Methods: Simulations and Real Data analysis

Simulations

Data

We simulated $N = 20$ DNA sequences of length $L = 100$ from an i.i.d. model for each site. We assumed a uniform distribution for the four bases. In each sequence, an experimentally verified binding site for a particular transcription factor was inserted. Functional binding sites for transcription factors are determined by DNA footprinting studies and can be obtained from several different databases [36]. We have 5 test sites in total: three from *Saccharomyces cerevisiae* (*gal4, abf1, pho4*) and two from *Escherichia coli* (*crp, purR*). In addition, in each sequence, we inserted a permuted version of the site, using the same permutation of columns for each sequence. When permuting the bi-modal examples (*gal4, abf1, crp*), we fixed the first and last positions so that the width of the motif is the same for the real and permuted sites.

Starting points

Recall that the EM algorithm is an optimization algorithm, maximizing the likelihood of the data given the model, over the parameters. Starting points are very important for the EM algorithm. Depending on the starting point, the algorithm may get trapped in a local maximum. Thus, for each data set, we ran the algorithm to convergence with 100 different starting points and picked the final motif with the highest value for the likelihood of the data given the model. In practice, for this data size, this procedure was adequate for discovering either motif.

Starting Point Selection

The starting points are matrices of the multinomial parameters, in which the columns are picked at random from the prior distributions. For each position, we must sample from 4-dimensional multinomials ($p_j, j = 1, 2, 3, 4$), according to the specified prior and subject to the constraints $\sum_{j=1}^{4} p_j = 1$, $1 \geq p_j \geq 0$ (4-d simplex). The normalization factor for our prior cannot be determined analytically. To avoid an iterative sampling procedure, such as rejective sampling [37], we discretize the 4-d simplex by increments of 0.1, as a simple, yet coarse alternative, thus each $p_j \in A = \{0, 0.1, 0.2, \ldots, 1\}$. This corresponds to sampling from 235 possible vectors weighted by the prior. For 3-dimensional multinomial parameters, ($p_j, j = 1, 2, 3$), the discretization results in a grid on the 3-d simplex surface. The number of points on the grid is $\sum_{i=1}^{k} i$, where $k = |A|$. For the 4-d simplex, the number of possible points is $\sum_{m=1}^{k} \sum_{i=1}^{m} i$.

Ignoring base identity, there are only 23 unique vectors. Finding the number of unique multinomial parameters vectors, in units of .1, and ignoring base identity is equivalent to finding the number of partitions of n objects into at most m parts, where $n = 10$ and $m = 4$ in this example. This number $p_m(n)$ can be determined by a recursive relationship. The values for different n and m are listed in “Handbook of Discrete and Combinatorial Mathematics” [38].

Evaluation

After completion, the algorithm outputs the final updates of $P[Y_{ik} = 1|\mathcal{P}^r, \mathcal{X}]$ in the E-step. These quantities give the probability of each position in a sequence being a motif start site, given the final parameter estimates and the data. We take the position with the maximum posterior probability for each sequence, $\arg\max_k P[Y_{ik} = 1|\mathcal{P}^r, \mathcal{X}]$, and tally the number that are within 2 positions of the real site or permuted site. We look within 2 positions because sometimes a shifted version of a motif is found. We found that looking beyond 2 positions does not alter the results.

If more than 50% of the real site positions in the data are correctly identified, then the motif discovered is considered to be the real motif. Likewise, if more than 50% of the permuted site positions are correctly identified, then the motif discovered is considered to be the permuted motif. In the simulations, we found that if a motif is found, at least half of the sites are usually identified.

Instead of using the updated posterior probabilities, the positions can also be scored by turning the estimated motif matrix into a log odds scoring matrix [39], but final results are similar.

Real Data

Sequences

When elongating the sequences by 100 base increments, the amount of flanking sequence on each side of the site(s) is not necessarily equal. For the upstream sequences, we used the beginning of the open reading frame as a natural cutoff.

Motif Width

For the sites from uni-modal motifs, the overall width is not as clearly defined as with bi-modal motifs, which have the most conserved positions at the boundaries. In general, for the uni-modal motifs that we examined, the center regime 1 block widths are usually 6-7 positions. This is expected due to the nature of protein-DNA interactions. From the information content plots, it also appears that there are 3-4 positions of moderate conservation at each side of the regime 1 block. Therefore, for uni-modal motifs, such as *rap1* and *reb1*, we set the length to 13. We do not expect our results to be sensitive to this choice.

