A conservative parametric approach to motif significance analysis

Uri Keich, Patrick Ng
Department of Computer Science, Cornell University, Ithaca, NY, USA

We suggest a novel, parametric, approach to estimating the significance of the output of motif finders. Specifically, we rely on the good fit we observe between the 3-parameters Gamma family and the null distribution of motif scores. This fit was observed across multiple motif finders, background models and scoring functions. Under this parametric assumption we compute and show the utility of a conservative confidence interval for the p-value of the observed score. Since our method relies on the 3-parameters Gamma fit it should be applicable to a variety of finders.

1. Introduction

The identification of transcription factor binding sites, and of *cis*-regulatory elements in general, is an important step in understanding the regulation of gene expression. To address this need, many motif-finding tools have been described that can find short sequence motifs given only an input set of sequences. The motifs returned by these tools are evaluated and ranked according to some measure of statistical over-representation, the most popular of which is based on the information content or entropy [19] (see [3] for a recent comparative review).

Unfortunately, the area of motif significance analysis has lagged considerably behind the extensive development of tools for motif finding. Consider for example the popular profile or PWM (position weight matrix) based finders such as MEME [2], CONSENSUS [5] and the various approaches to Gibbs sampling (e.g. [8], [12], [6]). Many of these tools do not offer any significance evaluation at all [*] while others, notably MEME and CONSENSUS, rely on the E-value. Introduced originally in this context as the "expected frequency" [5] it is the expected number of random alignments of the same dimension that would exhibit an entropy score, or information content [19], that is at least as high as the score of the given alignment.

While a step in the right direction the E-value has significant shortcomings. First the E-values are currently only computed for the entropy score. The latter tacitly assumes the somewhat unrealistic, albeit popular, iid (independent and identically distributed) model for the background. The problem is that if the background sequences are generated using a more realistic model one quickly encounters

[*] Note that this problem differs from that of scanning sites with a known PWM, e.g., [20].

examples where ranking by the entropy score, or equivalently by the E-value, consistently yields sub-optimal motifs. Second, even when computed correctly [10] it can at times be conservative to the point where it is of no value [13]. This paper offers an alternative, parametric, approach to evaluate the significance analysis. Unlike the E-value, which only works for the entropy score assuming an iid model, our method works reasonably well for all combination of scores and background models we looked at.

Following the recommendation of [13] our new significance estimation takes into account the finder's specific performance. It hinges on an observation we previously made that the 3-parameter Gamma[†], or 3-Gamma for short, appears to fit very well F_f, the empirical distribution of the entropy score (under an iid background model) of several finders f. Here we first verified that this good fit extends to *every* combination of motif finder (MEME, CONSENSUS, many versions of Gibbs including Markov aware ones) and background model (iid or genomic samples) that we looked at.

Once we identify the parametric family of the null distribution of the score, F_f, we can use a parametric approach to evaluate the significance of the score. More precisely, by running our finder f on, say $n = 20$, random datasets (of the same dimension as the original input but from the assumed null distribution) we get a sample of size n from F_f. We can then fit a 3-parameter Gamma to this sample of scores and use the estimated parameters to obtain an estimation of the p-value of the observed score s. For the kind of parameters we are looking at (the shape parameter $a > 1$) this point estimator can be shown to be consistent [18] (i.e. it converges to the estimated p-value as $n \to \infty$). However this consistency is an asymptotic result and for small a sample size such as $n = 20$ this statistic can grossly over-estimate the significance of the observed score. One might be tempted to simply increase n to get a more reliable estimate, however that would increase the runtime of the finder by a factor of n. Thus in many cases we cannot realistically assume that we have access to a much larger sample size.

