Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 1183–1187, February 1989
Biochemistry

Identifying protein-binding sites from unaligned DNA fragments

(specificity/regulatory sites/pattern recognition/information theory)

GARY D. STORMO AND GEORGE W. HARTZELL III

Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309

Communicated by William B. Wood, November 14, 1988

ABSTRACT **The ability to determine important features within DNA sequences from the sequences alone is becoming essential as large-scale sequencing projects are being undertaken. We present a method that can be applied to the problem of identifying the recognition pattern for a DNA-binding protein given only a collection of sequenced DNA fragments, each known to contain somewhere within it a binding site for that protein. Information about the position or orientation of the binding sites within those fragments is not needed. The method compares the "information content" of a large number of possible binding site alignments to arrive at a matrix representation of the binding site pattern. The specificity of the protein is represented as a matrix, rather than a consensus sequence, allowing patterns that are typical of regulatory protein-binding sites to be identified. The reliability of the method improves as the number of sequences increases, but the time required increases only linearly with the number of sequences. An example, using known cAMP receptor protein-binding sites, illustrates the method.**

Gene expression is often controlled by protein factors that interact with DNA regions to affect transcription. Understanding the regulation of the expression requires knowing both the protein factors and the DNA sites at which they act. The sites have traditionally been determined by isolating cis-acting mutations that affect expression and then determining the changes in the DNA that accompany the mutant phenotypes (1). More recently, regulatory proteins have been used to affinity purify the DNA regions to which they bind, and "footprinting" techniques have further delimited the binding sites (2). Each of these methods is time consuming and gives only partial information about the binding site. The final determination of the binding site pattern usually includes a comparison of many example sites. A method to determine the binding sites from the DNA sequences alone could greatly facilitate the process. Current sequencing technology is rapid enough that the most efficient means of determining the binding specificity of a protein may be to sequence a collection of regions known to contain binding sites. These may be a set of restriction fragments that are shown to bind the protein or a collection of DNA segments to which binding sites have been mapped. Since each of the fragments contains a binding site, the pattern of bases recognized by the protein should be discernible as the most significant pattern in the collection.

Regulatory Patterns

The difficulty arises that binding site patterns are not usually simple strings of bases. For example, *Escherichia coli* promoter sequences have two highly conserved parts, called the −35 and −10 regions (3, 4). The consensus sequences for those are TTGACA and TATAAT, respectively. The consensus spacing between those regions is 17 bases, but other spacings are also allowed. An individual promoter may match the consensus at only a few positions and, while some positions are more conserved than others, no position is absolutely conserved. The most conserved bases in the −10 region are TAnnnT, but only ≈65% of all promoters even match this limited criterion (4). This means that methods of identifying the binding pattern that rely on common substrings, or "words," will likely fail.

A better representation of a protein's binding specificity than a consensus sequence is a matrix that has an element for each possible base at each position of the site (5). The matrix elements represent contributions of the individual bases to the protein–DNA interaction. The affinity of the protein for any site depends on the sum of all the interactions between the DNA and the protein. The individual interactions may or may not be independent; the simplest representation assumes independence. If the sequences of several binding sites are known, they can be used to construct a matrix representation of the protein that will give scores to individual sites that correlate well with the relative binding affinities of those sites (5–8).

The sequences of 23 identified binding sites of cAMP receptor protein (CRP) are shown in Fig. 1*A* (8, 9). Fig. 1*B* is a matrix whose elements are the frequency that each base occurs at each position within the CRP-binding sites. Fig. 1*C* is the matrix representation of CRP specificity, based on the information at each position of the site (5, 7, 8). The matrix is a representation of the specificity of the binding protein and can be used to search for new sites (5, 11–13). It can also be used to rank the affinities of different sites, and matrices of this type usually do well as predictors of quantitative activity (5–8). Fig. 1*D* shows the "information content" at each position of the site (10), derived from the formula

$$I_{seq} = \sum_{b=A}^{T} f_b \log_2\left(\frac{f_b}{p_b}\right), \qquad [1]$$

where f_b is the observed frequency of each base in the collection of sites and p_b is the fraction of each base in the genome. Note that the f_b terms are the elements of the matrix of Fig. 1*B*, and the $\log_2(f_b/p_b)$ terms are the elements of the matrix of Fig. 1*C*. Therefore, the "information content" plotted in Fig. 1*D* is the dot product of the position (column) vectors from those two matrices. The "information content" is a measure of how constrained the choice of bases is at each position in the binding sites (5, 7, 8, 10).

