From: Journal of Computational Biology, **4**(45–59), 1997.

Score distributions for simultaneous matching to multiple motifs

Timothy L. Bailey* Michael Gribskov

San Diego Supercomputer Center
P.O. Box 85608
San Diego, California 92186-9784

Email: {tbailey, gribskov}@sdsc.edu
Phone: (619) 534-8350
Fax: (619) 534-5113

*To whom correspondence should be addressed.

Abstract

Several computer algorithms now exist for discovering multiple motifs (expressed as weight matrices) that characterize a family of protein sequences known to be homologous. This paper describes a method for performing similarity searches of protein sequence databases using such a group of motifs. By simultaneously using all the motifs that characterize a protein family, the sensitivity and specificity of the database search are increased. We define the p-value for a target sequence to be the probability of a random sequence of the same length scoring as well or better in comparison to all the motifs that characterize the family. (The p-value of a database search can be determined from this value and the size of the database.) We show that estimating the distribution of single motif scores by a Gaussian extreme value distribution is insufficiently accurate to provide a useful estimate of the p-value, but that this deficiency can be corrected by reestimating the parameters of the underlying Gaussian distribution from observed scores for comparison of a given motif and sequence database. These parameters are used to calculate an "reduced variate" which has a Gumbel limiting distribution. Multiple motif scores are combined to give a single p-value by using the sum of the reduced variates for the motif scores as the test statistic. We give a computationally efficient approximation to the distribution of the sum of independent Gumbel random variables and verify experimentally that it closely approximates the distribution of the test statistic. Experiments on pseudorandom sequences show that the approximated p-values are conservative, so the significance of high scores in database searches will not be overstated. Experiments with real protein sequences and motifs identified by the MEME algorithm show that determining an overall p-value based on the combination of multiple motifs gives significantly better database search results than using p-values of single motifs.

Keywords: protein motifs; profiles; score p-values; score normalization; extreme-value distributions; sum statistics

1 Introduction

SEARCHING DATABASES OF PROTEIN OR DNA SEQUENCES for known patterns has become a matter of course in molecular biology. Many methods for doing this involve calculating a similarity score between the pattern and each sequence in the database. Commonly used patterns include sequences [Pearson, 1990; Altschul *et al.*, 1990; Smith and Waterman, 1981], aligned blocks of sequences [Pietrokovski *et al.*, 1996], profiles [Gribskov *et al.*, 1990] and regular expressions (Prosite signatures) [Bairoch, 1995]. Another type of pattern is the motif model [Bailey and Elkan, 1994]. A motif model describes a (gapless) sequence pattern by specifying the probability of each letter in the alphabet at each position of the motif. For instance, a protein pattern of width w would have a motif model with w columns each containing 20 entries, one for each of the 20 amino acid residues that could occur at each position in a protein pattern. Motif models are thus a kind of weight matrix. They are usually converted into "log-odds" matrices by dividing each frequency in each column by the background frequency of the corresponding letter and taking the logarithm. Motifs are a generalization of consensus sequences [Stormo, 1990] and a specialization of profiles to cases where the gap opening costs are infinite.

The main goal of searching a sequence database with a pattern is to sort the database according to the degree to which the sequences match the pattern. Most often, this similarity is interpreted as evidence of homology–common evolutionary ancestry. In other cases it can be viewed as an indication of common function or convergent evolution. A secondary, but extremely desirable, goal of a sequence database search is to assign to each sequence an estimate of the likelihood that the degree of match to the pattern is truly significant. One way of answering this question is to assert that a score is (very) significant if it is (very) unlikely to have arisen by chance. This is generally taken to mean that we want to know the p-value of score x–the probability that a random sequence would have score equal to or greater than x. The model for a random sequence is usually the independence (iid) model–all positions in the sequence have the same letter distributions and are independent of each other.

For the p-value of a score to be well defined, we must decide whether or not we wish to consider the length of the sequence. Longer random sequences have higher average scores than shorter sequences using search patterns of

the types we have mentioned because longer sequences have more positions. However, if we assume that each sequence in the database is equally likely *a priori* to match the pattern, then the definition of the p-value of score x becomes "the probability that a random sequence *of the same length* would have score equal to or greater than x." Experience has shown that scores compensated for sequence length generally sort the database more accurately, so this latter definition is used in this paper.

The p-values of sequence scores can be used to achieve both the above goals of sequence database searching. We define the p-value of a sequence to be the p-value of its match score to the pattern in question. Since the p-value of a sequence is a measure of the degree to which the sequence matches the pattern, it is reasonable to sort the database according to p-values, satisfying our first goal. To satisfy the second goal of determining the signficance of a match between a target sequence and a pattern in a database search, we must take into account the number of sequences in the database. One way to do this is to multiply the p-value of the target sequence by the number of sequences in the database. This gives the expected number of sequences which would have as good or better a p-value in a random database of the given size. Alternatively, we can compute p_{db}, the probability of at least one sequence having as good or better a p-value in a database of random sequences. If p is the p-value of the target sequence and n is the number of sequences in the database being searched, then p_{db} is given by

$$p_{db} = 1 - (1 - p)^n.$$

Either of these methods can be used to evaluate the significance of the observed scores in a database search.

The focus of this paper is on calculating p-values for multiple motif scores. Multiple motifs are groups of motifs that together define a pattern. For example, a multiple alignment of a family of distantly related proteins will often show a few regions of high similarity separated by regions were insertion, deletion and mutation events have been more frequent. Each of the regions of high similarity describes a motif for which a model can be constructed.[1] The multiple motifs present in a family of sequences can be viewed as a

[1]Computer algorithms such as MEME [Bailey and Elkan, 1995], the Gibbs sampler [Lawrence *et al.*, 1993] and Protomat [Henikoff and Henikoff, 1991] exist to assist in the automatic construction of motif models.

pattern that defines the family. The presence or absence of each motif in a target sequence is evidence for or against its membership in the family. Since each motif gives an independent measure of membership in the family, combining the scores for each of the motifs defining a family will often be more effective at separating the members of the family from all other sequences in a database search.

We first develop a method for calculating p-values of single motif scores and test it on simulated sequence data. We then extend the method to the case of multiple motifs and test it on simulated sequence data. Finally, we validate the method on actual sequence and motif data.

2 Approximate distribution of single motif scores

We want to know what the distribution of scores is for random sequences of different lengths when compared to a given motif model. We define the score of a sequence as the maximum score of any of its subsequences. To calculate this *sequence score*, we imagine sliding the motif log-odds matrix along the sequence. At each position, we calculate the *subsequence score* by summing one value from each column of the matrix. The value to be summed is determined by the letter at that position in the sequence. These definitions are summarized below in Eqns. (1) and (2).

Suppose we have a motif log-odds matrix S of width W and a sequence X of length L. The score given for letter a appearing at position i of an occurrence of the motif is $S_{a,i}$. Let $X(i)$ be the letter at position i in the sequence. For each position $1 < i < L - W + 1$ along the sequence, we compute the subsequence scores

$$S(X,i) \;=\; \sum_{j=1}^{W} S_{X(i+j-1),j}. \tag{1}$$

$S(X,i)$ is thus the comparison score for the match between the motif model and the subsequence of X beginning at position i. The sequence score for sequence X is its maximum subsequence score,

$$M(n) \;=\; \max_{1 \le i \le n} S(X,i), \tag{2}$$

where $n = L - W + 1$ is the number of positions where an occurrence of the motif would fit without overhanging one of the ends of the sequence.

