Finding regulatory sequences in DNA: motif discovery

The purpose of motif discovery as discussed here is to find binding sites of DNA binding regulators. We assume that the binding sites are short segments of DNA, not necessarily contiguous, to which a specific (family of) transcription factors can bind. While such sites may appear in genic as well as inter-genic regions, we focus here on finding binding sites only in the promoter region of each gene. It's worthwhile to note that the existence of a binding site is no guarantee that the site (the corresponding TF) plays a role in regulating the gene in question; the site may not be accessible due to chromatin structure, may never be occupied due to resource constraints and higher affinity sites elsewhere, and so on. Nevertheless, knowing who could participate in regulating each gene, and finding where they bind at a base-pair resolution, is a useful source of information.

There are several strategies we could follow to try to find such sites. For example, if we have genomes of two related species, simply aligning the promoter regions of orthologous genes (or aligning the whole genomes) would reveal interspersed segments of DNA that are highly conserved across the two species. The binding sites of DNA binding regulators are likely to fall within the conserved segments because of the evolutionary pressure to maintain the regulatory programs. The fraction of conserved segments that are interpretable as binding sites within, say, a promoter region, depends on the evolutionary distance between the species (time that they have evolved independently). Sufficient time is required for inessential portions of the sequences to diverge. Considering more than two related species would help emphasize the signal. For more information about this approach, see, e.g., Kellis, 2003.

We will follow here another approach, searching for binding sites of regulators within a single genome. We start with a set of genes that are likely to share regulators (bindings sites of regulators); a random subset of the genes are unlikely to share any regulators of interest. Note that it is not necessary to know who the common regulators are, only that they are likely to exist. Of course, knowing something about the common regulators, e.g., the protein families, can be extremely useful

Simple motifs, analysis

To set the problem a bit more formally, let $S_1, \ldots, S_n$ be n promoter sequences of interest. For simplicity we assume also that the sequences are of the same length, L, typically something like $500 - 1000$ bases (yeast). The simplest possible motif we could try to find is a w-mer, a sequence of length w. In other words, we are looking for a common w-mer (exact match) across the promoters $S_1, \ldots, S_n$. We'd like to understand first how L, w, and n relate to each other if we wish to claim that a common w-mer across the promoters is significant, unlikely to arise at random. Longer promoters (larger L) would increase the chances of finding a random match; increasing w, on the other hand, would make the random match less likely, as would having more sequences (larger n) so long as we require a match in all (or most of) the promoters.

Suppose each promoter sequence is sampled independently at random from a background distribution of bases, $B(x)$, $x \in \{A, G, T, C\}$. In other words, all the bases in all the promoter sequences are samples from the same distribution B which we take here to be uniform $B(A) = B(G) = B(T) = B(C) = 1/4$. Now, what is the probability that we find a common w-mer across n such promoters? Let's say first that we are looking for a match to a fixed w-mer x_w. The probability (over the random sampling of the promoters) that our probe x_w matches the first w bases of the first promoter sequence S_1 is simply $1/4^w$; this is the same for any w segment of any of the promoters. More formally, $P(x_w = S_i(j : j + w - 1)) = 1/4^w$, where $S_i(j : j + w - 1)$ is the w-mer in promoter S_i starting at position j. Now, using the fact that the probability of at least one event occuring out of many is bounded by the sum of probabilities of individual events (union bound), we get

$$P(S_i \text{ contains } x_w) \leq \sum_{i=1}^{L-w+1} P\Big(x_w = S_i(j : j + w - 1)\Big) = (L - w + 1) \cdot \frac{1}{4^w}$$

Since the promoters are sampled independently we can simply multiply the probabilities of finding a match in each promoter:

$$P(\text{ all } S_1, \ldots, S_n \text{ contain } x_w) \leq \left(\frac{L - w + 1}{4^w}\right)^n$$

Finally, since there are 4^w possible probes we could search,

$$P(\text{ a common } w\text{-mer in all } S_1, \ldots, S_n) \leq 4^w \cdot \left(\frac{L - w + 1}{4^w}\right)^n$$

This probability should be less than 0.05 in order for us to claim that any match we do find in n real promoters is significant. When $L = 500$ and $w = 5$, we would need to find an exact match in $n = 15$ promoters. If we are searching over a set of n promoters but find a common w-mer in only m of them, then the probability of a random match would be

$$P(\text{ a common } w\text{-mer in } m \text{ of } S_1, \ldots, S_n) \leq \binom{n}{m} \cdot 4^w \cdot \left(\frac{L - w + 1}{4^w}\right)^n$$