The Algorithm

The problem of identifying the binding sites from a collection of unaligned sequences is equivalent to finding an alignment that maximizes the "information content," at least within a local "window" that is the width of the binding site. That is, when the sequences are aligned by their binding sites there will be a peak of "information content," as in Fig. 1*D*, that

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: CRP, cAMP receptor protein.

A

Col E1 site 2	T	T	T	T	G	T	G	G	C	A	T	C	G	G	G	C	G	A	G	A	A	T
Col E1 site 1	T	T	T	T	T	T	G	A	T	C	G	T	T	T	T	C	A	C	A	A	A	A
ara site 2	A	A	A	A	G	T	G	T	C	T	A	T	A	A	T	C	A	C	G	G	C	A
ara site 1	T	T	A	T	T	T	G	C	A	C	G	G	C	G	T	C	A	C	G	C	T	T
bgl R mut	A	A	C	T	G	T	G	A	G	C	A	T	G	G	T	C	A	T	A	T	T	T
crp	G	T	A	T	G	C	A	A	A	G	G	A	C	G	T	C	A	C	A	T	T	A
cya	A	G	G	T	G	T	T	A	A	A	T	T	G	A	T	C	A	C	G	T	T	T
deo P2 site 2	T	T	A	T	T	T	G	A	A	T	C	A	G	T	T	C	G	C	A	T	T	A
deo P2 site 1	A	A	T	T	G	T	G	A	T	T	C	G	T	A	T	C	G	A	A	G	T	G
gal	T	A	A	T	T	T	A	T	T	C	C	A	T	C	T	C	A	C	A	C	T	T
ilv B	A	A	A	C	G	T	G	A	T	G	A	A	C	C	T	C	T	C	A	A	T	T
lac site 1	T	A	A	T	G	T	G	A	G	C	A	A	G	C	T	C	A	C	T	C	A	T
lac site 2	A	A	T	T	G	T	G	A	G	C	T	G	A	T	A	C	A	C	A	C	T	T
mal E	T	T	C	T	G	T	A	A	C	A	G	A	G	A	T	C	A	C	A	C	A	A
mal K	T	T	C	T	G	T	G	A	A	C	T	A	A	A	T	C	A	C	A	C	G	T
mal T	A	A	T	T	G	T	G	A	C	A	C	A	G	T	G	C	A	A	A	T	T	C
omp A	A	T	G	C	C	T	G	A	C	T	G	C	T	G	T	C	A	C	A	T	T	T
tna A	G	A	T	T	G	T	G	A	T	T	G	C	G	G	A	C	A	C	A	T	T	T
uxu AB	T	G	T	T	G	T	G	A	T	T	G	T	G	G	T	C	A	C	C	C	A	A
pBR P4	C	G	G	T	G	T	G	A	A	A	A	T	A	C	C	G	A	C	A	A	T	A
cat site 2	A	C	C	T	G	T	G	A	A	C	G	G	A	A	G	T	C	G	C	T	C	T
cat site 1	A	A	A	T	G	A	G	A	C	G	T	T	G	T	A	C	G	A	C	A	T	C
tdc	A	T	T	T	G	T	G	A	G	T	G	G	T	C	G	C	A	C	A	T	A	T