Goldstein and Waterman [1994] showed that the maximum score of a motif model against a single *random* sequence has a limiting Gaussian extreme value distribution (GEV). This can be understood intuitively in the following manner. We make two simplifying assumptions. First, we assume that the subsequence scores at each position along the sequence are independent.[2] Second, we assume that the subsequence scores are normally distributed with mean μ and standard deviation σ. Under these assumptions, the sequence score $M(n)$ is a random variable with a Gaussian extreme value distribution. Since $M(n)$ is the maximum of n iid $N(\mu, \sigma)$ random variables, the formula for the reduced variate $T(n)$ for $M(n)$ is [Kinnison, 1985]

$$T(n) = \frac{M(n) - u(n)}{a(n)} \tag{3}$$

where[3]

$$u(n) = \mu + \sigma(\sqrt{2\ln(n)} - \frac{\ln(\ln(n)) + \ln(4\pi)}{2\sqrt{2\ln(n)}}), \tag{4}$$

and

$$a(n) = \frac{\sigma}{\sqrt{2\ln(n)}}. \tag{5}$$

We can now write an expression for the average maximum score for a sequence of length L. It is the expected value of $M(n)$,

$$E[M(n)] \approx u(n) + \gamma a(n), \tag{6}$$

where gamma is Euler's constant $0.5772156649\ldots$. This formula gives the expected score of a sequence of length $L = n + W - 1$.

[2]For a random sequence, this is strictly true only for the subsequence scores of positions in the sequence separated by W or more.

[3]Throughout this paper, $\ln(x)$ is the natural logarithm of x, whereas $\log(x)$ refers to the base-10 logarithm of x.

All that is necessary to compute $E[M(n)]$ is the mean μ and variance σ^2 of the subsequence scores $S(X,i)$. Suppose we estimate μ and σ^2 by the sample mean

$$\hat{\mu} = \sum_{i=1}^{N} \sum_{j=1}^{L_i-W+1} \frac{S(X_i,j)}{m}, \tag{7}$$

and sample variance

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \sum_{j=1}^{L_i-W+1} \frac{(S(X_i,j)-\hat{\mu})^2}{m-1}, \tag{8}$$

of subsequence scores for a given motif model on a given dataset, where each X_i is a sequence, L_i is its length, and m is the total number of subsequence scores $S(X_i,j)$ summed in the calculation of $\hat{\mu}$.

Empirical studies using many different motif models and many different pseudorandom sequence datasets show that, using (7) and (8), respectively, as estimates of the mean and variance, Eqn. (6) is a poor estimate of the score of an average sequence even when all the sequences are very long. This is because the results of Goldstein and Waterman [1994] are asymptotic in the width, W, of the motif, not just in the lengths of the sequences being searched. However, if we are willing to adjust our estimates of μ and σ slightly, the expression for $E[M(n)]$ in Eqn. (6) can be fit to the observed data. This suggests that the sequence scores are following a Gaussian extreme value distribution with slightly different underlying mean and variance.

Figure 1 illustrates the discrepancy between the mean sequence score predicted by Eqn. (6) and the observed mean sequence score, as a function of motif width. The error in the mean sequence score is normalized for motifs of different widths by dividing by the observed standard deviation of the sequence scores. That is, we define the error as

$$\frac{E[\hat{W}(n)]-\mu_{seq}}{\sigma_{seq}},$$

where $E[\hat{W}(n)]$ is the predicted mean score using the *subsequence* score sample mean and standard deviation, $\hat{\mu}$ and $\hat{\sigma}$, and μ_{seq} and σ_{seq} are the *sequence* score sample mean and standard deviation. The graph illustrates that the error in the predicted mean sequence score using the GEV approximation is

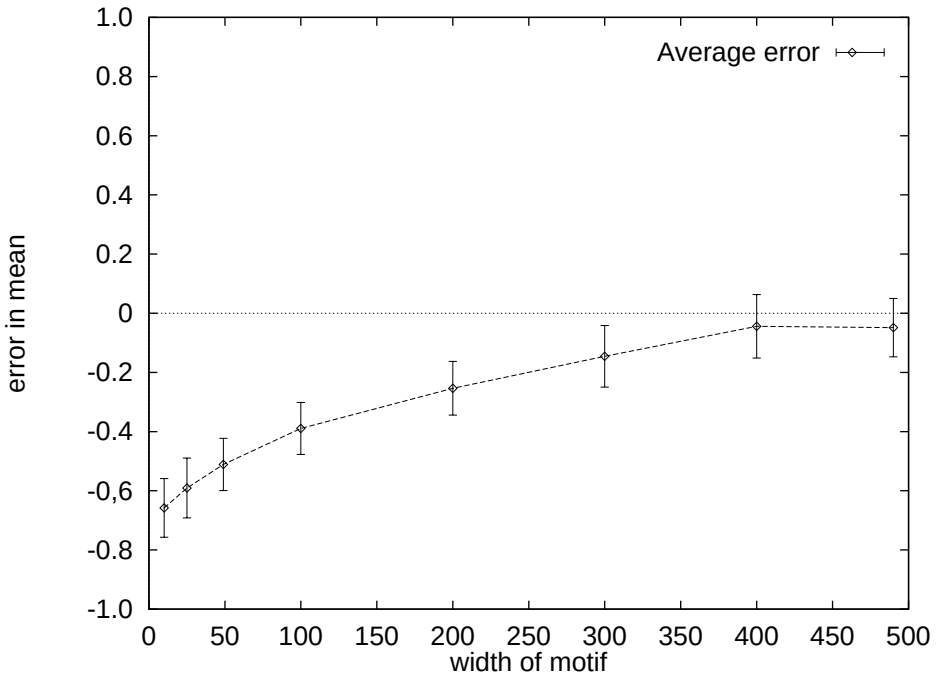

Figure 1: **Error in the mean sequence score predicted by the GEV approximation as a function of motif width.** The error is in units of the sample standard deviation of the sequence score for a given motif width. Each point represents the average discrepancy between the predicted and observed mean sequence score for 100 pseudorandom sequences each of length 2000 in comparison with motifs of different widths. The averages and 1 standard deviation error bars are sample statistics from 100 repetitions of the experiment with distinct sets of pseudorandom sequences. The source of the motifs is described in the text.

quite large for motifs narrower than 50 residues, which is the range where most biologically interesting motifs are found. The error is negative indicating that the GEV approximation consistently underestimates the mean sequence score. As predicted by theory, the error approaches zero as the motif width increases.[4] The data shown in Figure 1 are typical, and were generated using 100 pseudorandom sequences each of length 2000 with residue frequencies the same as in SWISS-PROT release 31. The motifs were created by truncating or duplicating and shuffling the columns of the motif used in Goldstein and Waterman [1994].

The idea of fitting the equation for the mean of a GEV (Eqn. 6) to the observed sequence scores illustrated in Figure 2. Each data point represents the observed mean score of twenty randomly generated sequences of the given length when compared to a motif model of width 13. The lower curve is $E[M(n)]$ using the initial estimates (sample mean and variance) for μ and σ^2 given in Eqns. (7) and (8). It badly underestimates the observed mean scores. The upper curve is the result of fitting the formula for $E[M(n)]$ in Eqn. (6) to the observed data by adjusting the values of μ and σ^2. This was done using the Levenberg-Marquardt non-linear least squares curve-fitting algorithm [Press *et al.*, 1986]. The goodness-of-fit for the fitted curve is 0.366, indicating that Eqn. (6) models the data well when we adjust the estimates of the mean and variance of the subsequence scores to agree with the observed distribution. As can be seen in the figure, the magnitude of the error in the predicted mean sequence score for a motif of typical width is substantial.

