BIOINFORMATICS

Vol. 22 no. 14 2006, pages e393–e401
doi:10.1093/bioinformatics/btl245

Apples to apples: improving the performance of motif finders and their significance analysis in the Twilight Zone

Patrick Ng[1], Niranjan Nagarajan[1], Neil Jones[2] and Uri Keich[1,*]

[1]Department of Computer Science, Cornell University, Ithaca, NY, USA and [2]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA

ABSTRACT

Motivation: Effective algorithms for finding relatively weak motifs are an important practical necessity while scanning long DNA sequences for regulatory elements. The success of such an algorithm hinges on the ability of its scoring function combined with a significance analysis test to discern real motifs from random noise.

Results: In the first half of the paper we show that the paradigm of relying on entropy scores and their E-values can lead to undesirable results when searching for weak motifs and we offer alternate approaches to analyzing the significance of motifs. In the second half of the paper we reintroduce a scoring function and present a motif-finder that optimizes it that are more effective in finding relatively weak motifs than other tools.

Availability: The GibbsILR motif finder is available at http://www.cs.cornell.edu/∼keich

Contact: Uri Keich, keich@cs.cornell.edu

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 [17] (see [19] for a recent comparative review).

Keich and Pevzner [10] define a *twilight zone* search as one in which there is a non-negligible probability that a maximally scoring random motif would have a higher score than motifs that overlap the ''real'' motif (in the model considered there, a ''real'' motif is implanted into randomly generated background sequences). In such cases, even if one had access to a hypothetically ideal finder that was guaranteed to return the highest scoring alignment in the dataset, the motif might remain unfound. Locating the twilight zone is necessary in deciding whether or not the current state of the art in motif finding is good enough: if existing tools find the correct motif for datasets all the way into the twilight zone in a reasonable time, further improvement will yield at best marginal returns. Of course, improving motif finding tools to be effective into the twilight zone is not merely a theoretical exercise: a biologist searching for regulatory motifs in DNA sequences would generally prefer to choose longer rather than shorter regions in order to avoid missing regulatory elements that are far away from the transcription start site of a gene. The longer the input sequences are, the more likely they are to contain high scoring random motifs, pushing the biologically valid motifs into the twilight zone.

Most existing motif finders can be divided into two classes depending on how they model a motif. Tools that rely on a combinatorial model of a motif define a motif to be a consensus sequence with an associated distance (usually Hamming distance), as described in [14]. Under this definition, the problem of finding motifs in random sequences is mostly solved. The statistics of optimal random motifs are well understood in this context, which led to the characterization of the twilight zone [10]. Moreover, the PatternBranching tool [16] exhibits good performance, even in the twilight zone for reasonable choices of parameters leaving little motivation for further improvement.

On the other hand are tools that describe a motif as a profile, a probabilistic distribution generally modeled with a position weight matrix (PWM). Prominent examples of this class are MEME [1], CONSENSUS [6] and the various approaches to Gibbs sampling (e.g. [11],[13],[7]). Under this definition of a motif, there has been no definitive demonstration of any particular tool's dominance. Moreover there is no reliable characterization of the distribution of optimal random motifs[1], nor is the twilight zone completely understood.

In most applications of a motif finder, the user must decide whether or not a motif reported from a motif finder warrants further biological investigation based on its statistical significance. The first half of this paper deals with the significance analysis of the ubiquitous entropy score. We begin by showing that the common practice of using the *E*-value of the entropy score (defined below) to evaluate the significance of an alignment reported by a motif finder can lead to undesirable results in twilight zone searches. We then discuss two additional intuitively motivated measurements of statistical significance and some pitfalls in their application to motif finding. The second half of the paper discusses an alternative scoring scheme. This is motivated by the observation that comparing entropy scores across different motif finders often leads to inconsistent results regarding the identification of the implanted motif. We reintroduce the Incomplete (data) Likelihood Ratio (ILR) and show it is a better classifier

*To whom correspondence should be addressed.

[1]Some progress was made recently by Frith, *et al*. [5], but the analysis presented there only holds for a small number of sequences.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

when it comes to predicting overlap with implanted motifs. This motivates GibbsILR, a new variant of the Gibbs sampler that attempts to maximize the ILR rather than the entropy score.

2 ARE MOTIF FINDERS PSYCHIC? THE CONUNDRUM OF *E*-VALUES

One of the key measurements in determining if a motif finder has identified an important motif is the *E*-value of the entropy score defined as follows. The entropy score or information content of the reported alignment is defined as [17]:[2]

$$I := \sum_{i=1}^{w} \sum_{j=1}^{A} n_{ij} \log \frac{n_{ij}/n}{b_j},$$

where w is the motif width, n_{ij} denotes the number of occurrences of the jth letter in the ith column of the alignment, b_j is the background frequency of the jth letter[3], n is the number of sequences in the alignment, and A the alphabet size ($A = 4$ in this paper). Introduced originally in this context as the ''expected frequency'' [6], the *E*-value is the expected number of random alignments of the same dimension that would exhibit an entropy score that is at least as high as the score of the given alignment. When the *E*-value is high, one can have little confidence in the motif prediction, and conversely when the *E*-value is low, one can have more confidence in the prediction. It is computed by multiplying the number of possible alignments by the *p*-value of the alignment. The latter is defined as the probability that a single *given* random alignment would have an entropy score $\geq$ the observed alignment score. Assuming the customary iid (independent identically distributed) random model the *p*-value can be computed accurately using techniques we previously described [12].