Hyper-geometric distribution

Another simple way to evaluate a motif is to see if it occurs preferentially in promoters of a functionally coherent set of genes (e.g., genes known to participate in the same biological process). Let N be the total number of promoters in the genome, and n the number of genes whose promoters contain the motif. If our relevant (fixed) set of genes has m members, we'd like to evaluate the probability that a completely unrelated motif, motif whose pattern of occurence has little to do with the functional category, would nevertheless occur in at least k of the m relevant promoters. Put another way, if the set of n promoters that the motif occurs in is a random sample from N possible promoters, what is the probability that this set would contain at least k members from the relevant set? The probability of overlap of exactly k members is given by the hyper-geometric distribution:

$$P(\text{ overlap is exactly } k) = \frac{\binom{N-m}{n-k}\binom{m}{k}}{\binom{N}{n}}$$

and thus

$$P(\text{ overlap of at least } k\ ) = \sum_{l=k}^{\min\{n,m\}} \frac{\binom{N-m}{n-l}\binom{m}{l}}{\binom{N}{n}}$$

This probability should again be smaller than 0.05 for us to claim that the association between the motif and the relevant genes is significant. Note that if the motif is chosen on the basis of the same relevant genes, this probability will no longer be valid as a measure of significance. It could nevertheless still be used as a criterion to weed out irrelevant motifs (see, e.g., Hughes et al., 2000)

Motif models, three estimation problems

Binding sites of each regulator can show considerable variation from site to site highlighting the fact that a regulator may be able to bind to different but related sequence elements, possibly with different affinities. For example, the following aligned sites are examples of putative GAL4[1] binding sites

[1] A DNA binding regulator of the yeast Galactose system.

```
CGGTCAACAGTTGTCCGAGC
CGGCGGCTTCTAATCCGTAC
CGGAGGGCTGTCGCCCGCTC
***           ***
```

where the relevant signature is CGG . . . CCG (the length of the gap may also vary slightly from site to site). GAL4 binds as a dimer (has two DNA binding domains) and the conserved signature of the binding site represents how the two parts make contact with DNA.

We have to be able to somehow capture this variation so as to find other instances of the binding site. A possible strategy, and one that we will follow here, is to build a statistical model from examples of bindings sites. The advantage of such a model is that we might be able to capture the manner in which the sites vary based on relatively few examples. The difficulty in general is that the model has to be estimated in conjunction with discovering examples of the binding sites!

The problems we have to solve in this context can be categorized in terms of the available data.

1. We have a set of pre-aligned binding sites. We need to build a statistical motif model that captures the variation in the sites. This is a sub-problem we have to solve in any case.

2. We have a set of promoter sequences that are known to contain (at least) one copy of the binding site. In this case we have to find where the sites are in conjunction with estimating the motif model.

3. We only have a set of promoter sequences, some subset of which contain at least one copy of the binding site.

Problem 1: Position specific weight matrix

We begin here with a simple example where the binding site is a contiguous 4-mer. The aligned binding sites could be, for example,

```
TGAC
TGAC
CGAG
```

Suppose we assume that the variation of bases in these sites is independent across the columns (relative positions). Then we are left with estimating a distribution over the bases for each relative position. For example, based on the above 4-mers, we would estimate $P_1(A) = 0$, $P_1(G) = 0$, $P_1(T) = 2/3$ and $P_1(C) = 1/3$, where the subindex 1 refers to the relative position. Similarly, $P_3(A) = 1$. We can collect these probabilities or frequencies of bases in each position into a position specific weight matrix:

$$\hat{M} = \begin{bmatrix} P_1(A) & P_2(A) & P_3(A) & P_4(A) \\ P_1(G) & P_2(G) & P_3(G) & P_4(G) \\ P_1(T) & P_2(T) & P_3(T) & P_4(T) \\ P_1(C) & P_2(C) & P_3(C) & P_4(C) \end{bmatrix} = \begin{array}{c} A \\ G \\ T \\ C \end{array} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1/3 \\ 2/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 2/3 \end{bmatrix} \tag{1}$$

This is how we try to summarize the variation in the binding sites. The frequencies in the above matrix are maximum likelihood estimates of the base frequencies. To understand this, consider evaluating the probability, given the weight matrix, of the first 4-mer:

$$P(\texttt{TGAC}|M) = P_1(T) \cdot P_2(G) \cdot P_3(A) \cdot P_4(C) = 2/3 \cdot 1 \cdot 1 \cdot 2/3 = 4/9$$

The likelihood of all three sites is then

$$P(\texttt{TGAC}|M) \cdot P(\texttt{TGAC}|M) \cdot P(\texttt{CGAC}|M)$$

The numerical values in the M matrix are chosen to maximize this likelihood – probability that the model reproduces the data.