B

A	0.48	0.48	0.39	0.04	0.00	0.04	0.13	0.83	0.26	0.22	0.13	0.48	0.22	0.31	0.09	0.09	0.65	0.26	0.65	0.17	0.30	0.26
C	0.04	0.00	0.13	0.09	0.04	0.04	0.00	0.04	0.30	0.35	0.17	0.04	0.17	0.17	0.09	0.87	0.09	0.65	0.13	0.26	0.09	0.13
G	0.09	0.13	0.13	0.00	0.78	0.00	0.83	0.04	0.17	0.26	0.35	0.26	0.44	0.26	0.17	0.00	0.22	0.04	0.17	0.17	0.00	0.09
T	0.39	0.39	0.35	0.87	0.17	0.91	0.04	0.09	0.26	0.17	0.35	0.22	0.17	0.26	0.65	0.04	0.04	0.04	0.04	0.39	0.61	0.52

C

A	0.94	0.94	0.64	-2.64	-2.75	-2.63	-0.94	1.73	0.07	-0.18	-0.94	0.94	-0.18	0.31	-1.47	-1.47	1.38	0.07	1.39	-0.54	0.26	0.06
C	-2.64	-2.75	-0.94	-1.47	-2.63	-2.63	-2.75	-2.64	0.28	0.49	-0.56	-2.64	-0.56	-0.56	-1.47	1.80	-1.47	1.39	-0.93	0.07	-1.47	-0.94
G	-1.47	-0.94	-0.94	-2.75	1.66	-2.75	1.73	-2.64	-0.54	0.06	0.49	0.06	0.82	0.06	-0.56	-2.75	-0.18	-2.63	-0.54	-0.54	-2.75	-1.47
T	0.64	0.64	0.49	1.80	-0.54	1.88	-2.64	-1.47	0.07	-0.56	0.49	-0.18	-0.56	0.06	1.38	-2.64	-2.64	-2.63	-2.63	0.66	1.29	1.06

D

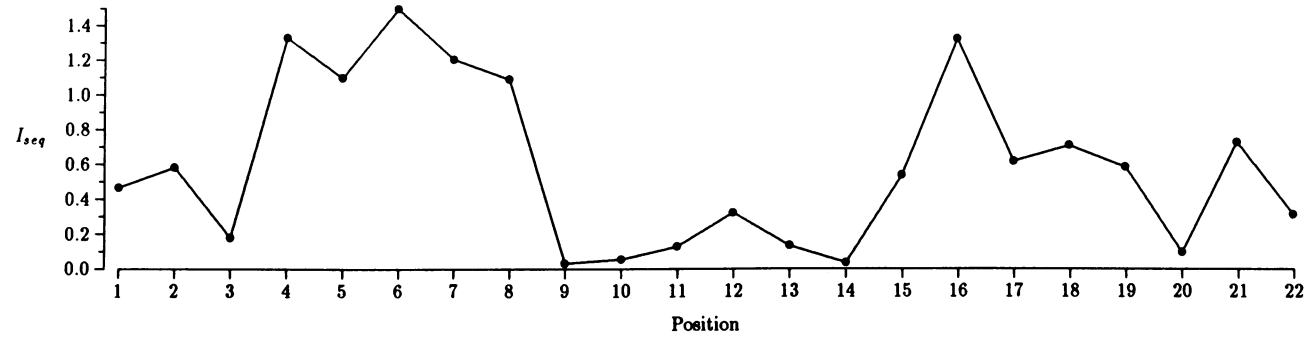

FIG. 1. CRP binding sites. (*A*) Twenty-three sites identified as binding to CRP (8, 9). (*B*) The frequency with which each base occurs at each position in the CRP binding sites. (*C*) The specificity matrix for the protein, based on the binding sites, which is calculated as $\log_2(f_b/p_b)$, where f_b is the observed frequency of each base (from the matrix above) and p_b is the *a priori* probability of obtaining base *b*. In this example p_b = 0.25 for all *b*, approximating the *E. coli* genomic composition. At positions for which $f_b = 0$, an estimated frequency of 0.5/23 is used in the calculation. (*D*) The "information content" at each position of the CRP binding sites is plotted (10). The sum of all positions is 13.06 bits.