Here we complement the naive estimator by providing a conservative confidence interval for the estimated p-value. Conceptually we do so by first finding Θ, a 3-dimensional confidence set of the three estimated parameters of the distribution. We then maximize $1 - F_\theta(s)$ over all $\theta \in \Theta$ where $F_\theta(s)$ is the distribution function of a 3-Gamma with parameters θ and s is the observed score. While the idea itself is rather straightforward its implementation faces several difficulties which we had to overcome. In particular, how does one finds such 3-dimensional confidence sets and how can one guarantee a reasonably good maximization. These issues are discussed in more details below. The bottom line is that our significance evaluation is quite

[†] The distribution function of a 3-parameters Gamma with $\theta = (a, b, \mu)$ is a given by $F_\theta(s) = F_{\Gamma(a,b)}(s - \mu)$ where $F_{\Gamma(a,b)}$ is the Gamma distribution with it usual shape and scale parameters. We previously referred to it as a shifted-Gamma but a more standard definition is a 3-parameters Gamma where μ is the location parameter [7].

robust and should be applicable to many other combinations of scoring functions, motif finders and background models. To date no such general method is available except for a naive Monte Carlo estimation of the significance by directly comparing the observed score s to a random sample generated the same way we generate it. In the Section 7 below we compare our method with this as well as other significance evaluation methods.

2. The parametric family of the distribution of a motif finder's score

In [13] we showed that, especially in twilight zone motif searches, taking into account the performance of the specific motif finder that is considered yields a more reliable significance analysis. Thus, our new significance analysis attempts to estimate the distribution of the optimal score reported by the specific finder f which we denote F_f (or simply F). While in principle F can be estimated through extensive Monte Carlo simulations it is not in general a feasible solution. In particular F is not really a single distribution but rather an infinite number of distributions F^α, one per each set of values α of several relevant parameters such as: the size of the input sample, the finder's search parameters and the background model (whose parameters might be estimated from the sample). In practice this could mean we need to estimate from scratch a different F^α every time we run our finder which would present a generally unacceptable cost. However, if you know all the different distributions F^α belong to some parametric family then you only need to estimate the parameters of F^α rather than the entire distribution.

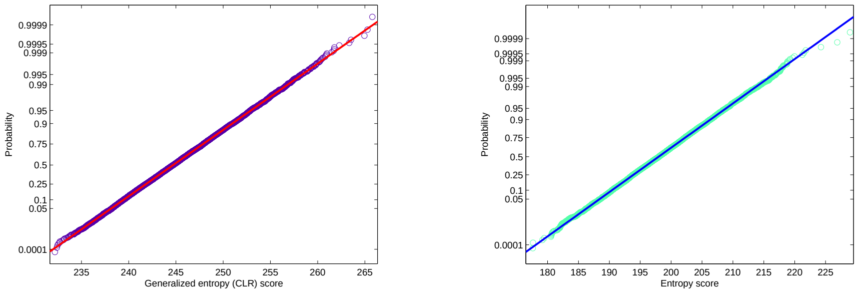
We previously mentioned that under the iid background model the null distribution of the optimal entropy score reported by several motif finders follows a 3-Gamma curve [13]. Interestingly, additional experiments we conducted using other scores[‡] as well as using the fairly realistic genomic background[§] indicate this result is fairly general: for all combinations of finders/scores and samples that we tested we found that the 3-Gamma family provides a fairly good fit to the empirical data. A couple of those fits are displayed in Figure 1.

At this point we do not know why the 3-Gamma family offers such good fits to these finder-specific distributions. Regardless, in what follows we assume these distributions do indeed belong to the 3-Gamma family. This is somewhat analogous to BLAST's assumption of a Gumbel distribution when analyzing gapped alignments. For while there exists an asymptotic theory for ungapped alignments no such theory can fully explain the observed Gumbel distribution in the gapped case [1]. Nevertheless, people rely on that assumption when assessing the significance of a gapped local alignment.

[‡] Specifically, ILR [13] and a generalization of the entropy score that adjusts for a Markov background.

[§] Where sequences were sampled out of a filtered human chromosome 22.