To assess the performance of motif finders in twilight zone searches, we designed the following experiments containing 400 data sets (see the COMBO experiments in the Methods section). Each randomly generated data set contained a deliberately implanted profile motif in such a way that for a non-trivial percentage of datasets, the motif finders we considered would pick motifs that would not overlap the implants. Thus, it is not surprising that the *E*-value of the implanted motif is relatively high. However, with a median *E*-value of 8×10^{15} it seems this problem is way beyond the twilight zone. Indeed, one would suspect that in this case even the ideal finder would not be able to pick out an alignment with significant overlap to the implanted motif from the large number of background alignments with better entropy score. Rather startlingly, exactly the opposite is true: of 400 data sets, the Gibbs sampler [11] found an alignment overlapping more than 30% of the implanted sites in 288 cases[4]. It is important to note that these data sets are constructed exactly according to the model used in computing the *E*-values, thus we can safely assume the *E*-value is quite accurate [12].

How can our motif finders be so lucky that they pick a ''real'' motif out of such a huge haystack? A partial answer to this riddle is obtained by noting that when a motif is implanted into a set of long sequences, there is a good chance that a random string in one of the sequences will slightly improve the entropy score. Of the 288 data sets for which the Gibbs sampler found an overlapping alignment (above the 30% threshold), the median *E*-value of the reported motif was 8.7×10^{11} or 4 orders of magnitude better than the initial motif. Still, it is a very impressive haystack and a more complete answer probably lies in what we do not see: how many alignments that overlap with our implant have a score as good as the one found? These high scoring ''satellite'' alignments define some ''domain of attraction'' for a motif that is difficult to characterize analytically. Presumably, its size has to be of the order of the *E*-values as sampling optimization procedures such as Gibbs somehow find it. We remark that characterizing this domain of attraction is a potential way to describe the twilight zone of a profile-based motif.

Whatever the explanation is, it is clear that the *E*-value offers little benefit in analyzing the significance of twilight zone search. We next explore alternative approaches to this problem.

3 ALTERNATIVE SIGNIFICANCE ANALYSES

One alternate measure of significance suggested by Hertz and Stormo [6] is that of the ''overall *p*-value''—or $OPV(s)$—of an entropy score s. It is defined as the probability that a random sample of the same size as the input set will contain an alignment of the same dimensions that scores at least as high as s. While this statistic is intuitively appealing, its use faces two hurdles. On the one hand, at present it is all but impossible to calculate $OPV(s)$ for moderately large datasets: even generating an empirical estimate of the OPV would necessarily require the ability to reliably find the highest scoring alignment in any given sample, which cannot be guaranteed for realistic problem sizes. On the other hand, even if an accurate method for calculating $OPV(s)$ were known, the evidence presented next suggests that this significance measure would impose too high a barrier on the entropy score for functional motifs to be distinguishable from noise.

The value of $OPV(s)$ may be conservatively estimated by the probability that at least one of several motif finders would find an alignment of score $\geq s$ in the random data. The point is the latter is amenable to Monte Carlo estimation. Using 1600 randomly generated datasets with no motif implanted we obtain an empirical estimate of the 0.95 quantile of the latter distribution; this is the minimal value s_0 such that for 95% of the datasets all our finders report a top alignment of score $\leq s_0$. We then use s_0 as an empirically derived conservative estimate of the threshold s_1 such that $OPV(s_1) = 0.95$. That is, 95% of the top scoring noise alignments have entropy less than or equal to s_1 and $s_1 \leq s_0$ with high probability. When this derived 5% significance level was applied as a threshold for significance of the 400 data sets in the COMBO experiment, nearly 90% of the correct runs of the Gibbs sampler (i.e., those runs that overlapped the implanted motif by more than 30%) were classified as noise. Since s_0 the conservatively estimated 0.95 quantile is very likely to be greater than the true quantile s_1, this should become more pronounced with better approximations of $OPV(s)$ suggesting it is also too conservative.

One can see that $1 - OPV(s)$ is the distribution function of the ideal motif finder. This raises the natural extension of using a finder-specific OPV: $1 - F_f(s)$ where F_f is the null distribution of the score of the optimal alignment detected by the particular finder. That is, we ask for the probability that the finder will find an

[2]Strictly speaking, relative entropy is defined as I/n.

[3]Typically estimated from the entire sample.

[4]See the Methods section for the parameter setting.

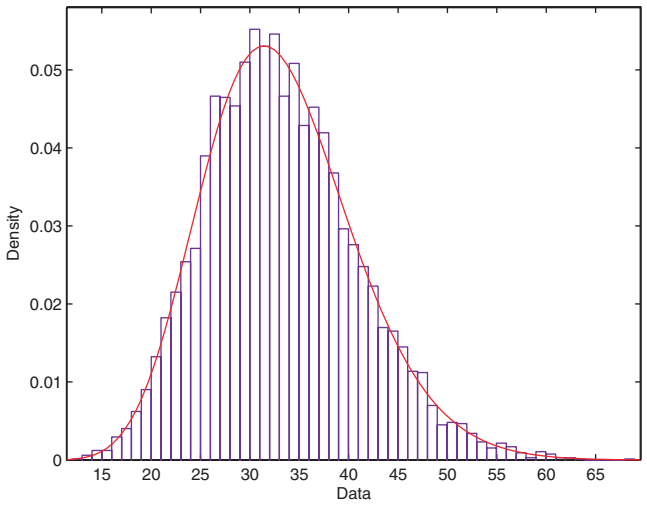

Fig. 1. Shifted gamma fit to 6400 runs of Gibbs with parameters `13 -t100 -L100` on 40 random sequences of length 750, uniformly distributed with no implanted motif.

alignment scoring $\geq s$ in a random dataset (of the same dimensions). Again, we can estimate the quantiles of this distribution function through a Monte Carlo generated empirical distribution. In this case we found that the .95 quantile threshold of Gibbs, estimated from 1600 datasets, yields 13 false positives (FP) and 228 true positives (TP) when applied to the same 400 data sets of the COMBO experiment.[5]

While the empirical distribution can be extremely useful in analyzing the significance of a motif finder's output, generating it *a priori* is typically impossible due to the large number of combinations of parameters. Similarly, generating even a rough estimate of a 0.95 quantile per problem instance is impractical as it would require at least 100 additional runs of the motif finder on a dataset of the same size as the input.

