The Information Content of Binding Sites on Nucleotide Sequences

Thomas D. Schneider*†, Gary D. Stormo*, Larry Gold*, and Andrzej Ehrenfeucht‡

version = 1.12 of schneider1986.tex 2002 Oct 16

J. Mol. Biol. (1986) **188**, 415-431[1]

I have not finished transforming this document into LaTeX. Still remaining to be done:

1. **check equations**
2. **read the text for errors.**
3. **put all references into bibtex so they hyperlink**
4. **Do figs 3 - 10**
5. **Fig 1 is missing the 0 lines!**

If you have corrections, please email them to me at toms@ncifcrf.gov. By releasing this paper now, Figures 1, 11, 12, 13 and the Appendix (which are finished) are available for people who would like to learn about the small sample correction.

NEWS:

- version 1.11, 2001 July 5: Tables 1 and 2 are done!
- version 1.08, 2001 June 22: Figs 11, 12, 13 now are in the html appendix.
- version 1.08, 2001 June 22: Figs 1 and 2 sizes fixed in html. Figs 11, 12, 13 now function.

---

*Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309, U.S.A.

†Present address: National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Experimental and Computational Biology, P. O. Box B, Frederick, MD 21702-1201. (301) 846-5581 (-5532 for messages), fax: (301) 846-5598, email: toms@ncifcrf.gov. http://www.lecb.ncifcrf.gov/~toms/

‡Department of Computer Science, University of Colorado, Boulder, Colorado 80309, U.S.A.

- version 1.07, 2001 June 8: Fig 2 (ribosome Rs curve) done.

Repressors, polymerases, ribosomes and other macromolecules bind to specific nucleic acid sequences. They can find a binding site only if the sequence has a recognizable pattern. We define a measure of the information ($R_{sequence}$) in the sequence patterns at binding sites. It allows one to investigate how information is distributed across the sites and to compare one site to another. One can also calculate the amount of information ($R_{frequency}$) that would be required to locate the sites given that they occur with some frequency in the genome. Several *Escherichia coli* binding sites were analyzed using these two independent empirical measurements.

The two amounts of information are similar for most of the sites we analyzed. In contrast, bacteriophage T7 RNA polymerase binding sites contain about twice as much information as is necessary for recognition by the T7 polymerase, suggesting that a second protein may bind at T7 promoters. The extra information can be accounted for by a strong symmetry element found at the T7 promoters. This element may be an operator. If this model is correct, these promoters and operators do not share much information. The comparisons between $R_{sequence}$ and $R_{frequency}$ suggest that the information at binding sites is just sufficient for the sites to be distinguished from the rest of the genome.

1. Introduction

When studying molecular binding sites in DNA or RNA, it is conventional practice to align the sequences of several sites recognized by the same macromolecular recognizer[2] and then to choose the most common bases at each position to create a consensus sequence (e.g. Davidson *et al.*, 1983). Consensus sequences are difficult to work with and are not reliable when searching for new sites (Sadler *et al.*, 1983b; Hawley and McClure, 1983). In part, this is because information is lost when the relative frequency of specific bases at each position is ignored. For example, the first position of *E. coli* translational initiation codons has 94% A, 5% G, 1% U and 0% C, which is not represented precisely by the consensus "A". To avoid this problem, four histograms can be made that record the frequencies of each base at each position of the aligned sequences. Such histograms can be compressed into a single curve by the use of a χ^2 function (Gold *et al.*, 1981; Stormo *et al.*, 1982a). Although these curves show where information lies in the site, they have several disadvantages: the χ^2 scale is not easily understood in simple terms; it is difficult to compare the overall information content of two different kinds of sites, such as ribosome binding sites and restriction enzyme sites; and χ^2 histograms are not directly useful in searching for new sites (Stormo *et al.*, 1982b).

We present here a method for evaluating the *information content* of sites recognized by one kind of macromolecule. The method begins with an alignment of known sites, just as with the evaluation of consensus sequences or χ^2 histograms. However, the calculation of the information content (called $R_{sequence}$) does not ignore variability of individual positions within a set of sites, as do consensus sequences. Furthermore, $R_{sequence}$ is a measure that encourages direct comparisons between sites recognized by different macromolecules, which is an improvement over χ^2 histograms. $R_{sequence}$ has units of bits per site. The values obtained precisely describe how different the sequences are from all possible sequences in the genome of the organism, in a manner that clearly delineates the important features of the site.

[2] We use the term "recognizer" to mean a macromolecule which locates specific sites on nucleic acids. These include repressors, activators, polymerases and ribosomes.

An independent approach is to measure the information needed to find sites in the genome. This relies on the size of the genome and the number of sites in the genome rather than nucleotide sequence information. There is at least one *lac* operator in *E. coli*, while there are thousands of ribosome binding sites. We have defined another measure, $R_{frequency}$, that is a function of the frequency of sites in the genome. More information would be necessary to identify a single site than any one in a set of thousands. Thus $R_{frequency}$ is greater for the *lac* operator than for ribosome binding sites. $R_{frequency}$, like $R_{sequence}$, is expressed in bits per site.

$R_{sequence}$, which measures the information in binding site sequences, should be related to the specific binding interaction between the recognizer and the site. $R_{frequency}$, based only on the frequency of sites, is related to the amount of information required for the sites to be distinguished from all sites in the genome. The problem of how proteins can find their required binding sites among a huge excess of non-sites has been discussed (Lin and Riggs, 1975; von Hippel, 1979). $R_{sequence}$ and $R_{frequency}$ give us quantitative tools for addressing this problem. Thus we compare $R_{sequence}$ and $R_{frequency}$ and come to the pleasing conclusion that the values are similar for each site studied. This result was not necessarily expected.

2. Materials and Methods

(a) *Calculation of* $R_{sequence}$

(i) Formula for $R_{sequence}$

Data for calculating $R_{sequence}$ comes from two sources. One is the nucleotide sequences at which a recognizer has been shown to bind. The other is the nucleotide composition of the genome in which the recognizer functions. The sequences are aligned by one base (the zero base) to give the largest possible homology between them (see figure 9 for an example). Some positions have little variation, while others have more. We tabulate the frequency of each base B at each position L in the site, to make a table called $f(B, L)$. Focusing on one position at a time, we want to measure the possible variations. For this we have chosen the "uncertainty" measure introduced by Shannon in 1948 (Shannon, 1948; Shannon and Weaver, 1949; Weaver, 1949; Abramson, 1963; Singh, 1966; Gatlin, 1972; Sampson, 1976; Pierce, 1980; Campbell, 1982; Schneider, 1984).