Because the expression for mean sequence scores as a function of sequence length fits the observed data well if we adjust the underlying values of mean and variance, we hypothesize that the true distribution of sequence scores is approximately the distribution of a GEV with the *adjusted* values of mean and variance. If this is true, then we would expect using this adjusted GEV distribution to give good estimates for the p-values of sequence scores.

The approximate p-value of a GEV can be calculated as follows. If $M(n)$ has an extreme value distribution and $T(n)$ is the corresponding reduced

[4]Eqn. (6) is almost perfectly accurate at predicting the mean of a true GEV. Sampling 10000 times from a GEV distribution with $n = 2000$ gave an observed error (normalized by the observed standard deviation of the GEV, as above) of less than 0.0001. This shows that the error in the estimate is due to the non-gaussian nature of subsequence scores, not the size of n.

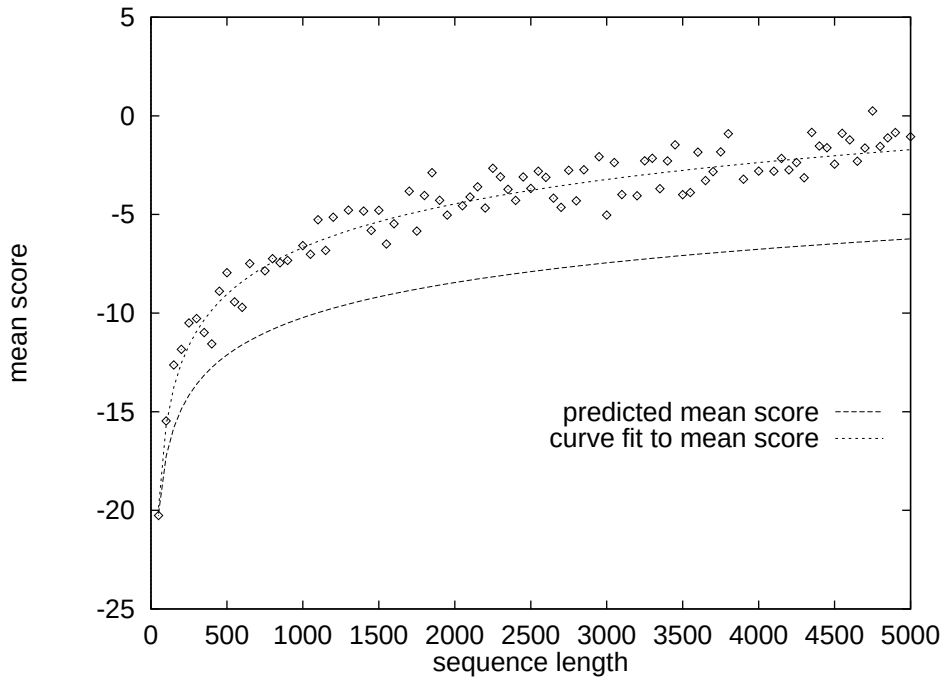

Figure 2: **Mean sequence score as a function of sequence length.** The data points are the mean score of twenty pseudorandom sequences of the given length using a typical motif of length 13. The lower curve is the theoretical mean of a GEV using the sample mean and sample variance of the subsequence scores as μ and σ^2, respectively. The upper curve is the result of a least-squares fit of the theoretical curve to the data, using μ and σ as the parameters to fit.

variate, then, when n is large,

$$Pr(M(n) \geq m(n)) \approx 1 - \exp(-e^{-t(n)}), \tag{9}$$

where $m(n)$ is the observed value of $M(n)$, and $t(n)$ is the observed value of $T(n)$ [Kinnison, 1985].

Eqn.(9) is known to converge only slowly with increasing n [Hall, 1980]. The speed of convergence is studied in Figure 3. Each curve in the figure was created by sampling 100,000 times from a GEV distribution with the stated value of n. For each value of n tried, the p-value of each GEV sample was calculated according to Eqn.(9), and the number of scores whose estimated p-value was less than or equal to $1 \cdot 10^{-6}, 2 \cdot 10^{-6}, 4 \cdot 10^{-6}, \ldots, 1.0$ was counted. Each curve is the average of ten random repetitions of this procedure for a particular value of n. We would expect all the curves to lie on the line $x = y$ if Eqn.(9) were exact. For instance, in 100,000 random GEV samples, we would expect to see about ten scores with p-values of less than or equal to 10^{-4}, but only about one is observed (observed frequency is approximately 10^{-5}) on average in this experiment when $n = 1000$. All the curves lie above $x = y$ which indicates that the approximation yields p-values that are too large. This is preferable to the opposite situation wherein that relatively unsurprising events would be assigned high significance. It is clear from the figure that GEV p-value estimates are fairly conservative even when n is quite large compared to the length of typical DNA and protein sequences.

We can use Eqn. (9) to compute the approximate probability (p-value) of observing a sequence score of at least $m(n)$ for a random sequence of length $n = L - W + 1$. We will refer to the p-value approximation based on the sample mean and standard deviation as the "unadjusted GEV p-value approximation", and that based on the values of μ and σ calculated by fitting the formula for $E[M(n)]$ to the observed mean sequence scores as the "adjusted GEV p-value approximation".

To test and compare the accuracy of the two p-value approximations, we conducted tests on pseudorandom sequences. We created a dataset containing $N = 100,000$ sequences of lengths varying uniformly from 10 to 1000 characters where each position was iid with the residue frequencies of SWISS-PROT release 31. The comparison score for each sequence in the dataset and a motif model was calculated, the $E[M(n)]$ curve fit to the scores to get adjusted values for μ and σ, the p-value of the score estimated using Eqn. (9),

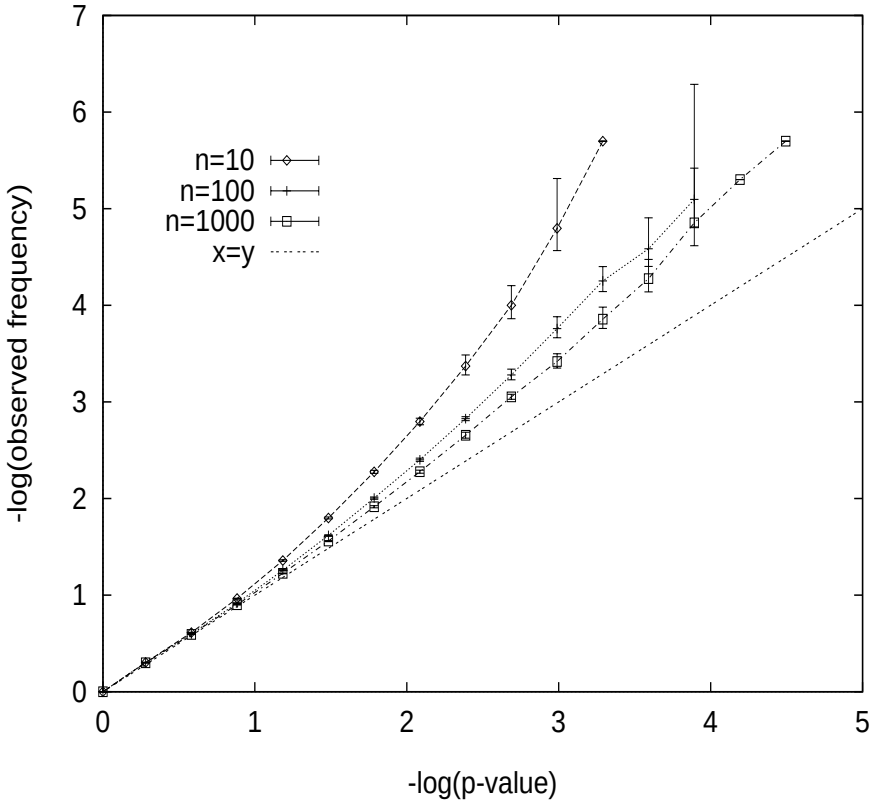

Figure 3: **Accuracy of the extreme value distribution for various values of n.** Each curve plots the observed frequency of randomly generated GEV samples whose p-value according to Eqn. (9) was less than or equal to the value on the x-axis. A true GEV with parameter n was sampled 100,000 and the number of p-values less than or equal to chosen values counted. This was repeated ten times for each value of n. Error bars are one standard deviation above and below the mean results of the ten trials. Error bars are not shown when the standard deviation exceeds the mean observed frequency. If Eqn. (9) were perfect, all curves would lie on the line $x = y$.

