Gene expression is often regulated transcriptionally through the action of protein factors that bind to DNA and affect the rate of initiation of the (usually) nearby promoter, either increasing it (activators) or decreasing it (repressors). In order for this process to be promoter-specific, rather than affecting the expression of all genes, at least some of the transcription factors must recognize and bind to specific DNA sequences. The purpose of this article is to define a quantitative measure of specificity and describe various models to represent the specificity of a particular protein that can be used to predict binding sites in genomic DNA. It also briefly describes how those models can be used in pattern recognition methods to identify regulatory sites from sets of coregulated genes.

**QUANTITATIVE SPECIFICITY**

Consider a DNA-binding protein that binds to sites that are \( l \)-long. There are \( 4^l \) such sequences, and we refer to each one as \( S_i \) with \( 1 \leq i \leq 4^l \) (\( i \) and \( l \) are integers). Each sequence binds to the protein with some free energy, denoted by \( H_i \). Of course that will depend on the binding conditions, but for simplicity it is assumed there is some standard condition that is always used. (It is very interesting to study changes in \( H_i \) as a function of the reaction conditions, but that issue is not addressed in this article.) Figure 4.1a represents the list of all \( 4 \)-long sequences and the binding energy to each one for some hypothetical protein. The energies are shown relative to a sequence with average affinity, so that those with negative values are preferred and those with positive values are discriminated against by the protein. In general we
Figure 4.1 Specificity of a hypothetical DNA-binding protein. (a) The complete list of binding free energies to all 256 4-long sequences. In this example the units are kcal/mole, and the binding energies are relative to the site with average affinity. (b) The weight matrix model that provides the same binding energies for all sequences. The energy for an individual sequence is the sum of the values for the bases that correspond to that sequence in each position (see figure 4.2).

care only about the differences in binding energies, because those determine the probabilities of binding to different sequences. For example, the ratio of binding affinities, $K_i$, $K_j$, for two sequences, $S_i$, $S_j$, is just

$$\frac{K_i}{K_j} = e^{-H_i + H_j}$$  \hspace{1cm} (4.1)

At equilibrium the distribution of sites bound by the protein, $f_i$, can be obtained from the list of binding energies, $H_i$, and the distribution of potential binding sites, $g_i$, from the Boltzmann equation (Heumann et al., 1994):

$$f_i = \frac{g_i e^{-H_i}}{\sum_{j=1}^{4^4} g_j e^{-H_j}} = \frac{g_i e^{-H_i}}{Z}.$$  \hspace{1cm} (4.2)
Z is the partition function and assures that $\sum_i f_i = 1$. The temperature is not included because it is assumed to be one of the conditions that is held constant. $g_i$ is the number of occurrences of each sequence, $S_i$, in the set of possible binding sites. In vivo, $g_i$ would be the composition of the genome in words of length $l$, or at least the composition of sites that are available for binding to the protein. As stated above, it is only the difference in binding energies that matters, because $f_i$ is unaffected by replacing all values of $H_i$ by $H_i + c$.

One common choice for the baseline of energy is to set the energy for the preferred sequence to 0, so that all other sequences have positive energy (Berg and von Hippel, 1987). We often choose the average affinity as the baseline for the energy, such that $Z = G$, the total number of possible binding sites, which is the total number of available sites in the genome when considering the in vivo situation. We can substitute $p_i = g_i/G$, where $p_i$ represents the probability of each sequence $S_i$ in the set of possible binding sites.