Starting points

The main text describes the MEME starting point procedure that is used on the real data. However, for the *rap1* results in Table 3, the algorithm was run with random starting points, which are selected in the manner described above in "Starting Point Selection". Because the *S. cerevisiae* genome is A/T rich and the *rap1* test sets extend to very long lengths, we found that most of the MEME starting points for *rap1* consisted of A/T repeats. There was not much variety in these starting points. Therefore, we generated random starting points according to our prior with $\lambda = 20$. With these starting points, we were able to explore more of the likelihood surface and find the motif for longer lengths. By using the MEME starting points, the motif was not detected for lengths longer than 1400. Although we sampled from our prior, we used the same starting points for each λ in the L data set, so our results were not biased against $\lambda = 0$. This example shows that the type of starting point also plays a critical role in the performance of the method.

In the simulations, we used random starting points because it is much faster to generate random motif matrices according to our prior motif distribution than to scan through each of the many simulated data sets in this manner. We found that alternative starting point selection procedures did not qualitatively change the results from the simulations.

Evaluation

We use the same convergence criterion and evaluation measure as with the simulations. We take $\arg\max_k P[Y_{ik} = 1|\mathcal{P}^r, \mathcal{X}]$, for each sequence k, and tally the number that are within 2 positions of the real site. This count is the number of correctly identified sites for the data set. For sequences with more than one site, if one of them is detected by this criterion, this is counted as a correct identification.

We also examined the predicted motif for the cases where at least one site was correctly identified. We evaluated whether the consensus of the predicted motif matched the consensus of the true motif. In all our results, we found that if less than 20% of the sites were identified, the correct motif was not predicted. These sites were correctly predicted by chance, with a spurious motif. For our test sets, less than 20% of the sites roughly corresponds to one or two sites. Therefore, if more than 20% of the sites are correctly identified, we declare that the real motif in the data was discovered. The cases in which one or two sites were predicted by the incorrect motif will be identified in the tables with an asterisk.

BioProspector

In BioProspector, we used the default algorithms options and then tried several values for the model options. For the bi-modal examples (*crp, abf1*), we first fixed the gap between the blocks and then varied the block widths. Next, using the block widths with the best results, we used different ranges for the gap. For the uni-modal examples (*rap1, reb1*), we tried different values for the width of the single block. For *abf1*, the regime specifications should be [**1**(4), **2**(5), **1**(3)]. The minimum block width in BioProspector is 4, therefore we could not specify 3 for the second block. We ran the program with both block widths set to 4 first, but also specified block widths 5 and 4 because of the asymmetry in the original specifications.

To compare BioProspector with our method, we also specified that there is a motif occurrence in each sequence, matching the assumption of one occurrence per sequence (OOPS) in our model, and that the motif is not palindromic.

The output from BioProspector shows the top motif according to their criteria and the predicted motif occurrences in each sequence. For some sequences, there are multiple predictions. We take their first prediction and then tally the number of all predictions that are within two positions of the real site. For the uni-modal motifs, we tally the number that are within six positions. BioProspector uses a single block to model a uni-modal motif. In contrast, in our method, we use three blocks to model the uni-modal motifs and the overall width is much longer than for BioProspector. Therefore, we need to look in a larger window to make sure that we can identify whether a site has been predicted by BioProspector.

Gibbs Motif Sampler (GMS)

We ran GMS with the default algorithm options. To compare with our method, we also specified that the motif is not palindromic, that the maximum number of sites in a sequence is 1 and that the expected number of motif occurrences is the number of sequences. The choices for the last two options make the model equivalent to the OOPS model.

We selected several different model options to evaluate the performance of GMS for our test sets. The basic motif model in GMS is a conserved block of width W, similar to a uni-modal motif type. The GMS model also has the option (fragmentation model) to pick J positions in a larger window of W as more important. There is no specification of where the J positions within the W are located. To handle sequences with transcription factor binding sites, the latest version of GMS has the option to specify that the J positions should fall at the edges of the W block and not the middle (fragmentation from center). This is analogous to our bi-modal model. We used both the fragmentation model and the “fragment from center” option for the bi-modal motifs. For the uni-modal examples, we first ran GMS without the fragmentation model, but with several motif widths. Next, we specified the fragmentation model, with $W = 13$ and different values for J. These are the options that are closest to our model specification.

The output for GMS, according to the selected options, lists the predicted site for each sequence. For the bi-modal motifs, we tally the number of all predictions that are within two positions of the real site. For the uni-modal motifs, we tally the number that are within six positions for the same reason as with BioProspector.

References

36. Latchman D: *Eukaryotic Transcription Factors*, San Diego, CA: Academic Press. 3rd edition 1998 chap. Methods for Studying Transcription.

37. Ripley B: *Stochastic Simulation*. New York: Wiley 1987.

38. Rosen K, Michaels J, Gross J, Grossman J, Shier D (Eds): *Handbook of Discrete and Combinatorial Mathematics*, New York: CRC Press 2000 chap. Sets, Relations, and Functions.

39. Tatusov R, Altschul S, Koonin E: **Detection of Conserved Segments in Proteins: Iterative Scanning of Sequence Databases With Alignment Blocks**. *Proc. Natl. Acad. Sci., USA* 1994, **91**:12091–12095.