Bull. Inst. Chem. Res., Kyoto Univ., Vol. 69, No. 3, 1991

Analysis of DNA Functional Sites by Information Contents

Tomonori IIJIMA* and Minoru KANEHISA*

Received July 9, 1991

Functionally equivalent sites of DNA usually exhibit common sequence patterns. When such patterns are weak and localized, it usually requires manual efforts to derive at a consensus pattern. We present an automatic procedure to detect such a consensus, given a set of sequences known to contain a certain functional site somewhere in each sequence. The method is based on maximizing a measure of the information content for alignments of segments chosen from the sequences. The procedure was applied to the sequences containing CRP binding sites, promoters, terminators, and ribosome binding sites.

KEY WORDS: DNA sequence analysis/ Multiple sequence alignment/ Consensus sequence/ Information theory

1. INTRODUCTION

DNA is an information-rich polymer encoding the genetic information in the sequence of four types of nucleotides. Biologically speaking, the information content is not uniform along the DNA sequence; some regions contain important signals, such as for gene expression, while others seem to have little meaning. In the field of computer science, the information content is related to the entropy of the information source, which is determined from its probabilistic properties. Suppose that the sequence of four letters A, C, G, and T is generated as a stochastic process with probabilities P_b's (b=A,C,G,T), then the entropy is defined as:

$$H=-\sum_{b=A}^{T} p_b \log_2 p_b \qquad (1)$$

The information content of an actual DNA sequence is then measured by the decrease in this entropy.

In general, functionally equivalent sites of DNA exhibit conserved sequence patterns. The purpose of the present paper is to search for such consensus patterns, given a set of sequences known to contain functionally equivalent sites, according to the concept of information content. Several authors utilized similar concepts in the past. For example, Schneider et al.[1] calculated the information content of binding sites on DNA sequences, and claimed that it was sufficient for the site to be distinguished from the rest of the genome. Stormo and Hartzell[2] and Hertz et al.[3] constructed a multiple alignment of subsequences of a given length from a set of unaligned sequences by maximizing the information content for the subsequences. The aligned regions turned out to contain common functional sites.

* 飯島智徳, 金久 實 : Institute for Chemical Research, Kyoto University Uji, Kyoto 611

Our approach is similar to Stormo and Hartzell's, but the algorithm for maximizing the information content is different. Stormo and Hartzell's method for searching maximal information containing regions has intrinsic defects and our method always identifies better ones. Stormo and Hartzell applied their method to the sequences containing CRP-binding sites. We have, in addition, applied our method to promoter sites, terminator sites, and ribosome binding sites.

2. MATERIALS AND METHODS

2.1. Functional sites

1) CRP binding site

cAMP receptor protein (CRP) is involved in regulation of certain types of messenger RNA transcription. Stormo and Hartzell[2] collected 18 sequences, each with 105 bases, containing 23 CRP binding sites of Escherichia coli and its plasmids. We used the same data set in order to make comparison with their results.

2) Promoter

RNA polymerase complexed with the sigma factor binds to this site for initiation of transcription. It is widely known that the prokaryotic promoter contains two conserved regions around 35 and 10 bases upstream from the transcription initiation site, called −35 and −10 regions. They contain consensus sequence patterns TTGACA and TATAAT, respectively. We used 272 promoter sequences from E. coli, phages, and plasmids compiled by Harley and Reynolds[4].

3) Terminator

There are two types of transcription termination sites, rho factor dependent and independent. It is known that rho factor independent terminators contain common sequence features: an inverted repeat sequence which has a potential to form a hairpin loop structure, followed by a T rich region. There is also a similar functional site called an attenuator sharing similar sequence features. An attenuator is the transcription termination site within an operon, while a terminator is at the end of an operon. We collected from literature and GenBank database 21 sequences of 120 bases long containing terminators and attenuators of E. coli and phages.

