BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 16–23

DNA binding sites: representation and discovery

Gary D. Stormo

Department of Genetics, Washington University Medical School, St. Louis, MO 63110, USA

Abstract

The purpose of this article is to provide a brief history of the development and application of computer algorithms for the analysis and prediction of DNA binding sites. This problem can be conveniently divided into two subproblems. The first is, given a collection of known binding sites, develop a representation of those sites that can be used to search new sequences and reliably predict where additional binding sites occur. The second is, given a set of sequences known to contain binding sites for a common factor, but not knowing where the sites are, discover the location of the sites in each sequence and a representation for the specificity of the protein.
Contact: *stormo@ural.wustl.edu*

At least since the discovery of the *lac* operon, and the realization that its expression was regulated by a protein factor, a major objective in molecular biology has been to understand sequence-specific binding of transcription factors. The original sequencing of the *lac* operator (Gilbert and Maxam, 1973; Maizels, 1973; Dickson *et al.*, 1975) and of the λ operators (Maniatis *et al.*, 1974, 1975a; Walz and Pirrotta, 1975) as well as some other promoter regions (Pribnow, 1975) were significant accomplishments, especially when one considers the laborious methods required in the days before rapid DNA sequencing. After efficient DNA sequencing methods were introduced (Maxam and Gilbert, 1977; Sanger *et al.*, 1977), there was a rapid increase in the number of examples of binding sites. Footprinting methods (Galas and Schmitz, 1978) and efficient methods of synthesizing DNA of any desired sequence (Goeddel *et al.*, 1977) were other technological breakthroughs that helped provide rapidly increasing amounts of data. By the late 1970s there were many sequenced examples of DNA binding sites, including many mutated sites with altered activities. The total amount of DNA being sequenced annually was increasing rapidly (as is true even now) and it was clear that computer programs were needed to help identify important features in the sequences. In parallel with the experimental work, theoretical analyses were undertaken to describe the amount of information necessary for a regulatory system to function properly (Lin and Riggs, 1975; von Hippel, 1979), and those analyses also contributed to the perspectives on representing the specificity of protein-DNA interactions.

Representing sites

One distinction became clear early on, which is that the specificity of a regulatory protein is quite different from that of a restriction enzyme. The recognition sequence for a restriction enzyme can be written as a simple DNA sequence, such as GAATTC for the enzyme *Eco*RI, or a sequence that allows ambiguity, such as GTYRAC for the enzyme *Hinc*II. All sites that match those sequences will be cut (unless modified) and only matching sites will be cut. Sites that mismatch at only one position will be cut less well by several orders of magnitude (except under unusual buffer conditions where they can be more tolerant of changes). Regulatory sites, on the other hand, often have differences between any two sites. For example, in the λ operators, which contain 12 half-sites that one expects will interact with the proteins nearly equivalently, only two of the eight positions are conserved among all the sites, and the other positions have a range of variability (Maniatis *et al.*, 1975b). The collection of the first six *Escherichia coli* promoter −10 regions had only two conserved positions out of six, and differed at as many as four positions between two sites (Pribnow, 1975). Despite the variability, the common features of the −10 sites were visible because each example site was similar, with no more than two mismatches, to a 'consensus sequence'. It makes biological sense that regulatory sites should be variable, whereas restriction sites should not. Restriction enzymes are used as defense mechanisms to protect against viral infection, and they need to have an all or none activity. They should cut any DNA sites that are not protected by the cell's own modification system, and at the same time be so specific so as to not make any cuts in the cell's own DNA. But regulatory systems can take advantage of the variability in the sites to better control transcription. Not all promoters should have the same activity because some proteins are required by the cell at much higher levels than other proteins. The variability in expression can be partially attained by having promoters with different intrinsic affinities for the RNA polymerase, which implies different sequences in the binding sites. Likewise, regulatory proteins often control the expression

of several genes, but those genes may need to be expressed at different levels, or may need to be regulated to different extents. That too can be accomplished by having sites with different sequences and different affinities for the protein. Of course the variability in the sites leads to the complication that regulatory proteins have non-neglible affinity for DNA at positions in the genome besides their functional sites. Such non-specific DNA competes for binding to the protein with the sites, and requires that more protein be synthesized than might otherwise be necessary. Considerations such as these, and how they influence the specificity of the proteins and the amount of protein needed for proper functioning of the regulatory system were the subject of theoretical analyses (Lin and Riggs, 1975; von Hippel, 1979).

