theoretical constraints. PSI-BLAST has specified α using the formula

$$\alpha = \frac{m}{m + n - 1}, \qquad \mathbf{2}$$

where m is an empirically determined constant, although it was originally suggested that m might be chosen to grow proportionally with $\sqrt{n}$ (21,29).

We ran baseline PSI-BLAST on our training set using several different values of m. We found empirically the best integral value on the training set is 11. With $m = 11$, baseline PSI-BLAST attains a ROC$_{5000}$ of 0.2407 ± 0.0006 on the test set (Table 1).

Table 1. PSI-BLAST retrieval efficiency

PSI-BLAST program version	ROC_{5000} (± 0.0006)	$\mathrm{ROC}_{10\,000}$ (± 0.0004)
Baseline ($m = 11$)	0.2407	0.2572
New calculation of n ($m = 28$)	0.2419	0.2584
MDL principle ($m_0 = 5.5$)	0.2453	0.2628
Relative entropy formula ($m_0 = 5.5$, $a = 0.061$, $b = 0.8$)	0.2456	0.2631

Number of independent observations

In any application that adds pseudocounts to observed counts, one must first estimate the number of observed counts. Just as one must weigh sequences to account for correlations among them, so one must analyze a set of sequences to determine how many effectively independent observations they represent. For example, ten aligned sequences may all have a leucine in a particular column. If the ten sequences, although related, have low mutual sequence similarity, then the uniform appearance of leucine speaks strongly to the importance of this amino acid in this column. On the other hand, if the ten sequences are identical, then the evidence for the importance of leucine in this position is really no greater than the evidence from a single sequence.

To estimate accurately the number of effectively independent observations a column of data represents, it is valuable to have alignment data not just from the column in question, but from other columns as well. As described elsewhere (18), for a specific column from a multiple alignment produced by PSI-BLAST, we base our calculation on a reduced multiple alignment. Omitting a few details, this alignment is constructed from the set A of all sequences that participate in the alignment at this column, including those that have gaps there. Given A, we then consider the maximal set C of contiguous columns in which all sequences in A are aligned. Confining attention to the sequences in A, the baseline PSI-BLAST's estimate for the effective number n of independent sequences is the average number of distinct amino acids observed in the columns C, with nulls counted as a twenty-first amino acid. This method yields $n = 1$ independent observation for all columns from an optimal alignment of identical sequences, and n increases as the sequences aligned become mutually dissimilar. A clear disadvantage of the method, however, is that it saturates at a maximum of $n = 21$. Accordingly, we seek an improved method for estimating n.

Assume a model in which the amino acids in a column of size n are unconstrained, and have probabilities of occurrence $p_1, p_2, \ldots, p_{20}$. If the amino acids are chosen independently and at random, then the probability that an amino acid of type i occurs at least once is $1 - (1 - p_i)^n$. Therefore, the expected number N of distinct amino acids observed is given by

$$N = f(n) = 20 - \sum_{i=1}^{20}(1 - p_i)^n. \tag{3}$$

Table 2. Mean number of distinct amino acids N as a function of the number of independent sequences n

n	N	n	N	n	N	n	N
1	1.000	11	8.334	25	13.672	100	19.367
2	1.942	12	8.862	30	14.820	120	19.595
3	2.828	13	9.362	35	15.723	140	19.730
4	3.664	14	9.835	40	16.440	160	19.814
5	4.452	15	10.283	45	17.014	180	19.869
6	5.195	16	10.708	50	17.477	200	19.906
7	5.897	17	11.110	60	18.163	250	19.957
8	6.559	18	11.492	70	18.631	300	19.979
9	7.184	19	11.853	80	18.959	350	19.989
10	7.775	20	12.197	90	19.194	400	19.995

Because f is monotonic in n, and because it may be applied to nonintegral n, we may invert f to estimate the effective number of independent draws that correspond to N distinct observed amino acids. Table 2 lists some values of $f(n)$ calculated from the p_i implicit in BLOSUM-62.

As described above, although one may apply f^{-1} to the number of distinct amino acids N observed in a single column, it is generally better to calculate N as the average over as many columns as available. In calculating N, we count aligned gap characters as a twenty-first amino acid, but if there are 21 amino acids in a column, we set N for that column to 20.

