Factoring local sequence composition in motif significance analysis

Patrick Ng, Uri Keich*

Department of Computer Science, Ithaca, NY, USA 14853

We recently introduced a biologically realistic and reliable significance analysis of the output of a popular class of motif finders [16]. In this paper we further improve our significance analysis by incorporating local base composition information. Relying on realistic biological data simulation, as well as on FDR analysis applied to real data, we show that our method is significantly better than the increasingly popular practice of using the normal approximation to estimate the significance of a finder's output. Finally we turn to leveraging our reliable significance analysis to improve the actual motif finding task. Specifically, endowing a variant of the Gibbs Sampler [18] with our improved significance analysis we demonstrate that de novo finders can perform better than has been perceived. Significantly, our new variant outperforms all the finders reviewed in a recently published comprehensive analysis [23] of the Harbison genome-wide binding location data [9]. Interestingly, many of these finders incorporate additional information such as nucleosome positioning and the significance of binding data.

1. Introduction

Much of the recent progress in the area of motif finding can be attributed to leveraging additional pieces of data that are increasingly becoming available. These include quantitative binding assays (p-values) from ChIP-on-chip technology ([5], [31], [11], [8]), phylogenetic ([17] [32] [30] [21]), transcription factor structural class ([24], [22]), and nucleosome positioning information [23]. It has been convincingly demonstrated that finders incorporating such additional information can significantly outperform de novo finders† ([24], [23]). It is therefore somewhat surprising that we can report here on a de novo motif finding tool that outperforms all other finders reviewed in a recently published comprehensive analysis [23] of the Harbison genome-wide binding location data [9]. We stress that many of those finders incorporate additional data as described above suggesting that de novo finders can perform significantly better than has been perceived.

Local base composition has long been taken into consideration in sequence analysis. For example, isochores are taken into account in the GENSCAN gene finding tool [4]. A considerable effort was made into incorporating sequence composition in pairwise local alignment significance analysis (e.g., [1]). Another example is the

*to whom correspondence should be addressed

†A de novo motif finder is one that uses only the given sets and possibly a null reference set.

motif finder NestedMICA incorporating a "mosaic background" model. The latter is a mixture of several, differently parametrized, low order, Markov chains which allow one to factor in local composition [7]. Regardless of whether or not our finder incorporates such mixture models, we argue here that the local composition should be taken into account when analyzing the significance of its results. Intuitively, imagine a set of sequences containing stretches made only from A. In this case a motif such as AAAAAAAA should not be too surprising.

A reliable significance evaluation should be considered an essential component of any motif finder. Indeed, it is often the only information available to the users before they decide on whether to invest significant resources in further exploration or verification of the reported motifs. We recently introduced a reliable method to estimate "confidence" p-values from a small sample of the empirical null distribution of a motif finder's results [16]. In this paper, we naturally extend our confidence p-value approach to incorporate local base composition information. As the original confidence p-value estimate was rather robust and applicable to a wide range of finders and scoring schemes, we expect this extension to be fairly widely applicable as well. We demonstrate the ability of our local composition aware significance evaluation to reliably predict significant motifs in real biological setting.

Our confidence p-values are derived assuming the finder's null score follows a 3-parameter Gamma, or 3-Gamma, distribution[‡] [16]. An often used alternative in this context is to derive the p-value using a point estimator assuming a normal distribution (e.g., [19], [9], [21], [23]). We provide multiple evidence that such an estimation tends to inflate the significance of the reported motif. In particular, using an FDR analysis [3] we show that our p-values are significantly better calibrated than the normal derived ones mentioned above.

Finally, we leverage our significance analysis to improve de novo motif finding. Specifically, we introduce GibbsMarkov, a new variant of the Gibbs Sampler [18], which relies on our p-values to choose between multiple suggested motifs of different widths. The result is a de novo finder that attains the surprising results mentioned above.