The concept of the consensus sequence has been widely used to represent the specificity of transcription factors. But exactly how one is defined is somewhat arbitrary. In general it refers to a sequence that matches all of the example sites closely, but not necessarily exactly. There is a trade-off between the number of mismatches allowed, the ambiguity in the consensus sequence, and the sensitivity and precision of the representation. Consider the six −10 regions shown in Figure 1. If we use TATAAT as a consensus sequence and allow no mismatches, we would only identify two of six sites, and there would be about one match per 4000 bp of genomic sequence. If we allowed one mismatch, we would identify three of six sites, and there would be about one match per 200 bp of genomic sequence. We have to allow two mismatches to identify all six of six sites, but then there would be a match about every 30 bp in genomic sequence. If instead we use TATRNT as the consensus and allow no mismatches, we would identify four of six sites and only have one match per 500 bp in the genome. Allowing only one mismatch would identify all six of six sites, but again with about one match per 30 bp in the genome. Other possible consensus sequences can be defined as well, including some in which the allowed mismatches are confined to certain positions rather than allowing them at all positions, although this approach is uncommon. So while it is easy to write a consensus sequence to represent a collection of sites, it is not so straightforward to find one that is optimal for predicting the occurrence of new sites. Day and McMorris (1992) compared several methods for generating consensus sequences and outlined their strengths and weaknesses.

An alternative to consensus sequences is a weight matrix representation of the sites. Figure 2 shows a weight matrix to represent the −10 region. There is a matrix element for all possible bases at every position in the site. The score for any particular site is the sum of matrix values for that site's sequence. For example, the score for the consensus sequence TATAAT is the sum of the boxed elements, 85.

```
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

TATAAT  consensus sequence
TATRNT  alternate consensus sequence
```

Fig. 1. The −10 region of the six promoters from Pribnow (1975), and two possible consensus sequence representations.

A	−38	**19**	1	**12**	**10**	−48
C	−15	−38	−8	−10	−3	−32
G	−13	−48	−6	−7	−10	−48
T	**17**	−32	**8**	−9	−6	**19**

Fig. 2. Weight matrix represenation for −10 region of *E.coli* promoters. The boxed elements correspond to the consensus sequence TATAAT.

Any sequence that differs from the consensus will have a lower score, but the decrease depends on the differences. This is a convenient way to account for the fact that some positions are more highly conserved than others, and presumably are more important for the activity of the site. It is important to note that a consensus sequence can always be converted into a weight matrix such that the same set of sites will be matched, but the converse is not true. There is still the issue of what threshold one would use to predict sites, and the same concerns for sensitivity and precision need to be addressed. Also, the major issue with weight matrix methods is how to pick the elements of the matrix to represent the sites. The matrix in Figure 2 does not come from the six examples shown in Figure 1, but rather is based on a much larger collection of −10 regions that were published several years later (Hawley and McClure, 1983; Stormo, 1988). Several methods have been proposed to determine the weights for any particular collection of sites, as described below. But regardless of how the matrix weights are determined, there are efficient methods for calculating the distribution of scores that can be used to determine statistically significant matches (Staden, 1989; Claverie and Audic, 1996).

The first usage of weight matrices was actually not for DNA sites, but for RNA sites that function as translation initiation sites in *E.coli* (Stormo *et al.*, 1982b). Shine and Dalgarno (1974) had sequenced the 3′ end of the 16S rRNA and found that it was complementary to a short

