GIMSAN: a Gibbs motif finder with significance analysis

Patrick Ng and Uri Keich*

Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

Received on June 16, 2008; revised on July 25, 2008; accepted on July 29, 2008
Advance Access publication August 14, 2008

Associate Editor: Limsoon Wong

1 INTRODUCTION

A reliable significance evaluation should be considered as 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 demonstrated inherent flaws in the significance analysis based on the E-value of the information content (Ng et al., 2006) as well as on the empirical normal approximation (Ng and Keich, 2008). In contrast, we recently introduced a biologically realistic and reliable method to estimate the reported motif’s statistical significance based on a novel 3-Gamma approximation scheme (Keich and Ng, 2007). Exploiting the robustness of this technique, we showed how we can further improve its reliability by factoring in local GC content (Ng and Keich, 2008).

Here, we present a novel de novo motif finding tool called GIMSAN (GibbsMarkov with Significance ANalysis). GIMSAN combines GibbsMarkov, our variant of the Gibbs Sampler, described here for the first time, with our recently introduced significance analysis.

Availability: GIMSAN is currently available as a web application and a stand-alone application on Unix and PBS (Portable Batch System) cluster through links from http://www.cs.cornell.edu/∼keich.

Contact: keich@cs.cornell.edu

2 METHODS

2.1 3-Gamma-based P-values

GIMSAN reports two figures that indicate the significance of the reported motif as outlined next. Based on the user selected reference set, GIMSAN generates null sets of sequences that preserve the dimensions and local GC content of the input set. It then runs GibbsMarkov with the user selected parameters on these null sets thereby creating a small sample of the finder’s null distribution. Assuming this sample comes from a 3-Gamma distribution, GIMSAN reports a maximum likelihood point estimator of the P-value of the reported motif. Since the latter can significantly over-estimate the significance of the motif, GIMSAN augments it with a, roughly, 95% confidence interval of the P-value of the motif. For more details see Keich and Ng (2007) and Ng and Keich (2008).

2.2 Hybrid Gibbs sampler

By GibbsMarkov we refer here to our variant of a Gibbs Sampler finder (Lawrence et al., 1993). Currently it handles an OOPS (one occurrence per sequence) or a ZOOPS (zero or one occurrence per sequence) model (Bailey and Elkan, 1995). Its scoring function and sampling steps follow the techniques presented in Liu et al. (1995) and Jensen et al. (2004). There are a couple of distinctions between these works and our implementation which merits the following presentation. First, neither of the above papers specifically addresses the ZOOPS model described here. Second, these papers 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 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 (Jensen et al., 2004), it is not exactly the same. The ZOOPS model is specifically used in Narlikar et al. (2007) but, again, there are some differences between the functions optimized there and ours.2

Our generative probabilistic ZOOPS model of the input set is defined as follows. The given input to the model is: the number of sequences N, their lengths li, the (Markov) background model B and the motif modeled by a 4 x w PSFM (position-specific frequency matrix) Θ. We denote by p the probability that sequence Si contains a site. We determine p by randomly drawing from a prior p(a, b)-distribution. In practice, we choose a = b = αN where α is a parameter that reflects the strength of your prior. Note that this choice indicates our prior belief that on average half the sequences should amounts to multiple testing and a necessary correction should be employed by the user.

The performance of GIMSAN on real biological data is demonstrated in (Ng and Keich, 2008).

*To whom correspondence should be addressed.