However, if we can characterize the distribution as belonging to some parametric family of distributions, we might do better to estimate the parameters of the distribution rather than directly estimating the quantiles of the distribution. The (limiting) distribution of a maximal ungapped pairwise alignment between two sequences is a Gumbel Extreme Value Distribution (EVD) [9]; the same distribution is encountered empirically in the gapped case and it is presumed to underly the distribution of scores when local multiple alignments are scored according to a presumed phylogeny [15] and in Frith *et al*. [5], which specifically discusses motif finding. Oddly, the empirical null distribution of the reported entropy score for several motif finders exhibited a better fit to a (shifted) Gamma distribution than to the intuitively more appealing Gumbel distribution (see Fig. 1 and Fig. 2 for an example involving Gibbs).[6]

[5]That we expect $5\% \cdot 400 = 20$ false positives and see only 13 is reasonable since some of those random datasets containing high-scoring alignments are masked by higher-scoring motifs that overlap the implant.

[6]To fit a shifted gamma distribution for each shift we find the likelihood of the shifted data by applying a standard maximum likelihood gamma fit to it, and then use a simple one dimensional search of the shift that yields the highest likelihood.

Naturally, the parameters of the distribution could depend on the size and background distribution of the dataset, as well as the parameters used in the motif search. It is surprisingly easy to get a good approximation of how the empirical distribution of CONSENSUS behaves with respect to these variables. The key is to consider the distribution of the *E*-value of the best reported entropy rather than the entropy itself. This *E*-value is fairly stable for a wide range of sequence lengths (375–1500), motif widths (10–50), and with different sequence composition (uniform background *versus* a biased background of (0.2, 0.2, 0.3, 0.3)), see Figure 3 and Table 1. That is, the empirical distribution depends primarily on N, the number of sequences; and q, the number of sub-alignments that CONSENSUS keeps at each stage. While the parameter estimates for the shifted gamma distribution are not perfectly stable, they could be readily improved by dividing the range of parameters (e.g., motif width) into several segments and using one shifted gamma distribution per interval.

We demonstrated above that the OPV or equivalently the distribution function of the ideal finder seems too conservative for estimating the significance of a motif finder's output. Nonetheless it is useful in delineating the twilight zone, which in turn is important for understanding to what extent existing tools might be theoretically improved upon. Indeed, by comparing the empirical distribution of a motif finder with that of the ideal one for a given set of parameters, we can assess the efficiency of the finder for these parameters. It is thus interesting to determine whether this distribution can be approximated by a parametric family. As above we find the surprising result that a shifted gamma distribution gives a better fit to the empirical distribution than a Gumbel distribution. One might expect that the result of maximizing over all possible alignments would naturally result in an EVD but according to our observations this is not the case (see Figure 4). One reason is that the high scoring alignments are heavily dependent, an observation made by Frith *et al*. [5] when trying to explain the less-than-perfect fit they got to a Gumbel distribution.

4 INCOMPLETE (DATA) LIKELIHOOD RATIO AND GIBBSILR

A good scoring function should separate as much as possible real motifs or, in the context of our model, alignments that have overlap with the implant, from purely random ones. The entropy score is the one chosen by popular motif finders such as MEME [1], CONSENSUS [6] and Gibbs Sampler [11]. The latter two specifically try to optimize this scoring function, while MEME uses it only to rank and analyze the significance of its output. It is thus tempting to assume that if we run, for example, both CONSENSUS and Gibbs and take the higher scoring motif we would do better than if we ran each one of them separately. Amazingly, this might not be the case, especially in twilight zone searches. In particular, in our COMBO experiment we find that in 380 of the 400 datasets CONSENSUS finds a motif with higher entropy score than Gibbs, yet Gibbs reports more motifs that have $\geq$30% overlap with the true implant (290 of the sets for Gibbs compared to 208 for CONSENSUS). Comparing the entropy score from different motif finders is thus not an apples to apples comparison as one would expect—somehow it matters how the entropy is maximized. This led us to ask if other scoring functions would possibly capture better the nature of real (implanted) twilight zone motifs. One