is greater (statistically more significant) than other peaks that occur by chance. A rigorous approach to identifying the binding sites would be to calculate, for every possible alignment of the fragments, the "information content." However, if there are N sequences, each of length L, then there are L^N possible alignments to consider, impractical for any interesting case. Methods have been developed for finding significant patterns in approximately aligned sequences (14, 15), but these are not useful when no alignment information is available. An alternative approach is to assume that the specificity of the protein can be represented as a matrix and to search for a matrix that gives at least one high-scoring site on each fragment. We have developed an algorithm that adds sequences to the analysis one at a time, at each step keeping many matrices (23). As more sequences are added to the analysis, the matrix representing the binding sites of the protein emerges from the background of possible alignments as the one with the highest "information content."

The basic algorithm can be outlined as follows:

(*i*) Each of the *k-words* (*k*-long substrings, with *k* a variable chosen by the user) of the first sequence are used as an initial set of "interesting" matrices. These initial matrices contain one element of value 1.00 in each column, corresponding to the base at that position of the sequence, and all other elements are of value 0. Each of these words is a potential binding site, and with no further information they are equally likely.

(*ii*) The next sequence on the list is added to the analysis. Each of the set of interesting matrices is compared with each position of the new sequence.

(*iii*) A new set of interesting matrices is generated by updating some of the previous matrices by the addition of sites from the new sequence.

(*iv*) Steps 2 and 3 are repeated until all sequences have been included in the analysis.

The set of new interesting matrices may be selected in several ways. For instance, a threshold of "information content" could be used, and all new matrices that exceed that threshold would be kept for the next step. Alternatively, if computer memory is adequate to store X matrices, then the X-best might be kept for the next step. We have used a simpler criterion in the example below. Each existing matrix is updated with its best scoring site on the new sequence. When there are ties, which occur when two (or more) sites score highest from a particular matrix, each site is kept. This can lead to a slight increase in the number of matrices at each successive step, but the total number is approximately equal to the length of the first sequence.

The choice of the matrix width is not especially critical; the size of the matrix need not match the binding-site size for the method to work. When the matrix is larger than the binding

```
cole1      taatgtttgtgctggtTTTTGTGGCATCGGGCGAGAATagcgcgtggtgtgaaagactgtTTTTTTGATCGTTTTCACAAAAatggaagtccacagtcttgacag
ecoarabop  gacaaaaacgcgtaacAAAAGTGTCTATAATCACGGCAgaaaagtccacattgaTTATTTGCACGGCGTCACACTTtgctatgccatagcatttttatccataag
ecobglr1   acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaatAACTGTGAGCATGGTCATATTTttatcaat
ecocrp     cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatGTATGCAAAGGACGTCACATTAccgtgcagtacagttgatagc
ecocya     acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaAGGTGTTAAATTGATCACGTTTagaccatttttcgtcgtgaaactaaaaaaacc
ecodeop    agtgaaTTATTTGAACCAGATCGCATTAcagtgatgcaaacttgtaagtagatttccttAATTGTGATGTGTATCGAAGTGtgttgcggagtagatgttagaata
ecogale    gcgcataaaaaacggctaaattcttgtgtaaacgattccacTAATTTATTCCATGTCACACTTttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr  gctccggcggggttttttgttatctgcaattcagtacaAAACGTGATCAACCCCTCAATTttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
ecolac     aacgcaatTAATGTGAGTTAGCTCACTCATtaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggAATTGTGAGCGGATAACAATTTcac
ecomale    acattaccgccaaTTCTGTAACAGAGATCACACAAagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk    ggaggaggcgggaggatgagaacacggcTTCTGTGAACTAAACCGAGGTCatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt    gatcagcgtcgttttaggtgagttgttaataaagatttggAATTGTGACACAGTGCAAATTCagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa    gctgacaaaaaagattaaacatacctttatacaagactttttttttcatATGCCTGACGGAGTTCACACTTgtaagttttcaactacgttgtagactttacatcgcc
ecotnaa    ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacGATTGTGATTCGATTCACATTTaaacaatttcaga
ecouxu1    cccatgagagtgaaatTGTTGTGATGTGGTTAACCCAAttagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4     ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgCGGTGTGAAATACCGCACAGATgcgtaaggagaaaataccgcatcaggcgctc
trn9cat    CTGTGACGGAAGATCACTTCgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaAAATGAGACGTTGATCGGCACG
(tdc)      gattttatactttaacttgttgatatttaaaggtatttaattgtaatаacgatactctggaaagtattgaaagttaATTTGTGAGTGGTCGCACATATcctgtt
```