sequence upstream of the initiation codon for several genes (Steitz, 1969). We were studying translation initiation and wondered whether the Shine/Dalgarno sequence (as the complementary regions became called) and the initiation codon, usually an AUG, were sufficient to identify ribosome binding sites. As more sequences were published many examples of sites that had Shine/Dalgarno sequences upstream of AUGs appeared and yet, to the best of our knowledge, did not function as translation initiation sites. Because these were mRNA sequences there could be a contribution of the secondary structure in determining whether a particular site could function as a ribosome binding site, and indeed there are examples where the structure can act to block ribosome binding (de Smit and van Duin, 1990). But we wondered whether there were other sequence features that could be used to distinguish true ribosome binding sites from other sites with similar sequences. We collected all of the available *E.coli* and coliphage sequences into a database (Schneider *et al.*, 1982) and attempted to find sequence patterns that would distinguish the true sites from 'non-sites'. We tried many different consensus sequences, including a rule-based system that incorporated several different consensus sequences into a single predictor (Stormo *et al.*, 1982a). While this approach gave improvments over simple consensus sequences, it was still not completely reliable. We also noticed that sequences within and around the ribosome binding site, besides the initiation codon and Shine/Dalgarno sequences, were highly biased (Gold *et al.*, 1981). This led us to the hypothesis that maybe many bases in the ribosome binding region of the mRNA could interact with the ribosome and that the probability that it would bind sufficiently well to initiate translation was the sum of all of the contributing interactions. Some interactions were more important than others, and therefore more highly conserved, but many positions around the start codon could influence its activity. Sites whose total contribution exceeded some threshold would be *bona fide* translation initiation sites, and those below the threshold would not. Thus was born the idea of a weight matrix as a representation of a collection of functional sites and the specificity of the protein that bound to them.

The first major task, of course, was to determine the appropriate values of the weights for the matrix. We wanted a matrix that was capable of distinguishing true sites from non-sites and we had many examples of each in our database. It was then that Andrzej Ehrenfeucht, a professor of computer science, suggested we try a 'Perceptron' algorithm (Stormo *et al.*, 1982b). This is a simple neural network that learns from examples. In our case, the weights of the matrix are the same as the weights of the network, and we train it on our example sites and non-sites to find a matrix and a threshold that distinguishes the two sets. We were able to find such matrices, a result that is not too surprising since we had so many free parameters, all of the weights in the matrix, and a relatively limited amount of training data. However, the result that convinced us that the idea had real merit was that when we searched new sequences not included in the set we used for training the weights, the matrix method was both more sensitive and more precise than the best consensus method available (Stormo *et al.*, 1982b).

In the next 2 years three papers were published that used alternative methods of obtaining weight matrices from purely statistical analyses of example *E.coli* promoters (Harr *et al.*, 1983; Mulligan *et al.*, 1984; Staden, 1984). The method introduced by Staden is very similar to current methods. In fact, except for not allowing insertions and deletions within the sites, it is the same method commonly used in hidden Markov models of protein families (Eddy, 1998). In this method the weights are simply the negative logarithms of the frequencies of each base at each position. So the sum for any particular site is just the negative logarithm of the probability of observing that particular sequence in the collection of known sites (assuming the positions are independent).

In Mulligan *et al.* (1984) it was shown that there was a strong correlation between the score for any particular sequence and its activity as a promoter. This demonstrated that the model of interaction implied by the matrix approach was, at least, not unreasonable. If the weights really correspond to features involved in the recognition process, then having more 'good' features should lead to higher activity. We then realized that if one had sufficient quantitative data, in the form of many sequences and the functional activity of each one, you could simply solve for the matrix weights that gave a best fit to that quantitative data (Stormo *et al.*, 1986). One advantage of this approach is that one can also determine if the best fit is actually good. It might not be if the underlying model is not appropriate. For example, in the standard weight matrix the scores for each position are added together to get the total score, which implies that each position contributes independently to the activity. If that assumption is not a good approximation then even the best fit will not be very good. In that case one can make more complicated models, for instance where the elements of the matrix correspond to di-nucleotides at the positions in the sites, rather than mono-nucleotides (Stormo *et al.*, 1986; Zhang and Marr, 1993). So this method not only obtains the best matrix for the available quantitative data, but can indicate something about the mechanism. The limitation is that quantitative data for many example sites is laborious to obtain, so this approach has been used only rarely (e.g. see Barrick *et al.*, 1994).

At the same time Tom Schneider was examining several different regulatory systems for which many binding

sites were known. He was primarily interested in the 'information content' of the sites, and how that compared with their frequency in the genome (Schneider *et al.*, 1986). The information content at a position in a site was defined to be

$$I_i = 2 + \sum_{b=A}^{T} f_{b,i} \log_2 f_{b,i} \tag{1}$$

where i is the position within the site, b refers to each of the possible bases, and $f_{b,i}$ is the observed frequency of each base at that position. I_i is between 0, for positions that are 25% of each base, and 2 $bits$ for positions completely conserved as one base. Furthermore, for most of the *E.coli* sites studied, the total information content matched very closely their frequency in the genome (Schneider *et al.*, 1986).