The model yielding Equation (3) assumes that the amino acids in a column are under no evolutionary constraint. For most positions in real proteins this is unlikely to be the case, and many columns will be more highly conserved. Therefore, for an alignment with $|C|$ columns, we calculate N as the average over the $\lceil |C|/2 \rceil$ columns with the greatest number of distinct amino acids. For $N \geq 19.995$, we cap the estimated n at 400. Finally, it is possible that an alignment involving r actual sequences will yield an estimated n that is greater than r. When this happens, we set n equal to r.

As described in Equation (2), PSI-BLAST uses the number of independent sequences n in conjunction with the empirical pseudocount parameter m to construct its position-specific scores. With the new way of calculating n, the optimal value of m changes; using our training set, we estimate its value at 28. The new calculation of n, with $m = 28$, yields a ROC_{5000} score on the test set of 0.2419 ± 0.0006 (Table 1). This is better than the baseline calculation of n, with $m = 11$, but the improvement is of marginal significance. The new calculation of n is used in all other variants of PSI-BLAST described below.

The minimum description length principle and protein profiles

The minimum description length principle has an extensive literature but has been little applied in the field of protein and DNA sequence alignment. In general, it proposes an answer to the question of which model to choose to describe a set of data, when various models of varying complexity are available. Models with a greater number of parameters will, in general, fit the available data better. However, once a certain level of complexity is reached they begin to overfit the data—they describe the observed data more precisely but at the cost of describing underlying regularities less well. Thus models that are too complex, as well as models that are too simple, do a relatively poor job of predicting new data.

Informative reviews of the MDL principle can be found in references (23) and (30). To simplify matters somewhat, given a set of data, we generally choose a ‘theory’ that best fits the data from among a parameterized set of theories, called a ‘model’. There may, however, be various models available, such as, in certain applications, the set of all linear functions, the set of all quadratic functions, etc. How complex should the model be from which to choose the best theory? The MDL principle observes that a description of the data can usually be divided into two parts: a description of the theory used to describe the data, and a description of the data given the theory. It proposes that the best or most predictive theory will be that which minimizes the sum of these two description lengths. To apply this principle, one needs to be able to quantify the description lengths.

Generally, the easier description length to calculate is that of the data given the theory. If the set of possible outcomes is discrete, a theory will assign them probabilities $P_1, P_2, \ldots$. From information theory, the description length of each data point corresponding to outcome i is $-\log_2 P_i$ bits (31). The description length of the data is then just of the sum of this quantity over all data points. If there are n data points, and they follow the probability distribution $\mathbf{f}$, for large n their minimum description length, using $\mathbf{f}$ as a theory, approaches

$$-n\sum_i f_i \log_2 f_i = nH(\mathbf{f}) \text{ bits}, \tag{4}$$

where $H(\mathbf{f})$ is the entropy of $\mathbf{f}$ (31). If instead, one were to describe the data using the distribution $\mathbf{q}$ as a theory, the description length would be $-n\sum_i f_i \log_2 q_i$ bits. In other words, the description length of the data would be increased by

$$n\sum_i f_i \log_2 \frac{f_i}{q_i} = nD(\mathbf{f}||\mathbf{q}) \text{ bits}, \tag{5}$$

where $D(\mathbf{f}||\mathbf{q})$ is the relative entropy of $\mathbf{f}$ and $\mathbf{q}$ (31).

It is harder to quantify the description length of the theory. This is best taken as a number attached to the complexity of the model from which the theory is chosen (23,30). Because theories with nearly identical parameters are not independent, a model can be understood to encompass a certain number of effectively independent theories. For a parameterized set of theories, in the limit of a large number of observations n, the density of independent theories can be thought of as proportional to the square root of the model’s Fisher information (23,30,32). Integrating this quantity over the parameters’ range yields a measure of the number of effectively independent theories a model contains. The description length of a model is the log of this number, i.e. the amount of information required to specify a particular theory within the model.