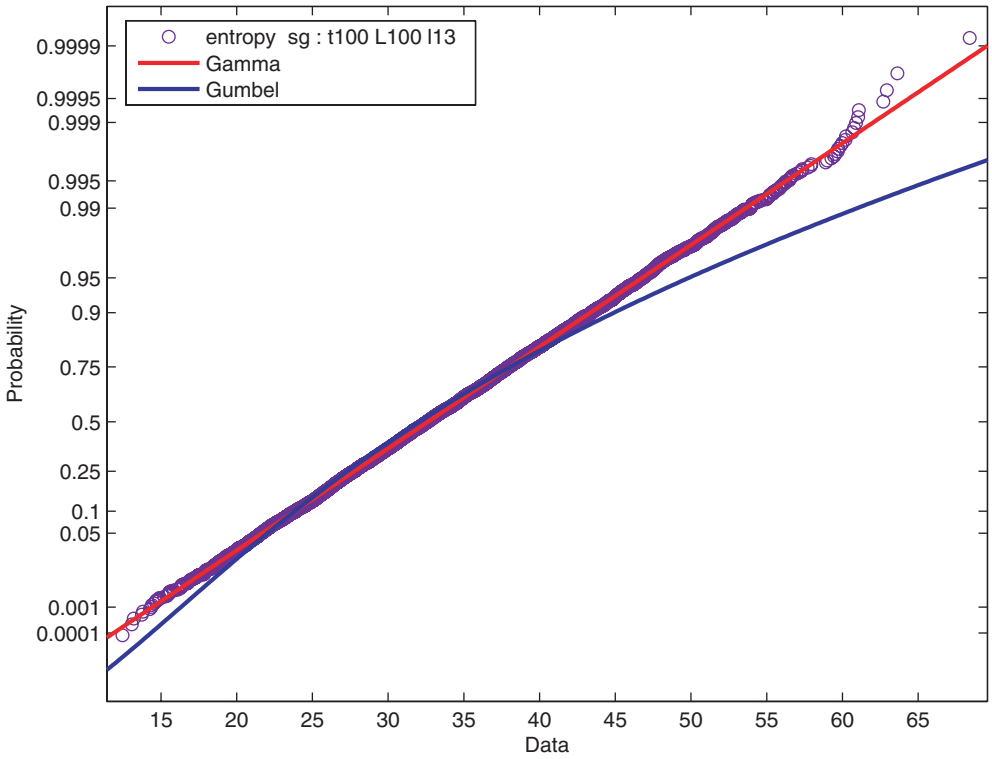

Fig. 2. The probability plot of the fit of a shifted gamma distribution and of a Gumbel Extreme Value Distribution to the data collected in Fig. 1.

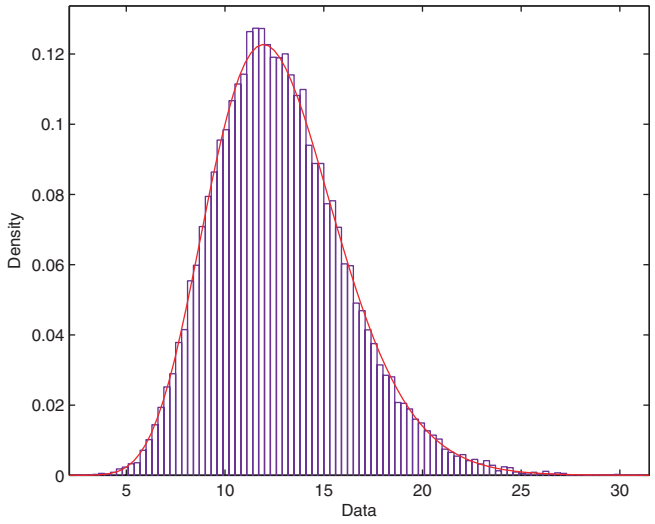

Fig. 3. A (shifted) gamma distribution fitted to the empirical null distribution of $-\log E$-value of the best motif. The data was compiled from 6400 datasets of 40 uniformly distributed sequences of length 750. For each $w \in \{10, 13, 16, 20, 30, 50\}$ CONSENSUS was ran on each dataset with $q = 1000$.

scoring function presented next shows consistent, and at times considerable improvement over the entropy score.

Given a set of N sequences $S = \{S^1, S^2, \ldots, S^N\}$, the formal definition of the Motif Finding problem is to find the set of starting positions within each sequence that corresponds to the location of an implanted motif. We assume there is a profile matrix $\Theta = (\Theta_{ij})$ of length w that represents the implanted motif model, and another profile matrix $\Theta_0 = (\Theta_{0j})$ that represents the background. We define the Incomplete (data) Likelihood Ratio as follows:

$$ILR(\Theta) = \prod_{n=1}^{N} \left[\sum_{m=1}^{|S^n|-w+1} \frac{P(S^n_{m:m+w-1}|\Theta)}{P(S^n_{m:m+w-1}|\Theta_0)} \cdot \frac{1}{|S^n| - w + 1} \right] \quad (1)$$

Intuitively, $ILR(\Theta)$ is the likelihood ratio between two competing hypotheses. The null hypothesis is that the data was entirely generated under the null model Θ_0. The alternative hypothesis is that the data was generated under the OOPS (one occurrence per sequence) model [2] using the motif Θ and the background Θ_0. Unlike the standard entropy score, the ILR scores a motif by taking into account all of the data in S, rather than only the data within a particular alignment. The EM algorithm optimizes the ILR [3], and by extension MEME does as well. However, MEME ranks motifs by entropy and assesses the reported motif's significance through the E-value of the entropy score. In particular, the ILR score has not been previously used to score and rank motifs.

Our tests (described in the Results section) demonstrate that for twilight zone searches ILR is a consistently better classifier than the entropy score for identifying motifs that overlap the implant. Since this holds for all the finders that we tested, most of which optimize the entropy score, it motivates the design of a new finder that tries to optimize the ILR: GibbsILR is based on the Gibbs-sampling technique described by Lawrence *et al.* [11]. Here we modify the original Gibbs sampling strategy by using a hybrid optimization procedure. The Gibbs-sampling motif finder begins each run by picking a random starting position in each sequence in the data

Table 1. Comparison of empirical distributions with the fitted gamma distribution in Figure 3. A column with heading $r(x)$ contains the ratio of the CDF of the empirical distribution to that of the fitted shifted gamma distribution at the x^{th} quantile of the fitted distribution. Note that when considering E-values, small quantiles are good

Test Set	w	r(0.1)	r(0.05)	r(0.01)
40 sequences of length 750	10	0.91	0.83	0.89
to which the gamma was fitted	13	1.13	1.21	1.22
	16	1.06	1.11	1.19
	20	1.10	1.19	1.44
	30	1.04	1.12	1.37
	50	0.65	0.67	0.87
40 sequences of length 1500	10	0.34	0.30	0.44
	13	0.73	0.67	0.94
	16	0.98	1.10	1.37
	20	0.94	0.91	1.19
	30	0.85	0.96	1.25
	50	0.51	0.59	0.75
biased composition	10	0.55	0.56	1.00
20 sequences of length 375	13	0.82	0.82	0.75
and 20 sequences of length 750	16	0.96	0.95	1.19
	20	1.13	1.24	2.12
	30	0.88	0.94	1.44
	50	0.49	0.51	0.37

set. The algorithm then iterates between two steps, commonly referred to as the predictive update step and the sampling step. In the k-th iteration, the predictive update step computes a motif model Θ^k based on the current chosen set of starting positions.[7] The sampling step in turn randomly selects new candidate starting positions with probability proportional to the likelihood ratio of the position given the current model Θ^k.