Berg and von Hippel (1987), employing statistical mechanics theory, showed that the logarithms of the base frequencies should be proportional to the binding energy contribution of the bases. This idea fits nicely with the information content analysis and suggested that the information content was related to the average binding energy for the collection of sites. However, equation (1) and the analysis presented by Berg and von Hippel are only appropriate for genomes with 25% of each base. Some species have very biased genome compositions, such as 64% A+T for *Saccharomyces cerevisiae*. In such a genome equation (1) would indicate positive information content, and therefore specific binding energy, for any collection of randomly chosen 'sites'. However, a more general form of equation (1) accounts for the genomic base probabilities:

$$I_{\text{seq}}(i) = \sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \tag{2}$$

where p_b is the frequency of base b in the whole genome. Equation (1) is a special case of this formula with $p_b = 0.25$ for all b (Schneider *et al.*, 1986; Stormo, 1988, 1990). I_{seq} is also known as the relative entropy and the Kullback–Liebler distance. It is also a normalized log-likelihood ratio statistic and so can be used to estimate the statistical significance of the pattern (Stormo, 1990). However, we came upon the best justification for using equation (2) as the estimate for binding energy contributions a few years ago. Consider an example where we have a collection of functional binding sites that are each known to have high affinity for the protein, although the exact affinity is not known. Suppose we also know the complete genome sequence of the organism that the protein and the sites are from. Following the additivity assumption that each position contributes independently to the total binding energy, there is some matrix $H(b, i)$ that contains those binding energy contributions as its elements. Given any particular sequence S_α, its total binding energy is then given by $H(b, i) \cdot S_\alpha$, where the dot-product of the matrix and the sequence is as shown in Figure 2. The probability that the protein would be bound to the site with sequence S_α, considering all of the possible binding sites in the whole genome, is

$$P(S_\alpha \texttt{ is bound}) = \frac{\mathrm{e}^{-H(b,i)\cdot S_\alpha}}{Z} \tag{3}$$

where Z is the partition function, the sum of the binding affinities over all the sites in the genome. Since we know that the functional sites in our collection must have high binding probabilities, we are justified in finding the matrix that maximizes the probability of binding to all of those sites. If we make one further assumption this is easy. If we assume that the genome is essentially random, then Z can be calculated analytically (Heumann *et al.*, 1994). Genomes are not random sequences, but the assumption is valid if short subsequences, the length of the binding site, occur with frequencies expected from the base composition of the genome. At that level, the assumption of random genomes is often not too bad an approximation. Also under that assumption it is easy to show that the elements of $H(b, i)$ that maximize the probability of binding to the collection of known functional sites is simply (Heumann *et al.*, 1994)

$$H(b, i) = -\ln \frac{f_{b,i}}{p_b} \tag{4}$$

Therefore, if one has only a collection of known binding sites for a particular protein, $(-\ln f_{b,i}/p_b)$ is a maximum probability estimate for the binding energy contribution of each base at each position, and I_{seq} is the average binding energy of all the known sites (Stormo and Fields, 1998).

One limitation of the weight matrix approach is the assumption that the positions in the site contribute additively to the total activity. More complicated models are possible within the framework of the matrix method, as mentioned above (Zhang and Marr, 1993; Stormo *et al.*, 1986). However, those require some prior information about what positions are non-independent. An alternative is to use a more general neural network than the simple Perceptron described earlier. A neural network that contains hidden layers, not connected to either the input sequence or the output score, are capable of discovering correlations in the data and taking advantage of them for purposes of discrimination (Horton and Kanehisa, 1992). Because promoter prediction in *E.coli* remains a difficult problem, with even the best matrix methods being less sensitive and less precise than desired, several groups have used general neural networks to try and get better discrimination (Demeler and Zhou, 1991; O'Neill, 1991; Horton and Kanehisa, 1992).

These methods were able to show improved discrimination on the training data, but when tested on new, independent data did not show significant improvements over the simple weight matrix methods (Horton and Kanehisa, 1992).

Discovering sites

The same division of approaches, between consensus sequences and weight matrices, can be used to classify methods for pattern recognition. In this problem one has a collection of sequences that are known to contain binding sites for a common factor, but neither the positions of the sites nor the specificity of the factor are known. Such data might be obtained through genetic or biochemical means, but the recent invention of expression array techniques provides a new, rapid method of identifying sets of genes that appear to be coregulated (Spellman *et al.*, 1998; Lashkari *et al.*, 1997; DeRisi *et al.*, 1997). Hence there have recently been several papers describing methods for finding transcription factor patterns for collections of genes.