Rearranging equation 4.2 provides a conceptually simple, but technically difficult, method to determine the binding energies of each sequence:

$$H_i = -\ln \frac{f_i}{p_i}$$

(4.3)

That is, you can imagine mixing all possible sequences together in known amounts, $p_i$, and then measuring the fraction of each sequence bound to the protein at equilibrium, $f_i$. The logarithms of those ratios give exactly the energy values desired. Furthermore, at equilibrium the average binding energy is simply

$$\langle H_i \rangle = -\sum_i f_i \ln \frac{f_i}{p_i},$$

(4.4)

which is the relative entropy, or Kullbach-Leibler distance, between the two distributions of the potential binding sites and the bound sites. The derivation of this simple, but useful, relationship has been achieved through the use of the Boltzmann equation, but there is an equivalent derivation using Bayes's theorem for conditional probabilities (Heuermann et al., 1994).
MODELS FOR SPECIFICITY

While in theory we could measure $H_i$ for every sequence and simply put the measurement into a lookup table to use whenever needed, as in figure 4.1(a), the task is impractical for typical sizes of $l$. Instead, one usually employs some sort of model that provides an estimate of $H_i$ for all sequences. The simplest model is to assume that the protein binds to some set of sequences (perhaps only one) and not at all (or at least very much worse) to any other sequence. If it binds equally well to all of its binding sites, each site gets some energy $x$, and every other sequence has energy $+\infty$ (or some large increase over the energy of the sites). For example, the specificity shown in figure 4.1a could be modeled as a negative free energy for the preferred sequence, AGCT, and a large positive value for all other sequences, using the assumption that only AGCT shows significant binding to the protein. Such an assumption might be valid for some proteins, such as restriction and modification enzymes that are extremely sequence-specific in their activities. However, most regulatory proteins are much less specific and show significant binding to sequences other than their preferred sites. The model could be modified to allow for binding to sequences similar to AGCT, perhaps those sequences with one mismatch (there are 12 such sequences). Any method that searches for binding sites using a consensus sequence, even allowing for some number of mismatches, implicitly assumes such a model for binding energies.

The next, more complicated model employs a "weight matrix" $H(b, m)$ that contains an energy for each base, $b \in \{A, C, G, T\}$, at each position, $1 \leq m \leq l$, in the site (Stormo, 1990, 2000). Figure 4.1b shows such a model for the hypothetical protein. The binding energy to any sequence is $H_i = H(b, m) \cdot S_i$. In this notation $S_i$ is a matrix with 1 for the base that occurs at each position and 0 elsewhere, so the dot product selects the energies that correspond to the sequence, as shown in figure 4.2. The main advantage of weight matrix methods over consensus sequence methods is that they allow different mismatches from the preferred sequence to have different effects on the binding site predictions. For the protein in figure 4.1, mismatches to the $T$ at position 4 would have dramatic effects on binding, whereas all other mismatches would have smaller effects. In fact, many double mismatches and even the triply mismatched sequence CTTT bind with higher affinity than a single
mismatch at position 4. Such differential effects are taken into account naturally by the weight matrix model, and they are not uncommon in the specificity of regulatory proteins, where some positions may be highly conserved and others highly variable.

The weight matrix model assumes that the total binding energy is the sum of the interactions at each position. For the hypothetical protein of figure 4.1 we assume complete additivity, so that all 256 values in figure 4.1a can be obtained precisely from the 16 parameters of the model in figure 4.1b. That additivity assumption may be reasonable for some proteins, but may not be good enough for others. The model can be made more complicated with a matrix that has a row for each dinucleotide rather than for each base (and only \( l - 1 \) columns). That would accommodate nonindependent interactions at adjacent positions in the binding site. If that is still not good enough, the matrix can be made with trinucleotide rows, or higher. (Of course this can be done as a Markov chain rather than an explicit matrix of all possible base combinations, but that only changes the notation.)

At some point the weight matrix method must provide a perfect "model" for the interaction because it becomes a matrix with \( 4^l \) rows and only one column, the lookup table of figure 4.1a. For practical reasons, primarily limited data, we usually assume that a simple weight matrix (bases by positions) will be a sufficiently good model, but we have to allow the data to drive us to more complicated models when necessary. Exactly when a particular model is "good enough" often depends on how it will be used.