A well-known property of Gibbs-sampling algorithms is that they are guaranteed to sample the global maximum given sufficient time, but this may take an unacceptably long time to happen. Instead, when the objective function is apparently not making any headway, we can ''restart'' the sampling procedure by initializing a new, independent, Gibbs-sampling run using a new set of random starting positions. Unlike previous Gibbs-sampling motif finders, GibbsILR runs an EM (Expectation-Maximization) algorithm that locally optimizes ILR on the final motif of each Gibbs-sampling run. GibbsILR then produces a motif that exhibits locally optimized ILR score by taking the highest ILR-scoring motif among all of the final motifs derived from the EM step. Finally, for each sequence in the dataset S, the motif instance corresponds to the position with the highest likelihood ratio with respect to the highest ILR-scoring motif profile.

5 RESULTS

The first group of results is based on extensive tests of the performance of six profile-based motif finders on synthetic data. Each of these randomly generated datasets contained a deliberately implanted profile motif (see the Methods section for more details). The output of each of the finders we considered (CONSENSUS, Gibbs, GibbsILR, GLAM, ProfileBranching, and MEME) was post-processed to yield both the entropy and ILR scores of the finder's top reported alignment. We then asked which of these two scores is a better predictor of overlap with the implant (which is a surrogate for a real motif).

We compare the entropy and ILR score by measuring the area under the ROC curve [18], or discrimination, for each finder under the two scoring functions. We classify a set of motif sites as negative if the overlap score is below 0.1; otherwise, we classify it as positive. Intuitively, given a random pair of positive and negative set of profile sites, the aROC tells us the probability of the test correctly identifying the pair's classification. The tests (Table 2) using ILR score have consistently better discrimination than the tests using entropy score. The reader should note that it is however unfair to compare the performance of the finders using aROC, because the number of negatives and positives differ across the finders. For example, GibbsILR has lower discrimination than MEME for both entropy and ILR in COMBO, but GibbsILR has 324 positives to discriminate whereas MEME has only 70 positives.

Similarly we can ask how many true positives (TP) are in the test set if we are willing to accept exactly 10 false positives (FP). Table 2 shows that ILR consistently has higher counts of such TPs than entropy. Moreover, if we would like to design a classifier that only accepts 10 FPs, this analysis shows that the combination of ILR score and GibbsILR would give us the highest number of TPs.

We next combine five motif finders: CONSENSUS, Gibbs_ss, GLAM, MEME, and ProfileBranching by choosing the set of motif sites from the finder with highest ILR. Likewise, we employ the same technique with the entropy. We found that the ILR variant of the combined-finder can perform better than any of its individual finders alone. In the COMBO experiment, the ILR variant found the implants in 311 datasets (i.e. overlap score greater than 0.1), whereas its best individual finder, which is Gibbs_ss, found the implants in only 302 datasets. In the same experiment, the entropy variant found the implants in 291 datasets, which is worse than its best individual finder. For a different approach to combining the output of multiple motif finding algorithms, see [8].

As an additional source of evidence for the utility of the ILR score we generated synthetic data sets implanted with motifs that were verifiably in the (entropy score) twilight zone. The branch and bound algorithm described in the Methods section was then used to find the motif with the optimal entropy score and the ILR score of that motif. Then, based on the results from 1000 such runs we asked the following question: which of the scores, entropy or ILR is a better predictor of overlap with the implanted motif? For the twilight zone data sets that we tested, ILR is consistently better than the entropy score as a predictor of overlap (as measured by the aROC score, with overlap being defined as an overlap score greater than 0.1). As a specific example, for $N = 14$, $L = 80$ and `SHORT` (see Table 3), the entropy score has an aROC score of 0.52 as compared to 0.60 for the ILR score. In practical terms, for a threshold that allows 50 false positives, the ILR score gives 143 true positives as opposed to 101 for the entropy score. Interestingly, in this example, while the ILR score has a positive

[7]The model Θ is inferred from the starting positions by the rule $\Theta_{ij} = \frac{c_{ij} + b_i}{N - 1 + \sum_j b_j}$ where c_{ij} is the count of letter j in the i-th sequence of the alignment and b_j is an *a priori* chosen pseudocount to avoid 0 probabilities.