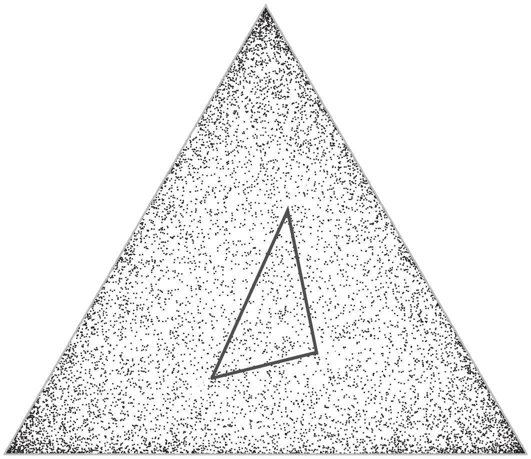

Figure 3. The effect of pseudocounts on the number of independent theories. A theory describing the frequency distribution of three amino acids can be represented as a point within an equilateral triangle. The density of independent theories, which is proportional to $(\prod p_i)^{-1/2}$, is represented by the density of points within the triangle, and increases as one moves away from its center. Using pseudocounts confines the theories one may consider to points within a simplex inside the frequency distribution space. This simplex has a smaller volume than the complete space, and also has a smaller average density of independent theories.

Let us examine how these considerations may be applied to the specification of protein profiles, and more specifically to the estimation of amino acid target frequencies $\mathbf{q}$ for a single profile position. One may model n data points using a multinomial distribution with expected value $\mathbf{q}$. The Fisher information of this model is

$$\frac{n}{\prod_i q_i}. \tag{6}$$

For a fixed number of degrees of freedom, a constant times the square root of this quantity may be integrated over the parameter space to estimate the number of independent theories described by the model. To illustrate, if one assumes a multinomial model with three rather than twenty amino acids, the density of independent theories is shown in Figure 3, where the points are chosen randomly, but with density proportional to the square root of the Fisher information. For a multinomial with k free parameters (corresponding to $k+1$ amino acids), it is possible to show (Supplementary Data A) that in the limit of large n, the description length of the model approaches

$$\frac{k}{2}\log_2\frac{ne}{k} - \frac{1}{2} \text{ bits}. \tag{7}$$

For the standard amino acid alphabet, $k = 19$.

The usual application of the MDL principle is to select a model with an appropriate number of parameters. We will, however, use it differently, as described below, to select an appropriate number of pseudocounts. This requires additional technical assumptions which are described in Supplementary Data B.

The MDL principle and pseudocounts

Given a column with n effective observations, and observed frequencies $\mathbf{f}$, the choice of $\alpha = 0$ in Equation (1) yields the maximum likelihood estimate $\mathbf{q} = \mathbf{f}$. The description length of the data using $\mathbf{q}$ as a theory is then $-n\sum_i f_i \log_2 f_i = nH(\mathbf{f})$. However, if $\alpha \neq 0$, the description length of the data increases to $-n\sum_i f_i \log_2 q_i$. In other words, the description length of the data increases by $nD(\mathbf{f}||\mathbf{q})$; see Equation (5). By the MDL principle, this can be advantageous only if the description length of the model decreases by a greater amount.

As discussed above, the effective number of independent theories corresponding to a multinomial can be estimated by integrating a constant times the square root of the Fisher information over the whole parameter space. As illustrated in Figure 3, when pseudocounts are employed with $\alpha \neq 0$, the volume of parameter space in which the predicted $\mathbf{q}$ can fall decreases. Subject to technical assumptions described in Supplementary Data B, the decrease in the model description length may be derived from an integral over this smaller volume. Specifically, this decrease, in bits, is the difference between the logarithm base 2 of the two integrals. We are unable to calculate the integral over the smaller volume analytically, but can do so numerically.

Figure 4 shows, on a logarithmic scale, the effect that applying pseudocounts, with the matrix $M'(\alpha)$ implied by BLOSUM-62, has on the description length of the multinomial model, as α ranges from 0 to 1. Positive values on the y-axis represent decreases in model description length with respect to $\alpha = 0$. Increasing α affects the description length in two ways: first, the volume in which $\mathbf{q}$ can fall decreases; second, the average density of independent theories within this volume decreases. The first effect, labeled 'volume effect' in Figure 4, can be calculated analytically. By Equation (1), using pseudocounts to calculate $\mathbf{q}$ is equivalent to multiplying the observed frequencies $\mathbf{f}$ by the matrix $M'(\alpha)$. This has the effect of multiplying the volume of parameter space in which $\mathbf{q}$ can fall by the absolute value of the determinant of $M'(\alpha)$. The second effect, labeled 'density effect' in Figure 4, must be computed numerically. We sampled 5×10^9 vectors uniformly from the set of all possible amino acid frequency vectors. For α ranging from 0 to 1, in increments of 0.002, we applied $M'(\alpha)$ to all of our sampled frequency vectors and calculated the average value of the square root of the Fisher information over the resulting points. The total decrease in the description length of the multinomial model is obtained by adding the volume and density effects, and is labeled 'total effect' in Figure 4. Note that the curves shown in Figure 4 are valid only in the limit of large n. Although this will constrain our ability to apply the MDL principle in detail, it will still allow us to draw several valuable conclusions.