When there are M possible symbols, with probabilities P_i (such that $\sum_{i=1}^{M} P_i = 1$) , the general formula for uncertainty is

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i \qquad \text{(bits per symbol).} \tag{1}$$

One bit of information resolves the uncertainty of choice between two equally likely symbols. For nucleotide sequences, there are $M = 4$ possible bases. Using the frequencies of bases as estimates for probabilities, the uncertainty is calculated as

$$Hs(L) \;=\; -\sum_{B=A}^{T} f(B, L) \log_2 f(B, L) \qquad \text{(bits per base).} \tag{2}$$

(B is either A, C, G or T). The formula gives sensible results for three simple cases: 1) If only one base appears in the sequences, such as an A, then $f(A, L) = 1$ while the other frequencies are zero. $Hs(L)$ gives zero bits ($0 \; log \; 0 \; = \; 0$), meaning that if we were to sequence another site, we would have no uncertainty that the next base will be an A. 2) If two bases appeared with equal frequency, [as in $f(A, L) = 0.5$, $f(C, L) = 0$, $f(G, L) = 0.5$ and $f(T, L) = 0$], our uncertainty would be 1 bit. 3) If all 4 bases appeared with equal frequencies, then $f(B, L) = 0.25$ and the uncertainty is 2 bits.

If we sequenced randomly in the genome, and aligned sequences arbitrarily, we would see all 4 bases, with probabilities $P(B)$ and our uncertainty about what base we would see next would be:

$$H_g \;=\; -\sum_{B=A}^{T} P(B) \log_2 P(B) \qquad \text{(bits per base).} \tag{3}$$

This number is close to 2 bits for the organism *E. coli*, considered in this paper. In contrast, when sequences are aligned at binding sites (as in typical consensus alignments) a pattern appears which decreases the uncertainty below that of randomly aligned fragments (equation (2)). For each position L the decrease would be:

$$R_{sequence}(L) = H_g - Hs(L) \qquad \text{(bits per base).} \tag{4}$$

This is a measure of the sequence information gained by aligning the sites. The total information gained will be the total decrease in uncertainty:

$$R_{sequence} = \sum_{L} R_{sequence}(L) \qquad \text{(bits per site).} \tag{5}$$

(By summing, we make the simplifying assumption that the frequencies at one position are not influenced by those at another position. It is also possible to calculate $R_{sequence}$ from dinucleotides or oligonucleotides [Shannon, 1951; Gatlin, 1972; Lipman and Maizel, 1982]. When dinucleotides were used for ribosome binding sites, the total information content was not different from that given in Results, [unpublished observation]. Unfortunately, sampling error prevents one from making the calculation in most cases.)

(ii) Graphs of $R_{sequence}$ and Correction for Sampling Error

In Fig. 1, we show the curve $R_{sequence}(L)$ for either 61 (a), 17 (b) or 6 (c) *Hin*cII sites (GTPyPuAC; Roberts, 1983) chosen from the left end of bacteriophage T7 (Dunn and Studier, 1983). Here, the G's in the *Hin*cII sites have been placed at position $L = 0$, and $R_{sequence}(L)$ was calculated for 20 bases on either side. There are two major 2-bit peaks of information content surrounding a 1-bit valley in curve (a). None of the curves go to zero (the solid straight line) outside the sites, although they come close at several points. This effect is not small: for six sites (Fig. 1c) the background is at 0.44 bits per base so that with sequences 41 bases long, $R_{sequence}$ will be overestimated by 18 bits. A sampling error correction for $Hs(L)$ ($e(n)$, Appendix I, page 19). can be joined with H_g to give the final formula:

$$R_{sequence} = \sum_{L} \left(E(H_{nb}) \;-\; Hs(L)\right) \qquad \text{(bits per site).} \tag{6}$$

With this correction, the information content measured at various positions of an aligned set of random sequences will vary above and below zero. On the average it should be zero outside a binding site. The information content inside a site will rise above zero. These features can be seen in all figures, where the corrected zero is shown as a dashed line.

The standard deviation reported for each $R_{sequence}$ is based on the variance of H_{nb} (Appendix I, page 19) which is sensitive to the number of sequence examples, but not to the actual sequences. It is only a measure of variance in the correction for small sample sizes; the variation in the information content of individual sites will be described elsewhere. The variance of the sampling correction is shown in all figures as a bar extending one standard deviation above and below the $R_{sequence}(L)$ curve.

(iii) Determining the Binding Site Size

The *range* is the nucleic-acid region over which the sum of $R_{sequence}(L)$ is taken. If the range is larger than the binding site, the $R_{sequence}(L)$ fluctuations outside the site will cancel each other on the average. On the other hand, if the range is too small information content will be lost. That is, one must be sure not to delete part of the site.

Determining the range of a site is difficult because experimental methods such as deletion analysis, chemical protection or footprinting, do not define the exact region contacted. It is dangerous to judge the range by eye from the sequences themselves or the $R_{sequence}(L)$ curves derived from a small sequence collection (note that some positions of Fig. 1c show the same information content as the 1-bit valley). To avoid these difficulties, we have added 5 bases to both sides of the largest range suggested by experimental data. Consequently, the results will be more variable than they may have been, but it is unlikely that part of a site will be lost. On the average the background will be cancelled, although in specific cases it may not be. In the cases where two sites are adjacent, we extend the range to just before the point of overlap. If adjacent sites do interpenetrate, then some of the information content is lost.

When it is likely that a site is symmetrical, both the sequence and its complement are used in the analysis. This doubles the number of sequences available, and refines the answer. If we had arbitrarily chosen an orientation for each sequence we might have biased the results.

(iv) Variable Spacing

When a recognition site has two or more parts with various spacings between them, alignment by one part may blur out information in the other part. For example, if the four variants of this site:

```
ACGTACGTACGTnnnnnnnnGGCC
nACGTACGTACGTnnnnnnnGGCC
nnACGTACGTACGTnnnnnnGGCC
nnnACGTACGTACGTnnnnnGGCC
   ●●●●●●●●●●
```

occurred with equal frequency, then the positions marked by dots would have zero information content, even though these sequences would give a large information content if they were aligned with each other. To handle this one may align each part separately and add the information contents together. However, this leads to an overestimate of the information because the variable spacing is not taken into account. To take it into account, one may calculate how uncertain the spacing is from a tabulation of the frequency of each spacing and subtract this from the total information of the two parts. (This is equivalent to increasing the uncertainty of the site, Hs.) For the example above, $R_{sequence} = 24$ (ACGTACGTACGT) + 8 (GGCC) - 2 (spacing) = 30 bits. When this was done for ribosome binding sites, the total information content was not different from that given in Results (unpublished observation).