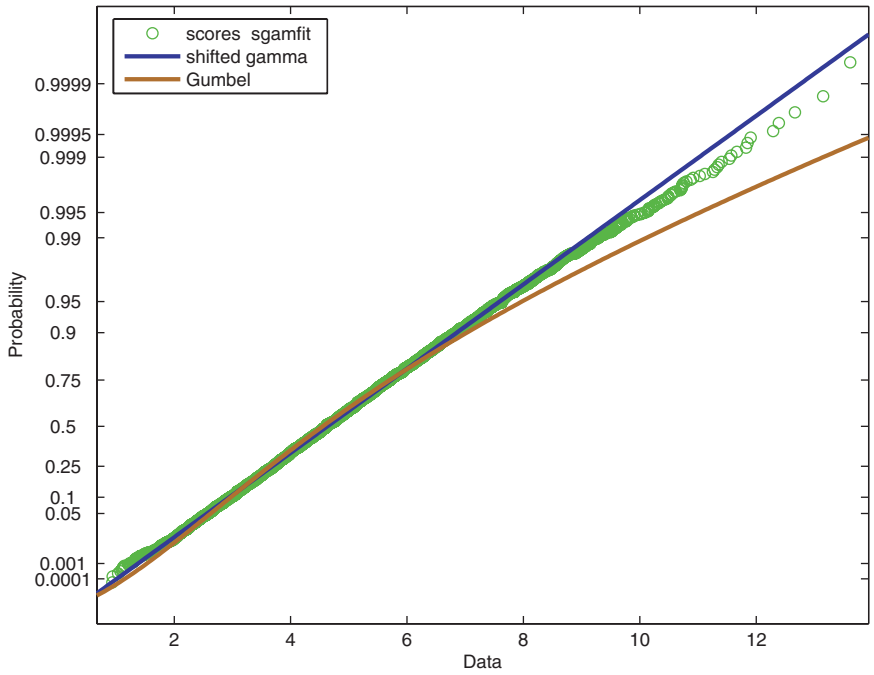

Fig. 4. The probability plot of a fit of the OPV distribution to a shifted gamma distribution and to an EVD distribution; the OPV distribution was generated from the output of the ideal motif finder (searching for motifs of width 7; see Methods section) run on 10000 datasets composed of 10 uniformly distributed sequences with length 100.

Table 2. aROC and only accepting 10 FPs. The column >10% contains the number of datasets that score above the 0.1 overlap threshold. The column TPs contains the number of true-positives in a test if it is willing to accept ≤10 FPs

Experiment	Finders		entropy		ILR	
		>10%	aROC	TPs	aROC	TPs
COMBO	CONSENSUS	223	0.88	154	0.93	169
	Gibbs_ss	302	0.88	208	0.91	231
	GibbsILR	324	0.85	254	0.90	258
	GLAM	170	0.90	117	0.94	127
	MEME	70	0.90	43	0.92	48
	ProfileBranching	222	0.95	183	0.96	190
FIFTY	CONSENSUS	27	0.73	5	0.85	13
	Gibbs_ss	87	0.94	70	0.96	76
	GibbsILR	186	0.96	171	0.96	171
	GLAM	116	0.94	91	0.89	84
	MEME	4	0.64	0	0.73	0
	ProfileBranching	8	0.60	1	0.72	1

Spearman correlation, the entropy score has a statistically significant negative correlation with the overlap score (Spearman correlation p-value of $5.2 \cdot 10^{-4}$)[8].

Finally while not an objective demonstration of the advantage of ILR, GibbsILR did show improvement in our experiments over the other five finders we tested. Fig. 5 shows the overlap distribution for the various finders. For example, the bars at 0.1 are the number of datasets that a particular motif finder found with overlap score between 0.1 and 0.2. GibbsILR finds the most datasets above 0.1 overlap score for both experiments. In the case of FIFTY (Fig. 5b), GibbsILR is significantly better. Note that we tried to equalize the running time of all the algorithms in the benchmark as described in the Methods section below.

[8]Recall that the detected motif was optimized for the entropy, rather than the ILR.

6 CONCLUSION AND FUTURE WORK

We have demonstrated several discouraging observations regarding the use of E-values to determine the statistical significance of a motif. However, the E-value of the entropy score has at least two redeeming qualities. First, it is always conservative, so a motif exhibiting a very low E-value is probably significant. Second, the distribution of the overall p-value for CONSENSUS was easy to characterize when considering the E-value of the entropy score, rather than the entropy score itself. We acknowledge the possibility that other motif finders may exist (that we have not yet tried) whose finder-specific OPV may be similarly quantified.

We have also presented an alternative scoring function to be used in place of entropy, along with a motif finder that uses this function to achieve demonstrably better results than existing algorithms. The motif finder described here is admittedly simplistic, so it seems likely that more effective algorithms could be developed.

We were surprised to discover that the empirical distribution of optimal alignment scores displayed by each algorithm, including the exhaustive motif finder, fit a gamma distribution far better than the intuitively more appealing EVD. It would be informative to determine why the gamma distribution, specifically, seems to model this problem accurately.

Table 3. The position weight matrices used in these experiments

	COMBO				FIFTY				SHORT			
Pos.	A	C	G	T	A	C	G	T	A	C	G	T
1	0.95	0.00	0.00	0.05	0.50	0.00	0.00	0.50	0.95	0.00	0.05	0.00
2	0.00	0.50	0.50	0.00	0.00	0.50	0.50	0.00	0.00	0.05	0.95	0.00
3	0.70	0.10	0.10	0.10	0.50	0.50	0.00	0.00	0.29	0.29	0.21	0.21
4	0.00	0.70	0.30	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.50	0.50
5	0.50	0.00	0.00	0.50	0.50	0.50	0.00	0.00	0.00	0.00	0.05	0.95
6	0.25	0.25	0.25	0.25	0.00	0.50	0.50	0.00				
7	0.95	0.00	0.00	0.05	0.00	0.50	0.00	0.50				
8	0.25	0.25	0.25	0.25	0.00	0.50	0.00	0.50				
9	0.70	0.10	0.10	0.10	0.50	0.00	0.50	0.00				
10	0.00	0.50	0.00	0.50	0.00	0.50	0.50	0.00				
11	0.00	0.70	0.00	0.30	0.50	0.50	0.00	0.00				
12	0.70	0.10	0.10	0.10	0.00	0.50	0.50	0.00				
13	0.00	0.50	0.50	0.00	0.00	0.50	0.00	0.50				

Finally, an extensive study of the distributional properties of the ILR is a necessary condition for it to become a widely adopted scoring scheme.

