From: Journal of Computational Biology, 5(211-221), 1998.

Methods and statistics for combining motif match scores

Timothy L. Bailey* and Michael Gribskov

May 7, 1998

Abstract

Position-specific scoring matrices are useful for representing and searching for protein sequence motifs. A sequence family can often be described by a group of one or more motifs, and an effective search must combine the scores for matching a sequence to each of the motifs in the group. We describe three methods for combining match scores and estimating the statistical significance of the combined scores and evaluate the search quality (classification accuracy) and the accuracy of the estimate of statistical significance of each. The three methods are: 1) sum of scores, 2) sum of reduced variates, 3) product of score p-values. We show that method 3) is superior to the other two methods in both regards, and that combining motif scores indeed gives better search accuracy. The MAST sequence homology search algorithm utilizing the product of p-values scoring method is available for interactive use and downloading at URL `http://www.sdsc.edu/MEME`.

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

---

San Diego Supercomputer Center, P.O. Box 85608, San Diego, California 92186-9784
Email: {tbailey, gribskov}@sdsc.edu
*To whom correspondence should be addressed.

1 Introduction

Protein sequence families can be characterized by one or more highly conserved, ungapped regions herein called motifs. Family members will contain some or all the motifs, usually with a highly conserved ordering and spacing. Motifs correspond to structurally and functionally important regions of the proteins. By ignoring less constrained regions of the sequences, they abstract the most salient sequence features of the family. The description of a set of motifs is a powerful tool for detecting distant family members (homologs) because there are fewer chances for spurious matches between non-family members and less conserved parts of the family member sequences. Motifs can also elucidate the important structural and functional features of the family and individual family members.

We describe a sequence motif using a position-specific scoring matrix (PSSM).[1] Each column in the matrix corresponds to a position in the sequence motif. The entries in a particular column are scores to be given to each letter in the sequence alphabet, $\mathcal{A}$, and are assumed to be *integer*. For a given PSSM, we calculate the *match score* of a sequence segment of length w, the width of the motif, by summing the appropriate entries in the matrix. In other words, the match score, f, of the segment beginning at position i in sequence s is

$$f(s,i) \;=\; \sum_{j=1}^{w} m_{s(i+j-1),j} \tag{1}$$

where $s(k)$ is the letter at position k in the sequence, and

$$M \;=\; \begin{vmatrix} m_{a,1} & m_{a,2} & \dots & m_{a,w} \\ m_{b,1} & m_{b,2} & \dots & m_{b,w} \\ \vdots & \vdots & \vdots & \vdots \\ m_{z,1} & m_{z,2} & \dots & m_{z,w} \end{vmatrix} \tag{2}$$

is the position-specific scoring matrix with $m_{a,k}$ being the score for letter $a \in \mathcal{A}$ at position k in the motif.

Computer algorithms exist for automatically constructing a characteristic set of sequence motifs from a family of biological sequences [Bailey and

[1] PSSMs can also be thought of as gapless sequence profiles [Gribskov *et al.*, 1990].

Gribskov, 1996; Neuwald *et al.*, 1995; Henikoff *et al.*, 1995]. Several computer algorithms for searching sequence databases using motifs also exist [Neuwald *et al.*, 1995; Tatusov *et al.*, 1994; Henikoff, 1992]. Unfortunately, none of the previously available motif-search programs estimate the statistical significance of simultaneous matches to multiple motif queries. The MAST (Motif Alignment and Search Tool) algorithm [Bailey and Gribskov, 1998] used in this paper

- accepts multiple motifs as the query,
- allows the motifs to occur in any order (or be missing) from the target,
- estimates the statistical significance of matches to the query.

This makes MAST a useful tool for the detection of distant family members and allows the statistical validity of matches to be assessed.

The main goal of searching a sequence database with a pattern is to sort the database according to the degree that the sequences match the pattern. Most often, a sufficiently high degree of 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 desirable, goal of a sequence database search is to estimate the significance of the observed degree of match of each sequence with the pattern. 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. In other words, we want to know the p-value of score x–the probability that the score of the match of a random sequence and the pattern would be greater than or equal to 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.

The p-values of sequence scores can be used to sort the sequences in the database to achieve both of the above goals. For the p-value of a score to be well defined, we must decide whether or not 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 potentially matching 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 used here.

2 Methods

We are interested in searching a database of target sequences for matches to a query consisting of one or more motifs. The match of a target sequence and the query is a function of the match scores of the sequence and each motif. For each motif in the query, we use the single best match score for any position in the target sequence (Eqn. 1). The best match score for each motif is computed independently so the best matching positions for a motif may overlap the best matching positions for other motifs.