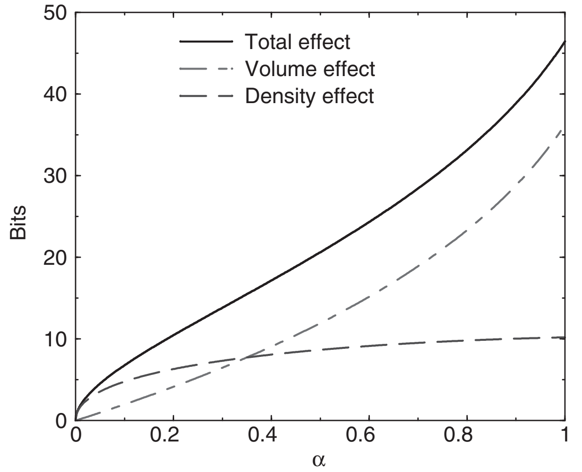

Figure 4. Decrease in model description length as a result of using pseudocounts implied by the BLOSUM-62 substitution matrix. For large n, one may calculate the decrease in the description length of the model for α between 0 and 1, compared to the description length of the model at $\alpha = 0$. The total decrease can be decomposed into a decrease in simplex volume, and a decrease in independent theory density. Half as many independent theories corresponds to a decrease of one bit. Positive values on the y-axis represent decreases in model description length.

To summarize, as α and the number of pseudocounts increases, the description length of the data $\mathbf{f}$, given the calculated theory $\mathbf{q}$, increases (Supplementary Data C). However, the description length of the multinomial model, as measured by the logarithm of the effective number of independent theories it comprises, decreases, as shown in Figure 4. The MDL principle claims that the optimal value of α will be that for which the sum of these two description lengths is minimized.

RESULTS

Pseudocounts as a function of the number of independent observations

To investigate the implications of the MDL principle, we first examine a toy hydrophobic protein column. Table 3 shows the observed amino acid frequencies $\mathbf{f}$ for this column and the background frequencies $\mathbf{p}$ implicit in the BLOSUM-62 matrix. In Figure 5, we plot, for these frequencies and $n = 500$ observations, the change in the data, model, and total description lengths with respect to a base at $\alpha = 0$. For the model and total description lengths, a positive value in the plot indicates a 'decrease' in the description length, whereas for the data description

Table 3. The observed frequencies f_i of a toy hydrophobic alignment column, and the background probabilities p_i of BLOSUM-62

	f_i	p_i		f_i	p_i
A	0.010	0.074	M	0.200	0.025
C	0.010	0.025	N	0.010	0.045
D	0.001	0.054	P	0.010	0.039
E	0.001	0.054	Q	0.010	0.034
F	0.050	0.047	R	0.001	0.052
G	0.010	0.074	S	0.003	0.057
H	0.010	0.026	T	0.003	0.051
I	0.200	0.068	V	0.200	0.073
K	0.001	0.058	W	0.020	0.013
L	0.200	0.099	Y	0.050	0.032

length a positive value indicates an ‘increase’. The data description length was computed using Equation (5), whereas the model description length was computed numerically as described in the previous section. The minimum total description length is at $\alpha = 0.0375$. Solving Equation (2) for m yields

$$m = (n - 1)\frac{\alpha}{1 - \alpha} \;, \qquad \textbf{8}$$

which implies the optimal number of pseudocounts is $m = 19.4$.

We can perform the same calculation for varying numbers n of independent observations. While the curve in Figure 5 showing change in the description length of the model remains fixed, the curve representing the increase in the description length of the data shifts upward with increasing n. Accordingly as n grows, the optimal α will decrease, although it is not obvious how the implied optimal *number* of pseudocounts m will behave. In Figure 6, we plot the calculated value of m as a function of n. It is evident that while not precisely constant, the value of m is almost unchanging for n between 300 and 1000.