(b) *Formula for* $R_{frequency}$

If a genome contains G bases, there are $M{=}G$ ways that its sequence can be aligned or G potential binding sites. If these are all equally likely, then $P_i = 1/G$ and formula (1) reduces to:

$$H_{gf} = \log_2 G \qquad \text{(bits)}. \tag{7}$$

If the genome contains γ sites, we assume that the probabilities of binding to each site are equal and that the probability of significant binding to other sequences is zero. This allows formula (1) to be reduced to:

$$H_{sf} = \log_2 \gamma \qquad \text{(bits)}. \tag{8}$$

(One property of H is that it is at a maximum when the probabilities are equal. Thus both H_{gf} and H_{sf} are maxima.)

The decrease in positional uncertainty during binding or alignment is the difference:

$$\begin{aligned} Rfrequency &= H_{gf} - H_{sf} = -\log_2 \frac{\gamma}{G} \\ &= -\log_2 f \qquad \text{(bits per site)} \end{aligned} \tag{9}$$

where f is the frequency of sites in the genome.

$R_{frequency}$ is the amount of information needed to pick γ sites out of G possible sites. As the number of sites in the genome increases, the information needed to find a site decreases. As long as the simplifying assumption for equation (8) holds and γ is restricted to the number of known sites (that is, γ is not an estimate), equation (9) gives an upper bound on $R_{frequency}$ since some sites may exist that are not now known. A second property of this formula is that $R_{frequency}$ is insensitive to small changes in G or γ. The frequency of sites must change by a factor of two to alter $R_{frequency}$ by only one bit. The largest possible value of $R_{frequency}$ occurs for a single site in the genome: $\log_2 G$. (For *E. coli*, $R_{frequency} = 22.9$ bits in this case.) On the other hand, if all positions in the genome were sites, one would not need any information to find them, and $R_{frequency}$ would be zero.

The number of potential binding sites (G) is twice the number of base pairs in a DNA genome because there are two orientations for a recognizer to bind at each base pair. A symmetrical recognizer on DNA has two ways to bind each base pair, and both ways are used at a binding site. Here, γ is twice the number of conventional binding sites. An asymmetric recognizer on DNA will use only one orientation at any particular base pair. In this case, γ is equal to the number of binding sites. On RNA there is only one possible orientation. Thus G and γ reflect not only the genome size and number of binding sites but also the symmetries of the recognizer and nucleic acid.

(c) *Skewed Genomes*[3]

This paper considers the relationship between $R_{sequence}$ and $R_{frequency}$. For restriction enzymes cutting genomes with equal numbers of the four bases randomly distributed, $R_{sequence}$ and $R_{frequency}$ are equal. For example, one commonly assumes that *Hae*III (GGCC; Roberts, 1983; $R_{sequence} = 8$ bits) cuts once in 256 bases ($R_{frequency} = 8$ bits). This is not true for "skewed" genomes, in which the frequencies of each base are significantly unequal. For example, in a genome like that of bacteriophage T4 which is two-thirds A-T, $R_{sequence}$ for any tetramer is 7.7 bits. Yet GGCC should occur once in every 1296 bases ($(1/6)^4$; $R_{frequency} =$ 10.3 bits) and conversely AATT should occur once in every 81 bases ($(1/3)^4$; $R_{frequency} =$ 6.3 bits). An alternative formula,

$$R^*_{sequence}(L) = \sum_{B=A}^{T} f(B,L) \log_2 \frac{f(B,L)}{P(B)}, \tag{10}$$

matchs $R_{frequency}$ in examples of this type. When the genomes are equiprobable, as they are in this paper, the two $R_{sequence}$ formulas give the same values. We suggest that both be tried for sites in skewed genomes.

(d) *Programs and Computers*

All programs used for analyses were written in Pascal (Jensen and Wirth, 1978; Schneider *et al.*, 1982, 1984). The major programs used were:

Name	*Version*	*Purpose*
CalHnb	2.15	calculate statistics of H_{nb}: $E(H_{nb})$, $AE(H_{nb})$ and $Var(H_{nb})$ (generates Fig. 12).
Rseq	4.46	information content of sequences, $R_{sequence}$ as calculated in this paper (with correction for sampling error).
RsGra	2.45	a non-standard FORTRAN program using device independent graphics (Warner, 1979) for drawing the figures on microfilm.

Most work was performed on a CDC Cyber 170/720 computer. Figures were generated on a CDC 280/284 microfilm recorder.

(e) *Sequence Data*

We used two large procaryotic sequence data bases called LIB1 (bacteriophage) and LIB2 (*E. coli* and *S. typhimurium*) (Stormo *et al.*, 1982a) for the sequences of ribosome binding sites. Twenty-five new sites were included: T4 gene 67, (Völker *et al.*, 1982), T4 lysozyme, IPIII (Owen *et al.*, 1983); *E. coli* genes: *thrB*, *thrC* (Cossart *et al.*, 1981), *rpsT* (Mackie, 1981), *rpsB*, *tsf* (An *et al.*, 1981), *ndh* (Young *et al.*, 1981), *aroH* (Zurawski *et al.*, 1981), *alaS*

[3] I inserted this section at the insistance of my colleagues. I never thought that was correct. The main reason is that the units are no longer bits since the maximum value for selection amongst 4 objects may be arbitrarily larger than 2 by this measure, but clearly only 2 bits are required. For further discussion see references (1, 2, 3) — TDS.

(Putney *et al.*, (1981), *rpoD* (Burton *et al.*, 1981), *tufA* (Yokota *et al.*, 1980), *unc1, unc6, uncC, uncB, uncdelta, uncA* (Gay and Walker, 1981a,b; Kanazawa *et al.*, 1981), *tufB* (An and Friesen, 1980), *lexA* (Horii *et al.*, 1981; Miki *et al.*, 1981; Markham *et al.*, 1981), *ampC* (Jaurin and Grundström, 1981; Jaurin *et al.*, 1981), *Eco*RI endonuclease, methylase (Greene *et al.*, 1981; Newman *et al.*, 1981), DHFR (Swift *et al.*, 1981; Zolg and H*aumlaut*nggi, 1981). Sequences other than ribosome binding sites were stored in a library called SITELI. The corresponding Delila instructions were stored as modules in a single file called SITEIN and the Module program was used to extract the instructions for each analysis. The sequences for *carAB*, *argI* and *argR* were from Cunin *et al.* (1983). The *lacZ* "pseudo"-operator sequence was from Kalnins *et al.*, (1983). The remaining SITELI sequences described in Results were from the GenBank (TM) magnetic tape, release 14.0, (November 1983) which is available from Bolt Beranek and Newman Inc., Cambridge, Mass.

3. Results

(a) *Ribosomes and Ribosome Binding Sites*

We aligned the sequences of 149 *E. coli* and coliphage ribosome binding sites by their initiation codons because the process of initiation requires that the fmet-tRNA$_{\mathrm{F}}$ bind there. Since ribosomes search mRNA, we used the composition of the transcript library (Stormo *et al.*, 1982a) to calculate H_g: A=29526, C=25853, G=27800, T=28951 for which H_g=1.99817 bits/base. The frequencies of bases at each position of the sites were used to find the information content, $R_{sequence}(L)$, as a function of position (equations 2, 3 and a.8). Fig. 2 shows that the largest peak is for the initiation codon. The second largest peak represents the "Shine and Dalgarno" sequence (Shine and Dalgarno, 1974). There are at least five other distinct peaks.