The simple weight matrix model is unable to represent many types of variation. For example, we couldn't capture the conserved parts of the GAL4 binding sites if the gap length varies from site to site. Similarly, we couldn't capture the variation introduced by a change in the orientation of an asymmetric but perfectly conserved site: `AAAGGG` and `GGGAAA` (artificial example). The maximum likelihood estimate of the weight matrix based on these two sites would be

$$\hat{M} = \begin{array}{c} A \\ G \\ T \\ C \end{array} \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Thus, according to the model, `AGAGAG` is just as likely as one of the original binding sites.

Another (related but actually useful) feature of the weight matrix is that it can be very sensitive to the correct alignment of the binding sites. For example,

```
cTGAC
TGACa
```

where the lowercase letters indicate bases around the site, introduced here due to an alignment error and the fact that we consider 5-mers. The correct weight matrix would capture the perfectly conserved sequence `TGAC` but the maximum likelihood estimate of the two incorrectly aligned sites involves much more uncertainty about the bases:

$$\hat{M} = \begin{array}{c} A \\ G \\ T \\ C \end{array} \left[\begin{array}{ccccc} 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 1/2 \end{array} \right]$$

(note that the model in this case is over 5 bases). The increased uncertainty about the bases decreases the likelihood of reproducing the data (the probability mass assigned to each observed base is lower). So, the better we can align the binding sites, the higher the likelihood of the data.

Problem 2: motif score, search

Here we dispense with the requirement that we have pre-aligned binding sites. We assume only that the available promoter sequences contain at least one copy of the binding site. In other words, we have to find where the binding sites are in addition to estimating the position specific weight matrix.

To get started we need to be able to score (evaluate the likelihood of) whole promoter sequences assuming they contain binding sites in specific locations. Let $S_1, \ldots, S_n$ be the promoter sequences with binding sites in locations $i_1, \ldots, i_n$, respectively. This is similar to the previous problem since specifying the binding site locations is equivalent to aligning them. Once we have a score for promoters with motifs in fixed locations, we can search over the possible locations to increase the score (log-likelihood).

In addition to the motif model, we need a background model, $B(x), x \in \{A, G, T, C\}$, to generate the bases in the promoters around the binding sites. The background frequencies of bases are no longer assumed to be uniform; the frequencies may be estimated either on the basis of the available promoter sequences or from inter-genic regions more generally. For our purposes here the background model remains fixed.

We imagine generating the bases in each promoter sequence S (of length L) as follows. We start with the background model and generate the bases until we hit the binding site at position i. At that point we switch to using the position specific base frequencies from the matrix, and sample a base for each successive position in the motif. The remaining bases in the promoter, after the binding site, are again sampled from the background model. The probability that we correctly generate the bases in S, assuming a binding site at i, is therefore:

$$P(S|M,i) = \overbrace{\prod_{j=1}^{i-1} B(S(j))}^{\text{bases before the site}} \cdot \overbrace{\prod_{j=1}^{w} P_j(S(i+j-1))}^{\text{bases within the site}} \cdot \overbrace{\prod_{j=i+w}^{L} B(S(j))}^{\text{bases after the site}}$$

So, for example, if we use the weight matrix from eq (1), and assume that the promoter sequence $S =$ `ACGTGACT` contains a binding site in position 4, we would evaluate the above probability as

$$\begin{aligned} P(S|M,i=3) &= B(A)B(C)B(G)\cdot P_1(T)P_2(G)P_3(A)P_4(C)\cdot B(T) \\ &= B(A)B(C)B(G)\cdot 2/3\cdot 1\cdot 1\cdot 2/3\cdot B(T) \end{aligned}$$

It is instructive to compare the probability $P(S|M,i)$ to the probability of generating all the bases in the promoter from the background model or $P(S|B)$. To facilitate the comparison we can decompose $P(S|B)$ in terms of the binding site location so long as we use the background model to generate all the bases. In other words,

$$P(S|B) = \prod_{j=1}^{i-1} B(S(j)) \cdot \prod_{j=1}^{w} B(S(i+j-1)) \cdot \prod_{j=i+w}^{L} B(S(j))$$