2. Factoring local base composition in motif significance analysis

2.1. *Background: 3-Gamma and the finder's null distribution*

In [25] we argue that the finder's null distribution is well suited for estimating the significance of a finder's output. This null distribution is defined as the distribution of the score of the finder on a randomly drawn set, generated for example by resampling a large genomic file. Note that this distribution varies not only with the null model that generates the dataset (including the set's dimensions), but also with

[‡]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 and μ is the location parameter [14].

the parameters of the finder (e.g., width). Since there are typically infinitely many combinations of these problem-parameters (finder and dataset) it is impossible to precompute this distribution.

For any specific set of problem-parameters we can approximate the finder's null distribution with an empirical null distribution. The latter is obtained by applying the motif finder to a sample of randomly drawn null sets. Increasing the sample size improves the quality of our approximation but at a significant cost: each new sample point essentially takes as much running time as the original run whose significance we are trying to estimate. Thus, using this non-parametric approach to reliably estimate small p-values, as we often need to when correcting for multiple hypotheses, is typically forbiddingly expensive (e.g. Harbison dataset has over 300 experiments [9]).

If, however, we know that the finder's null distribution can be well approximated by some parametric family then we only need to estimate these parameters. While the normal distribution is often used in this context ([19] [9] [21] [23]), we find that it consistently offers a relatively poor approximation to the finder's null distribution. In particular, using the normal approximation tends to inflate the significance of high scores which are the ones we are interested in (see Figure 2 below). Instead we find that the 3-parameter Gamma [14], or 3-Gamma for short, appears to fit very well the empirical null distribution for many combinations of motif finders and null models including the biologically realistic, genomic resampling (see Figure 2).

The parameters of the (Gumbel EVD) distribution of the optimal pairwise ungapped local alignment can be computed analytically [15] based on the theory of [6]. In our case the problem is complicated further by the dependence on the finder: our null distribution is of the *finder's* optimal score rather than the optimal alignment score [25]. Thus, it remains a challenging open problem whether a theory can be developed to estimate the parameters of the 3-Gamma from those of the problem. In the meantime we can resort to parametric statistical estimation. For example, suppose we want to estimate the p-value of the observed score s, denoted by $p(s)$. We can generate a *small* sample $X = (X_1, \ldots, X_n)$ from the finder's null distribution and find the 3-Gamma MLE (maximum likelihood estimator) $\hat{\theta} = \hat{\theta}(X)$. We can then find the MLE of $p(s)$, $\hat{p}(s) = \hat{p}(s, X)$, by using the popular plug-in method: $\hat{p}(s) = 1 - F_{\hat{\theta}}(s)$, where F_θ is the 3-Gamma CDF (cumulative distribution function).

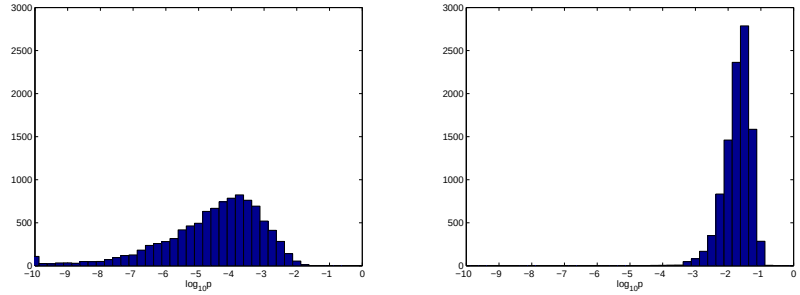
As noted in [16] for a realistically small sample size such as $n = 20$[§], $\hat{p}(s)$ can grossly over-estimate the significance of the observed score s. This type of MLE estimation, albeit using the normal approximation, is used in ([19] [9] [21] [23]). We suspect that it further inflated the significance of the observed scores beyond that due to the selection of the normal approximation (see Figure 1 and Section 4.2 for evidence).