4) Ribosome binding site

This is the site which is bound by ribosome before mRNA is translated into protein on ribosome. The consensus sequence pattern known as Shine-Dalgarno sequence exists; it is complementary to the 3'-terminus of 16S rRNA. We used 39 E. coli and phage sequences of 25 bases long covering the region upstream from the initiation codon, complied by Stormo et al.[5].

2.2. Definition of information content

Given an alignment of N segments of length L (Fig. 1), we define the information content of this alignment by:

$$I=\sum_{j=1}^{L} I_j/\mathrm{L} \qquad (2)$$

which is the average of the information content I_j at each position j:

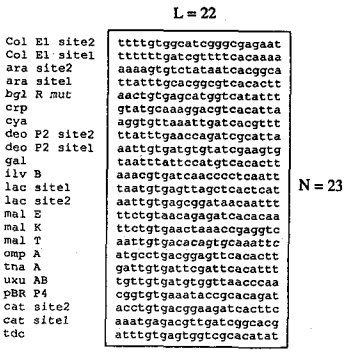

Fig. 1. The alignment of 23 CRP binding sites of 22 bases long determined from experiments[2]. N is the number of sequences in the alignment and L is the length of the alignment.

$$I_j = -\sum_{b=A}^{T} f_b \log_2 f_b + \sum_{b=A}^{T} p_b \log_2 p_b \qquad (3)$$

Here f_b (b=A,C,G,T) is the base frequency at position j of the alignment calculated by the following formula:

$$f_b = \frac{n_b + 1}{N + 4} \qquad (4)$$

where n_b is the observed base counts and N is the number of segments. For the estimation of probabilities p_b, we take the base composition of the entire sequences in the data set, from which maximal information containing alignments are searched. Alternatively, the probabilities may be estimated from the base composition of the entire genome.

2.3. Search algorithm

Suppose that there are N sequences in the data set. For simplicity, let us assume that the sequence length M is the same for all N sequences. Our task is to find the maximal information containing alignment of L-base segments, each taken from each of the N sequences. Since each sequence contains (M-L+1) segments, the maximal information containing alignment can be found by calculating Eq. (2) for $(M-L+1)^N$ combinations. However, this is impractical for moderate sizes of M and N. We have devised an iterative procedure to find optimal alignments as below:

1) Compare (M-L+1) segments of the first sequence against $(N-1)\times(M-L+1)$ segments of the remaining (N−1) sequences, and find the sequence containing the segment which gives the best scoring pairwise alignment according to Eq. (2). Between the first sequence and this sequence, make $(M-L+1)^2$ segment comparisons, and select 30 best scoring alignments.
2) Compare the 30 selected pairwise alignments against $(N-2)\times(M-L+1)$ segments of the remaining (N−2) sequences, and find the sequence containing the segment which gives the best scoring three-segment alignment. Using this

sequence and the 30 pairwise alignments, make $30\times(M-L+1)$ three-segment alignments, and select 30 best scoring alignments.

3) Repeat procedure 2) to the remaining sequences. Namely, compare the last 30 alignments against all the segments in the remaining sequences, and find the sequence which gives the best scoring alignment. Using this sequence and the 30 previous alignments, make $30\times(M-L+1)$ new alignments, and select 30 best scoring alignments.

4) Procedures 1) through 3) give 30 best scoring alignments of N segments. For each of the 30 alignments an iteration is performed to improve the information content as follows. Remove one segment from the N-segment alignment and compare the remaining $(N-1)$-segment alignment against all $(M-L+1)$ segments of the sequence containing the removed segment. If there is a segment which improves the information content, replace the removed segment by this segment and make a new N-segment alignment. Repeat this procedure in turn for all segments and until no replacement becomes necessary.

5) Procedures 1) through 4) are performed for a fixed choice of the first sequence. Examine $(N-1)$ other choices of the first sequence and find the final best scoring alignment.

