Technical Report IMB-TR0001: Details of MCAST Statistics

Timothy L. Bailey

June 3, 2003

Abstract

1 Statistics of match scores with negligible gap costs

We define a "hit" as a statistically significant (gapless) alignment of a motif and a sequence. A motif alignment score, b, is considered statistically significant if the probability of a randomly generated sequence of length w (the width of the motif) having alignment score b or higher is less than the user-defined p-value threshold p. Letting p_b be the p-value of alignment score b, we define the p-score, x, of a hit to be

$$x \quad = \quad -\log_2 \frac{p_b}{p}. \tag{1}$$

By this definition, the p-score of a hit will always be positive, and larger values of x correspond to more significant alignment scores. Note also that adding hit p-scores corresponds to multiplying the score p-values.

We define a match between a set of motifs and a sequence as one or more hits separated by gaps no longer than the user-specified maximum gap length, L. We compute match scores as the sum of hit scores (p-scores), minus the gap costs. Our gap costs are affine, consisting of a fixed cost for opening a gap (the in- and out-transition costs), g_o, and a fixed cost for extending a gap (the self-loop cost), g_e.

Alignment scores are computed in the standard manner for scoring sequences with PSSMs, and p-values are found by table lookup [Bailey and Gribskov, 1998]. We create a lookup table for each motif PSSM assuming that each alignment score is for an independent, length-w sequence generated by by a 0-order Markov chain. (This ignores the fact that hits come from overlapping windows on a single sequence, and, hence are not truly independent. It also ignores the fact that real biological sequences are often poorly modeled by 0-order Markov chains.) We first scale and round the motif PSSM to convert it to integer values. The lookup table is then filled in by summing the probabilities of all possible ways of achieving each integer score. This is feasible since we assume the contribution from each column of the motif is independent.

We are intrested in the statistical properties hit scores (p-scores), and of match scores found by the repeated match algorithm [Durbin *et al.*, 1998; p. 24-25] when the within-match gap costs are negligible. We set the repeat threshold in the repeated match to a tiny number ($1e-6$), and no gap can have a cost larger than the repeat threshold. This makes the gap costs neglible relative to the hit scores, and, hence, the overall match score. So match scores are essentially just the sum of hit scores. The properties of match scores are thus dominated by the statistics of hits and hit scores.

Next, we define q as the probability of a hit occurring in a (random) sequence of length L, the maximum gap length. Under the additional assumption that log-odds scores of the m different motifs (and their reversecomplements) are independent, we can estimate q by

$$q \quad \approx \quad 1-(1-p)^{mL}. \tag{2}$$

(This will obviously be an overestimate if the motifs are highly similar, since then there will be fewer independent chances for some motif to have a hit.)

We can use q, the probability of a hit in a sequence of length L, to estimate the expected number of hits per match. If q is small, the probability of two hits occurring within the maximum gap distance will also be small. Random matches will then tend to be very short (contain few hits). Conversely, if q is close to zero, hits will tend to occur close together, and matches will be long (contain many hits).

By definition, all matches contain at least one hit. The number of *additional* hits per match is well approximated by a geometric distribution with parameter q. This follows from assuming that matches are independent and considering the probability of a match containing a single hit being extended (in one direction) to contain more hits. The number of additional hits is analogous to the number of successes before the first failure in independent, Bernouli trials with probability of success q. Let h be the number of hits in a match. The expected number of hits per match is approximately

$$E[h] \quad \approx \quad 1 + \frac{q}{1-q}. \qquad (3)$$

Next, we estimate the expected gap length, $E[d]$, where d is the length of a gap. This depends on the conditional probability distribution of the distance between hits given that there is a hit within distance L from the previous one. Let $\hat{p}$ be the probability of a hit occurring at any randomly selected position in the sequence. We can approximate $\hat{p}$ by

$$\hat{p} \approx 1 - (1-p)^m.$$

Let $p(i)$ be the probability of the first hit occurring at distance i from the end of the previous hit. This has the geometric distribution

$$p(i) \approx (1-\hat{p})^i \hat{p}.$$

Using the definition of conditional probability

$$Pr(A|B) = Pr(A,B)/Pr(B),$$

we can write the probability of a gap having length i as

$$\begin{aligned} p_d(i) &= Pr(\text{gap length} = i | \text{hit within distance } L) \\ &= \frac{Pr(\text{gap length} = i, \text{hit within distance } L)}{Pr(\text{hit within distance } L)} \\ &\approx \frac{p(i)}{q}. \end{aligned}$$