Consensus approaches to the problem of site discovery really go back to the original papers that sequenced *E.coli* promoter regions. From those few sequences the −10 and the −35 consensus sequences were determined 'by eye' (Pribnow, 1975; Rosenberg and Court, 1979). That was possible because there were only a few sequences and they could be approximately aligned because the start of transcription was known, at least for many of the sequences. It was observed that all of the sequences had very similar sites at two locations, approximately 10 and 35 bases upstream of the start. However, as more sequences were collected and with less information available about how to align them, computer algorithms were required to locate the important features. The first such algorithm was by Galas *et al.* (1985) who looked for common words and their 'neighbors', which are approximate matches to those words, over a window of possible alignments. A similar algorithm was also developed by Mengeritsky and Smith (1987). Within the past several years a variety of algorithms designed to identify consensus sequences from unaligned DNA sequence have emerged (Staden, 1989; Pesole *et al.*, 1992; Roytberg, 1992; Frech *et al.*, 1993; Lefevre and Ikeda, 1993; Ulyanov and Stormo, 1995; Wolfertsteter *et al.*, 1996; Rigoutsos and Floratos, 1998). A review of these methods, and their use for both DNA and protein motifs, was published recently (Brazma *et al.*, 1998a). These consensus methods have been applied to collections of yeast genes known to be coregulated, or expected to be coregulated based on expression array analysis (van Helden *et al.*, 1998; Brazma *et al.*, 1998b). On control sets, where the correct patterns are known, the methods usually perform quite well. So it can be expected that patterns they identify that are not already known may also correspond to new transcription factor binding sites.

The alternative approach is to search directly for a weight matrix that serves to discriminate well between the sequences known to be coregulated and other, mostly unregulated sequences. We originally developed a method to do that using a greedy algorithm that builds up an entire alignment of the sites by adding in new ones at each iteration (Stormo and Hartzell, 1989; Hertz *et al.*, 1990). The criterion for identifying the best alignment of potential sites was to choose the one with highest information content I_{seq}. Recent advances allow us to calculate a p-value for the alignment and use that as the criterion to rank different alignments (Hertz and Stormo, 1999). An expectation–maximization (EM) method was developed for the same problem by Lawrence and Reilly (1990). The EM approach can be described briefly as an iteration between two steps. As shown in Figure 2, given a matrix one can calculate the score for all possible binding sites on a sequence. Using that score one obtains a weighted alignment of all the possible sites. The alignment is used to derive a new matrix representation for those sites. Those two steps are repeated until the method converges, which it is guaranteed to do. While not guaranteed to always find the optimal alignment of sites, it generally works quite well and often finds the correct sites and the matrix to represent them. Bailey, Grundy and Elkan have also developed an EM approach to this problem (Bailey and Elkan, 1994, 1995; Grundy *et al.*, 1996), which is implemented in the MEME program. The MEME method allows for the simultaneous identification of multiple patterns. Lawrence and colleagues also developed a 'Gibbs' Sampling' variation of the EM method (Lawrence *et al.*, 1993) which has also been used to define weight matrices for known transcription factors (Schug and Overton, 1997). Zhang has recently developed a version of the EM method and used it on sets of coregulated yeast genes (Spellman *et al.*, 1998; Zhu and Zhang, 1999) and a similar approach has been used on *E.coli* (Robison *et al.*, 1998).

In most of these methods to identify the weight matrix directly from the unaligned sequences, the criterion for the best alignment is the one with maximum information content. This compensates for the base composition of the genome and often identifies the proper sites. However, sometimes the assumption of a random genome is too poor an approximation, and instead of finding the proper sites the methods identify some other patterns that appear to be significant but are not discriminatory for the promoters in the collection. For example, many yeast promoters have unexpectedly common stretches of poly(A) or poly(T) sequences, and those can appear as the patterns identified by the programs. But those patterns occur in many promoters, not just the subset known to be coregulated,

and so cannot be the binding site of interest. Under such circumstances, the method using the partition function, briefly described above, can be used. It looks for the weight matrix that maximizes the probability of binding to the promoters in the collection, given the background of actual competing sites in the genome (Heumann *et al.*, 1994). This approach has been used on several sets of yeast genes and shown to work even when some of the other methods have failed (Workman and Stormo, 2000).