Our conservative "confidence p-value", $\hat{p}_c(s, X)$, presented in [16] corrects the

[§]A sample of size n increases the runtime by a factor of n.

tendency of the point estimator $\hat{p}(s)$ to over-estimate the 3-Gamma p-value, $p(s)$. It does so by constructing a confidence interval for the estimated $p(s)$. In principle, the confidence p-value can be applied whenever the 3-Gamma distribution is expected to offer a reasonably good fit to the finder's null distribution.

Fig. 1: Comparing the estimators $\hat{p}_n$ and $\hat{p}_c(s, X)$ of p-value$= 10^{-3}$



(a) $\hat{p}_n$ overestimates the significance (b) $\hat{p}_c(s, X)$ is mostly conservative

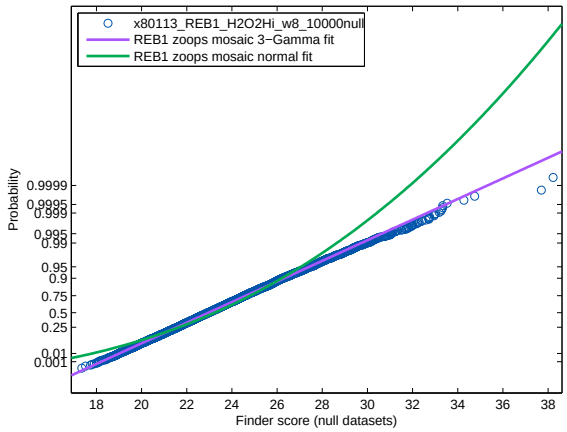
Histograms of 10^4 independent evaluations of the point estimator $\hat{p}_n(s)$and of the conservative $\hat{p}_c(s, X)$, where s was set to the empirical 0.999 quantile. $\hat{p}_n$ is the MLE plug-in estimator of the p-value assuming a normal approximation, and $\hat{p}_c(s, X)$ is our conservative "confidence p-value" assuming a 3-Gamma distribution. The quantile s was learned from the scores of GibbsMarkov on 10,000 resampled sets of 30 sequences each of length 1,000. The resampling was done from the human genomic file. This set of null scores was then used to create the 10,000 resamples X of size $n = 20$ drawn with repetitions. An ideal estimator of $p(s)$ should have all the mass concentrated on the point -3 because s was set to the 0.999 quantile. It is clear from the graphs that $\hat{p}_n$ has a considerably larger variance than $\hat{p}_c$ and that it can badly over-estimate the significance of the score s. GibbsMarkov was run in OOPS mode with the parameters `-l 23 -gibbsamp -best_ent -t 170 -L 100 -em 0 -markov 3 -p 0.10`. Statistical estimations were done in R [27].

2.2. *Incorporating local GC content in our confidence p-value*

We can factor local, or any other, composition information in our significance analysis in a rather straightforward manner. In principle, all we need to do is to condition our generated random sets on the relevant set of constraints. If the null distribution of the finder's score on these conditioned sets can be well approximated by the 3-Gamma distribution, then our confidence p-value method should be valid. Having no theory that could justify this approximation we resort to the empirical studies as we previously did. Indeed, we can simply think of our conditional generating model described below as just another null set generator. Figure 2 below compares the normal with the 3-Gamma approximation of such a conditional empirical null distribution.

Technically, our local GC-content adjusted resampling is done as follows. We first divide our genomic reference file into partially overlapping windows of a fixed size L (overlap size is $L/2$). We then place each window in one of K bins that uniformly cover the entire spectrum of GC-content. This preprocessing step need only be done once. Given an input set we generate local GC-content adjusted resampled images of it as follows. We first divide each sequence into non-overlapping windows of size L and determine their GC-content. We then replace each of the original windows with a randomly drawn genomic window from the appropriate bin. Note that within

Fig. 2: Approximating a finder's null distribution conditioned on local GC-content