(The approximation is due to the fact that $p(i)$ and q are both approximations.) We can then compute the expected value of the gap length as

$$E[d] \quad \approx \quad \sum_{i=0}^{L} i p_d(i).$$

We define the span of a match, s, as the distance from the start of the first hit to the end of the last. This is equal to the sum of the widths of the hits plus the lengths of the gaps between them. We would like to know the expected span of a match, $E[s]$. Let $\bar{w}$ be the average width of the motifs in the query. Assuming that each hit comes with equal probability from one of the m motifs–which is approximately true due to the definition of hits in terms of p-values–the average hit will have width $\bar{w}$. On average, matches contain $E[h]$ hits, so the contribution of the hits to the expected span will be $E[h]\bar{w}$. The expected number of hits in a match is $E[h]$, so the expected number of gaps is $E[h]-1$. The expected length of a gap is $E[d]$. We can thus estimate $E[s]$ as

$$\begin{aligned} E[s] &= E[\sum \text{width of hits} + \sum \text{width of gaps}] \\ &\approx E[h]\bar{w} + (E[h]-1)E[d]. \qquad (4) \end{aligned}$$

To verify the accuracy of our estimates of q, $E[d]$, $E[h]$ and $E[s]$, we conduct a series of simulation experiments. We use a fixed query, a different *random* database, and a different combination of p and L in each experiment. The query contains five motifs with average width $\bar{w} = 9.8$. Each database contains 10 randomly-generated DNA sequences, each of length 100,000bp. The random sequences are generated using a 0-order Markov model with ACGT frequencies (0.297, 0.203, 0.203, 0.297). We try all combinations of values of p in the set {0.001, 0.0001, 0.00001} and L in the set {50, 100, 150, 200, 250, 300, 400, 500}. Table 1 shows the (approximate) expected values and observed means of d (length of a gap), h (number of

p	L	q	$E[d]$	$\bar{d}$	$E[h]$	$\bar{h}$	$E[s]$	$\bar{s}$	μ
1e-5	50	0.005	25	32	1.01	1.01	10	11	1.13
1e-5	100	0.010	50	44	1.01	1.01	10	11	1.14
1e-5	150	0.015	75	58	1.02	1.02	11	11	1.16
1e-5	200	0.020	100	107	1.02	1.02	12	13	1.14
1e-5	250	0.025	125	135	1.03	1.03	13	15	1.16
1e-5	300	0.030	150	129	1.03	1.02	15	14	1.07
1e-5	400	0.039	199	214	1.04	1.04	18	18	1.15
1e-5	500	0.049	248	208	1.05	1.05	23	22	1.14
1e-4	50	0.049	25	25	1.05	1.05	12	12	1.41
1e-4	100	0.095	50	48	1.11	1.12	16	17	1.43
1e-4	150	0.139	74	74	1.16	1.17	23	24	1.38
1e-4	200	0.181	97	95	1.22	1.24	33	35	1.42
1e-4	250	0.221	120	120	1.28	1.31	47	50	1.41
1e-4	300	0.259	143	142	1.35	1.38	63	67	1.42
1e-4	400	0.330	187	188	1.49	1.54	107	117	1.39
1e-4	500	0.393	230	224	1.65	1.72	165	179	1.37
1e-3	50	0.394	23	23	1.65	1.69	31	32	1.56
1e-3	100	0.632	42	41	2.72	2.84	99	104	1.55
1e-3	150	0.777	57	56	4.49	4.81	242	259	1.57
1e-3	200	0.865	68	67	7.40	8.06	511	555	1.56
1e-3	250	0.918	77	76	12.20	13.88	986	1109	1.57
1e-3	300	0.950	84	82	20.12	22.37	1801	1966	1.56
1e-3	400	0.982	92	88	54.71	63.49	5481	6190	1.55
1e-3	500	0.993	96	92	148.78	154.95	15660	15727	1.56

Table 1: **Accuracy of estimates of number of hits per match and span.** Each line in the table reports the results of one experiment where a five-motif query searches a (distinct) random database using the give values of p and L. The probability of a hit within distance L from the previous one, q is shown. The expected values and observed means of the length of a gap, d, the number of hits per match, h, and the span of a match, s are reported. The last column gives the observed mean score per hit (match score divided by the number of hits in the match).

hits per match) and s (span of a match) in each experiment. The table also shows the value of q and the average score per hit, μ, for each experiment.