$R_{sequence}$, the total information content of the site, is found by adding together the individual information contents from each position (equation 6). Previous statistical analyses showed a range of -21 to +13 (zero is the first base of the initiation codon), which corresponds well to the regions of RNA protected by ribosomes from ribonucleases (Gold *et al.*, 1981). This range was extended by 5 bases on both sides. For this range, we calculate an $R_{sequence}$ of 11.0 bits per site. Alignment by the Shine and Dalgarno sequence gives less than 8.3 bits (data not shown), which suggests that this is not a good alignment.

A good estimate for the size of the *E. coli* genome is 3.9×10^6 basepairs (Bachmann and Low, 1980). In determining $R_{frequency}$, we assume that almost all of the genome is transcribed into messages and that for the most part only one strand is transcribed. The number of potential ribosome binding sites is therefore 3.9×10^6. Based on the coding capacity versus DNA insert size of 24 plasmids selected at random from the Clark-Carbon bank (P. Bloch, personal communication; F.C. Neidhardt *et al.*, 1983), and a genome size of 3.9×10^6 bp, we estimate the number of proteins encoded by *E. coli*, and therefore the number of ribosome binding sites, to be 2574. Equation (9) therefore gives an $R_{frequency}$ of 10.6 bits per site. The data for all analyses are gathered in Table 1. ⇐Table 1

(b) *LexA and SOS Boxes*

Organism	Recognizer	Type	n	Range	R_s	S.D.	γ	$G{\times}10^{-6}$	R_f	R_s/R_f	$R_s{-}R_f$
E. coli	**Ribosome**	**A**	**149**	**-26 to 18**	**11.0**	**0.1**	**2574**	**3.9**	**10.6**	**1.0**	**0.4**
E. coli	**LexA**	**E**	**14**	**-9 to 10**	**21.1**	**0.6**	**22**	**7.8**	**18.4**	**1.1**	**2.7**
E. coli	**TrpR**	**E**	**6**	**-18 to 19**	**23.4**	**1.9**	**6**	**7.8**	**20.3**	**1.1**	**3.0**
E. coli	**LacI**	**O**	**2**	**-21 to 21**	**19.2**	**2.8**	**2**	**7.8**	**21.9**	**0.9**	**-2.6**
E. coli	**ArgR**	**E**	**16**	**-9 to 10**	**16.4**	**0.5**	**22**	**7.8**	**18.4**	**0.9**	**-2.0**
λ	**cI/Cro**	**O**	**12**	**-9 to 9**	**17.1**	**0.7**	**12**	**7.8**	**19.3**	**0.9**	**-2.2**
T7	**RNA Pol**	**A**	**17**	**-29 to 12**	**35.4**	**0.7**	**83**	**7.8**	**16.5**	**2.1**	**18.9**
T7	**Symmetry**	**E**	**34**	**-6 to 7**	**16.4**	**0.2**	**34**	**7.8**	**17.8**	**0.9**	**-1.4**

Table 1: Information content of several molecular binding sites.
Type of site: A = asymmetric, E = symmetric without a central base (Even), 0 = symmetric with a central base (Odd). n is the number of sequenced sites (for symmetric sites, both strands are counted). The range is the region over which $R_{sequence}$ is calculated. R_s stands for $R_{sequence}$. S.D. is the standard deviation of $R_{sequence}$ owing to small sample size; the variance of information content for individual sites will be presented elsewhere. γ is the number of distinct binding sites in the genome. For symmetrical sites, there are two possible ways to bind, so γ is twice the number of conventional sites. G is the number of potential binding sites on the genome. Rf stands for $R_{frequency}$. Calculations were carried out to five decimal places and then rounded.

In response to DNA damage, a set of unlinked *E. coli* genes are expressed (Kenyon *et al.*, 1982; Little, 1983; for a review see Little and Mount, 1982). The genes of the SOS regulatory system are controlled in part at the level of transcription by the direct binding of the *lexA* gene product to the promoters. Five binding sites are well characterized. Two sites are linked to *LexA*, one is linked to each of *recA* (Little *et al.*, 1981; Brent and Ptashne, 1981; Uhlin *et al.*, 1982), *uvrA* (the same site as for *ssb*) (Sancar *et al.*, 1982a; Brandsma *et al.*, 1983; Backendorf *et al.*, 1983), and *uvrB* (Sancar *et al.*, 1982b). Two others have been reasonably well identified: at *sulA* (=*sfiA*) (Cole, 1983) and on the plasmid cloDF13 (van den Elzen *et al.*, 1982). Several plasmid promoters may have two deeply overlapping LexA sites (Ebina *et al.*, 1981; van den Elzen *et al.*, 1982; Morlon *et al.*, 1983). Since it is possible that one of these is not functional, which would confuse the analysis, we did not use these sites. Since there are two adjacent sites upstream from the *lexA* gene, the range was limited to 20 bases. This is approximately the region protected by LexA protein from digestion by DNaseI (Little *et al.*, 1981; Brent and Ptashne, 1981). For both the $R_{sequence}$ and $R_{frequency}$ calculations, we assumed that LexA repressor binds to its operators symmetrically (Little and Mount, 1982), and that the center of the symmetry is between bases 0 and 1 (Fig. 3). For the 14 example sequences, $R_{sequence} = 21.1$ bits per site. The nucleotide composition used for this and all remaining recognizers was from *E. coli* chromosomal DNA (LIB2): A = T = 21260, C = G = 21644 (Stormo *et al.*, 1982a). $H_g = 1.99994$ bits/base.

The damage-inducible (*din*) genes are spread around the *E. coli* genome (Little and Mount, 1982), so the size of *E. coli* DNA determines G. There are at least 11 chromosomal genes under *lexA* control (Little and Mount, 1982), giving a minimum estimate for the number of sites γ, and an upper bound on $R_{frequency}$ of 18.4.

(c) *Trp Aporepressor and Trp Operators*

At least three operons of *Escherichia coli* are transcriptionally controlled by the *trp* aporepressor: the tryptophan biosynthetic operon *trpEDCBA*, the aromatic amino acid biosynthesis operon *aroH* and the gene for *trp* aporepressor itself, *trpR* (Bennett *et al.*, 1976; Gunsalus and Yanofsky, 1980; Singleton *et al.*, 1980; Bogosian *et al.*, 1981; Zurawski *et al.*, 1981; Joachimiak *et al.*, 1983).