What can we make of the apparent divergence of m for small n? As stated above, the theory that allows us to calculate the description length of the model is valid only for large n. As seen in Figure 4, when $\alpha = 1$, we calculate that the description length of the model is about 46.3 bits less than when $\alpha = 0$. However, by formula 7, the total description length of the model is less than 46.3 bits when n is less than 212. Thus, our calculation cannot be accurate in this range of n. Not having a good way to apply the MDL principle for small n, we take the near constancy of m for n between 300 and 1000 as an indication that m can be assumed to be nearly constant for all n less than 1000. This conforms with the empirical result that a constant m outperforms one proportional to $\sqrt{n}$.

If our asymptotic theory can deal with small n only by implication, what does it say about very large n? Except when $\mathbf{f}$ is equal to the background frequencies, in which case applying pseudocounts has no effect, the description length of the data is a strictly increasing function of α. The calculated value of the description length of the model does not depend on n, but the curve representing the change in the data shifts upward with increasing n. Therefore, α must converge to zero as n grows large, though it converges at different rates for different values of $\mathbf{f}$.

It is possible to show that for small α, the decrease in model description length is approximately $k_1\sqrt{\alpha}$ (Supplementary Data D), whereas the increase in data description is approximately $k_2 n\alpha^2$. Thus the total decrease in description length is given by

$$\Delta(\alpha) \approx k_1\sqrt{\alpha} - \; k_2 n\alpha^2, \qquad \textbf{9}$$

which reaches a maximum at $\alpha = k_3 n^{-2/3}$. By Equation (8), this implies that the optimal number of pseudocounts should grow as $n^{1/3}$. We have seen that, for our toy column, m is almost constant for $n < 1000$, so the asymptotic growth of m with the cube root of n does not have appreciable effect until n is very large indeed. It would be hard to test this theoretical result empirically, because multiple alignments with which we are likely to deal will not have n sufficiently large.

Pseudocounts as a function of column composition

So far, the MDL principle has only confirmed the earlier empirical result that for practical alignment sizes the

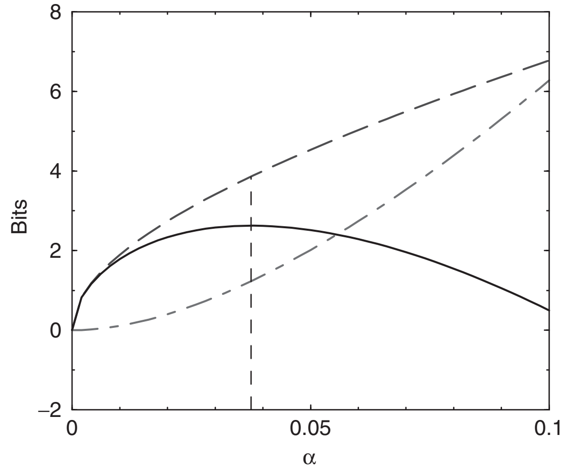

Figure 5. Selecting an optimal proportion of pseudocounts using the MDL principle. For $n = 500$ and the observed frequencies $\mathbf{f}$ listed in Table 3, we apply pseudocounts as implied by the BLOSUM-62 substitution matrix. We use Equation (5) to compute the change in the description length of the data, when compared to the description length of the data at $\alpha = 0$, for α between 0 and 0.1. The dot-dashed curve (in red) shows the increase in the description length of the data. The dashed curve (in blue) shows the decrease in the description length of the model. The total decrease in the description length, shown by the solid curve (in black), is maximized at $\alpha = 0.0375$, which corresponds to 19.4 pseudocounts.

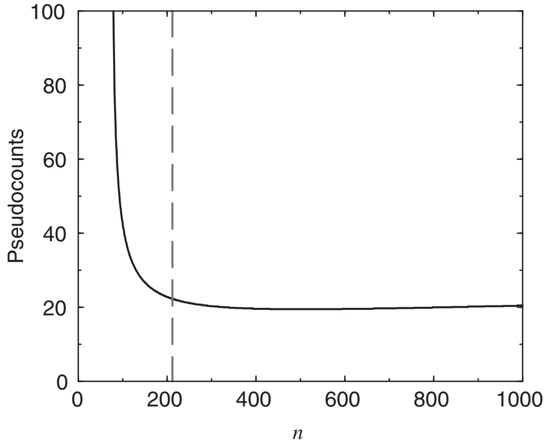

Figure 6. Optimal number of pseudocounts, m, as a function of the number of independent observations, n. Using the data listed in Table 3 and the method illustrated in Figure 5, we found the optimal number of pseudocounts for varying n. The method cannot be valid for $n < 212$ (vertical dotted line), because the calculated decrease in model description length for $\alpha = 1$ is greater than the description length of the model at $\alpha = 0$, but it is not possible for a model to have a negative description length. For n between 212 and 1000, the calculation suggests we use a nearly constant number m of pseudocounts, roughly 19.4. In the limit of very large n, the MDL principle suggests the number of pseudocounts should grow proportionately to $n^{1/3}$.