The best match score for each motif is evidence for (or against, if the score is low) the membership of the target sequence in the sequence family described by the query. This evidence is combined to compute the p-value of the combined match (combined p-value) using one of three methods. The first method computes the p-value of the sum of the single best match scores for each motif and the target sequence (sum of scores). The second method separately estimates a "reduced variate" with a Gumbel limiting distribution for the best match scores for each motif, and computes the p-value of the sum of the reduced variates as described in Bailey and Gribskov [1997] (sum of reduced variates). The third method calculates a separate p-value for the best match scores of each motif and the target sequence, and then computes the p-value of the product of those p-values (product of p-values). This combined p-value can be used to sort the target sequences.

We show that the product of p-values approach is superior to the other two methods in terms of accuracy and sensitivity. We measure accuracy in terms of how well the p-values computed according to each method estimate the true statistical significance of combined match scores. Our metric for sensitivity is how well sequences sorted by the p-value given by each method are separated according to known membership in the sequence family characterized by the group of motifs.

With each method, we compute the p-values of the match of the query and target sequences. This involves defining a "score combining function", g, that

combines the individual best match scores for each motif and determining the probability that a *random* sequence of the *same length* as the target would have as "good" or better a value of g.[2]

We use the following notation. Let the target sequence, s, have length l. The query, Q, consists of n motifs defined by position-specific scoring matrices M_i, $1 \leq i \leq n$. Let the width of motif i be w_i and let the match score, $f_i(s)$, for sequence s and motif i, be defined as the best segment score for the sequence and the motif,

$$f_i(s) = \max_{1 \leq j \leq l_i} f_i(s, j), \tag{3}$$

where $l_i = l - w_i + 1$.

Each of the three methods of combining scores defines a distinct function $g(s, Q)$ for which we compute the cumulative distribution function over random sequences s of length l. For convenience of exposition, we use the convention that the cumulative distribution function for random variable X is $Pr(X < x)$, rather than the more common convention of $Pr(X \leq x)$. Random sequences are formed by concatenating l independent samples from the alphabet, where sampling is done using the average letter distribution, b_a for $a \in \mathcal{A}$, observed in naturally occurring sequences. We make the simplifying assumption that the segment scores, $f_i(s, j)$, $1 \leq j \leq l_i$, are independent. Assuming the iid sequence model just described, $f_i(s, j)$ and $f_i(s, k)$ are only independent if $|j - k| > w$, so this independence assumption is not strictly true. However, we will show, this assumption introduces little error. We also assume that the match scores $f_i(s)$, $1 \leq i \leq n$, are independent. This assumption can introduce substantial error if two or more motifs in the query are extremely similar to each other since two match scores, $f_i(s)$ and $f_j(s)$, may correspond to the same (or overlapping) positions in the sequence. We will show that this error can be greatly reduced by applying a simple test to detect and remove similar motifs from the query. In what follows, we also assume that the PSSM entries are integers and that the range of match scores is $r_i \leq f_i(s) \leq R_i$, $1 \leq i \leq n$. This requirement can always be satisfied (preserving any desired degree of accuracy) by scaling the entries in the original PSSM.

[2]For the sum of scores and sum of reduced variates methods, good means larger values, and for the product of p-values method, good means smaller values.

2.1 Sum of scores

The score combining function for this method is the sum of the match scores for each of the motifs,

$$g(s,Q) = \sum_{i=1}^{n} f_i(s). \tag{4}$$

The distribution function for $g(s,Q)$ over random sequences of length l can be estimated in three steps by taking advantage of the fact that the position-specific scoring matrices are integer-valued, so the distribution functions are discrete. This makes it possible to use iterative formulas to calculate the distribution functions.

Step 1) Estimate $C_i(x)$, the probability of observing a match score $f_i(s) \leq x$ with a random sequence of the same length as sequence s using the method of Staden [1990].

Step 2) Estimate $D_i(x)$, the probability distribution of $f_i(s)$ using

$$\begin{aligned} D_i(x) &= Pr(f_i(s) = x) \\ &= \begin{cases} C_i(x) - C_i(x+1) & \text{if } x < R_i \\ C_i(R_i) & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

Step 3) Estimate $P(x)$, the cumulative distribution function for $g(s,Q)$ using an induction formula similar to that used in step 1). Let $p^{(k-1)}(x)$ be the distribution function of the sum of match scores for the first $k-1$ motifs. Then, assuming all scores are independent, the probability distribution for the sum of match scores of the first k motifs is

$$\begin{aligned} p^{(k)}(x) &= Pr\left(\sum_{i=1}^{k} f_i(s) = x\right) \\ &= \sum_{i=r_k}^{R_k} p^{(k-1)}(x-i) D_k(i). \end{aligned} \tag{6}$$

To start the induction, we set $p^{(1)}(x) = D_1(x)$ because the distribution of $g(s,Q)$ when there is just one motif is just the distribution of the match scores for that motif. The cumulative distribution is trivially computed by summing the appropriate elements of $p^{(n)}(x)$.