FIG. 2. The example data set. These are 105-base regions surrounding the sites in Fig. 1*A* (shown capitalized). They were obtained from GenBank Release 55, with the LOCUS name shown at the left. (The *tdc* gene is not in that release; the sequence was obtained from ref. 16.)

site, several matrices will be found that each include the binding sites within. When the matrix is smaller than the binding site, several overlapping matrices will be found that combine to give the entire site. The important consideration is that the matrix be large enough for the information in the binding sites to stand out against the background of other matrices. We have chosen to start with a matrix width of 20 bases; this is a fairly typical size for a binding site, at least for prokaryotic regulatory proteins (10).

An Example

Fig. 2 shows the data used in this example. Each sequence is 105 bases long and contains at least one CRP-binding site. In each case, the DNA strand shown is the one that appears in GenBank, and the order of sequences is alphabetical by GenBank LOCUS name. The regulated genes can occur at either end of these sequences. This example is typical of the type of data one might have from which to deduce binding sites for a protein. Were the orientation of the sites not known and no assumption of symmetry possible, both DNA strands would have to be compared.

The 86 20-long words of the first sequence constitute the initial set of matrices. As described above, each of these words is compared with each of the 20-long words of the next sequence and the best match for each matrix kept as a two-sequence matrix. As there was one tie, 87 matrices are kept at this step. Each of those is then compared with the 20-long words of the next sequence, and the best match to each matrix is kept again. This procedure is followed until all 18 sequences have been included; the total number of matrices at that point is 94. The "information content" of each matrix is calculated by *Eq.* **1**. Fig. 3*A* shows the distribution of "information content" of these 94 matrices. Three matrices stand out as clearly significant above the others. These are overlapping matrices that collectively cover 22 bases, essentially as in Fig. 1. The other cluster of 13 matrices below the most significant, but still above the main distribution, are variations on the binding sites. In some cases they overlap parts of the binding site but not all of it. In other cases they overlap it completely but have some additional sites that decrease the total "information content."

Examination of the three best matrices reveals two things. (*i*) The overlap shows that the site is better represented as a 22-long matrix. (*ii*) The most significant part of each matrix is contained in a central region of only 16 bases, positions 4–19 of Fig. 1. The entire analysis was repeated using each of these alternative matrix widths. In each case a single best matrix was found that stands out above the distribution (Fig. 3 *B* and *C*) and represents the known binding sites. The best scoring 16-wide matrix is shown in Fig. 4. This "information content" is very similar to that obtained from the collection of known binding sites (Fig. 1). The "information content" in Fig. 4*C* is actually greater and somewhat more symmetric than from positions 4–19 of Fig. 1*D*. This difference is due to the inclusion in Fig. 1*A* of several second, and presumably weaker, sites that lower the total information and that are least conserved on the 3′ side. The matrices of Fig. 4 come from only one site per sequence, thereby selecting for the most highly conserved collection of sites.

The matrices of Fig. 4 have the highest "information content" of all the 16-wide matrices obtained by our algorithm. To test whether the matrix of Fig. 4*B* represents the specificity of CRP, every position of each sequence shown in Fig. 2 was evaluated with the matrix (as described in refs. 5 and 11–13). All the identified CRP-binding sites were among the highest scoring sites. With two exceptions, the highest scoring site on each sequence was an identified CRP-binding site. These two exceptions are interesting to examine. The most highly conserved bases in the binding sites constitute a symmetric consensus sequence of TGTGAnnnnnnTCACA (Fig. 1 and ref. 8). The identified CRP-binding site for the *malK* gene has six matches to that, five on the 5′ side (Fig.