The experiments summarized in Table 1 explore the accuracy of our estimated expected values of h and s for a wide range of values of q: $0.005 \leq q \leq 0.993$. This covers extremely local searches, where the mean number of hits per match is close to 1.0, as well as highly non-local searches, where the mean match contains almost 200 hits. The table shows that our estimates of the expected values are quite good. The major divergence occurs with extremely large values of q, where the repeated-match algorithm will be finding global rather than local matches. The expected number of hits is accurate for $q < 0.95$; the expected span is accurate for $q < 0.78$. For values of q smaller than 0.78, the observed means are generally within 10% of the expected values. The average score per hit (μ) is essentially determined by p, as seen from the last column of Table 1.

We now can now consider how (random) match scores will be distributed. As we have seen, match scores are essentially the sum of a random number of p-scores. The number of p-scores in the sum follows (approximately) a geometric distribution with probability of success q. We expect p-scores to have the (approximate) density

$$f(x) \quad = \quad \mu \exp(-x/\mu) \tag{5}$$

and cumulative density function

$$F(x) \quad = \quad 1 - \exp(-x/\mu). \tag{6}$$

This can be derived as follows. Firstly, p-values of alignment scores are approximately uniform random variables on the interval $[0, 1]$ [Bailey and Gribskov, 1998]. (The approximation is due to the fact that alignment scores are discrete and finite.) Since we discard all scores with p-scores less than 0, this implies that p_b, the p-value of alignment score b, is distributed (approximately) uniformly on the interval $(0, p]$

$$p_b \quad \sim \quad U[0, p],$$

where p is the p-score threshold, as before. The derivation of the (approximate) distribution of p-scores follows algebraically. We can write, for all $x \in (0, p]$,

$$\begin{aligned} Pr(p_b < x) &= \frac{x}{p}, \\ Pr(\frac{p_b}{p} < \frac{x}{p}) &= \frac{x}{p}. \end{aligned}$$

Noting that $p_b/p \in (0, 1]$, we can write, for all $x \in (0, 1]$,

$$\begin{aligned} Pr(\frac{p_b}{p} < x) &= x, \\ Pr(-\log_2 \frac{p_b}{p} > -\log_2 x) &= x. \end{aligned}$$

We note now that the expression to the left of the the less-than sign in the last equality above is a p-score. This implies that, for $x > 0$,

$$\begin{aligned} Pr(-\log_2 \frac{p_b}{p} > x) &= 2^{-x}, \\ Pr(-\log_2 \frac{p_b}{p} > x) &= e^{-x \log 2}, \\ Pr(-\log_2 \frac{p_b}{p} \leq x) &= 1 - e^{-x \log 2}, \\ F(x) &= 1 - e^{-x \log 2}. \end{aligned}$$

Thus, p-scores are (approximately) distributed as an exponential random variable with mean $\mu = 1/\log 2 \approx 1.44$.

We note from Table 1 that the average score per hit is very close to 1.44 in the experiments with $p = 0.0001$. When $p = 0.00001$, the average score per hit drops below $1/\log 2$ because, with the short motifs used in the experiment, very small p-values cannot be achieved by *any* sequence. This causes the assumption that positive p-scores are uniformly distributed to become less accurate. The longer the motifs are, the smaller we can set p before this inaccuracy becomes noticable. On the other hand, when $p = 0.001$, the average hit score is larger than $1/\log 2$ in the experiments in Table 1. This can be understood by considering that the repeated-match algorithm assembles matches from the maximally-scoring

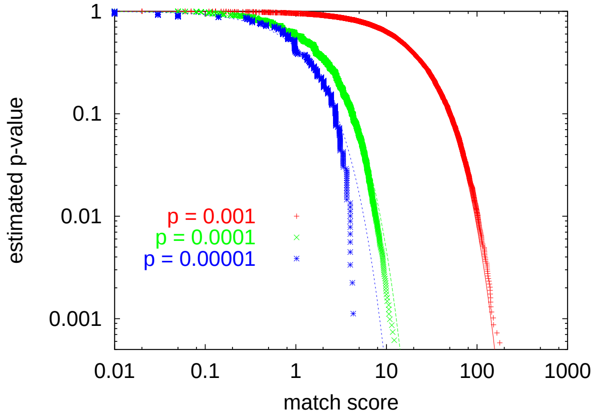

Figure 1: **Match scores follow exponential distributions.** The three sets of points in the figure correspond to three experiments from Table 1 with $L = 250$ and the given values of p. Each set of points plots score vs. estimated p-value for all the matches in one experiment. The lines in the figure are cumulative density functions fit to the match scores.