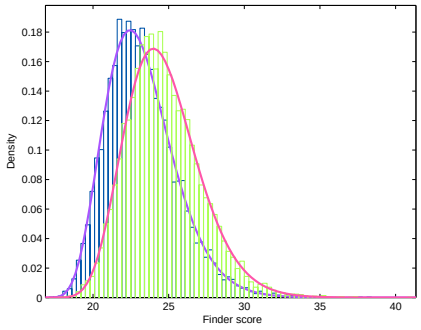
The figure demonstrates the difference between the quality of the normal and the 3-Gamma approximations to a finder's null distribution. In this example, GibbsMarkov was applied to 10,000 sets of GC-content adjusted resampled sequences ($L = 100, K = 20$). The sequences were resampled from the *S. cerevisiae* intergenic file. The mold, or input, set was the Harbison REB1_H2O2Hi dataset consisting of 48 sequences of average length 431bp [9]. The 3-Gamma seems to offer a reasonably good fit for this conditional null distribution while the normal does not. GibbsMarkov was run in ZOOPS mode with the parameters `-l 8 -gibbsamp -p 0.05 -best_ent -cput 300 -L 200 -em 0 -markov 5 -r 1 -ds -zoops 0.2`

a set we draw windows without replacement as repetitive elements can wreak havoc on motif finding. For the same reason we exclude overlapping windows within a set. The same kind of exclusion applies to our "uniform" resampling strategy.

Does factoring local GC content make a difference in the significance analysis? We give two different types of evidence that it does. First, Figure 3 compares histograms of our GibbsMarkov run on null sets that were generated according to the two models we are comparing. One model was generating sets using uniform resampling of a *S. cerevisiae* intergenic file while the other was using the local GC content framework described above. Notice that the two histograms are distinctly different. For example, a score whose p-value, when factoring in local GC content, is 0.0002 has a p-value of only 0.001 when assuming the uniform model.

As we just saw, taking into account the local GC-content can considerably impact the significance of an observed score s. Our original construction of the confidence p-value [16] did not account even for the *global* base composition of the sample as outlined above. Indeed we followed the common procedure of resampling a relevant genomic file. To demonstrate the potential difference between such a naive approach and our local GC-content adjusted one we devised the following experiments. This experiment is realistic in the sense that it emulates a real problem we encountered when analyzing DNA replication origins in *Saccharomyces kluyveri*. We first generated 200 random datasets by resampling from our human genomic file (see Section 5). To make these sequences look closer to the *S. kluyveri* sequences we were analyzing, we accepted only sequences whose AT-content is above 65%. We then implanted in each sequence exactly one site generated from the *Saccharomyces*

Fig. 3: Comparing the uniform and the local composition aware null generators



The data for "right" histogram was generated by applying GibbsMarkov to 10,000 sets that were resampled uniformly from the *S. cerevisiae* intergenic file. The "left" histogram was generated using the same local GC-content preserving scheme as described for Figure 2. To highlight the difference both histogram were ML-fitted with a 3-Gamma distribution. GibbsMarkov was run in ZOOPS mode with the parameters `-l 8 -gibbsamp -p 0.05 -best_ent -cput 300 -L 200 -em 0 -markov 5 -r 1 -ds -zoops 0.2`

cerevisiae AT-rich ACS profile (see Figure 4)¶. We next ran our GibbsMarkov in OOPS mode on each of these 200 datasets, and noted the score, as well as whether or not the finder succeeded in uncovering the implanted ACS motif. Finally, we computed confidence p-values for each of these 200 scores in two different ways. The first was derived from our previous approach of uniform genomic resampling‖. The second was derived from the new local GC-content preserving resampling scheme. Table 1 summarizes the results. Notably, the latter identifies 50% more TPs. The FPs are under control in both cases as expected.

Fig. 4: The ACS motif

3. A hybrid Gibbs Sampler

By GibbsMarkov we refer here to our variant of a Gibbs Sampler finder [18]. Currently it handles an OOPS (one occurrence per sequence) or a ZOOPS (zero or one) model [2]. Its scoring function and sampling steps follow the techniques developed by Jun Liu and his colleagues in [20] and [13]. There are a couple of distinctions between the original work of Liu et al. and our implementation. First, neither of the above papers specifically addresses the ZOOPS model. Second, Liu and his colleagues use a complete Bayesian framework which includes a prior on the matrices. Instead, we use a hybrid model which incorporates a prior on the percentage of

¶The ACS is a 17bp site to which the *S. cerevisiae* ORC (origin recognition complex) binds to initiate local chromosomal replication [28]. We expect its *S. kluyveri* analogue to be somewhat similar.