Summary and prospects

Over the last 25 years an enormous amount has been learned about transcription factor interactions with DNA sequences. Many of their structures have been solved by crystallography and there are many more known binding sites for a large collection of factors. Although not perfect, methods for representing the specificity of the factors are generally pretty reliable and can be used to search genomic DNA to predict new potential binding sites. The largest problem is that there tend to be many false positives in such searches. These are probably sites that would bind to the protein if they were available, but they are probably sequestered most of the time. The pattern recognition methods also work fairly well, and can usually be relied upon to uncover at least the sites involved in the coregulation of the collection of identified genes. However, we are still not able to reliably determine the complete set of regulatory interactions for complicated promoters typical of metazoans. These are usually regulated by multiple factors, often interacting cooperatively with one another. So there are still significant challenges to solve in order for us to take full advantage of the genomic sequences that are being determined increasingly rapidly, especially to be able to infer regulatory networks from the sequence alone, or even using both the sequence and expression information.

References

Bailey,T.L. and Elkan,C.P. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Intell. Sys. Mol. Biol.*, **2**, 28–36.

Bailey,T.L. and Elkan,C.P. (1995) The value of prior knowledge in discovering motifs with. *Intell. Sys. Mol. Biol.*, **3**, 21–29.

Barrick,D., Villanueba,K., Childs,J., Kalil,R., Schneider,T.D., Lawrence,C.E., Gold,L. and Stormo,G.D. (1994) Quantitative analysis of ribosome binding sites in *E. coli*. *Nucl. Acids Res.*, **22**, 1287–1295.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1998a) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, **5**, 279–305.

Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, **8**, 1202–1215.

Claverie,J.M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–439.

Day,W.H. and McMorris,F.R. (1992) Critical comparison of consensus methods for molecular sequences. *Nucl. Acids Res.*, **20**, 1093–1099.

Demeler,B. and Zhou,G.W. (1991) Neural network optimization for *E. coli* promoter prediction. *Nucl. Acids Res.*, **19**, 1593–1599.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

de Smit,M.H. and van Duin,J. (1990) Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci. USA*, **87**, 7668–7672.

Dickson,R.C., Abelson,J., Barnes,W.M. and Reznikoff,W.S. (1975) Genetic regulation: the Lac control region. *Science*, **187**, 27–35.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Frech,K., Herrmann,G. and Werner,T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucl. Acids Res.*, **21**, 1655–1664.

Galas,D.J. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding secificity. *Nucl. Acids Res.*, **5**, 3157–3170.

Galas,D.J., Eggert,M. and Waterman,M.S. (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. *J. Mol. Biol.*, **186**, 117–128.

Gilbert,W. and Maxam,A. (1973) The nucleotide sequence of the lac operator. *Proc. Natl. Acad. Sci. USA*, **70**, 3581–3584.

Goeddel,D.V., Yansura,D.G. and Caruthers,M.H. (1977) Binding of synthetic lactose operator DNAs to lactose represessors. *Proc. Natl. Acad. Sci. USA*, **74**, 3292–3296.

Gold,L., Pribnow,D., Schneider,T., Shinedling,S., Singer,B.S. and Stormo,G. (1981) Translational initiation in prokaryotes. *Ann. Rev. Microbiol.*, **35**, 365–403.

Grundy,W.N., Bailey,T.L. and Elkan,C.P. (1996) ParaMEME: a parallel implementation and a web interface for a and protein motif discovery tool. *Comput. Appl. Biosci.*, **12**, 303–310.

Harr,R., Haggstrom,M. and Gustafsson,P. (1983) Search algorithm for pattern match analysis of nucleic acid sequences. *Nucl. Acids Res.*, **11**, 2943–2957.

Hawley,D.K. and McClure,W.R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucl. Acids Res.*, **11**, 2237–2255.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hertz,G.Z., Hartzell,G.W.III and Stormo,G.D. (1990) Identification of consensus patterns in unaligned sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Heumann,J.M., Lapedes,A.S. and Stormo,G.D. (1994) Neural networks for determining protein specificity and multiple alignment of binding sites. *Intell. Sys. Mol. Biol.*, **2**, 188–194.