quantity	*mean*	(*sd*)
sequences per dataset	34	(36)
dataset size	12945	(11922)
sequence length	386	(306)
shortest sequence	256	(180)
longest sequence	841	(585)
pattern width	12.45	(5.42)

Table 1: **Overview of the 75 Prosite datasets.** Each dataset contains all protein sequences (taken from SWISS-PROT version 30) annotated in the Prosite database as true positives or false negatives for a single Prosite family. Dataset size and sequence length count the total number of amino acids in the protein sequences. The Prosite families used in the experiments are: PS00030, PS00037, PS00038, PS00043, PS00060, PS00061, PS00070, PS00075, PS00077, PS00079, PS00092, PS00095, PS00099, PS00118, PS00120, PS00133, PS00141, PS00144, PS00158, PS00180, PS00185, PS00188, PS00190, PS00194, PS00198, PS00209, PS00211, PS00215, PS00217, PS00225, PS00281, PS00283, PS00287, PS00301, PS00338, PS00339, PS00340, PS00343, PS00372, PS00399, PS00401, PS00402, PS00422, PS00435, PS00436, PS00490, PS00548, PS00589, PS00599, PS00606, PS00624, PS00626, PS00637, PS00639, PS00640, PS00643, PS00656, PS00659, PS00675, PS00676, PS00678, PS00687, PS00697, PS00700, PS00716, PS00741, PS00760, PS00761, PS00831, PS00850, PS00867, PS00869, PS00881, PS00904 and PS00933.

and the number of scores whose estimated p-value was less than or equal to $1 \cdot 10^{-6}, 2 \cdot 10^{-6}, 4 \cdot 10^{-6}, \ldots, 1.0$ counted. If the p-value estimate is good, we would expect the fraction of sequences having p-value x or less to be equal to x. We performed this experiment for 75 motif models generated by the motif discovery program MEME [Bailey and Elkan, 1995] on 75 distinct protein datasets. Each dataset consisted of all the SWISS-PROT sequences specified as belonging to a single Prosite family. The datasets are summarized in Table 1.

The results of these experiments are shown in Figure 4. Two of the curves show the results of using the unadjusted GEV and adjusted GEV p-value approximations. The third curve shows the result of sampling from true GEV distributions and calculating the p-values of the observations using the known mean and standard deviation of the underlying Gaussian distribution and Eqn. (9). For this curve, one true GEV sample was taken for each sequence in the pseudorandom dataset. The true GEV sample corresponding to a sequence of length n was generated by sampling n times from a standard normal distribution and taking the maximum of the observations. The line $x = y$ is also given in the figure for reference.

The unadjusted GEV p-value curve in Figure 4 lies entirely below the line $x = y$, indicating that the observed frequency of high scores is consistently higher than the p-value approximation would predict. This is undesirable because it means that the unadjusted GEV p-values tend to greatly overestimate the statistical significance of scores, making matches seem more surprising than they truly are. On the other hand, the adjusted GEV p-value curve lies entirely on or above the line $x = y$, indicating that the adjusted GEV p-value estimates are conservative. Finally, the curves for the adjusted GEV p-value approximation and true GEV sample p-values are almost identical. This confirms the hypothesis that the sequence scores have essentially a GEV distribution with different values of μ and σ.

Based on this large sample of protein family motifs, we conclude that the adjusted GEV p-values are a reliable and conservative way to estimate the statistical significance of sequence scores. It is clear from Figure 4 that the error in score p-values computed using the adjusted GEV estimate is entirely due to the fact that Eqn (9) is only the limiting distribution of a reduced variate, not the exact distribution. The adjustment procedure removes virtually all effect of the assumption that the (discrete) sequence scores have a Gaussian extreme value distribution. The adjusted estimate is

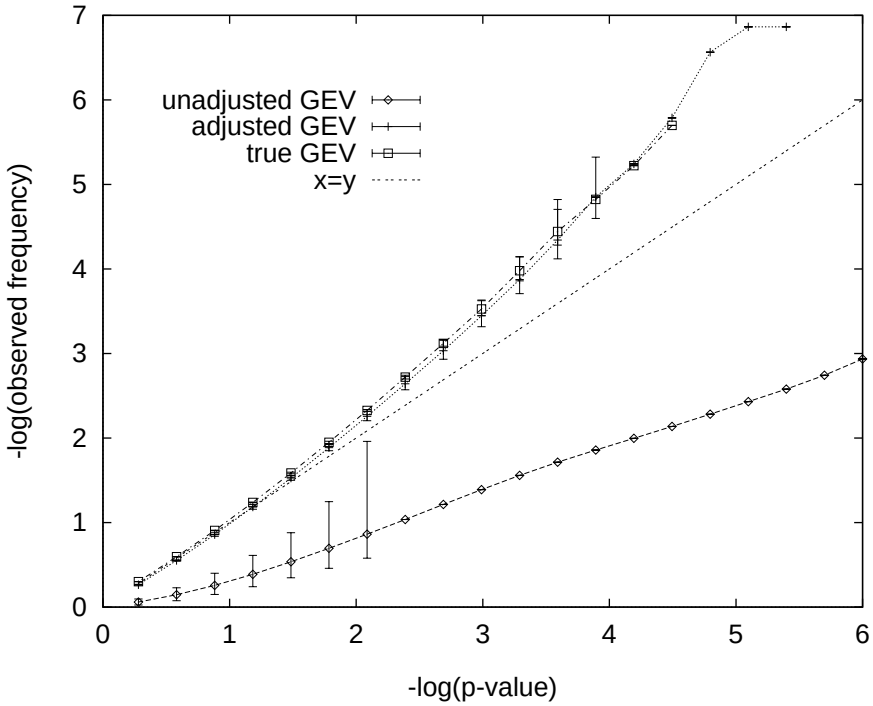

Figure 4: **Observed frequency of scores with p-values below a given value versus unadjusted GEV, adjusted GEV and true GEV p-values.** Points above the line $x = y$ underestimate the statistical significance of scores, below the line overestimate score significance. The unadjusted GEV and adjusted GEV curves show the result of 75 tests each of which used a single, different motif to score 100,000 pseudorandom protein sequences. The true GEV curve shows the results of ten trials, each of which sampled true GEV distributions once for each sequence in the dataset, using n equal to the length of the corresponding sequence in the dataset. Error bars are one standard deviation above and below the mean. Error bars are not shown when the standard deviation exceeds the mean observed frequency. The average (standard deviation) motif width was 13.57 (7.22). The smallest motif width was 5 and the largest was 49.

very accurate for p-values greater than 10^{-2} and becomes progressively more conservative for smaller p-values. In contrast, the unadjusted GEV p-values consistently overestimate the statistical significance of sequence scores.