A single dimer of aporepressor binds to the operator in the presence of L-tryptophan (Joachimiak *et al.*, 1983). Likewise, each binding site contains a two-fold symmetry protected by aporepressor from nucleases. We define the center of this symmetry to be between positions 0 and 1 (Fig. 4). A deletion ending at one end of the *trp* operator, *trp*ΔLC145, is thought to define the range of the sites, since it does not affect repression (Bertrand *et al.*, 1976; Bennett and Yanofsky, 1978). However, when *E. coli trp* aporepressor is bound to *trp* operator DNA of *S. typhimurium* and the methylation of unprotected purine residues is measured (Oppenheim *et al.*, 1980), the aporepressor protects the region -13 to +14 rather than -11 to +12. We used the range covering 5 bases on either side of this protected area, giving $R_{sequence} = 23.4$ bits per site. If one uses the exact range defined by deletion *trp*ΔLC145, $R_{sequence}$ would be 20.6.

Although non-physiologically high concentrations of *trp* aporepressor can regulate several other operons (Johnson and Somerville, 1983; Bogosian and Somerville, 1983), we calculated $R_{frequency}$ for only three sites. The relevant genome is that of *E. coli*, so $R_{frequency} = 20.3$ bits per site.

(d) *Lac Repressor and the Lac Operator*

One cannot measure an information content from a single sequence. Dyad symmetries in DNA (palindromes) are an exception because both the sequence of the palindrome and its complement are available. This allows us to estimate how much information appears in the *lac* operator (Beckwith, 1978; Goeddel *et al.*, 1978; Sadler *et al.*, 1983a). Gilbert and Maxam (1973) found that the tetrameric *lac* repressor protein protects 24 base pairs from DNase digestion. This is a region from -13 to +10, where the zero is the central base. More recently, exonuclease III digestion gave the range -14 to +16 (Shalloway *et al.*, 1980). To analyze the site we extended the range -16 to +16 by 5 bases on both sides (Fig. 5). This range includes the "extended operator" (Dickson *et al.*, 1975; Heyneker *et al.*, 1976). As with other operators, the sequence was compared to its complement using the program Rseq. The central position was included, giving $R_{sequence} = 19.2$ bits per site. Because there are only two examples, there is a large sampling error. If there is only one functional *lac* repressor binding site in the *E. coli* genome, then $R_{frequency} = 21.9$ bits per site. "Pseudo"-operator sequences exist for which there is no known function (Reznikoff *et al.*, 1974; Winter and von Hippel, 1981). If we include the strong secondary "pseudo"-operator, $R_{sequence} = 16.2 \pm 2.6$ and $R_{frequency} = 20.9$ bits.

(e) *ArgR and Arg Boxes*

The gene *argR* encodes a repressor that controls the synthesis of enzymes of arginine biosynthesis (Maas and Clark, 1964; Maas *et al.*, 1964). Several symmetrical binding sites have been tentatively identified by a few mutations and sequence similarities (Cunin *et al.*, 1983). Since some sites are adjacent, the range only covered 20 base pairs (Fig. 6). Also, we used an alignment for the ArgR sequence that was shifted one base to the left of that in Cunin *et al.* (1983). This is presumably a better alignment because it increased $R_{sequence}$ by 1.5 bits. (It would also improve the "consensus".) $R_{sequence} = 16.4$ while $R_{frequency} = 18.4$ bits per site.

By avoiding overlapping sites, we may have deleted part of the arginine boxes. It is possible that two neighboring sites can interpenetrate, if the recognizers bind to different faces of a DNA helix (Hochschild *et al.*, 1983). If the sites were extended to a range -15 to +16, $R_{sequence}$ becomes 18.6. In any case, the sites of the arginine regulon have not yet been characterized by DNase footprinting, chemical protection or other experiments and several more sites remain to be sequenced.

(f) *cI Repressor, cro and l Operators*

All six symmetrical operators of bacteriophage l are bound by both of the dimeric proteins repressor and cro (Ptashne *et al.*, 1976, 1980; Johnson *et al.*, 1981; Matthews *et al.*, 1983). Maniatis *et al.* (1975) originally suggested that the sites are 17 basepairs wide, separated by A-T rich "spacers". Since then it has been thought that these regions are not part of the sites. However, a nonrandom sequence contains information. Chemical protection experiments that probed for guanine residues (Humayun *et al.*, 1977a,b; Johnson *et al.*, 1978; Pabo *et al.*, 1982) did not address the issue since the region is almost completely devoid of G's and contacts in the region may not be directed to GC pairs. Adenine residues were unprotected either because the proteins do not cover that region or because the proteins bind to the opposite side of the DNA from the modifiable group. Two promoter mutations in these regions increase the A-T richness and do not affect repressor binding (Ptashne *et al.*, 1976; *prm116*, Meyer *et al.*, 1975; *sex1*, Kleid *et al.*, 1976). One mutation, *prm up-1*, decreases the A-T richness. The effect of *prm up-1* on repressor binding is said to be small (Johnson *et al.*, 1979; Meyer *et al.*, 1980). In contrast to this mutant, nuclease protection experiments show the sites to be 25 base pairs wide (Humayun *et al.*, 1977a). Thus it is possible that a portion or all of the "spacers" are part of the binding sites. However, in keeping with the rules defined in Materials and Methods, we used a range 19 bases wide to avoid overlap between O_L3 and O_L2 (Fig. 7). (This also avoids the *prm up-1* site.) Most information content of the "spacers" was lost by this procedure; $R_{sequence} = 17.1$, $R_{frequency} = 19.3$ bits per site. If overlaps are ignored, and the sites extended to the size protected from DNase (25 base pairs wide, -12 to +12), $R_{sequence}$ becomes 19.0.

(g) *T7 RNA Polymerase and T7 Promoters*

One of the early bacteriophage T7 proteins, encoded by gene 1, is a new RNA polymerase (Chamberlin *et al.*, 1970). This polymerase transcribes the middle and late genes of the phage genome. Concurrently, the T7 proteins encoded by genes 0.7 and 2 inactivate the host RNA polymerase so that transcription is directed to the T7 genome rather than that of the host (Hesselbach and Nakada 1977a,b; see Studier, 1969, 1972; Kr*uumlaut*ger and Schroeder, 1981; Dunn and Studier, 1983 for reviews on T7).

All 17 T7 RNA polymerase promoters have been sequenced (Dunn and Studier, 1983). *In vitro* deletion experiments and homology among the promoters suggest that a functional promoter is at least 32 base pairs long. Five bases beyond the range -24 to +7 was used to calculate $R_{sequence}$ (Fig. 8). (The zero base is thought to be the start of each transcript, see Fig. 9 for the alignment.) $R_{sequence} = 35.4$ bits per site.