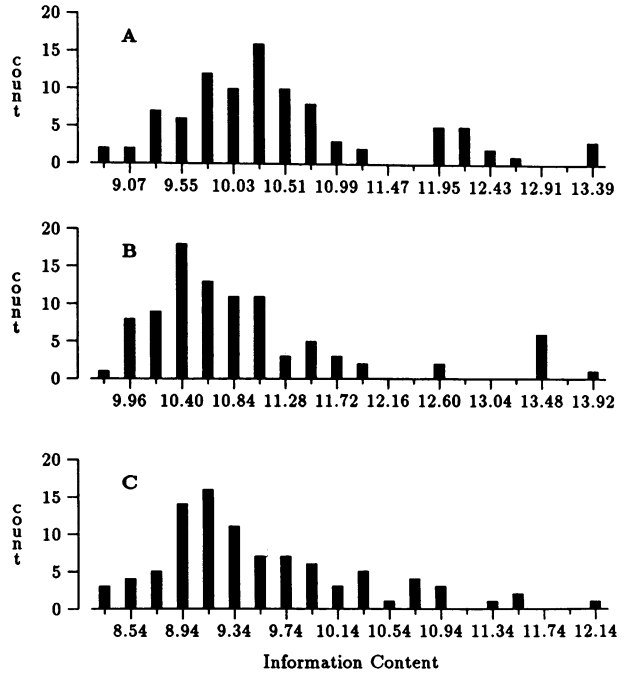

FIG. 3. Distribution of "information content" of the matrices at the end of each analysis, described in the text. Count is the number of matrices with "information content" in the interval shown. (*A*) The 94 20-wide matrices. (*B*) The 93 22-wide matrices. (*C*) The 93 16-wide matrices.

A

A	0.00	0.00	0.00	0.17	0.78	0.28	0.17	0.11	0.50	0.06	0.11	0.00	0.06	0.83	0.11	0.89
C	0.22	0.06	0.06	0.00	0.06	0.22	0.50	0.28	0.00	0.28	0.22	0.06	0.72	0.00	0.72	0.06
G	0.06	0.61	0.00	0.72	0.00	0.11	0.22	0.33	0.17	0.56	0.28	0.11	0.17	0.11	0.00	0.00
T	0.72	0.33	0.94	0.11	0.17	0.39	0.11	0.28	0.33	0.11	0.39	0.83	0.06	0.06	0.17	0.06

B

A	-3.43	-3.43	-3.43	-0.82	1.38	-0.10	-0.82	-1.45	0.74	-2.32	-1.45	-3.43	-2.32	1.47	-1.45	1.57
C	0.29	-1.59	-1.59	-9.97	-1.59	0.29	1.47	0.64	-2.70	0.64	0.29	-1.59	2.00	-2.70	2.00	-1.59
G	-1.81	1.54	-2.92	1.78	-2.92	-0.93	0.07	0.65	-0.31	1.42	0.42	-0.93	-0.31	-0.93	-2.92	-2.92
T	1.22	0.09	1.60	-1.50	-0.87	0.33	-1.50	-0.15	0.09	-1.50	0.33	1.42	-2.37	-2.37	-0.87	-2.37