2.2 Sum of reduced variates

We described this method of calculating the significance of simultaneous matches to multiple motifs in [Bailey and Gribskov, 1997] where we extended the work of Goldstein and Waterman [1994]. The random score distribution for a single motif is well approximated by a Gaussian extreme value (GEV) distribution. For each motif in the query, we estimate the parameters of this distribution using the sequences in the database being searched. We then transform the GEV random variables into reduced variates, each of which has a Gumbel limiting distribution [Kinnison, 1985]. The score combining function for this method is the sum of the reduced variates, $T_i(s)$, of each of the motifs,

$$g(s, Q) = \sum_{i=1}^{n} T_i(s), \tag{7}$$

where $T_i(s)$ is defined below (Eqn. 11). The p-value of the sum of reduced variates is then calculated using the formula given below for the cumulative distribution of the sum of independent Gumbel random variables.

This method assumes that the match score of motif i and a random sequence of length l has a limiting distribution which is the maximum of l_i independent samples of a Gaussian random variable with mean μ_i and standard deviation σ_i. We estimate μ_i and σ_i by fitting the curve of the mean of a GEV to the match scores of the query and the database. Curve fitting is done using the Levenberg-Marquardt non-linear least squares curve-fitting algorithm [Press *et al.*, 1986] on points $(l_i, f_i(s))$ after removing points where the match score, $f_i(s)$, is more than five sample standard deviations larger than the average match score for sequences of the same length as s. (Such sequences are presumed to be members of the family and, hence, not random with respect to the motifs in the query.) The mean of the a GEV can be approximated

$$E[f_i(s)] \approx u(l_i) + \gamma a(l_i), \tag{8}$$

where gamma is Euler's constant $0.5772156649\ldots$, and $u(l_i)$ and $a(l_i)$ are

defined as[3]

$$u(l_i) \;=\; \mu_i + \sigma_i(\sqrt{2\ln(l_i)} - \frac{\ln(\ln(l_i)) + \ln(4\pi)}{2\sqrt{2\ln(l_i)}}), \tag{9}$$

and

$$a(l_i) \;=\; \frac{\sigma_i}{\sqrt{2\ln(l_i)}}. \tag{10}$$

The reduced variate for the match score of motif i and sequence s, $T_i(s)$, is defined as

$$T_i(s) \;=\; \frac{f_i(s) - u(l_i)}{a(l_i)}. \tag{11}$$

For large values of $g(s, Q)$ in Eqn. 7, one minus the cumulative distribution is approximated by

$$Pr(g(s, Q) \geq x) \approx \frac{e^{-x}x^{n-1}}{(n-1)!}. \tag{12}$$

2.3 Product of p-values

The score combining function for this method is the product of the p-values of each of the match scores. The p-value of the match score for motif i and sequence s, $P_i(s)$, is defined as the probability of a random sequence of the same length as s having a match score as good or better than the observed match score. That is,

$$P_i(s) \;=\; Pr(f_i(s) \geq x), \tag{13}$$

where x is the observed value of the match score. The score combining function is, therefore,

$$g(s, Q) = \prod_{i=1}^{n} P_i(s). \tag{14}$$

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

To calculate the p-value of this product, we assume that the $P_i(s)$ are independent and uniform, and use the cumulative distribution function for the product of independent, uniform random variables.

To calculate each of the p-values in the Eqn. 14, we proceed exactly as in the first two steps of the first method. For motif i, this yields the probability distribution, $D_i(s)$, of the match score, $f_i(s)$. Summing $D_i(x)$ for all values up to x gives the cumulative distribution function for the motif i match scores, $P_i(s)$, as desired.

We say that the motifs in a query are independent if for all $1 \leq i, j \leq n$, $i \neq j$, match scores of random sequence s and motifs i and j are independent. Assuming motif independence, the p-values are independent and one minus the cumulative distribution function for $g(s, Q)$ is approximated by [Bailey and Gribskov, 1998]

$$Pr(g(s,Q) \geq p) \quad \approx \quad p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!} \tag{15}$$

for $0 < p \leq 1$, and is zero when p is zero.[4] The p-value of the combined score is trivially computed from the cumulative distribution function.