3. RESULTS

3.1. CRP-binding site

The data set complied by Stormo and Hartzell[2] consisted of 18 sequences each with 105 bases. The base frequency was:

$p_A=572/1890=0.30$
$p_C=345/1890=0.18$
$p_G=395/1890=0.21$
$p_T=578/1890=0.31$

This data set contained 23 CRP-binding sites in 18 operons identified by experiments, and each site was 22 bases long (Fig. 1). Figure 2(a) shows the information content calculated by Eq. (3) for the alignment of these 23 segments. The mean information content according to Eq. (2) was 0.36 bits/base. When three bases were removed from both 5' and 3' ends, the value was 0.43 bits/base for the middle region of 16 bases long.

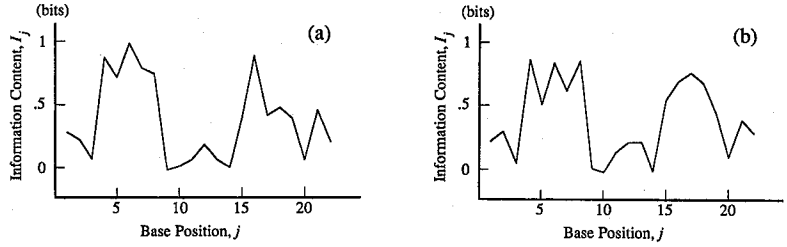

Fig. 2. (a) The information profile for the 22 base alignment of 23 CRP binding sites shown in Figure 1. (b) The information profile for the 22 base alignment of 18 segments obtained by our search procedure.

We applied the search procedure and obtained the maximal information containing alignment of 22 bases long from the 18 sequences. The information profile is shown in Figure 2(b). The mean value was 0.40 bits/base for the searched segments, compared with 0.36 bits/base for experimentally determined segments. Figure 3 shows the locations of segments found, which were mostly in agreement with experimental observations. However, there were three segments which were not identified as functional sites by experiments. The known consensus sequence of a CRP-binding site is (TGTGA……TCACA) which contains a dyad symmetry:

5' tgtga……tcaca 3'

3' acact……agtgt 5'

suggesting that this site interacts with a dimeric CRP protein. We examined the deviation from the consensus sequence in the missed sequences, ecogale, ecoilvbpr and ecomalk (Fig. 3).

```
colel     taatgtttgtgctggtTTTTGTGGCATCGGGCGAGAATagcgcgtggtgtgaaagactgtTTTTTTGATCGTTTTCACAAAAatggaagtccacagtcttgacag
ecoarabop gacaaaaacgcgtaacAAAAGTGTCTATAATCACGGCAgaaaagtccacattgaTTATTTGCACGGCGTCACACTTtgctatgccatagcatttttatccataag
ecobglr1  acaaatcccaataacttaattattgggatttgttatatataacttttataaattcctaaaattacacacaaagttaatAACTGTGAGCATGGTCATATTTttatcaat
ecocrp    cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatGTATGCAAAGGACGTCACATTAccgtgcagtacagttgatagc
ecocya    acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaAGGTGTTAAATTGATCACGTTTtagaccatttttttcgtcgtgaaactaaaaaaacc
ecodeop   agtgaaTTATTTGAACCAGATCGCATTAcagtgatgcaaacttgtaagtagatttccttAATTGTGATGTGTATCGAAGTGtgttgcggagtagatgttagaata
ecogale   gcgcataaaaaacggctaaattcttgtgtaaacgatttccacTAATTTATTCCATGTCACACTTttcgcatctttgttatgctatggttatttcatacccatgaagcc
ecoilvbpr gctccggcggggttttttgttatctgcaattcagtacaAAACGTGATCAACCCCTCAATTttccctttgctgaaaaatttttccattgtctcccctgtaaagctgt
ecolac    aacgcaatTAATGTGAGTTAGCTCACTCATtaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggAATTGTGAGCGGATAACAATTTcac
ecomale   acattaccgccaaTTCTGTAACAGAGATCACACAAagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk   ggaggaggcgggaggatgagaacacggcTTCTGTGAACTAAACCGAGGTCatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt   gatcagcgtcgttttaggtgagttgttaataaagatttggAATTGTGACACAGTGCAAATTCagacacatAaaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa   gctgacaaaaaagattaaacatacctatacaagactttttttttcatATGCCTGACGGAGTTCACACTTgtaagttttcaactacgttgtagacttacatcgcc
ecotnaa   ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacGATTGTGATTCGATTCACATTTaaacaatttcaga
ecouxul   cccatgagagtgaaatTGTTGTGATGTGGTTAACCCAAttagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4    ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgCGGTGTGAAATACCGCACAGATgcgtaaggagaaaatacccgcatcagggcgctc
trn9cat   CTGTGACGGAAGATCACTTCgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaAAATGAGACGTTGATCGGCACG
(tdc)     gattttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaATTTGTGAGTGGTCGCACATATcctgtt
```