non-overlapping hits. Thus, each hit is the result of an optimization over (potentially) competing, overlapping hits. When p is high relative to the number of motifs in the query, the number of competing, potential hits increases the value of μ noticeably.

We have shown that p-scores are distributed (approximately) exponentially, as long as p is not too small. The question remains how match scores are distributed. When the expected number of hits per match is close to 1, match scores are essentially p-scores, so we expect them to follow a distribution with mean equal to the average score per hit. The points labeled "p = 0.00001" in Figure 1 show that this is empirically true for the $p = 0.00001, L = 250$ experiment from Table 1. The estimated p-values of the match scores from this experiment, where the expected number of hits per match is 1.03, agree closely with an exponential distribution with mean approximately equal to the average score per hit.[1] The points lie quite close to (1 minus) the exponential cumulative density function

$$F(x) = 1 - \exp\frac{-x}{\mu},$$

with mean $\mu = 1.22$, and the average score per hit is 1.20 in this experiment (see Table 1). The values for the mean of the exponential distributions shown in Figure 1 are computed by the maximum likelihood method. For the exponential distribution, the maximum likelihood estimate for the parameter μ is simply the mean match score.

Somewhat surprisingly, match scores still appear to follow an exponential distribution even when matches contain many hits on average. The other two sets of points in Figure 1 labeled "p = 0.0001" and "p = 0.001", respectively, correspond to experiments with expected hits per match of 1.28 and 12.2, respectively. These match scores seem to follow exponential distributions as well. Naturally, the means of the exponential distributions are higher than the average score per hit, since the average match score in these experiments consists of the sum of several independent p-scores.

Perhaps even more surprisingly, match scores appear to follow an exponential distribution even when sequences in the database violate our assumption of being generated by a 0-order Markov chain. We searched shuffled versions of *Drosophila* non-coding regions with a query of five *Drosophila* transcription factors [Berman *et al.*, 2002] using MCAST. That the match scores from this search approximately follow an exponential distribution can be seen from the straight line formed by the match score vs. estimated p-value points labeled "shuffled sequences" in Figure 2. These points closely follow the exponential cumulative distribution function estimated from a similar search of synthetic sequences generated according to a 0-order Markov model derived from the real sequences ($\mu = 2.46$). That the match scores

[1] To estimate p-values in Figure 1, we use "rank p-values". The rank p-value of match score x, $R(x)$, is defined to be

$$R(x) = \frac{r(x) + 1}{n + 1},$$

where $r(x)$ is the rank of x when the match scores are sorted in decreasing order.

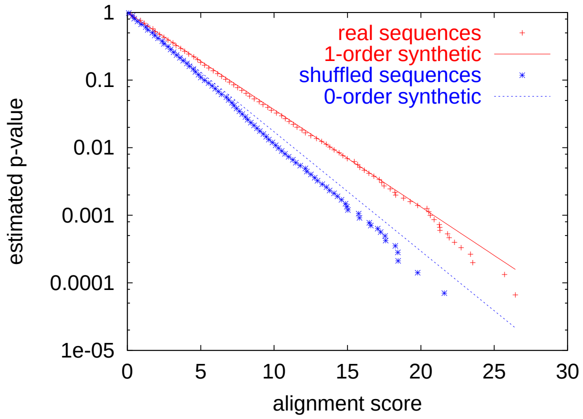

Figure 2: **Statistics of match scores of synthetic and real sequences.** Match scores vs. estimated p-values from MCAST searches of all noncoding regions of *Drosophila* ("real sequences"), and shuffled versions of these same regions ("shuffled sequences") are plotted. The lines are one minus the cumulative density functions fit to match scores from searches of synthetic sequences generated randomly according to 0-order and 1-order Markov models of noncoding *Drosophila* sequences. All four searches used the same query containing five motifs.

from this search approximately follow an exponential distribution can be seen from the (approximately) straight line formed by the match score vs. estimated p-value points labeled "real sequences" in Figure 2. When we search the *unshuffled Drosophila* sequences (points labeled "real sequences" in Figure 2), the average match score increases ($\mu = 3.06$). However, the scores still appear to closely follow an exponential cumulative distribution function. As a matter of fact, their distribution matches that estimated from a similar search of synthetic sequences generated according to a 1-order Markov model derived from the real sequences. The violation of the assumption that the real sequences are generated by a 0-order Markov chain appears not to affect the exponential form of the match score distribution.