‖For technical reasons we used the same human genomic file which has roughly the same AT-level as that of *S. kluyveri*.

Table 1: The effect of base composition on significance analysis

p-value threshold	TP	TN	FP	FN
0.1	26/49	78/77	0/1	96/73
0.05	21/33	78/78	0/0	101/89

The first number in each entry is the number of sets (out of 200) for which the p-value derived from sets generated by a uniform genomic resampling (57% AT-content). The second number is for the locally adjusted p-value. Notably, the latter identifies 50% more TPs. The overall high number of FNs is partly due to the conservative nature of the confidence p-value and partly due to the fact that these sets were designed as twilight zone ones [25].
Each of the 200 implanted sets consists of 30 sequences of length 2500 resampled from the human genomic file conditional on having an AT-content $\geq$ 65%. Each sequence was implanted with exactly one site generated by drawing from the ACS matrix. This ACS matrix (Figure 4) was generated by us from a compiled list of confirmed ARSs on OriDB [26]. GibbsMarkov was run in OOPS mode with the parameters `-l 17 -gibbsamp -p 0.05 -best_ent -cput 300 -L 200 -em 0 -markov 3 -r 1 -s 123`. The confidence p-values were derived from sets resampled in two different ways. Both resampled from our human genomic file but one conditioned the resampling on the local GC-content observed in the input dataset. Note that each one of these 200 input sets had a different local GC-content pattern.

sequences that include sites but we use a maximum likelihood approach for the matrix. While the latter is fairly similar to using the Stirling approximation to the full Bayesian model [13], it is not exactly the same. The ZOOPS model is specifically used in [24] and [23] but, again, there are some differences between the functions optimized there and ours**. A detailed account of GibbsMarkov's sampling step and scoring function will be described in another paper.

4. Results on the Harbison dataset

All the tests below refer to the Harbison dataset of 310 ChIP-chip, genome-wide location analysis, experiments of 203 yeast transcription factors [9]. By the "Narlikar test" we refer to the dataset consisting of the 156 sequence-sets from 80 TFs used in [23]. The literature consensus for each of these 80 TFs is published. We obtained these from [9], with the exception of DAL82, RTG1, and the modified CIN5 which we took from [21]. By the "MacIsaac test" we refer to the dataset consisting of 188 sequence-sets which include all 124 TFs whose matrices are reported in [21]. See more details in Section 5. In the following analysis our confidence p-values factor in the local GC-content as described in Section 2.2.

4.1. *GibbsMarkov performance on the Narlikar test*

We compared our motif finder GibbsMarkov with results from Table S1 in [23]. GibbsMarkov with fixed width $w = 8$ was run on the 156 sequence-sets. Using the same definition of success as defined in [23], GibbsMarkov successfully finds the correct motif in 71 of the 156 experiments. This is better than all other finders although PRIORITY-DN [23] which uses nucleosome positioning information is a close second with 70 successes. The next best *de novo* finder is PRIORITY-N [23] with 51 successes. The full list which includes many more finders can be found in [23].

**Our target function is different than theirs even in the case of uninformative prior they consider.

4.2. *How well calibrated are these p-values?*

If our p-values are well calibrated then the false discovery rate for any given threshold should be consistent with the rate guaranteed by the theory. To test that we applied the original FDR test [3] to find our p-value cutoff corresponding to an FDR of 5%. We applied this test separately to the p-values we assign to the 156 sets of the Narlikar test and then to the p-values we assign to the 188 sets of the MacIsaac test.