Fig. 1: Probability plots of 3-Gamma fits to the empirical score distribution



(a) GibbsMarkov

(b) MEME

The data was generated by running the mentioned program on 10^4 randomly generated datasets. In the case of MEME these were 20 sequences of each made of a block of 750 nucleotides sampled from a random location of a filtered human chromosome 22 whereas for GibbsMarkov we used 30 such sequences of length 1000 each

3. 3-Gamma based point estimator of the *p*-value

We previously showed that the parameters of the null distribution of CONSENSUS' entropy score can be reasonably grouped based only on two parameters of the finder [13]. In general however it is not clear how the parameters of the fitted distribution vary with the parameters of the problem (mentioned above as α). We therefore explore a different approach here relying on our ability to generate a *small* random sample $X = (X_1, \ldots, X_n)$ from F. Technically we generate this sample by first using the assumed background model to generate n independent random datasets of the same dimensions as the input dataset. We then run our finder on each of these n random datasets using the same settings as in the original application. Since this increases the runtime by a factor of n there is clear incentive to keep n as small as we can get away with.

Using this sample we can, for example, find $\hat{p}(s)$, the MLE (maximum likelihood estimator) of the p-value of the observed score s as follows. First we find $\hat{\theta} = \hat{\theta}(X)$, the MLE of the parameters of the 3-Gamma[¶], then we plug $\hat{\theta}$ into the definition of the p-value:

$$\hat{p}(s) = \hat{p}(s, X) = 1 - F_{\hat{\theta}}(s). \tag{1}$$

One problem is that with a small sample such as $n = 20$ one can only expect so much from fitting 3-parameters and indeed $\hat{p}(s)$ can badly over-estimate the significance of s (see Figure 2a). A standard way to account better for the variability of the sample is to introduce confidence intervals (e.g. [17]). For example, a 90%

[¶] In [13] we suggested fitting a 3-Gamma relying on our ability to fit the standard Gamma distribution. Here we adopt a different strategy relying on our ability to express the location parameter μ in closed form in terms of the shape and scale parameters at a critical point (calculation not shown).

confidence interval for the p-value is a random interval that contains the real p-value $p(s)$ with probability ≥ 0.9. We next construct such a confidence interval.

4. Confidence set for θ_0, the 3-Gamma parameters

Conceptually we construct our confidence interval for $p(s)$, the p-value of the observed score s in two steps. First, as described next, we use a generalization of the profile likelihood method (e.g., [14]) to construct a 3-dimensional confidence set for the 3-Gamma parameters θ_0. We then scan this confidence set to maximize the p-value of s at the prescribed confidence level.

Let $X = (X_1, \ldots, X_n)$ be our random sample and let $\hat{\theta} = \hat{\theta}(X)$ be the 3-Gamma MLE, so $\hat{\theta}$ is a 3-dimensional random vector. Let $L(X;\theta)$ be the 3-Gamma log-likelihood of the sample given the parameter $\theta \in \Omega$, where $\Omega \subset \mathbb{R}^3$ is the set of feasible parameters which would generally be a subset of ($a > 1$[‖], $b > 0$, $\mu \in \mathbb{R}$).

According to a general asymptotic result if the sample X is drawn according to the 3-Gamma distribution F_{θ_0} then

$$\Delta_L(X;\theta_0) = 2\left[L(X;\hat{\theta}) - L(X;\theta_0)\right]$$

converges in distribution to a $\chi^2(3)$ distribution as n, the size of the sample X, goes to infinity (e.g., [17]). *Assume* for now that this asymptotic result holds for our finite sample and define

$$\Theta_\gamma(X) = \left\{\theta \in \Omega \,:\, \Delta_L(X;\theta) \leq F^{-1}_{\chi^2(3)}(\gamma)\right\}, \tag{2}$$

where $F^{-1}_{\chi^2(3)}(\gamma)$ is the γ-quantile of the $\chi^2(3)$ distribution. Thus, $\Theta_\gamma(X)$ is random subset of $\Omega \subset \mathbb{R}^3$.

Claim 4.1. Assuming $\Delta_L(X;\theta_0)$ *is distributed under* θ_0 *as* $\chi^2(3)$, $\Theta_\gamma(X)$ *is a* γ-*confidence set for* θ_0.

