[22] Probing Information Content of DNA-Binding Sites

By Gary D. Stormo

Introduction

Information content is a convenient way to describe the specificity of a DNA-binding protein. In its simplest form information content is equivalent to the common description of a restriction enzyme as a "6-base cutter" or a "4-base cutter." That is, information content is a measure of the amount of specificity required for recognition, independent of the mechanism of recognition. Equivalently, it is a measure of the frequency of sites expected in a random sequence. We usually measure information content in bits, rather than bases. Since the choice of 1 base requires two bits, a "6-base cutter" has an information content of 12 bits. A restriction enzyme with multiple recognition sites, such as HincII, which cuts at GTYRAC (Y = C or T, R = A or G), recognizes 10 bits, or 5 bases, of information, even though its recognition site is 6 bases wide, because of the ambiguity in some of the positions.

The utility of information content as a measure of specificity is that it is applicable even to cases of high ambiguity, which is a typical characteristic of regulatory sites and which distinguishes them from restriction sites. Even though restriction sites can contain ambiguity they have essentially an all-or-none activity. That is, HincII cuts at every site that matches the sequence GTYRAC, and only such sites are cut. Regulatory sites, on the other hand, are usually similar to a "consensus sequence," but may differ at almost any position and at more than one position in any particular example. Furthermore, some positions within the consensus sequence may be more variable than others. The most studied example is Escherichia coli promoter sequences. They contain a two-part consensus sequence: a "-35" region of TTGACA and a "-10" region of TATAAT. In a compilation of over 200 promoters each of the 4 bases has been found at least once at each position of the consensus sequence, and none of the examples is a perfect match to it. But promoters are not random, because any particular promoter will likely match at most of the consensus positions, and some of the positions are highly, although not absolutely, conserved.


\[ I_{\text{seq}} = \sum_{b=A}^{T} f_b \log_2(f_b/p_b) \] (1)

where \( f_b \) is the fraction of each base \( b \) in the aligned sites, and \( p_b \) is the genomic proportion of each base.\(^1,2,4\) When the genomic proportions of

\[^4\text{G. D. Stormo, this series, Vol. 183, p. 211.}\]
ach base are equal ($p_b = 0.25$ for all $b$) the formula reduces to $I_{\text{seq}} = 0 + \sum f_b \log_2 f_b$, which can vary from 0, for a random position, to 2, for a position that is absolutely conserved.$^1$ The DNA-binding sites for several regulatory proteins have been examined.$^{1,2}$ The information content of ach is very close to that expected based on the number of sites in the genome. For highly ambiguous sites information content determined assuming independent positions, as in Eq. (1), is a lower bound on the value that would be determined using higher order analyses.$^5$ Therefore, the result of near equality between expected and observed frequencies$^1$ is not inconsistent with the idea that regulatory sites should be “overspecified.”$^{7,8}$ That result does allow one to estimate the frequency expected for particular type of site, and to gauge whether or not the pattern obtained or the specificity of a particular protein is appropriate. For example, the urge discrepancy we originally observed for the T7 promoters led to the experiments described below, from which we determined that there were additional constraints on phage promoters than just transcription initiation activity.$^5$

Experimental Determination of Information Content

While the information content determined from several examples of the binding sites of a protein can be a valuable tool in understanding its specificity, often a more detailed analysis is needed. For example, the known binding sites may have other constraints that contribute to their information content, but which are not required for function by the protein.

is also likely that not enough natural sites are known to have a reliable picture of the specificity of the protein from those data alone.

A variety of experimental procedures can also be used to determine the information content of a protein. The basic strategy is to randomize the binding sites and to select from the entire population those that retain functional activity. The definition of a functional site may be qualitative; that a particular site works or it does not, or it may involve a quantitative way that lets one use the relative affinity of the protein for different sites in the calculation of information content. Such experiments can either be carried out in vitro or by putting the randomized sites back into cells and determining function in vivo.


In Vivo Experiments

In our early studies of information content in regulatory sites the T7 promoters were an exception to the observation that information content is approximately equal to that expected from the frequency of sites.$^1$ In fact, T7 promoters had about twice the expected information content. We hypothesized then that perhaps only part of the information was required for promoter function, polymerase binding and transcription initiation, and that the remaining information was conserved for other, unknown reasons. In order to test this hypothesis we synthesized randomized variants of a 27-bp region that contained the promoter consensus sequence.$^5$

At each position of this synthesized promoter we used a mixture of nucleotides designed to give the consensus base 85% of the time and each of the other bases 5% of the time. In so doing we expected to get the complete consensus sequence about 1% of the time, assuring us that we would have a positive control of promoters that had to function. We also expected that there would be an average of about four changes per promoter. These randomized promoter sequences were then cloned into a plasmid that would serve to screen for functional promoters in cells that could be induced to make T7 RNA polymerase. The screen was that functional promoters killed the cells when induced, whereas nonfunctional promoters did not. Comparing the sequences that were recovered by the screen to analyses done by Chapman and Burgess,$^9$ we concluded that promoters had to be within two- or threefold the activity of the consensus promoter to pass the screen.