1A sequence logo is generated using the popular WebLogo Crooks et al. (2004).
contain a site and it can readily be changed. We next draw \( N \) independent samples \( \{Z_i\}_{i=1}^N \) from a Bernoulli(p) distribution. Each \( Z_i \) is the indicator function of the event ‘sequence \( i \) contains a site’. Sequences \( i \) for which \( Z_i = 0 \) are not containing sites. Therefore, they are generated according to a Bernoulli distribution. Alternatively, if \( Z_i = 1 \) we first choose \( Y_i \), the site location for sequence \( i \), uniformly from \( \{1, 2, \ldots, l-w+1\} \). We then generate the two background pieces \( S_i[Y_{i-1},l-w] \) and \( S_i[l-w+1,l] \) independently and according to the background model \( B \). The site itself, \( S_i[Y_{i-1},l-w+1] \) is generated according to the product of multinomials parametrized by the PSFM \( \Theta \): \( \prod_{i=1}^N \Theta_{0,j} \), where \( \alpha(j) := S_i[Y_{i-1}] \). Below we sloppily refer to the product of these last three probabilities/likelihoods as \( P_{B,\Theta}(S' | Y_i) \).

The score we are trying to optimize is the model’s joint likelihood:

\[
P_{B,\Theta}(S, Z, Y) = \frac{1}{\beta(a,b)} \prod_{i=1}^{w-1} \frac{(1-p)^a-p^{\sum Z_i}}{1-p} \prod_{i=0}^{N-\sum Z_i} P_B(S_i) dp
\]

\[
= \beta\left(\sum Z_i+a, N-\sum Z_i+a\right) \prod_{i=1}^{l-w+1} \frac{P_{B,\Theta}(S_i | Y_i)}{P_B(S_i)} \prod_{i=0}^{l-w+1} \frac{P_B(S_i)}{P_B(S_i) \cdot (l-i-w+1)}
\]

where \( \beta(a,b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx \) is the beta function. Specifically, we view \( Z \) and \( Y \) as missing parameters and we try to find:

\[
\arg\max_{a,b,Z,Y} P_{B,\Theta}(S, Z, Y) = \arg\max_{a,b,Z,Y} \arg\max_{a,b} P_{B,\Theta}(S, Z, Y).
\]

Note that since the sites are generated according to a product of multinomials, maximizing over \( \Theta \) given \( Z \) and \( Y \) is the standard multinomial MLE (maximum likelihood estimation) deduced from the sites’ letter counts.

It is convenient to divide by the constant \( \frac{1}{\beta(a,b)} \prod_{i=0}^{l-w+1} P_B(S_i) \) so that our target function simplifies to:

\[
\Psi(Z, Y, \Theta | S) = \beta\left(\sum Z_i+a, N-\sum Z_i+a\right) \prod_{i=1}^{l-w+1} \frac{P_{B,\Theta}(S_i | Y_i)}{P_B(S_i)} \prod_{i=0}^{l-w+1} \frac{P_B(S_i)}{P_B(S_i) \cdot (l-i-w+1)}
\]

As in the original work of Lawrence et al. (1993), our sampler iteratively resamples the site location, one sequence at a time. When resampling the location, \( \Theta \) is estimated from the sites chosen in the rest of the sequences.

We resample \( Y_i \) with probabilities proportional to:

\[
P(Y_i=j) \propto \beta\left(\sum Z_i+a, N-\sum Z_i+a\right) \prod_{i=1}^{l-w+1} \frac{P_{B,\Theta}(S_i | Y_i=j)}{P_B(S_i) \cdot (l-i-w+1)}
\]

for \( j \in \{1, 2, \ldots, l-w+1\} \), where \( \sum Z_i = \sum_j Z_i \). We allow for \( Z_i = 0 \) which, following the convention mentioned in Narlikar et al. (2007), we denote by \( Y_i = 0 \). Therefore, with the same proportionality constant as above:

\[
P(Y_i=0) \propto \beta\left(\sum Z_i+a, N-\sum Z_i+a\right)
\]

ACKNOWLEDGEMENTS

We thank Robert Bukowski for deploying GIMSAN as a web application. Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University which is partially funded by Microsoft Corporation. Finally, we thank anonymous reviewers for very constructive comments.

Funding: National Science Foundation (0644136).

Conflict of Interest: none declared.

REFERENCES