3 Approximate distribution of multiple motif scores

This approach for computing p-values can be extended to multiple motifs. Suppose we have r motifs and we compute scores $S_i(n)$, maximum scores $M_i(n)$ and reduced variates $T_i(n)$ as in Eqns. (1), (2) and (3). A natural test statistic is the sum of the reduced variates,

$$C_r(n) = \sum_{i=1}^{r} T_i(n). \tag{10}$$

This statistic combines the evidence from each of the motif scores and will be large when the sum of the scores is large. It is analogous to the sum statistics often used for evaluating multiple high scoring segments in pairwise sequence comparisons [Altschul and Gish, 1996].

To use Eqn. (10) for computing p-values, we need to know the distribution of $C_r(n)$. We will show that it is approximately that of the sum of independent Gumbel random variables.[5] A Gumbel random variable has density function

$$f(x) = \exp(-x - e^{-x}). \tag{11}$$

The distribution of the sum of r independent Gumbel random variables, $C = \sum_{i=1}^{r} x_i$, where $y = \sum_{i=1}^{r-1} x_i$ and $z = \sum_{i=1}^{r-1} e^{-x_i}$, can be written as

$$\begin{aligned} Pr(C \geq x) &= \int \ldots \int_{y+x_r \geq x} \prod_{i=1}^{r} f(x_i) dx_1 \ldots dx_r \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{i=1}^{r-1} f(x_i) \left(\int_{x-y}^{\infty} f(x_r) dx_r \right) dx_{r-1} \ldots dx_1 \end{aligned}$$

[5]The Gumbel distribution has cumulative distribution function $F(x; \alpha, \beta) = \exp(-e^{(x-\alpha)/\beta})$, where $-\infty < \alpha < \infty$ and $\beta > 0$. In this paper, all references to the Gumbel distribution are for $\alpha = 0$ and $\beta = 1$. In that case, the cumulative distribution function is, simply, $F(x) = \exp(-e^{-x})$.

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{r-1} f(x_i)[1 - \exp(-e^{y-x})]dx_{r-1} \ldots dx_1$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-y-z}[1 - \exp(-e^{y-z})]dx_{r-1} \ldots dx_1. \quad (12)$$

By rearranging Eqn. (9), it is easy to verify that the limiting distribution for reduced variates such as $T_i(n)$ has the Gumbel density of Eqn. (11). If the underlying motif scores $S_i(n)$ are independent, the limiting distribution of $C_r(n)$ is that of the sum of independent Gumbel random variables. Empirical evidence presented below using actual motifs and pseudorandom sequences will show that this independence assumption is justified in practice.

Eqn. (12) is expensive to compute for large r due to the multiple integrations required, but Figure 5 shows that it behaves as

$$Pr(C \geq x) \approx \frac{e^{-x}x^{r-1}}{(r-1)!} \quad (13)$$

when x, the observed value of C, is greater than $r-1$. For $2 \leq r \leq 6$ and $x > r-1$, when Eqn. (13) gives a p-value of less than 0.23, it is always within 39 percent of the correct value. This means that small p-values–those of most interest–are accurately approximated using Eqn. (13). The accuracy of the estimate improves as x increases (and the p-value decreases), and is extremely good for small values of r. Extrapolating from Figure 5, the estimate will remain approximately within a factor of two of the correct p-value for large x even when r is as large as ten.

We saw that p-values of observations of a single GEV random variable, calculated using the limiting distribution, become increasingly conservative with increasing observed values (see Figure 3). Figure 6 shows that the behavior of Eqn. (13) for the sum of reduced variates of independent GEV random variables is essentially identical to that of Eqn. (9) for a single GEV random variable. In particular, for sufficiently large observed values (x), the approximation in Eqn. (13) predicts the true p-value of the observed sum of reduced variates as well as Eqn. (9) predicts the p-value of a single GEV random variable. For example, the p-value of the sum of five independent GEV random variables is as accurately predicted as the p-value of a single GEV when the p-value is less than 0.1 ($-\log(\text{p-value}) = 1$ in Figure 6). Each curve in the figure was produced by sampling from the given number of independent GEV distributions, computing reduced variates according to

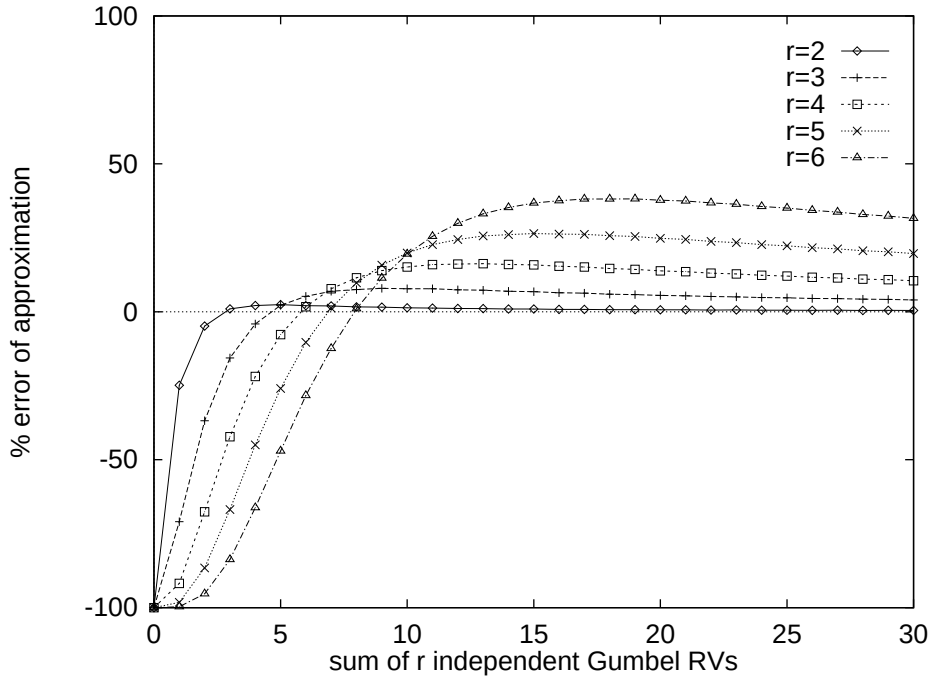

Figure 5: **Accuracy of the approximate distribution of the sum of independent Gumbel random variables.** Each curve shows the percent error in the approximate distribution of the sum of r independent Gumbel random variables given by Eqn. (13) for values of the the sum from zero to thirty. The correct value for the distribution was computed by numerically integrating Eqn. (12). Percent error is defined as $100 \times (Eqn.\ (13) - Eqn.\ (12))/Eqn.\ (12)$.