The log-likelihood ratio of the two probabilities is now given by

$$\begin{aligned} \log\frac{P(S|M,i)}{P(S|B)} &= \log\frac{\prod_{j=1}^{i-1} B(S(j))\cdot\prod_{j=1}^{w} P_j(S(i+j-1))\cdot\prod_{j=i+w}^{L} B(S(j))}{\prod_{j=1}^{i-1} B(S(j))\cdot\prod_{j=1}^{w} B(S(i+j-1))\cdot\prod_{j=i+w}^{L} B(S(j))} \\ &= \log\frac{\prod_{j=1}^{w} P_j(S(i+j-1))}{\prod_{j=1}^{w} B(S(i+j-1))} = \sum_{j=1}^{w}\log\frac{P_j(S(i+j-1))}{B(S(i+j-1))} \end{aligned}$$

So the advantage of specifying a binding site at i depends only on the bases within the site. Moreover, slightly higher probabilities of generating the successive bases in the binding site from the position specific probabilities can add up to a substantial advantage over the length of the binding site.

Now, given promoter sequences $S_1,\ldots,S_n$, binding sites in locations $i_1,\ldots,i_n$, and the motif model M, the likelihood of reproducing all the bases in the promoter sequences is given by

$$P(S_1|M,i_1)\cdots P(S_n|M,i_n)$$

This is the score we'd like to optimize in terms of both the binding site locations $i_1, \ldots, i_n$ and the motif model M. It is easier but equivalent to optimize the log-likelihood instead:

$$\log P(S_1|M, i_1) \cdots P(S_n|M, i_n) = \sum_{t=1}^{n} \log P(S_t|M, i_t)$$

We can perform the optimization iteratively by alternating the steps of finding the best locations $i_1, \ldots, i_n$ given the current (probably wrong) motif model M and subsequently updating the model M based on the new locations (alignment of binding sites) $i_1, \ldots, i_n$. More formally: start with some prior weight matrix M and, iteratively

(1) find the best binding site location for each promoter given M:

$$i_t \leftarrow \arg\max i \left\{ \log P(S_t|M, i) \right\}, \;\; \text{for each } t = 1, \ldots, n$$

(2) optimize the weight matrix M (as before) based on the aligned sites at locations $i_1, \ldots, i_n$

We can stop when the binding site locations $i_1, \ldots, i_n$ no longer change. Since the background model is assumed fixed, we could replace the first step by

$$i_t \leftarrow \arg\max_i \left\{ \log \frac{P(S_t|M, i)}{P(S_t|B)} \right\}, \;\; \text{for each } t = 1, \ldots, n$$

So the locations we find in the first step are the ones that the current motif model has the greatest advantage over the background model. The second step tries to reinforce this advantage by adjusting the motif model.

While this algorithm successively increases our score – log-likelihood of the bases in the promoter sequences – it is quite sensitive to the initial choice of the motif model. Put another way, the algorithm can easily get stuck in a bad solution if the initial choice of the motif model is not close to the correct one. Ideally we wouldn't want to commit to any specific alignment in step (1) above when we don't yet know what the appropriate motif model is.

Mixtures and the EM algorithm

We can improve the motif finding algorithm by rethinking the model a bit. What do we know about the binding site locations before running the algorithm? Suppose we have no idea where the binding site should be. We can express this ignorance by defining a uniform prior distribution over the $L - w + 1$ possible locations $P(i) = 1/(L - w + 1)$. So, since we are uncertain about the location of the binding site, we could generate the bases in S as follows: sample a location from the prior $P(i)$, introduce a binding site at i, and then generate the bases from $P(S|M, i)$, defined as before. The overall probability of generating the bases in S according to this model is given by

$$P(S|M) = \sum_{i=1}^{L-w+1} P(i)P(S|M, i)$$

This is called a mixture model since it "mixes" more specific models $P(S|M, i)$. Note that $P(S|M)$ is no longer a function of any specific binding site location; we are considering all possible locations. The model also depends on the prior probabilities $P(i)$ in addition to the motif model M. These prior probabilities can be adjusted on the basis of any information we have about where we would expect to find the binding sites (e.g., based on the degree of conservation of bases across species).