Given that we want to represent the specificity of a protein with a weight matrix, there are several approaches that might be used to ob-
tain one. Since the weight matrix has only $4l$ parameters (and since we care only about the differences in each column, just $3l$ degrees of freedom), one might simply measure the change in binding energy to all single base changes from the preferred sequence (Sarai and Takeda, 1989; Takeda et al., 1989; Fields et al., 1997). Alternatively, given a collection of sequences with known binding energy, one could find the matrix that provides the best fit to that quantitative data (Stormo et al., 1986; Barrick et al., 1994). This latter approach has the advantage of indicating whether the weight matrix provides a sufficiently good model of the interaction. If not, even the best fit will be fairly poor and more complicated models will be necessary (Stormo et al., 1986).

Probably the easiest and most common method is to collect a set of known binding sites and develop the matrix from a statistical analysis of those sites. The known sites may be naturally occurring or they may be selected from a random pool (Schneider et al., 1986; Fields et al., 1997). Because we assume a binding energy model of the form $H(b, m)$, we can treat the positions independently and convert the aligned collection of known sites into a frequency matrix $F(b, m)$ that counts the fraction of each base at each position in the aligned sites, as in figure 4.3. If we make the further assumption that the probability of a sequence, $p_i$, is determined by its composition, and each base $b$ has probability $P(b)$, then the best estimate of the energy matrix is

$$H(b, m) = -\ln \frac{F(b, m)}{P(b)}.$$ (4.5)

This method of estimating the binding energy for a protein based on a collection of known sequences originated with Staden (1984), except that he set $P(b) = 1$. Schneider (1997) uses the same form with $P(b) = 0.25$. Berg and von Hippel (1987) substituted $\max_i F(b, m)$ for $P(b)$ at each position $m$, so that the most common base is assigned energy 0 and all other bases get positive energy.

All of these approaches are equivalent if the $F(b)$ are all approximately equal, but can be inappropriate if they are not or if the $p_i$ are not well approximated by their composition. It is easy to show that if the $p_i$ are well approximated by the composition, then the formula of equation 4.5 provides the values of $H(b, m)$ that maximize the probability of binding to the known sites (Heumann et al., 1994). Then the average binding energy to the known sites at equilibrium is
Figure 4.3 Determining the matrix for a set of binding sites. (a) An example set of binding sites. These are taken from the list of possible binding sites, based on their probabilities. (b) The summary matrix of the 30 aligned sites, partially listed in (a). (c) The frequency matrix, using a small sample size correction of adding 1 to each value in the summary matrix, and dividing by the total (34). (d) The resulting energy matrix, using $P(b) = 0.25$ for all bases. These values have been converted back into units of kcal/mole, using $RT = 0.62$.

$$\langle H_i \rangle_{sites} = - \sum_b \sum_m F(b, m) \ln \frac{F(b, m)}{P(b)} ,$$ \hspace{1cm} (4.6)

the relative entropy between the base frequencies at the sites and the genomic base frequencies (Schneider et al., 1986; Stormo and Fields, 1998). This is a very useful objective function for the pattern recognition methods we employ because the set of sites that maximize the relative entropy would be those sites with the highest probability of binding to the protein under the assumptions of equilibrium, additivity, and a random background. Some of those assumptions can be removed without increasing the complexity of the problem too much, and we do so when it seems useful.
DISCOVERING BINDING SITES

The previous section described how we can represent the specificity of a DNA-binding protein as a weight (or energy) matrix, and how we can obtain the best estimate for that matrix, given a collection of known sites. However, often the problem of interest is how to discover the binding site pattern (energy matrix) for a transcription factor, given only a collection of genes that are known to be regulated by it. For example, an expression array experiment can indicate sets of genes that apparently are coordinately regulated (Spellman et al., 1998), and therefore may each have a binding site for a common transcription factor. We don’t know the energy matrix for the protein, or where the binding sites are, except that they are most likely to be in the promoter regions for the genes. The goal then is to find an energy matrix such that each of the coregulated sequences has at least one high-probability binding site.

Several approaches now exist to find such sites. There are word-based methods that try to find overrepresented words in promoter regions. This approach was first presented by Galas et al. (1985), has been modified several times, and most recently has been used effectively on sets of coregulated yeast genes (van Helden et al., 1998; Brazma et al., 1998). Although the algorithms don’t work this way, it is possible to think of them as dividing all potential binding sites into two classes, those that bind (with some energy $x$) and those that don’t (energy $+\infty$). Their objective is to maximize the probability of binding to the coregulated promoter regions, which will increase if they have more of the words that bind to the protein.