3 Results

To determine the best method of combining scores, we studied their classification and statistical accuracies. The classification accuracy of a score combining function $g(s, Q)$ is the degree to which it separates the true members of the family described by the query from all other sequences in the database being searched. We measure classification accuracy using the ROC_{50} metric [Gribskov and Robinson, 1996] because it combines measurements of the sensitivity and selectivity of a search method into a single number in a sensible manner [Swets, 1988]. The statistical accuracy of each method is the degree to which the calculated p-value estimates the true probability of a random sequence having a combined score as good or better than the observed combined score.

For each measurement, we used a set of 75 distinct sequence families from the Prosite database of protein sequence families [Bairoch, 1995]. (The

[4] Eqn. 15 is an approximation because the match score distributions are discrete rather than continuous, so the $P_i(s)$ are not truly uniform random variables.

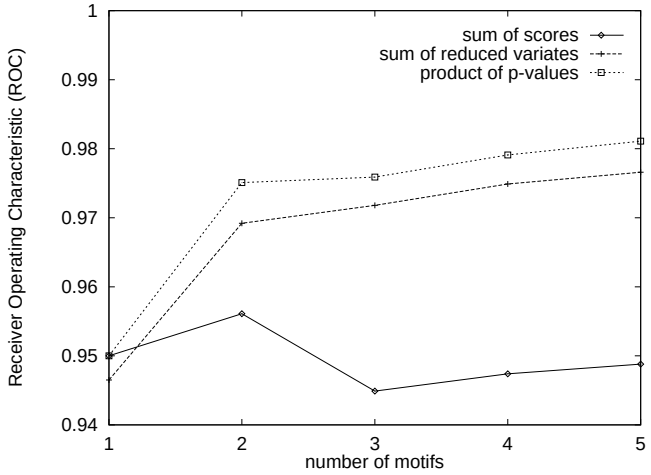

Figure 1: **Classification accuracy** The classification accuracy (ROC_{50}) versus the number of motifs in the query for each of the three methods of combining scores is shown. Each point represents the average result for 75 sequence families.

families chosen are listed in Bailey and Gribskov [1997]). The MEME motif discovery program [Bailey and Elkan, 1995] was used to generate a set of five motifs for each sequence family. This yielded a set of 75 distinct queries for which the correct classifications are known.

To measure classification accuracy of each method, we calculated the combined p-value of each sequence in SWISS-PROT release 28 [Bairoch, 1994]. Using the known family members for each query, we then computed the ROC_{50} classification accuracy of the method. We measured how well the score combining function utilizes the additional information in multiple-motif queries by using only the first, first two, first three, or first four motifs for a family as the query.

Figure 1 shows the classification for the three methods as a function of the number of motifs in the query. The product of p-values method is clearly superior to the other two methods, having better average classification accuracy for all multiple-motif queries. Surprisingly, the sum of scores method is extremely poor at utilizing the additional information in multiple motifs. On average, the classification accuracy of the sum of scores method is actually worse using all five motifs than with just the first motif in the query. The

motif number	*consensus sequence*
1	NYxVWxYR
2	PKNYQIWxHR
3	NxxAWxHR

Figure 2: **Similar motifs in one of the 75 queries.**

other two methods use the additional information to improve classification accuracy, with an especially large boost coming from including the second motif in the query. These two methods treat each motif as being more or less equally important in determining the classification because they normalize the raw scores before combining them. On the other hand, the sum of scores method allows wider motifs to dominate the combined score. This result suggest that classification is optimized when the contribution of each motif to the final score is normalized to be (essentially) independent of its width.

We measured the accuracy of the p-values computed by each method using the 75 sets of motifs and a database of pseudorandom sequences containing 100,000 sequences of lengths varying uniformly from ten to 1000 characters, where each position is iid with the residue frequencies of SWISS-PROT release 31. If p-values calculated by a method are exactly correct, we expect the fraction of sequences having a p-value less than or equal to x to be equal to x. For each query, we calculated the combined p-value according to one of the three methods for each of the sequences in the pseudorandom database and plotted the negative logarithm of the fraction of sequences whose p-value was less than or equal to various values against the negative logarithm of the p-value. In such a plot, points lying along the line $x = y$ indicate that the method is correctly estimating p-values. Points above (below) the line $x = y$ are caused by p-values being too large (small) on average.