Proof. We need to show that for any $\theta_0 \in \Omega$, $P_{\theta_0}(\theta_0 \in \Theta_\gamma(X)) \geq \gamma$. This follows immediately from the definitions:

$$P_{\theta_0}(\theta_0 \in \Theta_\gamma(X)) = P_{\theta_0}\left[\Delta_L(X;\theta_0) \leq F^{-1}_{\chi^2(3)}(\gamma)\right] = F_{\chi^2(3)}\left[F^{-1}_{\chi^2(3)}(\gamma)\right] = \gamma. \quad \square$$

When constructing confidence intervals it is a standard practice to assume as we did in the last claim, that an asymptotic distribution holds for a finite sample. Nevertheless, we would like to test what kind of errors does this assumption introduce in our case. This is particularly important since for practical reasons we restrict Ω, the set of feasible parameters, so that the shape parameter (coordinate) is restricted to a certain interval. We therefore conducted the following experiment.

[‖] We need $a > 1$ to guarantee the success of the estimation process as well as for the asymptotic result below [18]. Fortunately this is not a real restriction in our case as the distributions we are interested in have $a \gg 1$.

We first generated 44 different large sets of empirical scores. In this case all the scores came from one finder, GibbsMarkov, which is our version of Gibbs Sampler that uses a variant of the entropy score that accounts for a higher order Markov background model. GibbsMarkov will be described in detail in a following paper but for now it suffices to say it is similar to BioProspector [9]. Each such set of empirical scores contained 10^4 applications of GibbsMarkov on that many randomly generated datasets (with fixed dimensions per set). The sequences were randomly sampled from a filtered human chromosome and the dimensions ranged from 10 sequences of length 750 to 30 sequences of length 1000 each. Similarly, the width of the motif searched by GibbsMarkov varied from 8 to 28.

We then estimated the 3-Gamma parameters θ_0 for each of these 44 sets of empirical scores and defined $\Omega = \{(a, b, \mu) : a \in [10, 100], b > 0, \mu \in \mathbb{R}\}$. This range for the shape parameters was determined by taking a slightly larger interval than necessary to contain all 44 estimated shapes. Finally, we forget about the original sets and simply generate a large number (10^4) of random 3-Gamma samples of size $n = 20$ for each of these 44 estimated parameters θ_0. Since we know θ_0 in this case we can readily determine the proportion of times $\theta_0 \in \Theta_\gamma(X)$ and compare it to the theoretical rate of γ. The results of this test are summarized in Table 1.

Table 1: Actual confidence coefficients of parameter sets

γ in (2) set to:	0.85	0.90	0.95	0.99
actual confidence (%):	89.4-90.7	92.9-94.0	96.6-97.3	99.3-99.6

Range reported is of the percentage of time $\theta_0 \in \Theta_\gamma(X)$ for the specified γ observed across the 44 tests. Note how stable the observed ranges are, allowing us to correct for the slight conservative bias of the original $\chi^2(3)$ derived thresholds.

The table demonstrate our confidence sets are consistently slightly conservative. Consulting this table we can however adjust for the conservative nature of these confidence sets: for example, a nominal 85% confidence set is in fact a 90% one. Regardless of whether or not we adopt this adjustment we next show how we use our confidence set to generate a confidence interval for our real object of interest: the p-value of s, $p(s) = 1 - F_{\theta_0}(s)$,.

5. 3-Gamma based confidence interval for the *p*-value

Let

$$\hat{p}_c = \hat{p}_c(s, X) = \max\left\{1 - F_\theta(s) \,:\, \theta \in \Theta_\gamma(X)\right\}, \tag{3}$$

where F_θ is the 3-Gamma distribution with parameter θ.

Claim 5.1. *The random interval $[0, \hat{p}_c(s, X)]$ is a confidence interval for the p-value of s, $p(s)$, with confidence coefficient $\geq \gamma$.*

Proof. Since $\Theta_\gamma(X)$ is a γ-confidence set, $\theta_0 \in \Theta_\gamma(X)$ with probability $\geq \gamma$. In this case we clearly have

$$F_{\theta_0}(s) \geq \min\left\{F_\theta(s) \,:\, \theta \in \Theta_\gamma(X)\right\} \tag{4}$$

and therefore

$$p(s) = 1 - F_{\theta_0}(s) \leq 1 - \min\left\{F_\theta(s) \,:\, \theta \in \Theta_\gamma(X)\right\} = \hat{p}_c(s, X).$$

Thus, $p(s) \in [0, \hat{p}_c(s, X)]$ with probability $\geq \gamma$. □

Comment. Note that this estimate is conservative in nature as (4) might often hold even when $\theta_0 \notin \Theta_\gamma(X)$.