These methods can take the background probabilities, $p_i$, into account by comparing the frequencies of the words identified to their frequencies in other promoters (van Helden et al., 1998). These methods don’t really produce an accurate binding energy description, because most proteins do not bind with “all or nothing” affinities. But they can still be effective at identifying the regulatory sites, which can then be analyzed in other ways. This is especially true if the binding sites are relatively short and highly conserved. However, in general, matrix-based methods tend to be more reliable because they can also identify the short, highly conserved sites and can allow for more variability in the sites.
There are four general methods, and several specific programs, for identifying sites via weight matrix approaches: a greedy algorithm, expectation-maximization (EM), a Gibbs's sampling method, and a network method for maximizing specificity. The goal in each case is the same, so they often return the same solution. However, all of the methods are heuristic and may not find the true maximum of their objective. Each algorithm attempts to maximize the objective in a different way, and so may be susceptible to different local optima. It is usually advisable to try multiple methods to see if consistent results are obtained. Furthermore, there is usually a variety of parameters that can be set which affect how the programs run, and exploring various parameter values may be worthwhile.

Our CONSENSUS program is a greedy algorithm that builds up the total alignment of sites by adding new ones at each iteration (Stormo and Hartzell, 1989; Hertz et al., 1990; Hertz and Stormo, 1999). It is somewhat similar to CLUSTAL (Higgins et al., 1996) in that it develops the full multiple alignment of sites progressively, but has several important differences. First, it is specifically searching for ungapped local alignments of length \( l \). (The version WCONSENSUS does not require the length to be specified, and will search for the optimum length during the run.) Second, the program does not rely on a single best alignment at any step, but at each iteration keeps many (a user-defined parameter, typically set to 1000 or more) potential multiple alignments to be compared against the remaining sequences.

Third, the alignments are ranked by either their information content or their \( p \)-values. Originally we ranked alignments based on their information content because of its relationship to the average binding energy of the predicted sites. However, information content is always nonnegative and, by itself, not directly related to the significance of an alignment. But by taking into account the length and composition of the promoter regions, as well as some other relevant parameters, it is possible to compute the probability of finding by chance an alignment with an information content above some value, which gives us the \( p \)-value for each alignment (Hertz and Stormo, 1999). Ranking by \( p \)-value allows direct comparison of the significance of alignments of different lengths, different number of sites included, and even determination of whether a particular pattern is symmetric or not. This last issue, whether a binding site is a symmetric pattern, can indicate whether a
regulatory protein binds as a homodimer, as is fairly common. Given an alignment of sites that are intrinsically symmetric, it is always possible to obtain an asymmetric alignment with greater (or equal) information content, because each individual site has two ways to be aligned. Only by considering the reduced number of degrees of freedom imposed by the symmetry of the pattern is its symmetry apparent. Thus use of the \( p \)-value can provide improved estimates of the optimal binding energy parameters.

The EM algorithm and the Gibbs's sampling method are similar overall, but have one important difference (Lawrence and Reilly, 1990; Lawrence et al., 1993). Both are iterative algorithms that alternate between the two steps shown in figures 4.2 and 4.3. They usually start with an arbitrary alignment of sites, but can also start with an arbitrary matrix. Given a set of sites, a matrix is produced using the method in figure 4.3. Then, using the matrix, the probability of the protein binding to every site is determined from its predicted energy, as in figure 4.2. This procedure is iterated and is guaranteed to converge in the case of EM, and tends to higher values of information content with Gibbs's sampling.