Each method assumes that the match scores $f_i(s)$, $1 \leq i \leq n$ are independent. This requirement does not strictly hold for all 75 queries because some motifs in some queries are similar to each other. This is because some motifs represent sub-classes of a more general motif (Figure 2). The figure shows the consensus sequences for three motifs in one of the queries, aligned to emphasize the positions that are most similar. Clearly, sequences that

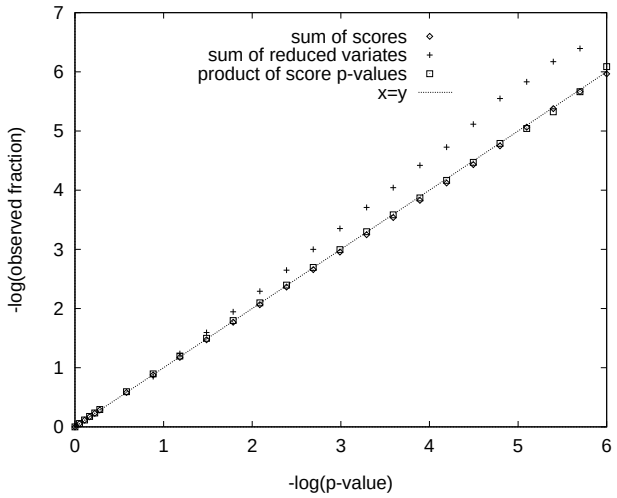

Figure 3: **Distribution accuracy with shuffled motifs.** The distribution of p-values predicted by each of the three methods is compared with the observed distributions. The graph shows the negative logarithm of the observed fraction of sequences with the given p-value or less versus the negative logarithm of estimated p-value. Each point is the average of 75 experiments where five motifs characteristic of a single protein family were used to search a pseudorandom sequence database of 100,000 protein sequences of varying lengths. The order of the columns in each motif was shuffled to remove possible dependencies among motifs in a query.

match the first motif will tend to have high scores with motifs two and three in this query.

Figure 3 shows the statistical accuracy of each of the methods with queries containing five motifs when the order of the columns of each motif is shuffled to reduce dependencies among the motifs in a query. Shuffling the columns of a motif does not affect its score distribution, but tends to make similar motifs dissimilar, reducing the dependence of match scores. Under these conditions, the combined p-values for both the sum of scores and product of p-values methods are accurate, lying very near the line $x = y$. (Compare this with the data for "all motifs" in Figure 4 where the motif columns are not shuffled– the combined p-values tend to overestimate the significance of the matches.) The combined p-value computed using the sum of reduced variates method is less accurate, showing an increasing tendency to underestimate the statistical

significance of rare events as the observed p-value decreases.

Actual database searches must use motifs whose columns have not been shuffled. To avoid correlated match scores, motifs which are highly similar to another motif in the query can be removed from the query. Removing motifs from queries will affect classification accuracy, but only very slightly as long as at least two motifs remain, as seen in Figure 1. We used the method of Pietrokovski [1996] to measure the correlation (similarity) between pairs of motifs in a query. This method aligns the two motifs (without gaps) as though they were sequences and computes the sum of the Pearson correlation coefficients for each pair of aligned columns. The measure is defined as the maximum of the sum over all possible alignments. It takes values from -1 to +1, with +1 indicating that one motif is identical to or contains the other motif.

Figure 4 shows the accuracy of the p-value estimates given by the product of p-values method when correlated motifs at different similarity levels are removed from the query. Each of the curves in the figure shows the average statistical accuracy of 75 distinct queries using, respectively, all motifs, motifs with no pairwise correlations above 0.75, and motifs with no pairwise correlations above 0.6. This required removing only two and ten motifs, respectively, from the 375 motifs in the 75 queries. Decreasing the maximum allowable pairwise correlation clearly improves the accuracy of the computed p-values, showing that the motif similarity metric works well in this application. When the queries are purged of all motifs sharing correlations above 0.6 statistical accuracy is excellent, except, perhaps, for events with p-values below 10^{-6} ($-\log(\text{p-value}) = 6$).[5] Removing only a small fraction of the motifs in the average query (ten motifs out of 375 total) is sufficient to insure that the p-values are reliable. It is clear from the figure that the statistical accuracy of the product of p-values method can be increased for extremely rare events (strong combined matches) by using a more stringent correlation cutoff than 0.6 at a very small cost in lost classification accuracy.