number of pseudocounts should not depend on n. In addition, the principle suggests that the number of pseudocounts should depend on the composition of the column: for different observed frequencies, the increase in data description length as a function of α differs, but the reduction in model description length as a function of α remains fixed (Figure 5).

To test this prediction, we implemented an MDL routine to optimize α, and thereby the number of pseudocounts m, for different observed frequency vectors $\mathbf{f}$. Our theory may be used properly only for a large number n' of independent observations, although then the number of pseudocounts it yields is effectively independent of n'. Therefore, whatever the actual value of n, we always apply the theory with $n' = 500$. The value 500 here is somewhat arbitrary, and a larger value of n' could be used with almost identical results.

We cannot, however, escape a certain degree of circularity, because when we assume that n' is large, it is not reasonable to claim as well that many of the observed frequencies are zero (as is frequently the case when n is small), and doing so yields poor results. Accordingly, before we apply the MDL theory, it is best to eliminate zero frequencies. We use Equations (1) and (2) with a small, fixed number m_0 of initial pseudocounts, to transform the observed frequency vector $\mathbf{f}$ to $\mathbf{f}'$ using the equation

$$\mathbf{f}' = M'\left(\frac{m_0}{m_0 + n - 1}\right)\mathbf{f}. \qquad 10$$

We then apply the MDL theory to $\mathbf{f}'$ and $n' = 500$ to obtain m, the estimated optimal number of pseudocounts. Finally, we estimate the target frequencies $\mathbf{q}$ using the equation

$$\mathbf{q} = M'\left(\frac{m}{m + n - 1}\right)\mathbf{f}. \qquad 11$$

We select m_0 to optimize retrieval on our training set, achieved at $m_0 = 5.5$. Using this value, and the procedure just described, we then study the effectiveness of the MDL method on our test set. The ROC_{5000} score is 0.2453 ± 0.0006, a significant albeit modest improvement on that of the constant pseudocount method (Table 1).

Pseudocounts as a function of column entropy

There is a strong correlation between the value of α chosen by the MDL principle and the relative entropy of a column's amino acid distribution to the background distribution. To study this relationship, we constructed a large set of columns representative of typical protein sequence alignments. Specifically, we ran PSI-BLAST, using the new method for calculating n and 28 pseudocounts, on the 103 queries of the aravind103 query set (18,33), against SWISS-PROT (34). For ease of implementation, column data were recorded on the fifth round for all non-X positions in those query sequences for which the search did not converge before the full five rounds were completed. We considered those 10 875 columns representing at least five independent observations. For each, we calculated the α implied by the MDL method described above. In Figure 7, we plot α versus $D(\mathbf{f}'||\mathbf{p})$ (abbreviated below as D) on a log-log scale, along with a linear regression line, shown in red. As can be seen, there is a good (correlation coefficient −0.87) linear correlation, implying an approximate power-law relationship