7 METHODS

To test the efficacy of any given motif finding algorithm, N independent sequences of length m were sampled by choosing symbols at random from the four letter DNA alphabet corresponding to an iid model for the background frequency. A position was chosen uniformly at random from each sequence and an instance of a profile Θ, generated as described below, was inserted in that position. Thus, the total length of each sequence is $L = m + w$ where w is the length of the motif. A profile is represented as a position weight matrix, a $4 \times w$ array of numbers where Θ_{ij} denotes the frequency of letter i in column j in all aligned instances of Θ. Since we wanted to have control over the implanted motifs the instance were essentially generated by permuting the columns of the alignment. Each column of the alignment matched the corresponding column of the profile up to discretizing effects.

The parameters N and L were chosen such that the motif finders we considered would have a non-trivial percentage of failures (i.e. datasets where they pick motifs with no overlap with the implants). As we allowed our finders to run for a fairly generous amount of time there is reason to suspect that at least some of those failures can be attributed to twilight zone searches [10], in which random alignments with no overlap with the implants score as high as the best motif that overlaps the implant.

Two of the experiments that we report here were generated according to the following rules:

(1) `COMBO`: The motif in this experiment has length 13 with two degenerate columns (6 and 8) as seen in Table 3. Each dataset has 40 sequences of length 1485 + 13.

(2) `FIFTY`: Each column in the motif consists only of two equally probable nucleotides. Each dataset has 40 sequences of length 1485 + 13.

In each experiment, 400 datasets were generated for a given profile, and various motif finding algorithms were run with parameter settings that allowed each motif finder to take from 8–10 minutes to place all motif finders on an equal footing. However, the MEME motif finder does not employ any parameters that allow the control of running time (MEME generally runs in much less than 8 minutes on each data set), so the generally poor performance of MEME compared to the other motif finders is not a reflection of MEME employing a bad algorithm but a reflection of a design decision to place a strict limit on the total amount of time MEME takes. The motif finders used in this study consisted of MEME [1] (`-mod oops -nmotifs 1 -w 13 -dna -text -maxsize 1000000`), the Gibbs Sampler run in Site Sampler (''Gibbs_ss'') of [11] (`13 -d -n -t280 -L200`), Gibbs altered to use the ILR scoring function (''GibbsILR'', `13 -t 250 -L 200 -p 0.05`), GLAM [5] (`-n50000 -r10 -1 -z -a13 -b13`), CONSENSUS [6] (`-L 13 -c0 -q 3000`), and ProfileBranching [16] (`-l 13 -verbose`). We note that Gibbs_ss is our version of the original algorithm optimized for site sampling mode, resulting in a three-fold improvement in running time. For this reason, the results of Gibbs_ss are better than the results of the original algorithm for a fixed running time. All experiments were run under Red Hat Enterprise Linux 4 on a cluster with nodes that have AMD 248 2Ghz 64-bit processors with 2GB RAM and 1GB swap.

The p-value of the entropy of the highest-scoring reported motif was computed by the sFFT algorithm described in [12]. An estimate of overlap for each data set and for each motif finder was computed in the following manner: Let a_n be the position of the implanted motif instance in S_n, and let $\hat{a_n}$ be the position of the motif reported by a motif finder. We define the *overlap* of a motif finder's prediction as:

$$\max_{|i| \le \frac{w}{2}} \left\{ \frac{w - |i|}{w} \cdot \frac{|\{n : a_n = \hat{a_n} + i\}|}{N} \right\}$$

All ILR scores in this paper were computed using a uniform pseudocount of 0.05.

7.1 Finding the optimal motif

For small datasets it is possible to employ a branch-and-bound algorithm for finding the motif with the optimal entropy (for a more elaborate approach addressing a similar problem see [4]). To see this, consider the space of alignments represented as a tree with the root representing the empty alignment and a node at depth n having $L - w + 1$ children corresponding to the choices of extending the alignment using the $(n + 1)^{\text{th}}$ sequence. A depth-first search (DFS) can then be employed on this tree to enumerate all the alignments at depth N and select the optimal motif. However, since complete enumeration is computationally expensive, even for very small datasets, we rely on pruning the search tree by not extending alignments that cannot possibly score better than the best score (s_{max}) that we have seen till now. This determination is made based on the following lemma:

LEMMA: *Let c_n denote the nucleotide counts for an alignment column of n sequences. Then*

$$\max_{c^N \ge c^n} I(c^N) = \max_{a \in \{A,C,G,T\}} I(c^n + (N - n)\delta_a),$$