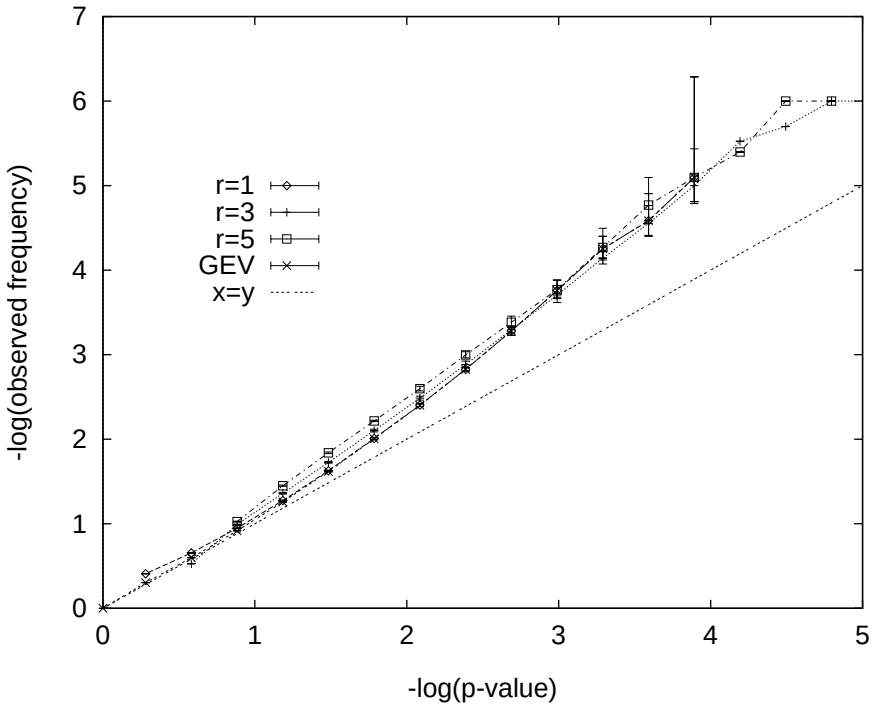

Figure 6: **Accuracy of the approximate distribution of the sum of independent Gumbel random variables for combining GEV random variables.** Each curve plots the observed frequency of the sum of the reduced variates of r independent GEV random variables whose p-value according to Eqn. (13) was less than or equal to the value on the x-axis. One, three or five GEV's with parameter $n = 100$ were sampled, their reduced variates were summed and the p-value of the sum was computed. This was repeated 100,000 times and the number of p-values less than or equal to chosen values counted. This procedure was repeated ten times and averaged for each value of r. For comparison, the curve labeled GEV shows the results for $r = 1$ when the p-value is computed using Eqn. (9). Error bars are one standard deviation above and below the mean results of the ten trials. Error bars are not shown when the standard deviation exceeds the mean observed frequency. If Eqn. (13) were perfect, all curves would lie on the line $x = y$.

Eqn. (3), taking their sum, computing the p-value corresponding to the sum (x) using Eqn. (13), and plotting the observed frequency of samples with p-values in different ranges as in Figure 3.

Replacing C with $C_r(n)$ in Eqn. (13) gives a computationally tractable way to estimate the p-value of the combined scores of matches of a single sequence to a group of motifs. We compute adjusted reduced variates $T_i(n)$ for each motif score independently, and use their sum as x in Eqn. (13). We will refer to this method of computing p-values as the "sum-of-reduced-variates" p-value approximation in what follows.

To test the accuracy of this method of estimating the p-values of multiple motif scores, we conducted tests analogous to the tests described earlier for single motif scores. We used the same database of 100,000 pseudorandom protein sequences of varying lengths and the first five motifs found by MEME in the 75 training sets. Each motif from a single run of MEME was used to independently score each sequence in the database. For each motif, its scores were used to calculate adjusted values for its mean and variance as in the single motif case. For each sequence, its scores for comparison to the five motifs and the five adjusted (μ, σ) pairs were used to calculate five reduced variates T_i, $i = 1, \ldots, 5$. The sums C_r for $r = 1$, 3 or 5 were then used with Eqn. (13) to calculate the p-values for the comparison scores of the sequence and the first, first three, and first five motifs reported by MEME for a given protein family.

The results of this test of using Eqn. (13) to estimate the p-values of multiple motif scores is given in Figure 7. The results are highly similar to those for single motif scores (compare with the adjusted GEV curve in Figure 4). All the curves lie above the line $x = y$, showing that the predicted p-values are conservative. The trends of the three curves indicate that the predicted p-values based on the GEV assumption become more conservative as they become smaller. As the number of motif scores being combined increases, the p-value estimates improve and become more accurate over a wider range of p-values. Predicted p-values near 10^{-5} are of particular interest since current protein sequence databases contain on the order of 10^5 sequences. At a predicted p-value of 10^{-5}, the observed frequencies of sequences in Figure 7 are approximately factors of 100, 10 and 5 smaller than predicted (more conservative) for one, three and five motifs, respectively.

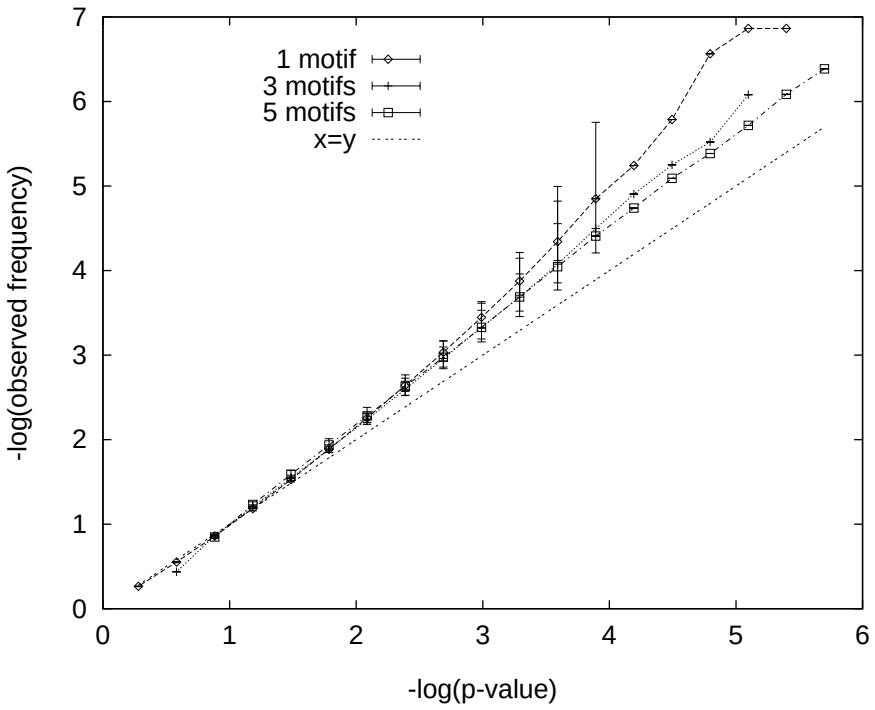

Figure 7: **Observed frequency of scores with p-values below a given value versus sum-of-reduced-variates p-values for multiple motif scores.** Points above the line $x = y$ underestimate the statistical significance of scores, below the line overestimate score significance. Curves show average results of 75 experiments using groups of one, three and five motifs. Each experiment used the motifs from a single of MEME on one of 75 datasets to score 100,000 pseudorandom protein sequences. Error bars are one standard deviation above and below the mean of the 75 experiments and are not shown if the standard deviation exceeds the observed frequency. The mean width of the motif models was approximately 15 with a standard deviation of about 8.