2 Non-negligible gap costs

So far we have always set the gap costs to be extremely small, essentially zero. This allowed us to ignore the gap costs in reasoning about the score function. However, it may be desirable to penalize longer gaps, as well as limiting the maximum gap size. The hope is that this will lead to more accurate searches, and, perhaps, reduce the sensitivity of the search to L.

We introduce a new parameter, α, that will be the *ratio* of the expected cost of a gap to the expected score of a hit:

$$\alpha = \frac{E[\text{gap cost}]}{\mu} \tag{7}$$

$$= \frac{E[d]g_o}{\mu}, \tag{8}$$

where $g_o = g_e$ is the gap opening and extension cost, μ is the expected score of a hit, and $E[d]$ is the expected distance between hits–the expected gap length.[2] Rearranging, we find the required gap costs for a given value of α,

$$g_o = g_e = \alpha \frac{\mu}{E[d]}. \tag{9}$$

[2] We use the expected distance between hits in a match *assuming zero gap costs* in defining α to prevent circularity.

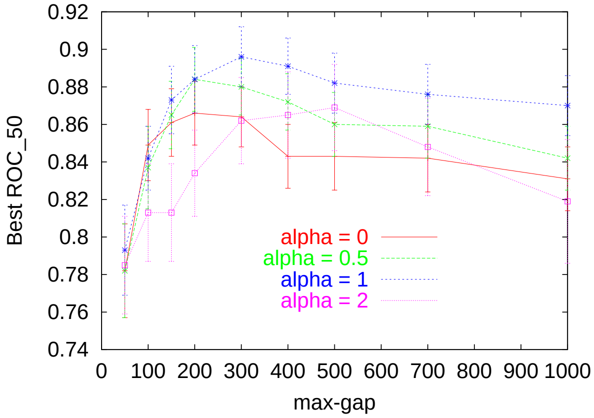

Figure 3: **Achievable search accuracy for fixed L and α.** For each query, different values of p were tried while holding L and α constant. Shows the average results for 100 distinct 5-motif queries. Each point is the mean of the best search accuracy (ROC_{50}) achieved using the given values of L and α. Error bars show the standard errors of the means.

(For convenience, when $\alpha = 0$, we revert to our previous method of setting the gap costs, ignoring the values of μ and $E[d]$.)

Figure 2 shows that search accuracy, on average, is best when $\alpha = 1$. This figure was generated using synthetic sequences and sets of real, TRANSFAC motifs as the queries. The default setting for α in MCAST is 1. Larger values of α can be used when the motifs in the query are highly dependent to counteract the tendency of $g = g_o = g_e$ to be set to small by Eqn.9.

3 Estimating Match Score E-values using Expectation Maximization

In order to calculate the E-values of match scores, we assume that random scores come from an exponential distribution and estimate the parameter of that distribution from the observed match scores. Of course, we hope that some of the scores (the high scores) do not come from the random distribution, so we actually assume that the match scores come from a mixture of two distributions. We use the expectation maximization (EM) algorithm to fit the parameters of this mixture distribution to the observed match scores. We then compute E-values of match scores using the cumulative density function for random scores whose parameters we have thus estimated.

We model the match scores as a mixture of two distributions, $f_1(x|\mu_1)$ and $f_2(x|\mu_2, \sigma_2)$. The first distribution, f_1, is intended to model the distribution of random scores and is assumed to be a member of the exponential family. It has density

$$f_1(x|\mu_1) = \frac{e^{-x/\mu_1}}{\mu_1},$$

and cumulative density

$$F_1(x|\mu_1) = 1 - e^{-x/\mu_1}.$$

The second distribution, f_2, will model the distribution of true match scores, and is assumed to be Gaussian:

$$f_2(x|\mu_2, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp(\frac{-(x-\mu_2)^2}{2\sigma_2^2}).$$

For simplicity, let $\theta_1 = \{\mu_1\}$ and $\theta_2 = \{\mu_2, \sigma_2\}$. If the fraction of random match scores in n total matches is $0 \leq c \leq 1$, then the mixture of the random and true match scores has probability density function

$$f(x|\theta_1, \theta_2, c) = cf_1(x|\theta_1) + (1-c)f_2(x|\theta_2).$$

We assume a prior distribution on c that is uniform between zero and one.