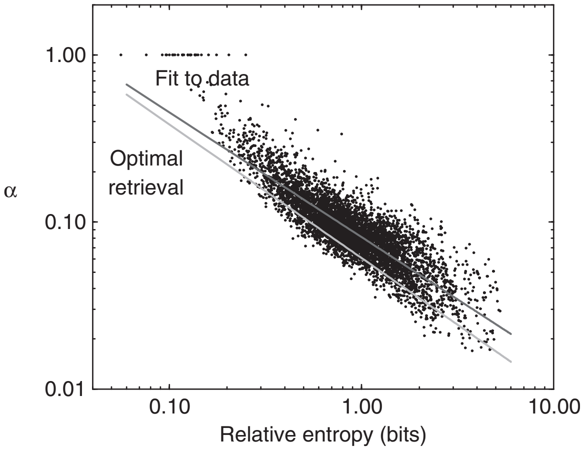

Figure 7. The relationship between the pseudocount proportion α implied by the MDL principle and column relative entropy. Each point represents a multiple alignment column constructed by PSI-BLAST from the aravind103 query set (18,33) run against SWISS-PROT (34). Only columns with $n \geq 5$ independent observations are considered. The x-axis represents the relative entropy $D(\mathbf{f}' \,||\, \mathbf{p})$, where $\mathbf{f}'$ is the observed frequency vector of the column after the addition of $m_0 = 5.5$ pseudocounts, and $\mathbf{p}$ is the background amino acid frequency vector implicit in BLOSUM-62. The y-axis represents the pseudocount proportion α calculated from the MDL theory. The upper diagonal line (shown in red) represents the best power-law fit to the data. The lower diagonal line (shown in green) represents the power-law relationship of α to $D(\mathbf{f}' \,||\, \mathbf{p})$ that empirically yields the optimal retrieval on the training set. Note that the background frequency vector $\mathbf{p}$ is the fixed point of the linear transformation M. Therefore, if $\mathbf{f}' = \mathbf{p}$, the increase in the description length of the data is identically zero for all α, implying that the MDL is optimized at $\alpha = 1$. For any finite n, vectors $\mathbf{f}'$ close enough to $\mathbf{p}$ also imply an optimal α of 1. A small number of such points are seen at the upper left of this graph.

$$\alpha = a\, D^{-b}, \qquad 12$$

with $a \approx 0.081$ and $b \approx 0.75$, when D is expressed in bits. Qualitatively, columns that are unlike the background frequencies imply a low α, and relatively few pseudocounts, while columns that are similar to the background frequencies imply a high α. In comparison to constant pseudocounts, this will tend to render the substitution scores implied by low relative-entropy columns even closer to zero, while it will tend to render the scores of high relative-entropy columns greater in absolute value. In other words, it will tend to increase the 'contrast' of the implied PSSMs.

One question that arises from examining Figure 7 is whether a simple power-law formula expressing α as a function of D might perform as well or better than the complicated MDL procedure for calculating α. Assuming an equation of the form (12), and $m_0 = 5.5$, we sought to optimize retrieval on our training set by varying a and b. The best values we could find were

$a = 0.061$ and $b = 0.8$, and the implied power law is shown as a green line in Figure 7. On our test set, the simple formula (12) yields a ROC_{5000} score 0.2456 ± 0.0006 (Table 1), statistically no different than that of the MDL theory.

DISCUSSION

We have shown that a new method for estimating the number *n* of independent observations represented by a multiple alignment column leads to improved retrieval accuracy. This method is now used in the PSI-BLAST code maintained and distributed by the NCBI (blastpgp release 2.2.18).

Nishida *et al.* (35), in a paper published jointly with this one, have made an empirical study of the use of pseudocounts in the construction of DNA position-specific score matrices. Like us, they conclude that the number of pseudocounts should be independent of the number of sequences in the source multiple alignment, at least for alignments of realistic size. Our papers also agree that the number of pseudocounts should decrease for alignments with greater relative entropy to the background distribution, although Nishida *et al.* (35) consider adjusting pseudocounts for complete PSSMs, whereas we consider varying them on a column-by-column basis. Henikoff and Henikoff (22) have also proposed that, in the protein alignment context, the number of pseudocounts should be decreased for columns with high relative entropy, although without any strongly argued motivation.

We improved the retrieval performance of PSI-BLAST by making the number of pseudocounts dependent upon a column's composition. The MDL principle provides one justification for this procedure, but it is possible to derive a similar PSSM score adjustment from a different theoretical perspective. For example, one might argue that because different protein positions evolve at different rates, the appropriate substitution matrix to use for a slowly evolving position is one corresponding to a lower PAM distance (4), or one with a greater relative entropy. Therefore, instead of varying the number of pseudocounts as a function of a the relative entropy of a column, one might instead vary the substitution matrix from which the pseudocounts are derived. If one examines Figure 2 and Equation (1), one may observe that decreasing the number of pseudocounts decreases α and moves the derived target frequency vector $\mathbf{q}$ closer to $\mathbf{f}$. Alternatively, decreasing the PAM distance of the underlying substitution matrix expands both inner simplexes and also moves $\mathbf{q}$ towards $\mathbf{f}$, without varying α or the number of pseudocounts. Thus, this alternative perspective has the same qualitative effect on PSSM scores as does the MDL principle.