Two hundred clones were screened and 58 (29%) of them were found to be functional. Five had multiple inserts, leaving 53 (including 33 with the consensus sequence) that were functional promoters. Figure 2 shows the composition by position of the functional promoters and of the T7 phage promoters for comparison. In determining the information content from these data two factors are important. One is that the population proportions of the bases, $p_b$, are not equimolar, but rather they have been biased to be 85% the consensus base, which varies from position to position. When these numbers are used in Eq. (1), the information content comes out at about 2.9 bits, which is not too far removed from the 1.8 bits expected from the 29% functional promoters. But the other important consideration to remember is that this is the information relative to the biased synthetic promoters. What we really want to know is the amount of information relative to a T7 infection, in which case the genomic proportions of the bases are approximately equal. We could have obtained that

Fig. 2. Tabulated results for strong T7 promoters. The numbers of each kind of base at each position (L) are given for both the functional promoters recovered by the screen and for sequences at promoters in wild-type T7 phage. The column labeled $\phi_{10}$ refers to a particular T7 promoter, the sequence of which is identical to the phage consensus. A symmetry element is indicated by boxed letters. (Reprinted with permission.)

number by completely randomizing the promoters instead of leaving a bias toward the consensus sequence. However, since we had only a screen, not a selection, for activity that would have required screening many thousands of clones to get a few that were functional. Instead, we can use the data we collected and "normalize" it to see what we would have obtained had we done the complete randomization. The assumption is that the data obtained accurately reflect the selectivity of the polymerase. We assume that the ratios of $f_b/p_b$ are equal to the ratios of the activities for each variant base. That is equivalent to assuming that the data reflect the differences in binding energies to each possible base, and that those

energy differences are independent of the base proportions in the population of possible binding sites.

The data we have are of $f_b$ and $p_b$ values. The normalization procedure is to find the $f_b$ values for new values of $p_b$. The constraints that must be satisfied are that the sums of $f_b$ and $p_b$ must equal one, and that the ratios of all the $f_b/p_b$ values must be unchanged. For the case of $p_b = 0.25$ this is straightforward, because then $\sum f_b/p_b = 4$. The fraction of each base calculated to be in binding sites if the nonsite bases are equimolar is

$$f_b = (f_b/p_b)/(\sum f_b/p_b)$$

We consider this situation, of equimolar nonsite bases, to be a "standard state," and the binding constant determined at this state, $K_b^* = 4f_b$, to be the specific binding constant. It can vary between 0 and 4, with a value of 1 for nonspecific binding. The formula for information content then becomes

$$I_{eq} = \sum_{b=A}^{T} f_b \log_2 K_b^*$$

which is easily seen to be proportional to the average specific binding energy.

When our randomized T7 promoter data are treated in this manner, the resulting information content at each position is as shown in Fig. 3. Also shown is the information content as calculated from the phage promoters. The most important result is that much of the information within the phage promoters is not required for promoter activity. In fact, only about half of the total conserved information is required, placing T7 promoters into the same class as other regulatory sites where the information content observed is about as expected from the frequency of sites. Careful examination of the data in Figs. 2 and 3 reveals several interesting features. First, there are some positions, such as $-7$, $-8$, and $-9$, that are absolutely conserved in both the phage and the functional promoters. There are several other positions, such as $-3$, $-4$, $-6$, and $-10$, that are absolutely conserved in the phage promoters, but at which any base can exist in a functional promoter. This finding alone confirms our hypothesis that there are constraints at T7 promoters other than those required by the polymerase for transcription activity. At several other positions there is information in the functional promoters, but less than in the phage promoters. The results at positions $-5$ and $-12$, where it looks like functional promoters contain more information than phage promoters, are not significant, but due to the small sample size. If we had seen even one nonconsensus base at those positions the information content would
by multiple regression using several alternative models. As shown in Fig.
4, the simple model that the context effect is determined by the 6 bases
surrounding the amber codon, three on each side, and that the effects of
each base are independent and additive does quite well in describing the
data. Figure 4b shows the fit by this model between the predicted and
observed values, which has a correlation coefficient of about 0.93. Figure
4a shows the values of relative $\ln K_b$ that were determined by the regressi-
on, in which we had arbitrarily set $K_T = 1$ at each position. (There
are only three independent parameters per position, so we set one and
determine the others relative to that one.)\(^9\) We can normalize those values
to a standard activity constant, as described above, and then calculate the
information content for suppression contexts, as in Eq. (3). The result is
a total of 0.70 bits, with 0.58 coming from the two positions 3' to the amber
codon, positions 3 and 4. While it was clear in our original paper that
these were the two most important positions for determining suppression
efficiency, this information content analysis lets us quantitate the relative
importance of each position.