We use the expectation maximization algorithm to estimate optimal values for the parameters of the mixture. Plugging the estimated value of the mean of the first component into the following equation gives the estimate for the E-value of score x:

$$E \approx \hat{n}e^{-x/\mu}, \tag{10}$$

where $\hat{n} = cn$ is the estimated number of match scores from the non-random distribution.

```
function EM($X$: list of scores)
  Find the values of $\mu_1, \mu_2, \sigma_2$ and $c$ that maximize the
  two-component likelihood function given the match scores.
  Compute initial estimates $\hat{\mu}_1, \hat{\mu}_2$, $\hat{\sigma}_2$ and $\hat{c}$.
  do
    E-Step: Compute the value of $z_i$ for each score $x_i$.
    M-Step: Maximize the two-component log likelihood function.
  until The log likelihood increases by less than $10^{-4}$.
end
```

Figure 4: Expectation Maximization Algorithm

The expectation maximization (EM) procedure [Dempster *et al.*, 1977] allows us to determine the values of θ_1, θ_2, and c that (locally) maximize the (log) likelihood function,

$$L(X|\theta_1, \theta_2, c) \quad = \quad \sum_{i=1}^{n} \log f(x_i|\theta_1, \theta_2, c),$$

where $X = \{x_i\}, i = 1, \ldots, n$, is the set of match scores found by the search. Initial estimates of these parameters must be supplied to the EM procedure. It then refines the estimates by alternately repeating two steps, the E-step and the M-step, until convergence. There is no guarantee of finding the global optimum, but EM seems to work very well in this application, as will be shown in the Results section.

The E-step (expectation step) computes the expected value of one auxiliary variable, z_i, for each score, given current estimates at the parameters of the model. The interpretation of the unary "missing information" variable z_i is that it is one if the match score x_i is from the random distribution, zero otherwise. The equation for the expected value of z_i is

$$\begin{aligned} \bar{z}_i &= E_{\hat{\theta}_1, \hat{\theta}_2, \hat{c}}[z_i] \\ &= \frac{\hat{c} f_1(x_i|\hat{\theta}_1)}{\hat{c} f_1(x_i|\hat{\theta}_1) + (1-\hat{c}) f_2(x_i|\hat{\theta}_2)}, \end{aligned}$$

where the "hatted" variables are the current estimates for the parameters.

The M-step (maximization step) maximizes the expected value of the augmented log likelihood function, $L(X, Z|\theta_1, \theta_2, c)$, where $Z = \{z_i\}, i = 1, \ldots, n$, is the set of missing information variables. This expectation is

$$\begin{aligned} E_{\hat{\theta}_1, \hat{\theta}_2, \hat{c}}[L(x, z|\hat{\theta}_1, \hat{\theta}_2, \hat{c})] \quad = \quad & \sum_{i=1}^{n} [\bar{z}_i \log f_1(x_i|\theta_1) + \\ & (1-\bar{z}_i) \log f_2(x_i|\theta_2) + \\ & \bar{z}_i \log(c) + \\ & (1-\bar{z}_i) \log(1-c)]. \quad (11) \end{aligned}$$

The values of the four parameters that maximize this expected likelihood can be computed from the data X and the missing data Z via the following equations:

$$\begin{aligned} \hat{\mu}_1 &= \frac{\sum_{i=1}^{n} x_i \bar{z}_i}{\sum_{i=1}^{n} \bar{z}_i}, \\ \hat{\mu}_2 &= \frac{\sum_{i=1}^{n} x_i (1-\bar{z}_i)}{\sum_{i=1}^{n} (1-\bar{z}_i)}, \\ \hat{\sigma}_2 &= \sqrt{\frac{\sum_{i=1}^{n} (1-\bar{z}_i)(x_i - \hat{\mu}_2)^2}{\sum_{i=1}^{n} (1-\bar{z}_i)}}, \text{ and} \\ \hat{c} &= \frac{1}{n} \sum_{i=1}^{n} \bar{z}_i. \end{aligned}$$

These new estimates are then substituted for the old ones in the E-step and the algorithm iterates until the expected likelihood converges.