We also tested the statistical accuracy of the product of p-values method when the motifs are created by the PROTOMAT algorithm [Henikoff *et al.*, 1995]

[5]This deviation from the ideal is within the precision of the experiment. Eleven pseudorandom sequences with combined p-values of 10^{-6} or less were observed in 7.5 million trials (75 queries times 100,000 sequences). Assuming a Poisson distribution for such rare events, there is an approximately 14% chance of observing eleven or more p-values below 10^{-6}.

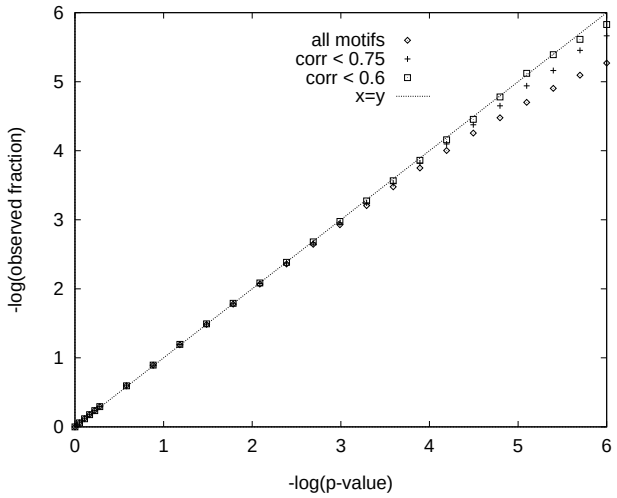

Figure 4: **Distribution accuracy with correlated motifs removed (product of p-values method).** The graph shows the negative logarithm of the observed fraction of sequences with the given p-value or less versus the negative logarithm of estimated p-value. The points labeled "all motifs" are each the average of 75 experiments where five motifs characteristic of a single protein family were used to search a pseudorandom sequence database of 100,000 protein sequences of varying lengths. The points labeled "corr < 0.75" are each the average of the 75 experiments with motifs removed so that all pairwise motif correlations are less than 0.75. The points labeled "corr < 0.6" are each the average of the 75 experiments with motifs removed so that all pairwise motif correlations are less than 0.6.

rather than MEME.[6] We used the 921 families of blocks in version 9.2 of the BLOCKS database [Pietrokovski *et al.*, 1996], which were created automatically from Prosite protein families using the PROTOMAT algorithm. The blocks were converted to position-specific scoring matrices using the blk2pssm program (Jorja Henikoff, personal correspondence). This resulted in queries for 921 protein families comprising between 1 and 20 motifs (average 3.65 motifs per family, and 760 queries with more than one motif).

[6]The output of the PROTOMAT web-server can now be sent directly to the MAST web-server in order to search protein databases using the motifs ("blocks") detected by PROTOMAT. The PROTOMAT algorithm can be used via a web-server at address `http://www.blocks.fhcrc.org/blockmkr/make_blocks.html`.

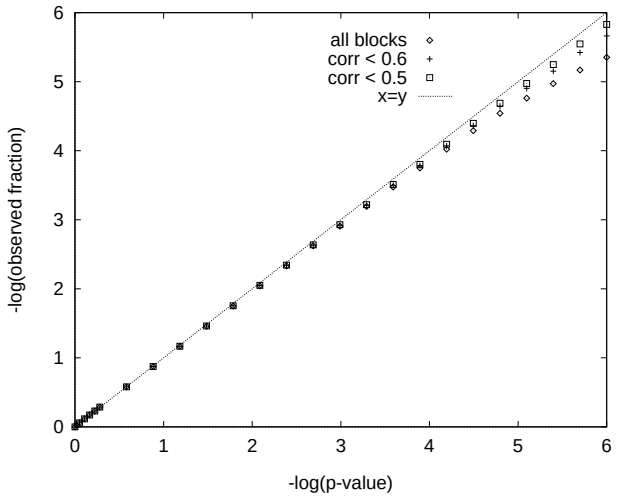

Figure 5: **Distribution accuracy on BLOCKS queries (product of p-values method).** The graph shows the negative logarithm of the observed fraction of sequences with the given p-value or less versus the negative logartihm of p-value. Queries, each comprising a BLOCKS version 9.2 database family converted from block to motif format, were searched against a pseudo-random sequence database of 100,000 protein sequences of varying lengths. The points labeled "all motifs" are each the average of 921 queries each consisting of all the motifs for a single BLOCKS 9.2 family. The points labeled "corr < 0.6" are each the average of 921 queries formed by removing motifs with correlations greater than 0.6 with other motifs in the same query. The points labeled "corr < 0.5" are each the average of 921 queries formed by removing motifs with correlations greater than 0.5.