In order to get an accurate classification of predicted motifs, we disregarded motifs where (1) the consensus sequence of the predicted motif is `AC`-repeat or `GT`-repeat, and (2) the predicted motif does not match the literature motif but has a statisically significant match to a motif in the MacIsaac set of motifs [21] (see Section 5 for details). Type (1) motifs which we found in ACE2_YPD, AFT2_H2O2Hi, ARR1_YPD, and SWI5_YPD were disregarded because `GT`-repeats are possibly functional in yeast ([8], [12]). Type (2) motifs were disregarded because TFs often have co-factors that are DNA-binding. Such detected motifs should therefore not be considered false positives as they could still be biologically relevant.

At a 5% FDR threshold, for the MacIsaac test, our observed FDR comes at about 6.67%: 4/60, while for the Narlikar test it is about 7.41%: 4/54. At a 10% threshold, the observed FDR of the MacIsaac test and Narlikar test were 11.4% and 10.2%, respectively. Hence it is reasonable to conclude that our confidence p-values are well calibrated.

We also looked at the observed FDR of the results of [23] which are based on the normal MLE of the p-value. Their results were already disregarding the `GT`-repeats (type (1) from above), but we could not disregard possible type (2) motifs because we do not have access to their predicted motifs. At the 5% threshold their observed FDR on the Narlikar test is about 48%: 63/132, which is significantly higher than the expected 5%. For comparsion, we repeated the FDR analysis on our confidence p-value by disregarding only the `GT`-repeats so that the comparison was on equal footing. At that 5% threshold, our observed FDR comes to about 12%: 7/57.

4.3. *Using the p-values to improve our results*

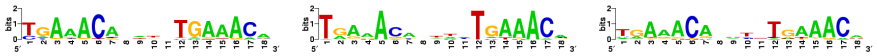
GibbsMarkov was run with multiple widths on the 156 sets of the Narlikar test, and a single predicted motif among the multiple widths was selected based on our confidence p-values. In the Narlikar test, our results improved from 71 successes with $w = 8$ to 76 with multiple widths. The improvement was more significant in the MacIsaac test: the multiple widths method correctly identified 114 motifs while GibbsMarkov using $w = 8$ found only 97 out of 188 sequence-sets.

To test our performance of using confidence p-values for multiple widths selection, we compare it against naively selecting widths according to average entropy. Thus instead of choosing a predicted motif among widths with the best confidence p-value, a prediction is chosen based on average entropy, which is simply the entropy score averaged over the width of a motif. In the MacIsaac test, width selection based

on average entropy found 99 while selection based on confidence p-values found 114 as reported above.

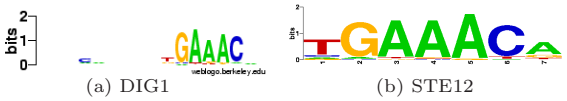
We have yet to throughly explore our predictions but one interesting dimer of width 18 caught our eyes. It appears essentially the same in three different experiments: DIG1 Alpha, TEC1 Alpha, and STE12 Alpha (see Figure 5). In all three cases width 18 exhibits the most significant p-value at: 3.7e-15, 1.3e-04, and 7.2e-08 respectively. A closer inspection shows the dimer is made of a repetition of the known motif common to DIG1 and STE12 (see Figure 6). This dimer was recently independently reported in [12].

Fig. 5: An interesting dimer picked up by GibbsMarkov



(a) DIG1 Alpha (b) TEC1 Alpha (c) STE12 Alpha

Fig. 6: Known motifs from [21]



(a) DIG1 (b) STE12

5. Methods

5.1. *Confidence p-value*

All confidence p-values were computed in R [27] using functions described in [16]. The necessary samples were derived from resampled data generated as described in the text.

5.2. *Genomic files*

For historical reasons we used two genomic files for resampling purposes. In both cases resampling was done by extracting contiguous sequences from a concatenated filtered genomic sequence. The "human genomic" contiguous sequence is from *Homo sapiens* chromosome 1 (HSA1). HSA1 was downloaded from the Ensembl Genome Browser v38 (NCBI build 36) [10]. RepeatMasker, TandemRepeatFinder, and DUST were applied to the data. The *S. cerevisiae* intergenic file was generated by removing from the *S. cerevisiae* genome downloaded from SGD [29] all protein and RNA coding sequences including tRNA, rRNA, snoRNA, snRNA, LTR, and other repetitive sequences.