Since we are no longer explicitly specifying where the binding sites are, how do we recover the binding site locations? We can evaluate the posterior probability of finding the binding site in any specific location:

$$P(i|M, S) = \frac{P(S|M, i)P(i)}{P(S|M)} \tag{2}$$

The posterior probability is proportional to the prior probability of finding the binding site at location i, or $P(i)$, and the ability of the motif model to generate the bases assuming the binding site at i, or $P(S|M, i)$. $P(S|M)$ above simply normalizes the posterior so it sums to one.

We can use the posterior probabilities to "align" putative binding sites for the purpose of updating the motif matrix M. The alignment is weighted: each $w-$mer in each promoter sequence is included in the alignment but with the weight equal to the posterior probability. So, most of the $w-$mers will have insignificant weights and, so long as the prior probability over the binding site locations is uniform, the highest weight (posterior probability) will be given to the site also identified by the previous algorithm.

Suppose we have two promoter sequences $S_1 =$ `CTGA` and $S_2 =$ `CGAC` and we have specified an initial $3-$mer motif model M. In this case we can put a $3-$mer binding site in only two possible locations within each promoter sequence. To get the posterior probabilities over the possible locations, say within S_1, we can start by evaluating

$$\begin{aligned} P(S_1|M, i = 1) &= P_1(C)P_2(T)P_3(G)B(A) \\ P(S_1|M, i = 2) &= B(C)P_1(T)P_2(G)P_3(A) \end{aligned}$$

These, together with the uniform prior $P(i) = 1/2$, $i = 1, 2$, determine the probability that the mixture assigns to S_1 or $P(S_1|M)$. The posteriors can be then computed directly from the Bayes rule in Eq (2). In terms of the weighted alignment, we will have

weight	3-mer site
$P(i = 1\|S_1, M)$	CTG
$P(i = 2\|S_1, M)$	TGA
$P(i = 1\|S_2, M)$	CGA
$P(i = 2\|S_2, M)$	GAC

The weights sum to 2 (the number of promoter sequences). This weighted alignment replaces the step (1) of our previous algorithm, and is known as the E-step of the EM (Expectation Maximization) algorithm. The M-step or finding the motif model in response to the alignment is analogous to the step (2) of our previous algorithm: we evaluate weighted frequencies of bases in each column. So, for example, if we rewrite the above table more concretely as

weight	3-mer
$P(i = 1\|S_1, M) = 0.1$	CTG
$P(i = 2\|S_1, M) = 0.9$	TGA
$P(i = 1\|S_2, M) = 0.6$	CGA
$P(i = 2\|S_2, M) = 0.2$	GAC

then we would estimate the first column of the new motif model M as

$$\begin{aligned} P_1(A) &= \frac{1}{2} \cdot 0 \\ P_1(G) &= \frac{1}{2} \cdot 0.2 = 0.1 \\ P_1(T) &= \frac{1}{2} \cdot 0.9 = 0.45 \\ P_1(C) &= \frac{1}{2} \cdot (0.1 + 0.6) = 0.35 \end{aligned}$$

In summary, the EM-algorithm for training a mixture motif model can be defined as follows: start from an initial choice of M,

> E-step: for each promoter sequence $S_1, \ldots, S_n$, evaluate the posterior locations of the binding site:
>
> $$P(i|S_t, M), \quad i = 1, \ldots, L - w + 1, \quad t = 1, \ldots, n$$

M-step: re-estimate the motif model M on the basis of the weighted alignment defined by the posteriors

The EM-algorithm is guaranteed to monotonically increase the (log-)likelihood of the promoter sequences according to the mixture model:

$$\sum_{t=1}^{n} \log P(S_t|M)$$

The solution is still iterative (we repeat the E and M-steps until convergence), we are not guaranteed to find the optimal solution, but the algorithm is nevertheless much less susceptible to getting stuck in a bad solution than the simpler version discussed above.

Problem 3: extensions

We can extend the basic mixture idea in various ways to incorporate structure into the motifs (e.g., gaps) or remove the assumption that we need to find a binding site in each promoter sequence. For example, suppose we expect roughly a fraction p of the relevant promoters to have the binding site of interest. A mixture model consistent with this assumption would be:

$$P(S|M,p) = pP(S|M) + (1-p)P(S|B)$$

In other words, with probability p we generate the bases in S using the mixture model discussed above that assumes a motif, and with probability $1-p$ we simply generate all the bases from the background model. This mixture can be estimated using a modified EM-algorithm. If you don't know the parameter p, it can be estimated along with the motif model M.

References

Manolis Kellis, MIT PhD thesis, 2003.

Hughes et al., Journal of Molecular Biology, 296(5):1205-14, Vol 2000