Transcription factor binding site characterisation by Bayesian network model averaging

Ankit Agrawal, Rajarshi Pal, Rahul Siddharthan*
The Institute of Mathematical Sciences, Chennai 600113, India
06 September 2014
*rsidd@imsc.res.in

Background

Transcription factor binding sites (TFBS) are loci where transcription factors (TFs), that is, proteins that regulate gene transcription, bind to DNA, typically upstream (sometimes quite distant) but often in introns or downstream.
The binding sites are weakly conserved and the most common description for them is position weight matrix (PWM), which is a $4 \times L$ matrices $W_{\alpha n}$ describing the likelihood of observing a nucleotide α (= A, C, G or T) at a position n within the binding sequence.

The PWM is the descriptive format used by standard databases such as JASPAR and Transfac, and *ab initio* motif finders such as MEME and various implementations of the Gibbs sampler also assume this description for TFBS.
PWMs are easily visualisable via a "sequence logo". This representation of TFBS motifs has become nearly ubiquitous in the literature.

Limitations of the PWM representation

It assumes independence of the columns and cannot take account of correlations among the positions. For example, if binding sites for a factor consistently contain the dinucleotide pattern AA or TT but rarely AT or TA, a PWM will be inadequate to describe this. Such is the case, for example, with nucleosome binding regions, but also with several transcription factors in various species.
Possibly because of these limitations, PWMs are poor predictors of binding sites *in silico*, offering numerous false positives as well as false negatives depending on the parameters used.

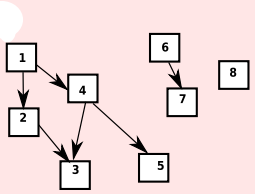
Over the years several alternatives to the PWM approach have been explored, such as Markov chains[1], biophysical approaches[2], feature-based approaches[3], and "TF flexible models"[4]. We previously described[5] a generalization of PWMs to take account of all pairs of dinucleotides – a brute-force approach that was effective in predicting binding sites in yeast.

Yet PWMs dominate

None of the above approaches has become widespread, and PWMs dominate. With the wide availability of high-quality high-throughput binding data, a better descriptive model should be possible!

Our approach

- We describe a description of TFBS as Bayesian networks, that is, directed acyclic graphs with links specifying conditional probabilities.
- The nodes in the network are individual positions in the binding sites.
- The network is sparse, often not fully connected (ie, may consist of multiple disjoint subnetworks), and several nodes may not be connected at all, in which case their contribution to the likelihood is the same as in the PWM model.
- The probability distributions for nucleotides at these nodes, as well as conditional probabilities for nucleotides at ends of links, are estimated by assuming a multinomial distribution with a Dirichlet prior.

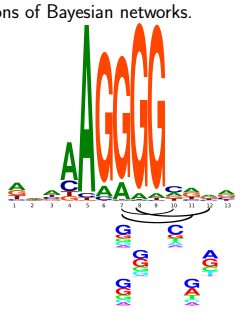


The network specifies a joint probability: in the figure, for example, the likelihood for a sequence $S = S_1 S_2 \ldots S_8$ given the network above would be

$$P(S) = P(S_1)P(S_2|S_1)P(S_4|S_1)P(S_3|S_2,S_4)P(S_5|S_4)P(S_6)P(S_7|S_6)P(S_8)$$

To each such network can be assigned a posterior probability given the data, and we average over possible networks.

Bayesian network descriptions have been previously attempted, by Barash *et al.*[6] and Ben-Gal *et al.*[7]. The chief difference in our approach is that while Barash *et al.* considered a single Bayesian network, and Ben-Gal *et al.* use "variable-order" Bayesian networks, we use a model-averaging approach. Barash *et al.* also considered collections of tree-like networks (each node has only one parent) for their computational tractability. But neither collections of tree networks, nor single Bayesian networks, are as general as model-averaging over collections of Bayesian networks.