5.3. *Is the predicted motif a known motif?*

Given a database of known motifs, we would like to determine whether a predicted motif has a statistically significant match to a known motif. For each predicted motif, we first obtained an empirical null distribution of maximal similarity scores

(a higher score implies more similar motifs). Each score from this null is the maximal similarity score over all database PFMs against a random permutation of positions/columns of the predicted motif. Then the p-value for similarity is simply estimated from the null distribution described above and the similarity score between the predicted motif and its most similar motif within the database. Note that this technique accounts for evaluating statistical significance at the extreme value case of choosing the most similar motif within the database. In our FDR analysis, the empirical null of each predicted motifs was generated with 10,000 randomly permuted motifs as described above and ignored cases where the predicted motif does not match the literature but has a p-value < 0.05 for similarity.

5.4. *Harbison dataset*

All the consensus sequences were converted to PFM by the same method as [9]. For the MacIsaac tests, we used the same definition of success as defined in [9]. Likewise, we used the definition of success defined in [23] for the Narlikar test with fixed width $w = 8$. For the Narlikar test with *multiple widths*, we slightly modified the average entropy constraint of inter-motif distance used in [23]. The average entropy of the predicted motif was taken over corresponding non-N positions of the literature consensus within an alignment, because predicted motifs such as GAL4 with literature consensus `CGGnnnnnnnnnnnCCG` should not be penalized for having degenerate positions at consensus positions with `n`.

GibbsMarkov was run with a fifth-order Markovian background estimated from the *S. cerevisiae* intergenic file. The strength of prior parameter in ZOOPS is $\alpha = 0.2$. The finder was allowed to run for 5 minutes with a plateau period of 200 iterations. 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 confidence p-values were computed from applying GibbsMarkov to 50 sequence-sets of local GC-content adjusted resampled sequences ($L = 100, K = 20$). For GibbsMarkov with multiple widths selection, GibbsMarkov parameterized with widths 8, 12, 15, and 18 were run separately on the input sequence-set, and then each were applied separately on the same 50 sequence-sets of local GC-content adjusted resampled sequences.

6. Conclusion & Future Research

We show that incorporating local base composition can improve the fidelity of our recently published confidence p-value method of estimating the significance of a finder's output [16]. We also demonstrate the practical advantage of this improvement over the previous method in identifying true motifs in a realistic experiment. We give evidence that the practice of using a normal approximation to estimate the significance of a finder's output is ill-advised on two counts. First, the normal distribution generally fits the finder's null distribution rather poorly. Second, the normal MLE point estimator of the p-value has a significant bias toward over-estimating

the significance of the observed score. To drive home this point we show that the use of this p-value on a real biological dataset creates a FDR which is significantly higher than the stated one. In contrast, a FDR analysis based on our confidence p-value is much closer to the declared rate. Our evaluation method is based on the validity of the 3-Gamma approximation of the finder's null distribution. As such, it is likely to be applicable to many more finders than the ones explored here.

We also develop GibbsMarkov, a variant of the Gibbs Sampler de novo motif finder. GibbsMarkov outperforms all the finders reviewed in a recent well designed study [23] of the Harbison genome-wide location analysis data [9]. Surprisingly, many of the finders that GibbsMarkov outperforms rely on additional information such as, the confidence of the binding, phylogenetic, and nucleosome positioning information [23]. Moreover, when we choose the best p-value among several GibbsMarkov runs using different widths, we get a roughly 10% increase in our TP rate.

As far as future issues, we could benefit from a more sophisticated alternative to the window based method that we currently use to track the local GC-content. HMM models naturally fit in this context. Regardless, note that, in principle, our method can be extended to factor any local composition feature that the user might be interested in accounting for. Eventually it all boils down to two things: do we have sufficient data to generate random sets that satisfy the required local conditions and is the associated finder's null distribution well approximated by the 3-Gamma distribution.