Our complete EM method (Figure 4) for estimating the distribution of random match scores begins by first computing initial estimates for the four parameters. First, we compute the sample mean, $\bar{\mu}$, and the sample standard deviation, $\bar{\sigma}$ of the match scores. Then, we estimate the expected maximum (random) score, x_{em}. No matter what the distribution is, the expected largest score will have p-value $\frac{1}{n+1}$. To estimate x_{em}, we assume the scores follow an exponential distribution with mean $\bar{\mu}$. Dividing Eqn. 10 by n, plugging in the expected smallest p-value and solving for x we get

$$x_{em} \quad = \quad -\bar{\mu} \log \frac{1}{n+1}.$$

Lastly, we estimate the number of matches from non-random sequences ("outliers"), n_o, by the number of scores larger than the expected maximum score. Letting x_m be the largest score in the set X, we set the initial estimates of the parameters of the mixture distribution to

$$\begin{aligned} \hat{\mu}_1 &= \bar{\mu}, \\ \hat{\mu}_2 &= x_{em}, \\ \hat{\sigma}_2 &= \frac{|x_m - x_{em}|}{2}, \text{ and} \\ \hat{c} &= \frac{n - n_o - 1}{n}. \end{aligned}$$

The initial estimate for $\hat{\mu}_1$ assumes that all scores are random, so the sample mean is the maximum likelihood estimate for μ_1. We set, $\hat{\mu}_2$, the mean of the non-random, high scores, to the expected largest score. Our initial estimate for the standard deviation of the non-random score distribution is one-half the distance between the mean score and the expected largest score. Finally, our estimate for the mixing component, $\hat{c}$, is the fraction of scores that are less than the expected maximum random score (minus $1/n$ to insure that $\hat{c} \neq 1$.)

Starting from the intitial estimates, our algorithm applies the E- and M-steps alternately to refine the estimates of the parameters. This stops when the the log of the likelihood function increases less than 10^{-4}. The final estimate of the mean of the first component, $\hat{\mu}_1$, is then used to estimate the E-values of match scores via Eqn.10.

3.1 Measuring the accuracy of E-values

We want to validate that our EM algorithm provides accurate E-value estimates from empirical match scores. We constructed five datasets of synthetic sequences in order to test this reliably. Each dataset is a mixture of 0-order Markov-generated sequences, and similar sequences with clusters of randomly-generated motif occurrences inserted into them.[3] We search each dataset using queries containing subsets of the five motifs used in the positive sequences. We also searched with a query containing the five motifs plus an additional 45, randomly selected TRANSFAC human motifs. We measure the accuracy of the E-values using "p-value slope error" (PSE) [Bailey and Gribskov, 2002]. We computed the p-values of the matches from negative sequences only, and measured how well they correspond to their rank p-values using the PSE metric. (The p-value of a score is gotten from its E-value by dividing Eqn.10 by $\hat{n}$.) For values of p at least 0.0005, the p-value accuracy is comparable (around 0.03) with that for Smith-Waterman alignment scores [Bailey and Gribskov, 2002] using the best empirical estimation methods (Tab. 2). Fig.5 illustrates how PSE is computed. Fig.5**a** shows that, for a typical value of PSE (around 0.03), the p-values of matches to negative sequences are extremely close to their expected values (their rank p-values). For very small values of p (e.g, 0.00001), the p-value accuracy suffers because the assumption that p-scores are uniformly distributed breaks down. (This happens because any motif has a *maximum* score that depends on its information content. This determines a *minimum* p-score, and smaller p-scores have zero probability.) This can be seen in Fig.5**b**. For a given value of p, the accuracy of the estimated p-values generally increases with the number of motifs in the query. With these randomly chosen TRANSFAC motifs, the minimum practical value for p is about 0.00005 for queries containing five motifs, but 0.00001 is usable for large, fifty-motif queries. Different ranges will be