C

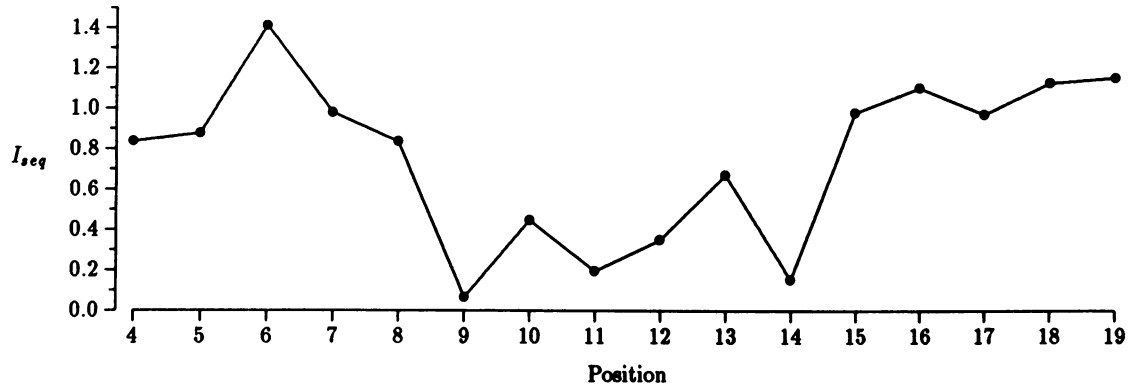

FIG. 4. The best 16-wide matrix. The positions are numbered 4–19, corresponding to the central positions of Fig. 1. (*A*) The frequency of each base for the sites included in the best matrix, as in Fig. 1*B*. (*B*) The specificity matrix determined from the frequency matrix, as in Fig. 1*C*. In this case the *a priori* values were determined from the data set shown in Fig. 2: $p_A = 0.30$; $p_C = 0.18$; $p_G = 0.21$; and $p_T = 0.31$. Analyses were also performed using $p_b = 0.25$ for all b. In that case, essentially the same matrix remained the best, but the distribution had more high-scoring matrices due to the high probability of adenine and thymine matches. The specificity matrix values for positions with $f_b = 0$ were estimated using $f_b = 0.5/18$ from the 18 sequences in the data set. (*C*) The "information content" at each position of the matrix is plotted. The sum from all positions is 12.15 bits.

1). The site with the highest score by the matrix of Fig. 4*B* occurs 3′ of the one identified, with seven matches to the consensus, cGTGAtgttgcTtgCA (Fig. 2). It is conceivable that the CRP protein binds to this site instead of the one identified or that it binds to both (17).

The other exception is in the *cat* regulatory region. Two binding sites have been identified (Figs. 1 and 2) and shown to interact with the CRP protein in footprinting experiments (18). Site 1 has six matches with the consensus, and site 2 has five matches. Site 1 demonstrated the tightest binding of the two sites. Although both of these sites score high with the matrix of Fig. 4, the highest scoring site on the fragment is between them, gGTGtccctgtTgAtA, also having six matches with the consensus. Binding to this site has not been reported. Site 2 is especially interesting because it has nine matches with the consensus when an additional base is allowed between the two highly conserved regions, TGTGAcggccgcTCACt. The fact that site 1 binds more tightly than site 2 suggests that either the variable spacing is not allowed or it causes enough strain in the interaction to limit the gain from additional consensus-base contacts. Variable spacing between the conserved regions of regulatory sites, excluding promoters, are not a common feature, although one has been verified (19). Our method could be amended to allow for a small number of gaps, perhaps one per site, but the complexity of the algorithm would increase substantially.

Conclusions

The complexity of the algorithm used in this study is one of its most appealing features. The memory required is independent of the number of sequences and linearly dependent on their lengths. The time required is linearly dependent on the number of sequences and the square of their lengths. This is in sharp contrast to rigorous methods that require comparing all possible alignments that require $O(L^N)$* time and space. The savings comes from selecting only a small subset of possible alignments to consider those that are "interesting," in that they are the best matches on each sequence to a set of possible matrix representations of the protein's specificity. The number of matrices that are considered at each step could be quite large, up to the limit of the computer memory available, although in the CRP example we needed to keep only one for each possible word in the first sequence. Reliability of the best matrix representation increases with the number of sequences. In the CRP example, in which we know the pattern of the binding sites, one pass through the data was sufficient to determine that pattern and identify the sites. If we hadn't known the answer, we would randomize the order of sequences several times and repeat the analysis to see whether we always got a similar answer or, if not, determine the variation among the most significant patterns.