Method

- We explore the space of possible Bayesian networks by Markov chain Monte-Carlo (MCMC) sampling on training data.
- The networks are sampled on their posterior probability given the training data.
- To avoid overfitting, this posterior is modified by a penalty per link, chosen to compensate for the expected standard deviation in the binomial distribution for a pair of nucleotides.
- A set of candidate networks is obtained by sampling, and then used to calculate the likelihoods of new sequences, each network being weighted by its prior.

onsider prior data D consisting of a set of known binding sites. Consider a putative Bayesian network G that specifies a model for these sites. G consists of L nodes, where L is the length of the binding site; and a set of links among the nodes, specifying conditional probabilities, with the requirement that the links form no cycles. This specifies a factorisation of the joint probability distribution.
We calculate $P(D|G)$, the likelihood of the data given the network, by assuming a multinomial distribution on the various probabilities. For the purposes of sampling and model averaging, one is interested in $P(G|D)$, which is proportional to $P(D|G)P(G)$. The prior $P(G)$ uses link penalties as described above, "strength" chosen to avoid finding spurious correlations in random test data of 40 or more sequences.
Given a new sequence S, its likelihood under the binding site hypothesis is

$$P(S|D) = \sum_G P(S|G)P(G|D)$$

. This is compared with the "background" (non-functional) hypothesis.

Results: Test data

On test data generated from a Bayesian network model, our method successfully recovers all embedded links, and no spurious links, from the model, whenever 40 or more sequences are supplied.

Results: ChIP-chip data (yeast)

From ChIP-chip data for 100+ TFs from Harbison *et al.*[8], as reanalysed by MacIsaac *et al.*[9], we chose 40 factors for which at least 60 binding targets could be identified stringently (as previously done in [5]). PWMs were used to identify binding sites in these targets, with 3 flanking nucleotides included on each side. For each factor,

- 2/3 of sequences were used for training, 1/3 for testing.
- An equal number of synthetic testing sequences was generated from the PWM model
- We asked which model (PWM or BN) is better able to distinguish the real from fake sites.

The results are as follows: overall (across all factors), the average BMAL was higher for real binding sites than for synthetic sites, by about 0.27 (which corresponds to 30% higher likelihood). The WML, on the other hand, was very slightly lower for real sites than for the synthetic sites, by about 0.007.
For 24 out of 40 factors, the BN gave a higher likelihood for real binding sites than for fake ones (and always by a greater margin than the WM, which often favoured the fake sites). For 16 out of 40 factors, the BN favoured the fake sites, but in 12 of those cases the WM favoured the fake sites even more strongly.

Results: ChIP-seq data (yeast)

With ChIP-seq data by Venters and Pugh, we used a slightly different appoach: for 25 of their 200 factors (more in progress), we compared performance of the BN and PWM approaches in distinguishing bound sequences from unbound sequences. For 7 factors, the BN scores the bound sequences higher but the PWM scores unbound sequences higher. In no case is the opposite true. In remaining factors, results are mixed. Further analysis, and work on data from other species, is in progress.

Visualising

We use arcs to show links within the motif, and "pairwise logos" below the usual sequence logo to show the dinucleotide patterns in those links. Unlike the usual "sequence logo" for PWMs, this is an incomplete representation of the mix of Bayesian networks, but nevertheless informative. An example is on the left.
For example, looking at positions 7 and 10 (counting with 1 on the left), the most common dinucleotide is GC, but the next most common is not GA as one would expect from the PWM logo.
An animated version with mouseover effects in SVG has been implemented.

Conclusion

These are preliminary results and further validation, using larger sets of synthetic data as well as more recent ChIP-seq data in various species, is ongoing.

[1] Kyle Ellrott, Chuhu Yang, Frances M. Sladek, and Tao Jiang. *Bioinformatics*, 18(suppl 2):S100–S109, 2002.
[2] Marko Djordjevic, Anirvan M Sengupta, and Boris I Shraiman. *Genome Res*, 13(11):2381–2390, Nov 2003.
[3] Eilon Sharon, Shai Lubliner, and Eran Segal. *PLoS Comput Biol*, 4(8):e1000154, 2008.
[4] Anthony Mathelier and Wyeth W. Wasserman. *PLoS Comput Biol*, 9(9):e1003214, 09 2013.
[5] Rahul Siddharthan. *PLoS ONE*, 5(3):e9722, March 2010.
[6] Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. In Proceedings, RECOMB, 2003.
[7] I. Ben-Gal, A. Shani, A. Gohr, J. Grau, S. Arviv, A. Shmilovici, S. Posch, and I. Grosse. *Bioinformatics*, 21(11):2657–2666, 2005.
[8] Christopher T. Harbison *et al. Nature*, 431:99–104, 2004.
[9] Kenzie MacIsaac, Ting Wang, D Benjamin Gordon, David Gifford, Gary Stormo, and Ernest Fraenkel. *BMC Bioinformatics*, 7(1):113, 2006.
[10] Bryan J Venters *et al. Molecular cell*, 41(4):480–492, 2011.