The main difference is in how the sites from which the matrix is made are chosen. In EM, every site contributes to the alignment but is weighted by its probability. In Gibbs's sampling, only one site is selected in one sequence, using the matrix derived from the predicted sites in the other sequences. The site chosen is based on its probability; the higher the probability, the more likely it is to be chosen, but any site may be chosen at each iteration. This stochastic aspect of the Gibbs's sampling procedure makes it able to escape from local optima, and therefore it is more robust than EM. The objective used in the Gibbs's sampling method is identical to information content, using a small size correction, as in figure 4.3 (see equations 1 and 2 in Lawrence et al., 1993). The EM algorithm has been implemented in the MEME package of programs (Bailey and Elkan, 1994; Grundy et al., 1996). Several newer versions of Gibbs's sampling algorithm exist (Spellman et al., 1998; Roth et al., 1998; Wasserman et al., 2000).

Each of the weight matrix methods described thus far considers only the promoters from the coregulated genes, although potential binding sites from the rest of the genome influence the probability of those sites being bound (equation 4.2). They assume that the rest of the genome can be modeled as a random sequence with a given composition, \( P(b) \).
However, for many sets of coregulated promoters this assumption can be misleading. For example, yeast promoters often contain strings of A/T-rich sequences that occur much more often than expected, even given the A/T-rich genome. Such patterns clearly are statistically significant, given the random genome assumption, but they are not specific to the set of coregulated genes. Rather, they are patterns that are common in many promoters, and so cannot be the sought-after binding sites responsible for the coregulation.

To address this issue we developed an alternative algorithm, implemented in the program ANN-Spec (Heumann et al., 1994; Workman and Stormo, 2000). It still uses a weight matrix model for the binding energy, but it determines the partition function explicitly using the entire genome, or at least all of the promoter regions for eukaryotic organisms. It starts with an arbitrary matrix and then selects binding sites from each sequence using Gibbs’s sampling approach. It also calculates the partition function and then determines the gradient of the site probabilities as a function of the weight matrix parameters. It follows this gradient until convergence.

This method is still heuristic, but tends to work as well as the other methods we’ve tried on both simulated sequences (Workman and Stormo, 2000) and real promoters from yeast and other eukaryotes (unpublished results). When the background is approximately random, its objective function is the same as information content (Heumann et al., 1994), so it should give the same results as the other methods. But its advantage is that it specifically identifies patterns that are able to distinguish the coregulated genes from other promoters, and therefore does not find the A/T-rich patterns that the other methods might (Workman and Stormo, 2000).

Of course there are other methods to avoid those common patterns in postprocessing steps. The AlignACE version of Gibbs’s sampling algorithm identifies many potentially significant patterns and then filters out the uninteresting ones based on a variety of criteria, including the frequencies of the pattern in the coregulated genes and the rest of the genome (Hughes et al., 2000).

**FUTURE DIRECTIONS**

One of the limitations of the approaches described is that they tend to view proteins binding independently of one another. It is easy to
accommodate situations in which two or more proteins compete for binding to mutually exclusive sites. More problematic are situations where proteins bind cooperatively. It is known that eukaryotic promoters are not usually regulated by the binding of a single transcription factor, as often happens in prokaryotes. Rather, it is common for multiple factors binding to multiple sites to be involved in regulating transcription. Provided each protein shows sufficient specificity on its own to be identified, the methods already described should be able to find all of the binding sites, either simultaneously or through iterated application of the program. But if the patterns themselves are significant only through their joint occurrence, then alternative methods are necessary that can specifically look for composite patterns that are significant jointly, even if not individually. Once combinations of individual patterns that lead to specific regulatory responses are identified, then the combinatorial network of gene regulation can be more fully modeled.

There has been some work in the area of defining complete, multi-component regulatory patterns (Frech et al., 1997; Hu et al., 2000; Liu et al., 2001; GuhaThakurta and Stormo, 2001), but more work is needed. Additional extensions would include knowledge of operon organization and regulatory proteins, as well as transcriptome experiments.

ACKNOWLEDGMENT

The work described from my lab has been supported by NIH grant HG00249.

REFERENCES


**SUGGESTED READING**


**URLs FOR RELEVANT SITES**

AlignACE. http://atlas.med.harvard.edu/cgi-bin/alignace.pl

CONSENSUS. http://ural.wustl.edu/consensus


MEME. http://meme.sdsc.edu/meme/website