As seen above, statistically indistinguishable performance is achieved by using the MDL principle to calculate an appropriate number of pseudocounts, and by using an empirical formula that calculates the number of pseudocounts as a function of column relative entropy. The PSI-BLAST code maintained and distributed by the NCBI (blastpgp release 2.2.18) now implements by default this empirical formula for calculating pseudocounts. The number of initial pseudocounts m_0 used to calculate $\mathbf{f}'$ and the parameters *a* and *b* in formula (12) may be modified based on further testing. A user may override the rule for calculating pseudocounts by specifying a fixed number of pseudocounts; for BLOSUM-62 we recommend 28–30. The new method of computing the effective number of observations is used, whether or not the number of pseudocounts is fixed.

As described, alternative theoretical formulations can yield adjustments similar to those implied by the MDL principle in the calculation of target frequencies and PSSM scores. Ultimately, it is the quality of retrieval that is important, not the theory behind target frequency construction. Nevertheless, theories are important in that they can suggest fruitful avenues for further investigation. The MDL theory supports the use of an essentially constant number of pseudocounts over the range of alignment sizes we are likely to encounter. It also predicts that pseudocount number should in general decrease with increasing column relative entropy. That these results are consistent with empirical retrieval performance suggests that the MDL principle provides a useful perspective when thinking about protein model construction.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Intramural Research Program of the National Library of Medicine of the National Institutes of Health. Funding for open access charge: Intramural Research Program of the National Library of Medicine of the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
2. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
3. Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.
4. Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, DC, pp. 345–352.
5. Schwartz,R.M. and Dayhoff,M.O. (1978) Matrices for detecting distant relationships. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, DC, pp. 353–358.
6. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
7. Altschul,S.F., Carroll,R.J. and Lipman,D.J. (1989) Weights for data related by a tree. *J. Mol. Biol.*, **207**, 647–653.
8. Sibbald,P.R. and Argos,P. (1990) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.*, **216**, 813–818.
9. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Protein Struct. Funct. Genet.*, **9**, 56–68.

10. Gerstein,M., Sonnhammer,R. and Clothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078. (Appendix: A method to weight protein sequences to correct for unequal representation.).
11. Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.
12. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.*, **10**, 19–29.
13. Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, **2**, 9–23.
14. Gotoh,O. (1995) A weighting system and algorithm for aligning many phylogenetically related sequences. *Comput. Appl. Biosci.*, **11**, 543–551.
15. Krogh,A. and Mitchison,G. (1995) Maximum entropy weighting of aligned sequences of protein or DNA. In Rawlings,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of the Third International Conference on Intelligent system for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 215–221.
16. Bailey,T.L. and Gribskov,M. (1996) The megaprior heuristic for discovering protein sequence patterns. In States,D.J., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R. (eds), *Proceedings of the Fourth International Conference on Intelligent system for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 15–24.
17. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
19. Brown,M., Hughey,R., Krogh,A., Mian,I.S., Sjölander,K. and Haussler,D. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter,L., Searls,D. and Shavlik,J. (eds), *Proceedings of the First International Conference on Intelligent system for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 47–55.
20. Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I.S. and Haussler,D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.
21. Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
22. Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, **12**, 135–143.
23. Grünwald,P.D. (2007) *The Minimum Description Length Principle*, MIT Press, Cambridge, MA.
24. Chandonia,J.-M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
25. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
26. Brenner,S.E., Chothia,C. and Hubbard,T.J. (1998) Assessing sequence comparison methods with reliable structurally identified-distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
27. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. and Federhen,S. (2008) Database resources of the National Centerfor Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
28. Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.
29. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
30. Grünwald,P.D. (2005) Minimum description length tutorial. In Grünwald,P.D., Myung,I.J. and Pitt,M.A. (eds), *Advances in Minimum Description Length: Theory and Applications*, MIT Press Cambridge, MA, pp. 23–79.
31. Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*, Wiley, New York.
32. Fisher,R.A. (1925) Theory of statistical estimation. *Proc. Cambridge Phil. Soc.*, **22**, 700–725.
33. Schäffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
34. Boeckmann,B., Bairoch,A., Apweiller,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
35. Nishida,K., Frith,M.C. and Nakai,K. (2008) Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.* (in press).