The accuracy of the product of p-values method p-value estimate using queries consisting of all the motifs for a family, motifs with no pairwise correlations above 0.6 (23 highly-correlated motifs removed from 14 queries) and motifs with no pairwise correlations above 0.5 (48 motifs removed from 27 queries) is shown in Figure 5. As in the previous test using MEME-generated motifs (Figure 4), highly correlated motifs within a single query cause small p-values to be underestimated. This problem can be greatly reduced by removing any highly correlated motifs from the queries, as shown by the improved statistical accuracy of the p-values in the two curves for queries with correlated motifs removed. In practice, very few queries tend to need

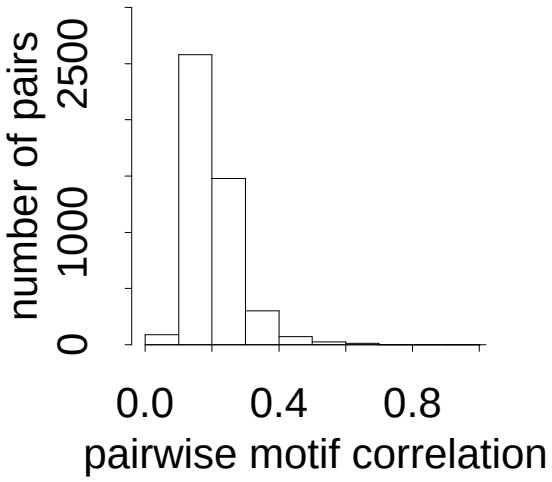

Figure 6: **Distribution of pairwise motif correlations in BLOCKS queries.** The 760 multi-motif queries in the BLOCKS datbase were examined for correlations between pairs of motifs in the same query.

such modification, as shown by the distribution of pairwise motif correlations in all 760 multi-motif queries in the BLOCKS database (Figure 6). Virtually all correlations between pairs of motifs in a single query are below 0.5, the threshold above which the accuracy of the p-value estimate begins to suffer.

Even with correlated motifs removed from the queries, the estimated p-values tend to be smaller than they should be (points below the line $x = y$ in Figure 5) due to small correlations among the scores for a single sequence. The error in the p-values is quite small, however, and should not be a problem in practice. For example, when searching a database containing 100,000 sequences with a p-value cutoff of 10^{-4}, if the p-value estimates were correct we would expect about ten random sequences ("false positives") to score below the cutoff. Figure 5 shows that if the query does not contain any motif-pairs with correlations above 0.6, about 1.3 times too many p-values below 10^{-4} will be observed on average. This corresponds to two or three additional false positives passing the cutoff. With a stricter cutoff, for instance, a p-

value of 10^{-6}, the curve for queries with no motif correlations above 0.6 indicates that about 2.2 times as many random sequences will have p-values below the threshold. This is of no consequence since the expected number of false positives if the p-value estimates were exactly correct would be $100,000$ times the p-value cutoff, or 10^{-2}. With the slight inaccuracy in the p-value estimate, the probability of a false positive is therefore still only about 0.022. We would, therefore, not expect any false positives in practice.

4 Discussion

The objective of this research is a scoring system for comparing protein sequences with a description of a sequence family. Such a scoring system forms the basis for a search algorithm which classifies sequences in a database with respect to membership in the sequence family described in a query. It is important that the statistics of the scoring system be known so that the statistical significance of search results can be evaluated.

Sequence motifs are a particularly powerful means of describing protein sequence families. However, many families are best described by groups of motifs. The motifs typically occur with a particular ordering and (approximate) spacing, but insertion, deletion and mutation events can cause motifs to be missing, duplicated or rearranged in some sequences of the family. For this reason, this research studied combining the single best match score for each motif in a query and a target sequence, without regard to the ordering of the motifs or possible overlap of motifs. This approach results in lower match scores if a motif in the query is missing from the target sequence, but does not preclude such a sequence from being classified as a family member. Similarly, multiple copies of a motif in the target sequence do not affect the combined match score, because only the highest match score for each motif is used in calculating the combined score.

We studied the accuracy of both classification and p-value estimates of three different ways of combining match scores. Each of these methods has intuitive appeal. The "sum of scores" method is reasonable because it is a direct extension of how individual motif scores are computed, and closely corresponds to scores for sequence comparisons with gaps. In this method, each column of each motif contributes a single, additive score to the combined match score. The "sum of reduced variates" method is analogous to the sum

statistics often used for evaluating multiple high scoring segments in pairwise sequence comparisons [Altschul and Gish, 1996]. Each motif contributes a single, scaled score that are summed to form the combined score. The "product of p-values" method treats each match score as the test statistic in a hypothesis test of whether the motif is present in the target sequence. The p-values of the scores are combined by multiplying them together, as suggested by Fisher's "omnibus" procedure for combining one-sided statistical tests [Fisher, 1970].