Fig. 3. The locations found according to our search procedure (underlined) in comparison with acutual CRP binding sites (capitalized) in the data set of Stormo and Hartzell[4].

	experimental sites	calculated sites
ecogale	tgtga……ataaa ** *	tgtaa……ccact * * *
ecoilvbpr	cgtga……cctca * * *	tctgc……gtaca * * **
ecomalk	tgtga……ccgag * ***	cgtga……ttgca * **

Here * indicates the deviation from the consensus. The experimental sites show good agreements with the left half of the consensus, while in the calculated sites mismatches are dispersed on both sides of the consensus. This can also be seen in Figure 2, where the calculated profile (Fig. 2(b)) is more symmetric than the experimental profile (Fig. 2(a)). Thus, it appears that the CRP protein can be functional if one of the binding sites is activated.

We have also used the segment size of 16 bases for the search of the maximal information containing alignment. The mean information content obtained was 0.47 bits/base compared with 0.40 bits/base for experimentally determined segments. There were again three segments which were not consistent with experimental sites, the same ones shown above.

3.2. Promoter

Using the alignment of 272 promoter sequences by Harley and Reynolds[4], the

information profile was calculated according to Eq. (3). The mean information contents for the −35 and −10 hexamers were 0.57 and 0.76 bits/base, respectively. Because the spacing between the two hexamers is variable, we divided the seqence data into two halves at the middle of the spacer region and applied our search procedure separately with the segment length of eight bases. In this case only, we assumed equal values for the estimation of baseline probabilities:

$$p_A=p_C=p_G=p_T=0.25$$

The information profile obtained is shown in Figure 4(a) and (b). The mean information contents were 0.57 and 0.28 bits/base for the octamers. Figure 4 also shows in dotted lines the information profile of the alignment by Harley and Reynolds. Extracting the hexamer from positions 2 through 7 of the left octamer and the hexamer from positions 3 through 8 of the right octamer, the mean information contents were 0.72 and 0.86 bits/base, respectively, which were better than the −35 and −10 hexamers. However, the actual location of the found octamers did not correspond in many cases to actual promoter sites. This must be largely due to the lack of constraint for the fixed spacer length between the two hexamers.

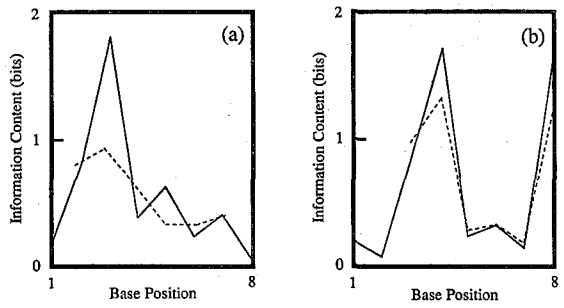

Fig. 4. The information profiles obtained for the eight base segment alignments containing (a) -35 and (b) -10 hexamers of promoter regions. The information profiles according to Harley and Reynolds'[4] six base alignments are shown in dotted lines, superimposed at most appropriate locations.

3.3. Terminator

The base frequency in the data set of 21 sequences was:

$$p_A=718/2520=0.29$$
$$p_C=558/2520=0.22$$
$$p_G=579/2520=0.23$$
$$p_T=665/2520=0.26$$

We applied our search procedure with the segment length of 10 bases. The obtained information profile is shown in Figure 5(a) and the mean information content was 0.71 bits/base. The sequence patterns detected correspond to the T rich region of rho factor independent terminators. The preceding dyad symmetry was not detected by our procedure. The locations found coincided with actual functional sites in 16 out of 21 sequences.