To calculate $R_{frequency}$, we must determine both G and γ. There are two genomes that can contribute to the potential binding sites: the host and the phage. The host DNA is destroyed by gene products 3 (endonuclease, Center *et al.*, 1970) and 6 (exonuclease, Sadowski and Kerr, 1970) which are synthesized from T7 RNA polymerase dependent transcripts. They are therefore made following the synthesis of the T7 RNA polymerase. This means that the gene 1 product may search both the *E. coli* and T7 genomes. The T7 genome is only one hundredth the size of the host genome, so it does not contribute much. The relevant genome is probably the host DNA. Because promoters are asymmetric, there are twice as many potential binding sites on the genome as there are base pairs, so G is twice the genomic size of *E. coli* (Table 1).

The transcriptional map of T7 is known in great detail (Carter *et al.*, 1981); there are almost certainly no more than 17 T7 polymerase sites (Dunn and Studier, 1983). The activity of T7 RNA polymerase on *E. coli* DNA is 4% of its activity on T7 DNA (Chamberlin and Ring, 1973; see also Summers and Siegel, 1970). Therefore the total number of sites on *E. coli* DNA could be (17 sites/39936 bp T7) x (3.9 x 10^6 bp *E. coli*) x 0.04 = 66. On infection by T7, there could be as many as 83 sites in the cell. This gives a lower bound for $R_{frequency}$ of 16.5 bits per site. If there are no sites in the *E. coli* genome, and thus only 17 sites in the cell, Rfrequency would be 18.8 bits per site. This is the first case for which $R_{sequence}$ is much bigger than $R_{frequency}$, so we studied the sequences more closely.

Oakley and Coleman (1977; Oakley *et al.*, 1979) observed that several of the T7 promoters contain a symmetric element centered between bases -3 and -2. The 17 promoter sequences are presented in Fig. 9. The extent of the symmetry in all 17 promoters was found by counting numbers of complementary matches between the two halves. For example, position -14 matches the corresponding position +9 in only 5 of the 17 sites. This number is likely to occur if the bases were not correlated. The rest of the complementary matches are tabulated in Table 2. 12 positions have a significantly high number of matches ($p < 0.005$), and these are taken to represent the symmetry. (The positions -6 and 1 are presumably not involved because they have exceptionally few complementary matches.) Several of the sites contain CTCnCTA:TAGnGAG, while in a few the GAG is shifted to the left by one position.

The information content of these palindromes was determined from the 17 sequences and their complements (34 sequences total) centered as described above (Fig. 10). The $R_{sequence}$ value given in Table 1. is for the 12 positions of the symmetry. $R_{sequence}$ is 16.4 bits per site. There are at least 17 sites in an infected cell, so $R_{frequency}$ is less than or equal to 17.8 bits per site.

(h) *E. coli RNA Polymerase and E. coli Promoters*

We also measured $R_{sequence}$ for sites recognized by *E. coli* RNA polymerase. Hawley and McClure (1983) compiled data on 112 well characterized *E. coli* promoters. For these

Left position	Right position	Number of matches	Probability of matches
-3	-2	12	8.8×10^{-5}
-4	-1	16	3.0×10^{-9}
-5	0	14	1.1×10^{-6}
-6	1	0	7.5×10^{-3}
-7	2	12	8.8×10^{-5}
-8	3	10	2.5×10^{-3}
-9	4	11	5.3×10^{-4}
-10	5	2	0.11
-11	6	4	0.22
-12	7	3	0.19
-13	8	4	0.22
-14	9	5	0.19
-15	10	3	0.19
-16	11	4	0.22
-17	12	3	0.19

Table 2: Matches between the left and right halves of the T7 promoter symmetry. The probability of each number of matches is calculated from a binomial distribution, where $p(\text{match}) = 0.25$ and $n = 17$.

promoters aligned by the -35 and -10 regions and using the range given by Hawley and McClure, $R_{sequence}$ is only 11.1 bits. There are two difficulties with this analysis. First, a variable gap was introduced between the two regions, which will increase the uncertainty Hs and decrease $R_{sequence}$ substantially, perhaps as much as 2 bits (unpublished observation). The other difficulty is that a reasonable estimate for the number of promoters in *E. coli* does not exist, so $R_{frequency}$ cannot be estimated. Nevertheless, promoters may be more frequent in *E. coli* (one per 500 bp) than is commonly assumed (see Discussion).

4. DISCUSSION

(a) *Measurement of* $R_{sequence}$

Many authors have estimated the frequency of a binding site by considering the site size (Gilbert and M*uumlaut*ller-Hill, 1970; Riggs *et al.*, 1970; M*uumlaut*ller-Hill *et al.*, 1977; Nei and Li, 1979; Pribnow, 1979; von Hippel, 1979; Harel, 1980). $R_{sequence}$, the sum of $R_{sequence}(L)$ over a binding site, is similar to counting the number of bases recognized by a macromolecule. In addition, it takes into account the variation of individual sequences. The sampling error correction prevents one from overestimating the amount of information in the sequences, but can lead to underestimation in some circumstances (see Fig. 1 and Appendix I, page 19).

$R_{sequence}$ does not tell us anything about the physical mechanisms a recognizer uses to contact the nucleic acid. For example, the ribosome prefers a particular base composition in the Shine and Dalgarno region. The mechanism is an RNA/RNA contact. *regA*, the translational repressor of bacteriophage T4 (Wiberg and Karam, 1983) uses protein/RNA contacts. It is possible for two such recognizers to have the same base preferences. Since we use sequences to estimate the probabilities of bases at each position, the analysis will give the same information content for two entirely distinct mechanisms. That is, not only is the mechanism irrelevant to the analysis, but one cannot infer anything about the mechanism from the sequence data, the frequency of bases or the information content because several mechanisms may give the same results. How physical and chemical contacts determine the preferred base frequencies is a separate question (Pabo and Sauer, 1984).

(b) $R_{sequence}$ *for Different Recognizers*

$R_{sequence}$ can be used to investigate relationships between different sites. First, one may ask which binding site has more information than another. For example, ribosome binding sites contain, on the average, less information (11 bits) than do *Eco*RI sites (12 bits). When repressors are compared, $R_{sequence}$ varies between 16 and 23 bits (Table 1), in every case representing an information content higher than for ribosome binding sites. Indeed, individual repressors regulate transcription at a subset of the *E. coli* genes.

Secondly, the information patterns are different for the various repressors. LexA and TrpR have high peaks 3 bases wide while ArgR has double spikes and cI/cro have single spikes. These distinctive morphological differences probably reflect the location and strength of structural contacts between the different repressors and their cognate sites.