Horton,P.B. and Kanehisa,M. (1992) An assessment of neural network and statistical approaches for prediction of *E. coli*

promoter sites. *Nucl. Acids Res.*, **20**, 4331–4338.

Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R.W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA*, **94**, 13057–13062.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lefevre,C. and Ikeda,J.E. (1993) Pattern recognition in DNA sequences and its application to consensus foot-printing. *Comput. Appl. Biosci.*, **9**, 349–354.

Lin,S. and Riggs,A.D. (1975) The general affinity of lac repressor for *E. coli* DNA: implications for gene regulation in procaryotes and eucaryotes. *Cell*, **4**, 107–111.

Maizels,N.M. (1973) The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **70**, 3585–3589.

Maniatis,T., Jeffrey,A. and Kleid,D.G. (1975a) Nucleotide sequence of the rightward operator of phage lambda. *Proc. Natl. Acad. Sci. USA*, **72**, 1184–1188.

Maniatis,T., Ptashne,M., Barrell,B.G. and Donelson,J. (1974) Sequence of a repressor-binding site in the DNA of bacteriophage lamda. *Nature*, **250**, 394–397.

Maniatis,T., Ptashne,M., Backman,K., Kleid,D., Flashman,S., Jeffrey,A. and Maure,R. (1975b) Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell*, **5**, 109–113.

Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, **74**, 560–564.

Mengeritsky,G. and Smith,T.F. (1987) Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. *Comput. Appl. Biosci.*, **3**, 223–227.

Mulligan,M.E., Hawley,D.K., Entriken,R. and McClure,W.R. (1984) *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. *Nucl. Acids Res.*, **12**, 789–800.

O'Neill,M.C. (1991) Training back-propagation neural networks to define and detect DNA-binding sites. *Nucl. Acids Res.*, **19**, 313–318.

Pesole,G., Prunella,N., Liuni,S., Attimonelli,M. and Saccone,C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucl. Acids Res.*, **20**, 2871–2875.

Pribnow,D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA*, **72**, 784–788.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284**, 241–254.

Rosenberg,M. and Court,D. (1979) Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu. Rev. Genet.*, **13**, 319–353.

Roytberg,M.A. (1992) A search for common patterns in many sequences. *Comput. Appl. Biosci.*, **8**, 57–64.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Schneider,T.D., Stormo,G.D., Haemer,J.S. and Gold,L. (1982) A design for computer nucleic-acid sequence storage, retrieval and manipulation. *Nucl. Acids Res.*, **10**, 3013–3024.

Schug,J. and Overton,G.C. (1997) Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. *Intell. Sys. Mol. Biol.*, **5**, 268–271.

Shine,J. and Dalgarno,L. (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA*, **71**, 1342–1346.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.*, **12**, 505–519.

Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.*, **5**, 89–96.

Steitz,J.A. (1969) Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature*, **224**, 957–964.

Stormo,G.D. (1988) Computer methods to identify recognition sequences. *Ann. Rev. Biophys. Biophys. Chem.*, **17**, 241–263.

Stormo,G.D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, **183**, 211–221.

Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Stormo,G.D. and Hartzell,G.W.III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, **86**, 1183–1187.

Stormo,G.D., Schneider,T.D. and Gold,L.M. (1982a) Characterization of translational initiation sites in *E. coli*. *Nucl. Acids Res.*, **10**, 2971–2996.

Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucl. Acids Res.*, **14**, 6661–6679.

Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982b) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucl. Acids Res.*, **10**, 2997–3012.

Ulyanov,A. and Stormo,G.D. (1995) Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions. *Nucl. Acids Res.*, **23**, 1434–1440.

van Helden,J., André,B. and Collodo-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol.*

Biol., **281**, 827–842.
von Hippel,P.H. (1979) On the molecular bases of the specificity of interaction of transcriptional proteins with genome DNA. *Biol. Regul. Develop.*, **1**, 279–347.
Walz,A. and Pirrotta,V. (1975) Sequence of the PR promoter of phage lambda. *Nature*, **254**, 118–121.
Wolfertstetter,F., Frech,K., Herrmann,G. and Werner,T. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.*, **12**, 71–80.
Workman,C.T. and Stormo,G.D. (2000) *Pacific Symposium on Biocomputing*, **5**, 112–123.
Zhang,M.Q. and Marr,T.G. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, **9**, 499–509.
Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.