As the amount of determined DNA sequence increases, methods that identify the important features of those sequences using only the information within the sequences will become increasingly important. Projects like the Human Genome Initiative (20) will generate such an enormous amount of sequence data that efficient methods of pattern identification will be essential to elucidate those features. We present a method to help in one such task—that of identifying regulatory protein-binding sites. Because the method relies on an information measure of similarity among sequences, which has been shown to give reliable representations of protein specificity (5, 7, 8), we expect the method to work for any sequence-specific DNA-binding proteins. Variations on the basic method should also be applicable to similar prob-

*The notation $O(L^N)$ is read "*order* L^N" and means that the calculation time and space (computer memory) required is $aL^N + b$, where a and b are constants. The algorithm described in this paper is $O(L)$ in space and $O(L^2N)$ in time.

lems in which functionally analogous sites can be identified by information within linear sequences. Functional domains of some proteins may be of this type (21, 22), and related methods may be useful in identifying them.

We thank Drs. Gerald Hertz and Charles E. Lawrence for stimulating discussions and helpful suggestions. We are grateful for comments on the manuscript from Drs. Susan Dutcher, Matt Scott, Bill Wood, and Mike Yarus. Sequences were extracted from GenBank using the EuGene software from the Molecular Biology Information Resource at Baylor College of Medicine (Houston). This work was supported by Grant GM28755 from the National Institutes of Health.

1. Gilbert, W., Gralla, J., Majors, J. & Maxam, A. (1975) in *Protein-Ligand Interactions,* eds. Sund, H. & Blauer, G. (de Gruyter, Berlin), pp. 193–206.
2. Kadonaga, J. T., Jones, K. A. & Tjian, R. (1986) *Trends Biochem. Sci.* **11,** 20–23.
3. Hawley, D. K. & McClure, W. R. (1983) *Nucleic Acids Res.* **11,** 2237–2255.
4. Harley, C. B. & Reynolds, R. P. (1987) *Nucleic Acids Res.* **15,** 2343–2361.
5. Stormo, G. D. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17,** 241–263.
6. Mulligan, M. E., Hawley, D. K., Entriken, R. & McClure, W. R. (1984) *Nucleic Acids Res.* **12,** 789–800.
7. Berg, O. G. & von Hippel, P. H. (1987) *J. Mol. Biol.* **193,** 723–750.
8. Berg, O. G. & von Hippel, P. H. (1988) *J. Mol. Biol.* **200,** 209–223.
9. de Crombrugghe, B., Busby, S. & Buc, H. (1984) *Science* **224,** 831–838.
10. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188,** 415–431.
11. Staden, R. (1984) *Nucleic Acids Res.* **12,** 505–519.
12. Staden, R. (1988) *Comput. Appl. Biosci.* **4,** 53–60.
13. Stormo, G. D. (1987) in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach,* eds. Bishop, M. J. & Rawlings, C. J. (IRL, Oxford), pp. 231–258.
14. Galas, D. J., Waterman, M. S. & Eggert, M. (1985) *J. Mol. Biol.* **186,** 117–128.
15. Mengeritsky, G. & Smith, T. F. (1987) *Comput. Appl. Biosci.* **3,** 323–328.
16. Goss, T. J. & Datta, P. (1985) *Mol. Gen. Genet.* **201,** 308–314.
17. Bedouelle, H. & Hofnung, M. (1982) *Mol. Gen. Genet.* **185,** 82–87.
18. Le Grice, S., Matzura, H., Marcoli, R., Iida, S. & Bickle, T. (1982) *J. Bacteriol.* **150,** 312–318.
19. Sadler, J. R., Sasmor, H. & Betz, J. L. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6785–6789.
20. Lewin, R. (1988) *Science* **240,** 602–604.
21. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.
22. Gribskov, M., Homyak, M., Edenfield, J. & Eisenberg, D. (1988) *Comput. Appl. Biosci.* **4,** 61–66.
23. Bacon, D. J. & Anderson, W. F. (1986) *J. Mol. Biol.* **191,** 153–161.