4 Validation of the method on real protein sequences

We conducted experiments to answer two questions. Firstly, does sorting by multiple motif p-value better separate homologs from non-homologs than sorting by "raw" score? Secondly, does using multiple motifs improve the sensitivity and selectivity of the homology search? By "raw" score in the first question we mean the sum of the sequence score for each of a group of motifs in comparison with a given sequence.[6] We therefore define the "raw" score of a sequence in comparison with a group of motifs to be

$$maxsum \;=\; \sum_{i=1}^{r} M_i(n).$$

Answering these questions requires sets of motifs for a number of protein families, a protein database to search, and a methodology for comparing the quality of an ordering of the sequences in the search database. We used the same 75 protein families as before and used the MEME program to discover five motifs for each family. As the search database, we used SWISS-PROT release 30 [Bairoch, 1994]. To measure the quality of a sort, we chose the ROC_{50} metric described in [Gribskov and Robinson, 1996]. *ROC* metrics have the virtue that they combine measurements of the sensitivity and selectivity of a search method into a single number. The ROC_{50} metric considers only the top of the sort down to the fiftieth non-family member. The metric has a value of 1 if all the true family members come before any non-family members in a sort of the sequences in the database. It has the value 0 if 50 non-family members appear before the first family member.

The results of these experiments shown in Figures 8 and 9 and in Table 2 demonstrate that using multiple motif p-values is superior to using *maxsum*, and using multiple motifs improves the quality of homology searches. The results do not support the conclusion that length normalization improves the quality of database searches using motif models. We explain how these conclusions are supported by the experimental results in the following paragraphs.

Figure 8 shows that, when a single motif is used to classify the sequence database, there is no difference between sorting by p-value and sorting by

[6] Since motif scores are log-odds scores it makes sense to add them.

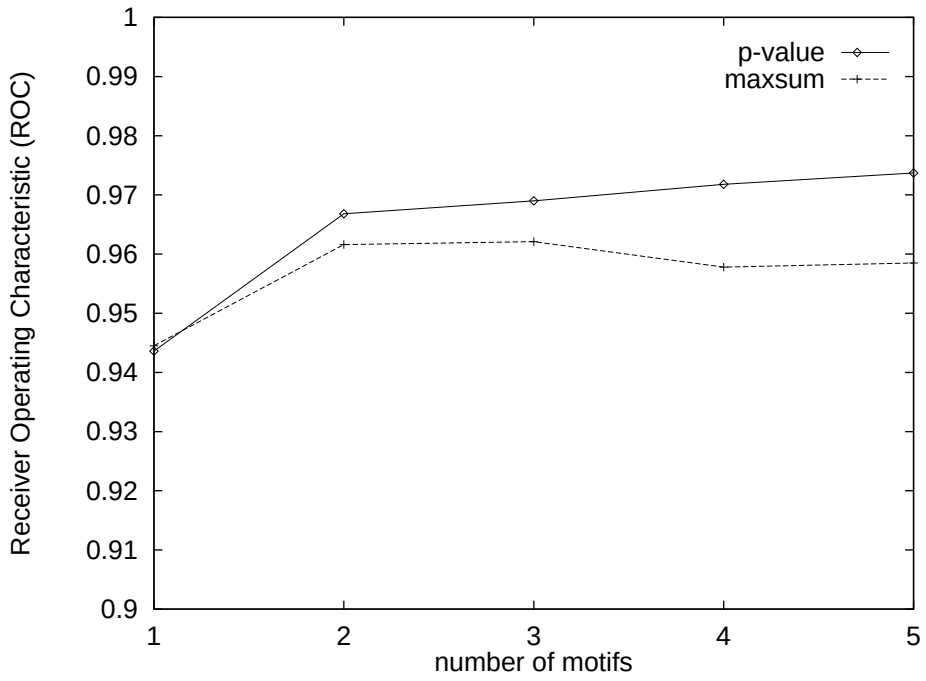

Figure 8: **Comparison of sorting sequences by p-value and raw score.** The plots show average ROC_{50} versus number of motifs when two different methods of scoring (p-value or *maxsum*) are used.

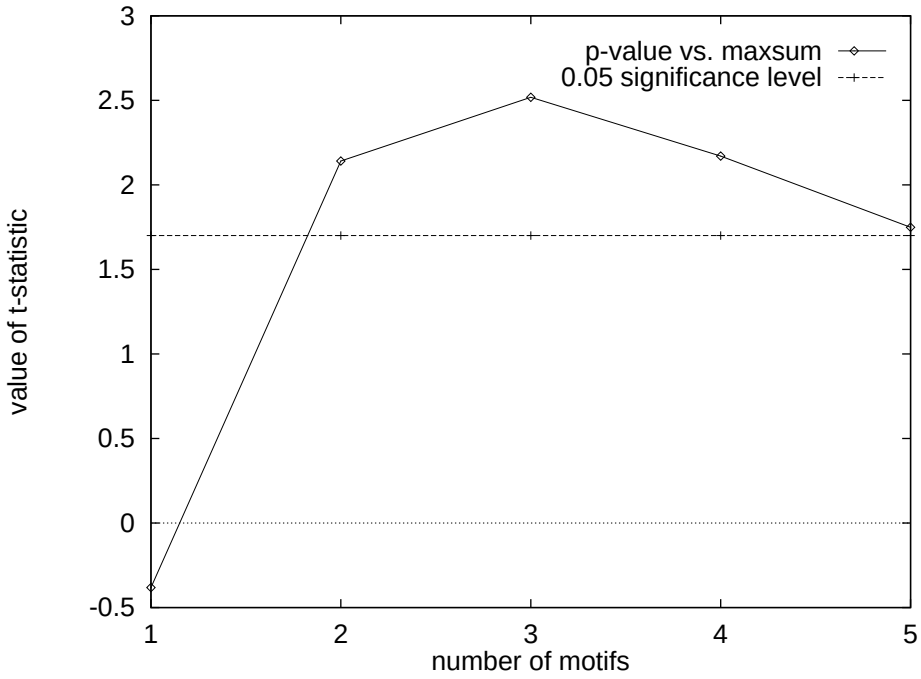

Figure 9: **Significance of improvement in sequence database sorting using p-values rather than raw score.** Each point shows the value of the paired t-statistic for 75 independent pairs of tests involving either p-values or *maxsum* and using the stated number of motifs per test.

	Number of Motifs				
	1	2	3	4	5
1		+	+	+	+
2			−	−	+
3				−	+
4					+

Table 2: **Statistical significance of the improvement in the ranking function using multiple motifs.** Each row compares searching using a given number of motifs (the row number) with using a larger number of motifs (the column number). A "+" indicates that using more motifs was statistically significant at the $P = 0.05$ level according to a paired t-test. A "−" indicates that the improvement was not statistically significant.

maxsum score. However, when multiple motifs characteristic of a protein family are used, the estimated p-value computed by the sum-of-reduced-variates method sorts the database better. This can be seen from the higher values of average ROC_{50} for the curve labeled "p-value" in the figure. The higher ROC indicates that the actual family members are appearing sooner in the sorted list of sequences when the sort is based on p-value than when it is based on *maxsum* score. This shows the clear advantage of using p-values rather than *maxsum* scores for homology searches using multiple motifs.

The difference in the quality of the sorting of the sequence database using the two methods (p-value vs. *maxsum*) is significant at the $P = 0.05$ level according to a paired t-test when more than one motif is being used. This is shown in Figure 9. The figure shows the measured value of the t-statistic comparing ROC_{50} using the two different sorting methods for each of the 75 datasets. The horizontal line shows the $P = 0.05$ significance level. When two or more motifs describing a family are used, p-value is significantly better than *maxsum* score at sorting the database of sequences.