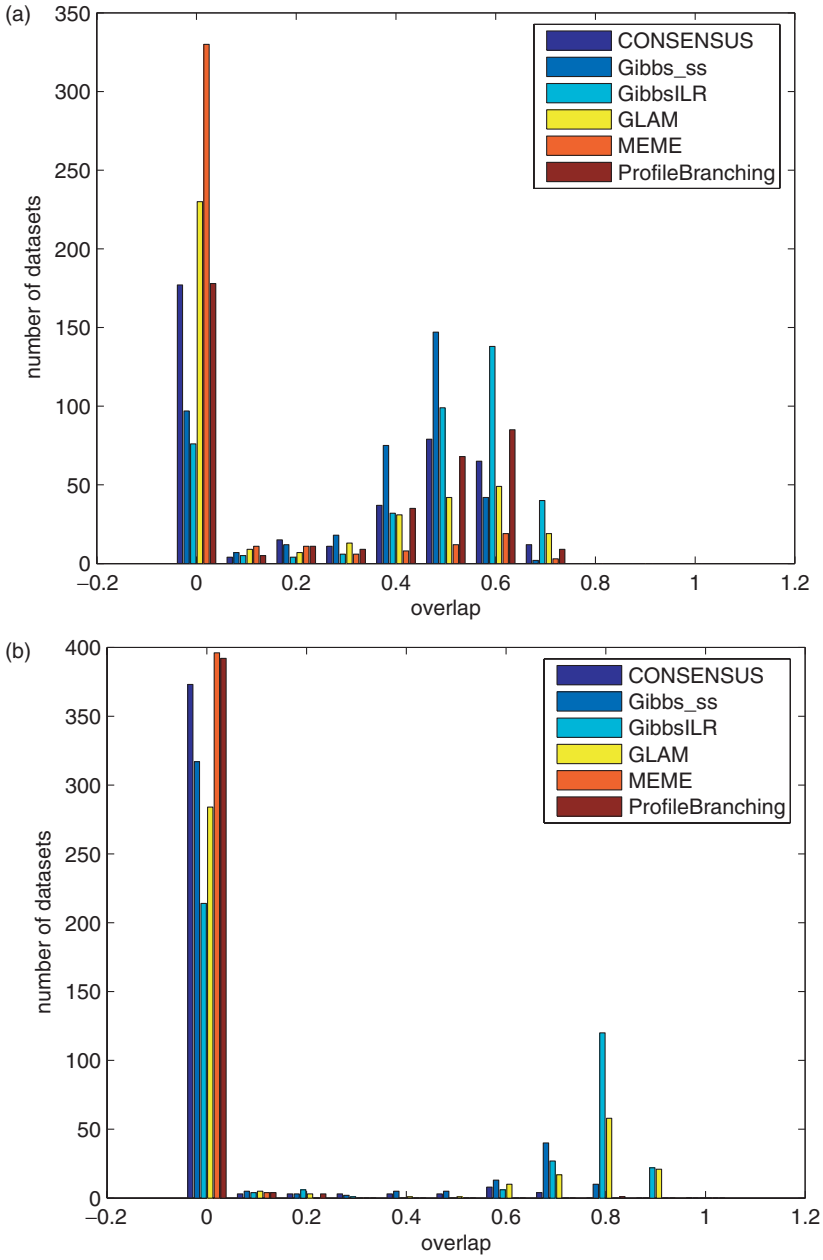

Fig. 5. Histogram of the number of datasets as a function of the amount of overlap with the implanted motif. (a) COMBO experiment (b) FIFTY experiment.

where δ_a *has a count of* 1 *for a and* 0 *otherwise*, $x \geq y$ *denotes pointwise inequalities*, $I_j(x) = x \log (x/\Theta_{0_j})$ *and* $I(c^n) = (\sum_j I_j(c_j^n)) - n \log n$ *is the entropy for a single column.*

PROOF OUTLINE: Suppose that there exists a maximally scored $c^N \geq c^n$ that is not of the form $c^n + (N-n)\delta_a$. Let $j = \text{argmax}_l I_l(c_l^N + 1) - I_l(c_l^N)$ and let $k \neq j$ be such that $c_k^N > c_k^n$ (there must exist such a k by definition of c^N). Then, since $I_l(x)$ is a monotonically increasing function, it can be shown that $I(c^N + \delta_j - \delta_k) > I(c^N)$, giving rise to a contradiction.

In words, the lemma says that an alignment column can be optimally extended in only four different ways and so we can quickly compute the optimal score that can arise out of the extension of a given alignment. In practice, this pruning strategy reduces the search space dramatically and allows us to find optimal motifs for moderate sample sizes (e.g. $w = 7$, $N = 10$ and $L = 100$). Note that we also provide the branch-and-bound algorithm with a good lower bound for s_{max}, as obtained from a motif-finding program such as CONSENSUS, to improve the initial pruning process and thus speed up the algorithm.

ACKNOWLEDGEMENTS

The last two authors are indebted to Pavel Pevzner for getting us hooked on these problems and for many discussions on the subject. This research uses computational resources funded by NIH grant 1S10RR020889. The last author would like to thank

the participants of the Second Barbados Workshop on Genomics and Gene Regulation hosted by McGill University Bellairs Research Institute in April 2005 for many interesting discussions. Finally, we wish to thank the anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] Bailey,T.L. and Elkan,C. (1994) 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*, Menlo Park, California, pp. 28–36.

[2] Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 3, 21–29.

[3] Dempster,A.P. Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B*, 34, 1–38,

[4] Eskin,E. (2004) From profiles to patterns and back again: A branch and bound algorithm for finding near optimal motif profiles. In *Proceedings of the Eight Annual International Conference on Research in Computational Molecular Biology, RECOMB 2004, San Diego, USA*.

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

[6] Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7–8), 563–577.

[7] Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–14.

[8] Jianjun Hu, Bin Li, and Daisuke Kihara. (2005) Limitations and potentials of current motif discovery algorithms. *Nucl. Acids Res.*, 33, 4899–4913.

[9] S. Karlin and S.F. Altschul. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA*, 87(6), 2264–2268

[10] U Keich and PA Pevzner. (2002) Subtle motifs: defining the limits of motif finding algorithms. *Bioinformatics*, 18(10), 1382–1390.

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

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

[13] AF Neuwald, JS Liu, and CE Lawrence. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4(8), 1618–1632.

[14] P.A. Pevzner and S.H. Sze. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol*, 8:269–78.

[15] Prakash,A. and Tompa,M. (2005) Statistics of local multiple alignments. *Bioinformatics*, 21 Suppl 1:i344–50.

[16] Alkes Price, Sriram Ramabhadran, and Pavel A. Pevzner. Finding subtle motifs by branching from sample strings. *Bioinformatics*, 19(90002):149ii–155, 2003.

[17] Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, 16(1), 16–23.

[18] Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

[19] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1), 137–44.