Acknowledgement

It is our pleasure to acknowledge Anand Bhaskar for processing the list of MacIsaac matrices. This research uses computational resources funded by NIH grant 1S10RR020889 and is supported by the National Science Foundation Grant No. 0644136 to UK.

References

[1] SF Altschul, et al. Protein database searches using compositionally adjusted substitution matrices. *FEBS J*, 272(20):5101–5109, Oct 2005.

[2] T Bailey and C Elkan. The value of prior knowledge in discovering motifs with meme. In *Proceedings of the Third ISMB*, pages 21–29, Menlo Park, California, 1995.

[3] Y Benjamini and Y Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS B (Methodological)*, 57(1):289–300, 1995.

[4] C Burge and S Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, Apr 1997.

[5] H Bussemaker, H Li, and E Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, Feb 2001.

[6] A Dembo, S Karlin, and O Zeitouni. Limit distribution of maximal non-aligned twosequence segmental score, 1994.

[7] TA Down and TJP Hubbard. Nestedmica: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, 33(5):1445–1453, 2005.

[8] E Eden, D Lipson, S Yogev, and Z Yakhini. Discovering motifs in ranked lists of dna sequences. *PLoS Comput Biol*, 3(3):e39, Mar 2007.

[9] CT Harbison et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.
[10] E Birney et al. Ensembl 2006. *Nucleic Acids Res*, 34:D556–61, Jan 2006.
[11] BC Foat, AV Morozov, and HJ Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–e149, Jul 2006.
[12] N Habib, T Kaplan, H Margalit, and N Friedman. A novel bayesian dna motif comparison method for clustering and retrieval. *PLoS Comput Biol*, 4(2), Feb 2008.
[13] S Jensen, X Liu, Q Zhou, and J Liu. Computational discovery of gene regulatory binding motifs: a bayesian perspective. *Statistical Science*, 19(1):188–204, 2004.
[14] NL Johnson, S Kotz, and N Balakrishnan. *Continuous Univariate Distributions, 2nd edition*. Wiley Series in Probability and Statistics, 1994.
[15] S Karlin and S Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS*, 87(6):2264–8, Mar 1990.
[16] U Keich and P Ng. A conservative parametric approach to motif significance analysis. In *The 18th International Conference on Genome Informatics*, Singapore, 2007.
[17] M Kellis et al. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.
[18] C Lawrence, et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, Oct 1993.
[19] X Liu, DL Brutlag, and JS Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes *PSB*, 127–38, 2001.
[20] JS Liu, A Neuwald, and C Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Amer. Stat. Assoc.*, 90:1156–1169, 1995.
[21] KD Macisaac, et al. An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1), March 2006.
[22] AV Morozov and ED Siggia. Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci U S A*, 104(17):7068–7073, Apr 2007.
[23] L Narlikar, R Gordân, and AJ Hartemink. Nucleosome occupancy information improves *e novo* motif discovery. In *RECOMB*, pages 107–121, 2007.
[24] L Narlikar, R Gordân, U Ohler, and AJ Hartemink. Informative priors based on transcription factor structural class improve *e novo* motif discovery. In *ISMB (Supplement of Bioinformatics)*, pages 384–392, 2006.
[25] P Ng, N Nagarajan, N Jones, and U 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.
[26] CA Nieduszynski, et al. Oridb: a dna replication origin database. *Nucleic Acids Res*, 35:D40–D46, Jan 2007.
[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.
[28] RA Sclafani and TM Holzen. Cell cycle regulation of dna replication. *Annu Rev Genet*, 41:237–280, 2007.
[29] SGD project. Saccharomyces genome database. http://www.yeastgenome.org/.
[30] R Siddharthan, ED Siggia, and EV Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, Dec 2005.
[31] A Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16(8):962–72, Aug 2006.
[32] T Wang and GD Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–80, Dec 2003.