While conceptually our method for generating the confidence interval for $p(s)$ works as described above, we found that technically it is better to combine the two steps into one. More precisely, we define a target function for maximization:

$$\varphi(\theta; X, s) = \begin{cases} 1 - F_\theta(s) & \Delta_L(X;\theta) \leq d \\ -\Delta_L(X;\theta) & \Delta_L(X;\theta) > d \end{cases},$$

where $d = d_\gamma = F^{-1}_{\chi^2(3)}(\gamma)$. The following claim guarantees it suffices to maximize $\varphi(\theta; X, s)$:

Claim 5.2. $\max_{\theta\in\Omega} \varphi(\theta; X, s) = \hat{p}_c(s, X)$

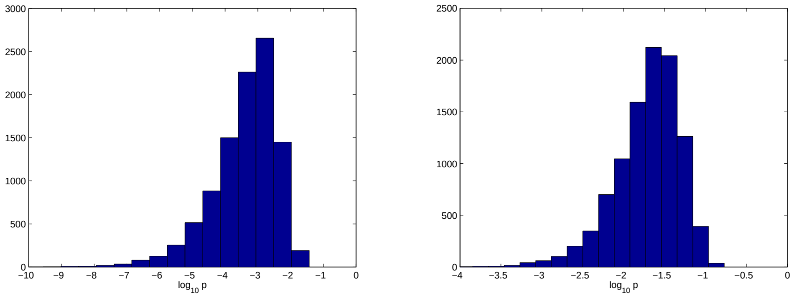
Proof. Immediate from the fact that $\varphi(\theta; X, s) < 0$ for $\theta \notin \Theta_\gamma(X)$. □

Mathematically, $\max_{\theta\in\Omega} \varphi(\theta; X, s) = \hat{p}_c(s, X)$ is a well defined statistic which, assuming the validity of the χ^2 approximation, defines a γ-confidence interval for the p-value, $p(s)$. However, in practice maximizing φ over Ω turned out to be somewhat tricky as the landscape of φ defined on $\Omega \subset \mathbb{R}^3$ is apparently quite complicated. This means that the actually computed version of $\hat{p}_c(s, X)$** might yield a confidence interval that would be smaller than it should be, i.e., its confidence coefficient would be $< \gamma$. In this case we would fail to achieve our goal here to get a confidence interval for $p(s)$ (we already have a reasonable point estimate in $\hat{p}(s)$ defined in (1) above).

In practice we used the Nelder-Meade [11] simplex based optimization procedure (implemented in the `constrOptim` function in `R` [15]) to maximize $\varphi(\theta; X, s)$ over $\theta \in \Omega$. Using Monte Carlo simulations for which we know the correct p-value we learned that simply relying on multiple random restarts is not satisfactory with this as well as with gradient based optimizations.We therefore used a pre-defined lattice of "reasonably good" starting points next to the boundary of Ω. Figure 2 shows a typical histogram of the computed conservative $\hat{p}_c(s, X)$ compared with the point estimator $\hat{p}$.

**We abuse the notations by not distinguishing between the mathematically defined statistic and its computed version.

Fig. 2: Comparing the estimators $\hat{p}$ and $\hat{p}_c(s, X)$ of $p = 10^{-3}$



(a) $\hat{p}$ can overestimate the significance

(b) $\hat{p}_c(s, X)$ is mostly conservative

Histograms of 10^4 independent evaluations of the point estimator $\hat{p}$ of $s = F_{\theta_0}^{-1}(10^{-3})$ and of the conservative $\hat{p}_c(s, X)$. The necessary 10^4 samples of size $n = 20$ were drawn with repetitions from a large set of empirical scores of the finder. In this case the set was made of 10^4 runs of GibbsMarkov on that many randomly generated datasets, each of which consisting of 30 sampled genomic sequences of length 1000. Since θ_0 is the 3-Gamma MLE of the set of empirical scores from which the samples of size n were taken, by definition, the real p-value of $s = F_{\theta_0}^{-1}(10^{-3})$ is essentially 10^{-3}.