Our results show that the classification accuracy of protein motif queries is highest using the product of p-values as the combined score. Classification accuracy is only slightly lower with the sum of reduced variates method, but the accuracy of the p-value estimate is much worse. Surprisingly, both of these methods give significantly better classification than the sum of scores method. A possible explanation is that scaling the scores to reduced variates or p-values makes each motif, regardless of width, have equal weight in the combined score. The sum of scores method, on the other hand, gives wide motifs more weight than narrow motifs. Judging from our results, it is clearly better to weight all motifs equally when classifying sequences. This is interesting in view of the fact that both the Smith-Waterman sequence alignment algorithm [Smith and Waterman, 1981] and hidden Markov models [Eddy, 1995; Krogh *et al.*, 1994], an alternative method of describing sequence families, use a scoring method which is closely akin to our sum of scores method.

Inaccuracies are introduced in the p-value estimate of combined scores because we allow two motifs to match the target sequence in an overlapping fashion. This introduces correlations in the (supposedly independent) match scores for the individual motifs, and causes the p-values to overestimate the significance of the combined match. However, our results show that this effect is negligible in practice, as long as no pairs of highly correlated motifs are present in a query, and we have presented a practical method for insuring this. It seems unnecessary, therefore, to change the algorithm for computing the individual match scores to preclude overlapping motifs (which would remove the correlations). To do so would be problematic, since several assumptions of our method would no longer hold as a result. In particular, it would no longer be true that there are $l_i = l - w_i + 1$ positions for a motif of width w_i in a sequence of length l due to the non-overlap constraint. A better approach might be to flag sequences where two or more match scores come from overlapping positions to notify the user that the combined p-value may

be biased.

The product of p-values method for combining motif scores emerges as the clear choice among the three methods tested for motif-based database searches due to its high classification accuracy, reliable p-values and ease of computation. For this reason, the latest version of the MAST algorithm computes scores using that method. MAST is available for downloading and interactive use on the web at URL `http://www.sdsc.edu/MEME`. This software also computes the pairwise correlations between all motifs in the query, allowing the user to modify the query if highly similar motifs were included.

5 Acknowledgements

We would like to acknowledge the helpful comments of the anonymous reviewer. 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.

References

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

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

[Bailey and Gribskov, 1996] Timothy L. Bailey and Michael Gribskov. The megaprior heuristic for discovering protein sequence patterns. In David J. States, Pankaj Agarwal, Terry Gaasterland, Lawrence Hunter, and Randall Smith, editors, *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 15–24, Menlo Park, California, 1996. AAAI Press.

[Bailey and Gribskov, 1997] Timothy L. Bailey and Michael Gribskov. Score distributions for simultaneous matching to multiple motifs. *Journal of Computational Biology*, 4:45–59, 1997.

[Bailey and Gribskov, 1998] Timothy L. Bailey and Michael Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14:48–54, 1998.

[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.

[Eddy, 1995] Sean R. Eddy. Multiple alignment using hidden Markov models. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120, Menlo Park, California, 1995. AAAI Press.

[Fisher, 1970] R. A. Fisher. *Statistical methods for research workers, 14th Edition.* Oliver and Boyd, Edinburgh, 1970.

[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.

[Henikoff *et al.*, 1995] Steven Henikoff, Jorja G. Henikoff, William J. Alford, and Shmuel Pietrokovski. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, 163:GC17–GC26, 1995.

[Henikoff, 1992] Steven Henikoff. Detection of *caenorhabditis* transposon homologs in diverse organisms. *New Biologist*, 4:382–388, 1992.

[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.

[Neuwald *et al.*, 1995] Andrew F. Neuwald, Jun S. Liu, and Charles E. Lawrence. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Science*, 4:1618–1632, 1995.

[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.

[Pietrokovski, 1996] Shmuel Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, 24:3836–3845, 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.

[Staden, 1990] Rodger Staden. Searching for patterns in protein and nucleic acid sequences. *Methods in Enzymology*, 183:193–210, 1990.

[Swets, 1988] John A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 270:1285–1293, 1988.

[Tatusov *et al.*, 1994] Roman L. Tatusov, Stephen. F. Altschul, and Eugene V. Koonin. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Sciences, USA*, 91:12091–12095, 1994.