(c) *The Relationship between* $R_{sequence}$ and $R_{frequency}$

We showed how to estimate the information contained in several binding sites ($R_{sequence}$), and we determined values for different kinds of sites. But what determines how much information is in a site? One way to approach this question is to make a different measurement, based on "how much information should be needed to locate the sites?" ($R_{frequency}$) and then compare this to the first measurement. The results of each analysis are summarized by the ratio of $R_{sequence}$ to $R_{frequency}$ and their difference (Table 1). For ribosomes, LexA, TrpR, LacI, ArgR and cI/cro, the ratio is close to 1. The *sum* of the differences for the same six systems is -0.7 bits (out of more than 100 bits of total $R_{sequence}$).

The large amount of information at T7 polymerase promoters is surprising. We cannot account for this result by using a different size genome, by changing the number of sites, by sampling error, by overspecification to avoid host sites, or by comparison to *E. coli* promoters. However, there is a simple explanation. The sites have twice as much information as is necessary to locate them in a genome the size of *E. coli*. Therefore, *a second recognizer could be using the extra bits*. The sites have symmetry elements that by themselves contain roughly half the information of the entire site. Since T7 RNA polymerase transcribes T7 DNA strictly in one direction (Chamberlin *et al.*, 1970; Summers and Siegel, 1970; Carter *et al.*, 1981; Zavriev and Shemyakin, 1982), it is surprising to find such strong symmetry elements in the promoter sequences. Because the polymerase acts asymmetrically, we assign it to the asymmetric portion of the site.

The symmetric elements could then be the binding site for the second recognizer. Symmetric elements in promoters suggest the presence of operators (Chamberlin, 1974; Dickson *et al.*, 1975; Dykes *et al.*, 1975; Smith, 1979; Ptashne *et al.*, 1980; Gicquel-Sanzey and Cossart, 1982; Joachimiak *et al.*, 1983). With this in mind, it is intriguing that wild type T7 bacteriophage decreases late mRNA synthesis around 10 minutes after infection, while an amber mutation in gene 3.5 prevents the shutoff; therefore gene product 3.5 is a candidate repressor of late T7 transcription (McAllister and Wu, 1978; McAllister *et al.*, 1981; Studier, 1972; Inouye *et al.*, 1973; Jensen and Pryme, 1974; Kerr and Sadowski, 1975; Silberstein *et al.*, 1975; Kleppe *et al.*, 1977; Miyazaki *et al.*, 1978; Kr*uumlaut*ger and Schroeder, 1981; Dunn and Studier, 1983).

The $R_{sequence}$ to $R_{frequency}$ ratio of 2 suggests that there are likely to be two sites at T7 late promoters. In almost all the examples other than T7, a ratio of 1 for $R_{sequence}/R_{frequency}$ suggested *one* site. The exceptional case now becomes the l operators, where we *know* that two different proteins bind: cI repressor and cro. (The effects of the third protein that binds these regions, *E. coli* RNA polymerase, are probably blurred out when $R_{sequence}$ is measured.) The existing biochemical and genetic data show that cI and cro bind to the same nucleotides (Johnson *et al.*, 1981). Both l repressor and cro are dimers that can bind symmetrically and so may share binding site information. If the two proteins used identical information, the ratio would be 1. If they had used different information the ratio could have been as high as 2, as occurs in the T7 promoter/operator sites. In T7, the proposed repressor would bind symmetrically, and so it could not depend only on information in the asymmetric promoter. Conversely, the polymerase could not depend entirely on symmetrical patterns. That is, asymmetric and symmetric sites must have some separate information.

(d) *How are Secondary Sites Avoided?*

Sequences that are "similar" to true sites might compete with the true sites for binding to the recognizer. For example, the *E. coli* genome should contain about 1,000 *Eco*RI restriction enzyme sites (GAATTC), but that same genome should also contain about 18,000 sequences one nucleotide removed from an *Eco*RI site. Site recognition by and action of *Eco*RI within *E. coli* must include enough discrimination against the more abundant similar sites to avoid a fragmented genome (Pingoud, 1985). Restriction enzymes have enough specificity to do this. It seems that many recognizers do not because operator mutations may decrease binding by only 20 fold (Flashman, 1978). Most single base changes in promoters and ribosome binding sites decrease synthesis by 2 to 20 fold (Mulligan *et al.*, 1984; Stormo, 1985). Binding to similar sites would degrade the function of the entire system. For repressors, binding to pseudo-operators would increase the chances of gratuitously inhibiting transcription and may also serve as a sink for the recognizer. For ribosomes, binding sites within mRNAs would lead to the expression of inactive protein fragments.

There are several solutions to the problem of avoiding many similar sites when the recognizer has limited specificity (Linn and Riggs, 1975). It is possible that similar sites are hidden so that they do not interfere. For example, mRNA secondary structure could prevent ribosomes from inspecting sites similar to ribosome binding sites (Gold *et al.*, 1981). Chromatin structure may occlude the DNA, so that repressors do not actually have as many potential binding sites as the number of base pairs. A related possibility is that similar sites do not exist in the genome. For example, if a repressor's binding site is composed of oligos that are relatively rare in the genome, the number of similar sites could be many fewer than expected just from mono-nucleotide information. Any such special effects constrain the genome to particular oligonucleotide patterns. Discrimination against some oligonucleotides might account for the observed non-random distribution of oligonucleotides in the genome (Grantham *et al.*, 1981; Stormo *et al.*, 1982a; Fickett, 1982; Nussinov, 1984). Finally, von Hippel (1979) pointed out that recognizers could enhance site selectivity by binding to longer sites. If a repressor were to recognize a fifteen base pair long sequence in *E. coli*, not only could its site be unique, but there might not be any sites with one mismatch. When this strategy is used, one expects $R_{sequence}$ to exceed $R_{frequency}$. The sampling error correction we made may have lead to an underestimate of $R_{sequence}$ (see Fig. 1). It is also possible that $R_{sequence}$ would be larger if it were calculated from longer oligos, rather than mononucleotides. We are usually prevented from doing that measurement because the sampling error variance increases rapidly. Still, our results suggest that $R_{sequence}$ is usually close to $R_{frequency}$.

(e) *Why is $R_{sequence}$ Approximately Equal to $R_{frequency}$?*

$R_{frequency}$ is a function of genome size and the number of sites. Both of these quantities are fixed by factors that have little to do with recognition: genome size is essentially invariant within a species, and the number of sites required by the organism is fixed by physiology and genetics. For example, a ribosome binding site must precede every gene and the number of genes is determined by physiology and evolutionary history. Unless the population of organisms is undergoing speciation or rapid change in a new environment (Gould, 1977), there is a reasonably fixed frequency of sites and thus $R_{frequency}$ is approximately fixed. To account for our results, we focus attention on $R_{sequence}$. Sequence drift will keep $R_{sequence}$ from being larger than is needed for the regulatory process to function properly. If an organism were to have a collection of sites that were more conserved in sequence than was required, mutations in some of the positions of the sites could be tolerated. This would mean an increase in the uncertainty Hs at those positions in the site and a decrease in $R_{sequence}$. Uncertainty is related to thermodynamic entropy (Shannon, 1948; Tribus and McIrvine, 1971). Just as the entropy of an isolated system tends to increase, excess binding site information content should tend to atrophy. The lower limit to the drift would be the point at which proper function of the regulatory circuit is diminished.