6. The fidelity and utility of the confidence interval for the *p*-value

It was reassuring to see above that our confidence sets for θ_0 attain their prescribed confidence level. However, especially in light of the difficulty in maximizing φ over Ω, the more important question is whether or not our computed confidence interval $[0, \hat{p}_c(s, X)]$ contains $p(s)$ with probability $\geq \gamma$. To test that we conducted the following experiment based on our 44 sets of empirical scores described above.

We chose a set of p-values ranging from 10^{-9} to 0.1 and computed the values $s = s(\theta_0, p_0)$ for which $F_{\theta_0}(s) = p_0$[††]. Again, θ_0 is the 3-Gamma MLE obtained by fitting a 3-Gamma to each of these 44 empirical scores sets. For each of these s_0 (one per p-value and empirical score set) we computed $\hat{p}_c(s_0, X)$ for 10^4 samples X of $n = 20$ scores drawn independently, with repetitions, from the appropriate empirical set of scores. Note that when computing $\hat{p}_c(s_0, X)$ one assumes a specific confidence coefficient γ. We could then find the percentage of time $p_0 \in [0, \hat{p}_c]$ and test whether or not it is bigger than the prescribed γ. Table 2 gives the positive summary for all the cases we looked at.

By construction $\hat{p}_c(s, X)$ is a conservative estimate of $p(s)$ so we should not be surprised that, as observed above, it tends to underestimate the significance of s. We should however expect that it would not loose all the information. To demonstrate the utility of $\hat{p}_c(s, X)$ we conducted two tests as extensions of the

[††] Except for $p_0 \geq 0.005$ for which we could fairly reliably estimate s_0 directly from the empirical distribution of scores.

Table 2: Fidelity of confidence coefficient of estimated p-values

p_0	0.1	0.01	10^{-3}	10^{-4}	10^{-6}	10^{-9}
median %	1.70	1.69	1.23	1.36	1.16	0.85
minimum %	1.00	0.65	0.30	0.40	0.28	0.23
maximum %	2.55	5.22	3.50	3.77	4.38	4.67

γ in (2) was set to 0.85 which per Table 1 should really be a 90% confidence coefficient. The first row gives the prescribed p-value p_0. Rows 2-4 give the median, minimum and maximum of the percentage of samples for which $p_0 \notin [0, \hat{p}_c]$ among all 44 sets (a percentage for each set was computed as described in the text). Note that the even the worst case scenarios still attain the prescribed γ. The results for $\gamma = 0.9$ were qualitatively the same.

previously described test. In the first of these tests we asked for the percentage of samples X above for which $\hat{p}_c(s, X) \leq 0.05$. The latter represent an analogue of the canonical 5% significance threshold. More interesting is the actual value of $\hat{p}_c(s, X)$ so we looked at its median value across all 10^4 samples X. Table 3 summarizes the fluctuations of these statistics across all 44 empirical sets as a function of the actual p-value. Note how, especially for small p-values, $\hat{p}_c$ conveys significantly more information than simply "I passed the 5% significance threshold". For example, for $p_0 = 10^{-6}$ the median value (over all 44 sets) of the median (over all samples X) of $\hat{p}_c$ is roughly 6e-4.