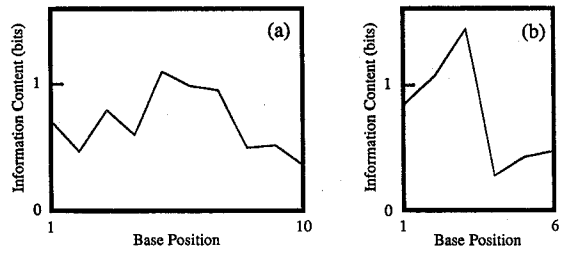

Fig. 5. The information profiles obtained for (a) the ten base segment alignment of terminators and (b) the six base segment alignment of ribosome binding sites.

3.4. Ribosome binding site

The base frequency in the data set of 39 sequences of 25 bases long was:

$p_A=366/975=0.38$

$p_C=158/975=0.16$

$p_G=205/975=0.21$

$p_T=246/975=0.25$

We applied our search procedure with the segment length of 6 bases. The obtained information profile is shown in Figure 5(b) and the mean information content was 0.76 bits/base. All the locations found matched at least three bases out of six to the complement of the Shine-Dalgarno sequence (5'-TCACCTCCTTA-3'). Five locations matched perfectly, and 17, 30, and 4 locations contained one, two, and three mismatches, respectively.

4. DISCUSSION

Stormo and Hartzell[2] and Hertz et al.[3] used Eq. (3) for identifying consensus patterns in unaligned DNA sequences. However, their algorithm for searching the highest information containing alignment was basically a one-path procedure, corresponding to procedures 1) through 3) in our algorithm. Their result was dependent on the order of sequences stored in the data set. Our algorithm performs a far more comprehensive search and always gives better alignments. For example, for the segment length of 16 bases, Stormo's algorithm gave 0.76 bits/base while ours gave 0.81 bits/base for the CRP binding site.

The segment length L is a parameter to be specified in the search procedure. It is desirable that the length is known experimentally for a functional site. Otherwise, by repeating the search with different values of L, it is possible to estimate an approximate value which covers peaks in the information profile. The number of alignments retained in each step of our procedure was set at 30 in the present analysis, but it can be variable. For the search of CRP binding sites with the segment length of 22, the information content of the final alignment was 0.32, 0.40, 0.40, and 0.40 bits/base when 1, 10, 20 and 40 alignments were retained in each step.

Once the data set is prepared, it is automatic to find the highest information

containing alignment according to our procedure. This alignment can be considered as a template for a specific functional site, and can be used for predicting actual functional sites in unknown sequences. However, as shown in Figure 3, the highest information containing segment is not always the actual functional site. It is possible that our optimization procedure did not detect additional information, for example, signals outside of the segment length L or hairpins and other higher-order structural properties. Thus, our procedure should be used to obtain a first approximation of the template, and the template should be refined according to manual inspection of experimental evidence.

ACKNOWLEDGEMENT

This work was supported by grants from the Ministry of Education, Science and Culture of Japan.

REFERENCES

(1) T.D. Schneider, G.D. Stormo, L. Gold and A. Ehrenfeucht, *J. Mol. Biol.* **188**, 415-431 (1986).
(2) G.D. Stormo and G.W. Hartzell, III, *Proc. Natl. Acad. Sci. USA*, **86**, 1183-1187 (1989).
(3) G.Z. Hertz, G.W. Hartzell, III, and G.D. Stormo, *CABIOS*, **6**, 81-92 (1990).
(4) C.B. Harley and R.P. Reynolds, *Nucleic Acids Res.*, **15**, 2343-2361 (1987).
(5) G.D. Stormo, T.D. Schneider and L.M. Gold, *Nucleic Acids Res.*, **10**, 2971-2996 (1982).