**In Vitro Experiments**

One can also determine information content from *in vitro* binding ex-
periments. In fact, these make it easy to use quantitative assays of binding
affinity in the determination. One method is to perform extensive binding
analysis on mutant binding sites, as has been done for the phage $\lambda$ proteins
Cro and Repressor.\(^11\)\(^12\) In these experiments every single base change was
made to the consensus $\lambda$ operator and the binding constant was determined
for each protein to each variant operator. The results were presented as
$\Delta \Delta G$ values, which can easily be converted to specific binding constants,
as described above. The information content determined by these calcula-
tions is about 15.5 and 14 bits for Cro and Repressor, respectively. These
numbers are low compared to that determined from the known binding site
sequences or from the frequency of known operators.\(^1\) These calculations,
based on quantitative measurements of independent changes, tend to under-
estimate the complete specificity of the proteins.\(^6\) Nonetheless, these
values are consistent with many experiments of determining the affinity
to sites with multiple changes, and serve to constrain the models of how
the proteins obtain their required specificity.\(^11\)\(^12\)

Information content can also be determined from experiments in which
the protein is bound to many different sites at once, providing there is a


means of determining the partitioning of each site between bound and unbound fractions. We have been developing such a method for the Mnt repressor of phage P22.12 Mnt is a repressor that binds as a tetramer to a symmetric operator sequence to regulate expression from the $P_{mt}$ promoter.13 We synthesized an operator with one randomized position, which could be cut with a different restriction enzyme for each different base in the variable position: ... AGGTCCACGTTGGAAGT ... The upper-case letters are the wild-type operator sequence, and the $n$ is the randomized position. Depending on whether the variable base is an A, C, G, or T, the oligonucleotide can be cut by SpeI, Sau96I, Alul, or MboI, respectively. Mnt repressor was added to the randomized synthetic operator in sufficient quantity to bind approximately half of the DNA, and the bound and unbound fractions were separated in a gel mobility shift assay.14 Both bands were eluted from the gel and the proportion of each base in each fraction was determined by cutting with the restriction enzymes. The results of this experiment were that the specific binding constants, $K_n^b$, were 0.71, 2.36, 0.28, and 0.64 for A, C, G, and T, respectively. The information content for this position is 0.41 bits.

The partitioning of each site into the two fractions could be determined by quantitative sequencing of each fraction. Then an equivalent experiment could be done at any position, regardless of whether an appropriate set of restriction enzymes could be found. It would also be possible to randomize several positions at once and determine the information content of each in a single experiment. Then only a few binding experiments might be sufficient to determine the complete specificity of the protein. So as to minimize the effects of nonindependent binding interactions, such an experiment should probably be done with nonadjacent positions randomized in a particular experiment.

Summary

An information content analysis of protein-binding sites gives a quantitative description of the specificity of the protein, independent of the mechanism of specificity. It gives useful information about the total specificity of the protein and about the individual positions within the binding sites. Information content is consistent with both thermodynamic and statistical analyses of specificity.24 When applied to a collection of known binding sites, the description provided may be limited by the sample size.

or by unknown constraints on those sites. Experimental procedures to
determine the information content can give much more reliable measures.
A large number of functional sites can be obtained from a much larger
pool of randomized potential sites. Quantitative assays for the activity of
different sites can be easily incorporated into the analysis, thereby
increasing its sensitivity. Both *in vitro* and *in vivo* experiments are amenable
to information content analysis.

Acknowledgments

The experiments on the Mt-binding specificity were performed using protein kindly
provided by Ken Knight and Robert Sauer. The work described in this chapter has been
supported by NIH Grant GM28755.

---

[23] Protein Chemical Modification as Probe of
Structure–Function Relationships

*By Kathleen S. Matthews, Artemis E. Chakerian, and
Joseph A. Gardner*

Chemical modification is a useful tool for examining the participation
of specific amino acid side chains in the structure and function of proteins.
In addition, chemical modification provides the potential for cross-linking
and, in the case of enzymes, for design of reagents that are activated by
the enzymatic mechanism. A number of reviews and books are available
for consultation in designing an experimental approach for chemical modifi-
cation of DNA-binding proteins.1–7 Figure 1 provides a flow chart for
developing an experimental design to modify a protein. In the ideal situation,
a specific reagent will react with only one type of side chain; in

1 R. L. Landblad and C. M. Noyes, "Chemical Reagents for Protein Modification, Volumes
5 A. N. Glazer, R. J. DeLange, and D. S. Sigman, "Chemical Modification of Proteins,"
Francisco, California, 1971.