Table 3: The utility of the confidence interval for the p-value

p_0	0.1	0.01	10^{-3}	10^{-4}	10^{-6}	10^{-9}
median %	0.10	32.33	86.67	99.40	100.00	100.00
minimum %	0.03	23.05	81.53	98.55	99.97	100.00
maximum %	0.30	47.60	92.90	99.85	100.00	100.00
median $\hat{p}_c$	0.25	0.068	0.022	0.0067	0.00059	1.2e-05
minimum $\hat{p}_c$	0.24	0.052	0.015	0.0035	0.00017	1.4e-06
maximum $\hat{p}_c$	0.26	0.079	0.027	0.0094	0.0012	3.9e-05

γ in (2) was set to 0.85 (90% in practice). The first row gives the prescribed p-value p_0. Rows 2-4 give the median, minimum and maximum of the percentage of samples for which $\hat{p}_c \leq 0.05$ among all 44 sets. Rows 3-6 yield the median, minimum and maximum among all 44 medians of $\hat{p}_c$. The results for $\gamma = 0.9$ were qualitatively the same.

One should keep in mind that with larger n the accuracy of $\hat{p}_c(s, X)$ can improve significantly. For example we compared using a sample of size $n = 20$ to $n = 40$ for $p_0 = 10^{-6}$. We found that while $n = 20$ yields a median (of medians) of roughly 5.9e-4, using a sample of size $n = 40$ cut the median to roughly 1.6e-4.

For a final practical test we went back to the Gibbs Sampler results on the `COMBO` experiment from [13] for which the E-value assessment failed miserably: the median of the *positive* examples was $\approx 10^{12}$. Using a set of 1600 runs on null iid datasets of the same dimension as in the original experiment we generated samples of size $n = 20$ and computed $\hat{p}_c(s, X)$ for each of the 400 scores s (each s is the entropy score of the Gibbs Sampler applied to a different, implanted dataset; see [13] for details).

We *predicted* a run as positive or successful if $\hat{p}_c(s, X) \leq 0.05$ and negative, or failure otherwise. We *labeled* a run as positive if the overlap between the reported and implanted alignment was $\geq 30\%$ and negative otherwise. Thus, we could count the number of TPs and FPs. To smooth out the results we repeated this process 100 times and averaged the number of TPs and FPs. Using $\gamma = 0.85$ our classifier defined above averaged 140.1 TPs and 6.3 FPs and an average of 55.6 of the scores s had a much more significant $\hat{p}_c(s, X) \leq 0.01$. Moving to a larger sample size of $n = 40$ we averaged 168.4 TPs and 8.5 FPs and an average of 80.6 had $\hat{p}_c(s, X) \leq 0.01$.

7. Discussion

We presented a novel approach for evaluating the significance of a motif finder results. It is important to keep in mind that as long as the fit of the finder's empirical scores distribution to a 3-Gamma is a reasonable one our method should be applicable to that finder. Since we have yet to see a case where that fit is not good we believe our method should apply to a wide variety of combinations of motif finders and scores and thus offer a unified parametric approach to estimating a finder's specific performance[‡‡].

We should point out that while our method suffers from a time penalty factor of n (the sample size), computing $\hat{p}_c(s, X)$ can be readily executed in parallel so that if sufficient additional CPU cores are available the effective time penalty reduces to only a factor of 2.

What are alternative significance evaluations? The authors of GLAM [4] assume that a scoring function they derive has a Gumbel distribution. They then try to evaluate its parameters analogously to BLAST. In particular, similarly to the E-value calculation they do not require the costly "on-the-fly" generation of a sample of null scores. While their method works reasonably well for a small number of sequences (≈ 5) it seems that the Gumbel assumption fails for a larger, more typical, number of sequences.

A more general alternative to the 3-Gamma distribution is the generalized extreme value (GEV) distribution [16]. While both of these distribution families offer fairly close fits and are difficult to distinguish at times, we found that typically the 3-Gamma offers a more reliable prediction of the right tail which is the one we are interested in. Additionally the 3-Gamma family was easier to handle in terms of

[‡‡] Technically, adjusting our method to a new finder would typically require some crude charting of the space of plausible values for the parameters: the more restrictive the range of feasible parameters Ω is, the more accurate will $\hat{p}_c(s, X)$ be.

fitting and predicting confidence sets. Finally, while the GEV might sound attractive as it is known to be the only possible asymptotic limit of a maximum of an iid sequence one should keep in mind that the motif finding problem is very different from the alignment problem where such extreme value theory applies.