Figure 8 also shows that the value of ROC_{50} when the sequence database is sorted by p-value improves as more motifs characteristic of a particular protein family are combined. We tested the statistical significance of the differences in ROC_{50} when various numbers of motifs were used for scoring the sequences. Using more than one motif is always significantly better ($P = 0.05$) than using a single motif (refer to Table 2). With these datasets, most of the information appears to be in the first two motifs generated by MEME because including the third or fourth motifs did not give significantly better results. Each motif however adds a little information, and using five motifs is significantly better than using only the first two.

This coincides with what one would expect given the nature of the MEME algorithm. MEME takes as input a set of sequences believed to be homologous and returns a collection of motifs each of which describes a pattern present in two or more of the sequences. The collection of motifs is ordered according to a statistical measure that combines the width, coverage (number of sequences containing the pattern) and strength (information content) of the pattern. In the protein families used here, the first motif returned by MEME is usually present in all or most of the sequences in the family. Later motifs may be present in only a small subset of the sequences. It is therefore to be expected that the first motifs are most descriptive of the family with later motifs adding relatively little information. This is confirmed by the decreasing improvement

in search performance using more than two motifs seen in Figure 8 and Table 2.

5 Discussion

The problem of assigning statistical significance to similarity scores is a recurring one in computational biology. It arises when searching sequence databases using a single sequence as a probe, as well as when comparing one or more sequences to a probe consisting of a set of motifs, blocks or patterns that characterize a sequence family. Raw similarity scores may be sufficient to order sequences with respect to similarity, but do not answer the important question of whether the apparent similarity is likely to be due to chance. One way to answer this question is to determine the distribution of similarity scores of random sequences scored against the probe.

For some types of similarity scores it has been possible to derive theoretical probability distributions for similarity scores of random sequences. Most notably, the distribution of the maximal segment pair (MSP) scores used by the BLAST algorithm [Altschul *et al.*, 1990] can be calculated directly from the scoring matrix and the assumed distribution of residues in the sequence database. The distribution of MSP scores is quite different than that of motif scores, so the theory is not directly applicable. As shown in Goldstein and Waterman [1994], motif scores can be expected (in the limit of motif width and sequence length) to have a Gaussian extreme value distribution. We have shown in this study that this theoretical distribution is not sufficiently accurate to be of use with motifs typical of protein sequence families, especially when the motifs are short (fewer than 50 columns). Furthermore, our results show how to combine multiple motif scores for a single sequence compared to a group of motifs that characterize a protein family.

When it has not been possible to derive a theoretical distribution for similarity scores, an empirical, curve-fitting approach is often used [Krogh *et al.*, 1994; Pearson, 1990; Gribskov *et al.*, 1990]. The general idea is to calculate the similarity scores of a large number of sequences compared with the probe and then fit a curve to the observed (sequence length, sequence score) pairs after removing outliers (which are presumed to be true positives and hence not "random".) This curve can then be used to estimate the parameters of the score distribution.

The approach we have taken in this study combines the theoretical and empirical methods described above. The method is based on fitting the formula for the mean of a Gaussian extreme value distribution to observed motif scores in order to estimate the parameters of the underlying distribution. These parameters are then used to calculate a "reduced variate" whose limiting distribution is the Gumbel distribution. Multiple motif scores are combined to give a single p-value by using the sum of the reduced variates for the motif scores as the test statistic. We have demonstrated an efficient way of estimating the distribution function of this statistic, as well as of computing the distribution of the sum of independent Gumbel random variables. Our method takes the length of the sequence being scored into account, and assigns a p-value to a sequence that is the probability of a random sequence of the same length scoring as well or better than the sequence in question.

The elimination of false positives in protein sequence database searches is a key concern. Our experiments on pseudorandom sequences show that the approximated p-values using our hybrid method tend to be conservative. This insures that the significance of database searches will not be overstated. These experiments also show that, as the number of motifs being used in a search increases, the accuracy of the p-value estimate improves while still remaining conservative. This shows that the p-values of (groups of) scores can reliably used to discriminate between true similarities and those that are likely to have occurred by chance.

We also have shown that the overall sensitivity and selectivity of a search using multiple motifs is improved when the sequences are sorted by p-value computed by our method rather than by the sum of the raw scores (*maxsum* score) for each motif. This effect increases as the number of motifs in the group whose scores are being combined increases. Converting raw scores to reduced variates before summing them has the effect of making all motifs "equal". When raw scores are summed, wider, more information-rich motifs contribute more to the total score. This means that true family members that do not contain a particular wide motif (possibly a motif characteristic of only part of the family) may have low *maxsum* scores but will still receive "good" (i.e., low) p-values.

Finally, our experiments show that using multiple motifs gives significantly better database search results than using single motifs. This is not surprising, since multiple motifs contain more information characteristic of the protein family than do single motifs. Additional information, which the

method we present here does not take into account, is contained in the ordering and spacing of the motifs. One would expect improved database search sensitivity and selectivity if the p-value took the probability of the observed motif spacing in a sequence (relative to the correct spacing for the family) into account. We intend to address this issue in future work.

6 Acknowledgements

This work was supported by the National Biomedical Computation Resource, an NIH/NCRR funded research resource (P41 RR-08605), and the NSF through cooperative agreement ASC-890285. Thanks to Dr. Charles Elkan at the Department of Computer Science, University of California (UCSD), and Dr. Michael Baker at the UCSD Medical school for helpful discussions during the course of this work.

References

[Altschul and Gish, 1996] Stephen F. Altschul and Warren Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

[Altschul *et al.*, 1990] Stephen F. Altschul, Warren Gish, Webb Miller, Eugen W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[Bailey and Elkan, 1994] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, Menlo Park, California, 1994. AAAI Press.

[Bailey and Elkan, 1995] Timothy L. Bailey and Charles Elkan. Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning*, 21:51–80, October 1995.

[Bairoch, 1994] Amos Bairoch. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Research*, 22:3578–3580, 1994.

[Bairoch, 1995] Amos Bairoch. The PROSITE database, its status in 1995. *Nucleic Acids Research*, 24:189–196, 1995.

[Goldstein and Waterman, 1994] Larry Goldstein and Michael S. Waterman. Approximations to profile score distributions. *Journal of Computational Biology*, 1:93–104, 1994.

[Gribskov and Robinson, 1996] Michael Gribskov and Nina L. Robinson. The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers and Chemistry*, 20:25–33, 1996.

[Gribskov *et al.*, 1990] Michael Gribskov, Roland Lüthy, and David Eisenberg. Profile analysis. *Methods in Enzymology*, 183:146–159, 1990.

[Hall, 1980] Peter Hall. Estimating probabilities for normal extremes. *Advances in Applied Probability*, 12:491–500, 1980.

[Henikoff and Henikoff, 1991] Steven Henikoff and Jorja G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, 19:6565–6572, 1991.

[Kinnison, 1985] Robert R. Kinnison. *Applied extreme value statistics*, page 53. Battelle Press, Richland, Washington, 1985.

[Krogh *et al.*, 1994] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[Lawrence *et al.*, 1993] Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[Pearson, 1990] William R. Pearson. Rapid and sensitive sequence comparision with FASTP and FASTA. *Methods in Enzymology*, 183:63–98, 1990.

[Pietrokovski *et al.*, 1996] Shmuel Pietrokovski, Steven Henikoff, and Jorja G. Henikoff. The BLOCKS database - a system for protein classification. *Nucleic Acids Research*, 24:197–200, 1996.

[Press *et al.*, 1986] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes*. Cambridge University Press, Cambridge, England, 1986.

[Smith and Waterman, 1981] Temple Smith and Michael Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[Stormo, 1990] Gary D. Stormo. Consensus patterns in DNA. *Methods in Enzymology*, 183:211–221, 1990.