[3] Each dataset contains 100 *positive* sequences containing random occurrences of selected motifs, and 1000 *negative*, completely random sequences. The positive sequences were generated by creating a META-MEME hidden Markov model from five, randomly selected TRANSFAC human motifs. Each dataset contains occurrences of a different set of five motifs. Sequences were generated that contain between 9 and 20 motif occurrences in a region about 1000 bp long. Each positive sequence was padded with 9000 bp to a total length of 10000 bp. Each negative sequence is likewise 10000 bp long.

query	p									
	0.00001		0.00005		0.0001		0.0005		0.001	
size	*mean*	*se*	*mean*	*se*	*mean*	*se*	*mean*	*se*	*mean*	*se*
1	0.371	0.177	0.195	0.103	0.063	0.025	0.035	0.012	0.033	0.013
2	0.207	0.074	0.117	0.020	0.085	0.028	0.038	0.012	0.027	0.008
4	0.142	0.032	0.066	0.016	0.061	0.012	0.031	0.007	0.027	0.005
5	0.168	0.032	0.070	0.009	0.049	0.006	0.029	0.002	0.024	0.003
50	0.040	0.006	0.021	0.003	0.017	0.003	0.023	0.001	0.010	0.002

Table 2: **Accuracy of estimated E-values.** The measured accuracy of E-value estimates for searches using various values of p and various numbers of motifs in the query is shown. $L = 50$ in each search. Mean and standard error of PSE for the negative sequences in five replicate experiments is shown.

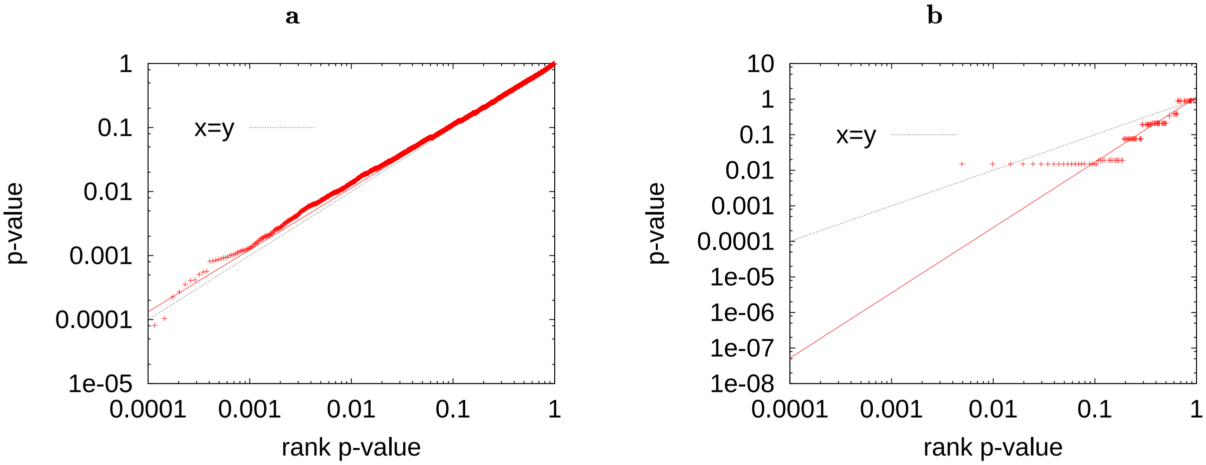

Figure 5: **Worst case and typical examples of p-value accuracy in synthetic data experiments.** Each panel plots the p-value (estimated using EM) vs. the rank p-value of each match to a negative sequence in a single experiment. Ideally, the points should lie along the line $x = y$ with slope one. The red line is computed by linear regression on the points, and its slope, m, determines p-value slope error: $PSE = |1-m|$. Panel **a** shows a typical experiment with PSE around 0.03 (0.032). Panel **b** shows the experiment with the worst PSE (0.84).

appropriate in cases where the motifs in the query are wider or shorter, or have higher or lower information content. The higher the information content of the query motifs, the lower p may be.)

References

[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.

[Bailey and Gribskov, 2002] Timothy L. Bailey and Michael Gribskov. Estimating and evalutating the statistics of gapped local-alignment scores. *J. Comp. Biol.*, 9:573–593, 2002.

[Berman *et al.*, 2002] B. P. Berman, Y. Nibu, B. D. Pfeifer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *PNAS*, 99:757–762, 2002.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[Durbin *et al.*, 1998] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis.* Cambridge UP, 1998.