Although we could not find any trace of this in the program itself, the BioProspector paper [9] suggests an approach which is similar in spirit to the computation of our point estimator $\hat{p}$ (1). One major difference is they suggest that the normal approximation should be used. We found no evidence supporting the use of a normal approximation and no such convincing evidence is presented in that paper. In all our studies the 3-Gamma family offered significantly superior fits at a cost of only one more parameter to estimate.

Alternatively we can resort to non-parametric tests. For example, we can use the generated sample to construct confidence intervals for the p-value the same way we estimate p of a binomial $B(n, p)$ distribution. The problem is these tend to be quite conservative so for $n = 20$ the best confidence interval for the p-value would be $[0, 0.11]$ while for $n = 40$ it would be $[0, 0.06]$.

A different kind of non-parametric test is to test, for example, if s is bigger than all the entries in a random sample of size $n = 20$ (this can be generalized using the Mann-Whitney statistic). As described this is a reliable test at the (roughly) 5% significance level whose main down side is that it offers very little information about the quality of significant results. In particular, it will not provide any more information about a score whose real p-value is 10^{-6} than it would about any other score s that passes the 5% test: all you learn is that you are 95% confident that the observed dataset is not a random one. Our method on the other hand, though conservative, does respond to differences between scores which among other things would make it more appropriate to compare motifs of different widths where the scores cannot be compared directly against one another. Moreover, this non-parametric method has a high "false positive" rates for scores s whose p-value is close to 0.05. For example, if $p(s) = 0.1$ then 12% of the time s will be declared significant at the 5% level and if $p(s) = 0.06$ this will happen 29% of the time.

There are many directions and questions that our paper opens up including: applying our significance analysis method to other finders and making them an integral option of GibbsMarkov as well as other finders, developing a better theoretical understanding of why the 3-Gamma offers such good fits to the optimal score distributions, and testing how good a fit remains once we start adding additional information such as ChIP-chip or phylogeny data to our motif finders. We plan on exploring these issues in future research.

References

[1] SF Altschul and W Gish. Local alignment statistics. *Methods Enzymol*, 266:460–80, 1996.

[2] T.L. Bailey and C. 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.

[3] Martin Tompa et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, Jan 2005.

[4] Martin C Frith, Ulla Hansen, John L Spouge, and Zhiping Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32(1):189–200, 2004.

[5] GZ Hertz and GD Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999.

[6] JD Hughes, PW Estep, S Tavazoie, and GM Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–14, Mar 2000.

[7] Kotz S. Johnson N.L. and Balakrishnan N. *Continuous Univariate Distributions, 2nd edition*. Wiley Series in Probability and Statistics, 1994.

[8] CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, Oct 1993.

[9] X Liu, DL Brutlag, and JS Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–38, 2001.

[10] Niranjan Nagarajan, Neil Jones, and Uri Keich. Computing the P-value of the information content from an alignment of multiple sequences. *Bioinformatics*, 21 Suppl 1(ISMB 2005):i311–i318, Jun 2005.

[11] J. A. Nelder and R. Mead. A simplex algorithm for function minimization. *Computer Journal*, 7:308?313, 1965.

[12] AF Neuwald, JS Liu, and CE Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4(8):1618–32, Aug 1995.

[13] Patrick Ng, Niranjan Nagarajan, Neil Jones, and Uri Keich. Apples to apples: improving the performance of motif finders and their significance analysis in the Twilight Zone. *Bioinformatics*, 22(14):e393–401, Jul 2006.

[14] Yudi Pawitan. A reminder of the fallibility of the wald statistic: Likelihood explanation. *he American Statistician*, 54(1):54–56, 2000.

[15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[16] Sidney I. Resnick. *Extreme values, regular variation, and point processes*. Springer-Verlag, New York, 1987.

[17] J.A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, second edition, 1995.

[18] Richard L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90, 1985.

[19] GD Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

[20] Jing Zhang, Bo Jiang, Ming Li, John Tromp, Xuegong Zhang, and Michael Q. Zhang. Computing exact P-values for DNA motifs. *Bioinformatics*, 23(5):531–537, 2007.