We are left with many puzzles. How does the information content of sites evolve to equal that needed to find the sites? How is binding energy related to information content? How are chemical contacts related to the base frequencies? What happens in skewed genomes? Lastly, are there situations in biology capable of sustaining large $R_{sequence}$ to $R_{frequency}$ ratios, similar to those observed for the T7 late promoters, but for which there is really only one macromolecular recognizer? That is, could a high information content be advantageous for reasons not encountered in the systems studied thus far?

We thank many friends and colleagues for their suggestions, criticisms and patience during the years that this work evolved. We also thank Phil Bloch for a current estimate of the coding capacity of *E. coli*; F.W. Studier for sending us the sequence of T7 before publication; Michael Perry for a general proof for formula (15); and Kathie Piekarski for typing the manuscript. Computer resources were generously provided by the University of Colorado Academic Computing Services. This work was supported by NIH grant GM28755.

APPENDIX

Calculation of Sampling Uncertainty and Variance

Thomas D. Schneider, Jeffrey S. Haemer and Gary D. Stormo

Using sampling frequencies in place of population probabilities leads to a bias in the uncertainty measurement H (Basharin, 1959). Here we discuss two methods to find the correction factor when estimating H from a few examples. The first method uses an exact calculation of the average uncertainty for small samples. The probability of obtaining a particular combination of n bases, nb, can be found from a multinomial distribution. The information for the combination, H_{nb}, is calculated and weighted by the probability of obtaining the combination. The weighted information summed for all combinations is the desired result, the expectation of H_{nb}, $E(H_{nb})$. The second method uses a formula to approximate the correction factor.

(a) *Exact method*

For the exact calculation of $E(H_{nb})$, there are four choices for each base at a position of a site. If one were to calculate H for each possible combination, and then average them, there would be 4^n calculations to perform, where n is the number of sites sequenced. The exact calculation would be impractical for all but the smallest values of n: note that $n = 17$ implies 10^{10} calculations.

Fortunately the formula for a multinominal distribution allows one to calculate many combinations at once (Breiman, 1969). If na, nc, ng and nt are the numbers of A's, C's, G's and T's in a site and Pa, Pc, Pg, Pt are the frequencies of each base in the genome, then the probability of obtaining a particular combination of na to nt (called nb) is estimated by:

$$P_{nb} = \frac{n!}{na!\ nc!\ ng!\ nt!} P_a^{na} P_c^{nc} P_g^{ng} P_t^{nt}, \tag{11}$$

where $n = na + nc + ng + nt$. The factorial portion on the left is the number of ways that each combination can be arranged. P_{nb} is the probability of obtaining the uncertainty H_{nb}:

$$H_{nb} = -\sum_{b=A}^{T} \left(\frac{nb}{n}\right) \log_2 \left(\frac{nb}{n}\right). \tag{12}$$

Finally, to obtain the average uncertainty as decreased owing to sampling:

$$E(H_{nb}) = \sum_{\text{all } nb} P_{nb} H_{nb}. \tag{13}$$

As a practical matter, one should note that equation (11) can be calculated quickly by taking the logarithm of the right side and spreading out all the components (including the factorials) into a set of precalculated sums (followed by exponentiation).

The catch in formula (13) is to avoid calculating all 4^n combinations. A nested series of sums will cover all the required combinations in alphabetical order:

$$\sum_{\text{all}} y = \sum_{b_1=A}^{T} \sum_{b_2=b_1}^{T} \sum_{b_3=b_2}^{T} \cdots \sum_{b_n=b_{n-1}}^{T} y. \tag{14}$$

At y, in the center of all these sums (nested loops in a computer program) the number of index variables that have value A must be tallied up to obtain na. This must also be done for nc, ng and nt. Several algorithms to simulate these sums are possible. In Fig. 3, we show an algorithm written in Pascal that uses only the variables na, nc, ng and nt to simulate nested loops. The algorithm begins with all A's by setting na to n and the other nb to zero. At each pass through the loop, the sum of $na + nc + ng + nt$ remains invariant. The loop is repeated until the variable DONE is set to true after the combination with all T's has been calculated. Since the combinations are covered in alphabetical order, two combinations such as $AAATCG$ and $TCGAAA$ will be counted only once. The factorial portion of equation (11) accounts for the actual number of combinations. It can be shown that the loop is entered only:

$$(n+1)\ (n+2)\ (n+3)\ /\ 6 \tag{15}$$

times. Since this is polynomial in n, the direct calculation of $E(H_{nb})$ is practical.

With large numbers of sites, the exact calculation of $E(H_{nb})$ still becomes enormously expensive. For ribosome binding sites, n varies with position in the site. Even if the entire sequence around the site were available, there are sites at the 5' end of a transcript, so there are regions in the aligned set that must be blank. It is not practical to calculate $E(H_{nb})$ exactly when n is between 108 and 149 (for the range -60 to +40).

(b) *Approximate method*

The second method to calculate the sampling error correction is from Miller (1955) and Basharin (1959) who derived an approximation for the expectation of a sampled uncertainty, $AE(H_{nb})$, that is good for large n:

$$AE(H_{nb}) = H_g - \frac{s-1}{2\ln(2)n} \qquad \text{(bits per base)} \tag{16}$$

where s, the number of symbols, is 4 for mononucleotides. Fig. 4 shows $E(H_{nb})$ and $AE(H_{nb})$ for several values of n. This table[4] helps one to choose between $AE(H_{nb})$ (a computationally cheap estimate that is inaccurate for small n but accurate for large n) and $E(H_{nb})$ (an exact calculation that is computationally costly for large n). We use $AE(H_{nb})$ above $n = 50$ because the cumulative difference between $E(H_{nb})$ and $AE(H_{nb})$ in a site 100 positions wide would be at most 0.078 bits. The exact $E(H_{nb})$ is used for n less than or equal to 50 since its computation is rapid in this range.

(c) *Use of the Correction Factor*

The two methods of calculation produce the expected uncertainty of n sample bases, $E(H_{nb})$:

$$E(H_{nb}) \ = \ H_g \ - \ e(n) \qquad (\textit{bits per base}). \tag{17}$$

When $Hs(L)$ is calculated from a small sample, it is too small by the amount $e(n)$, on the average. To correct $R_{sequence}(L)$, we use:

$$R_{sequence}(L) \ = \ H_g \ - \ [Hs(L) \ + \ e(n)] \qquad (\textit{bits per base}). \tag{18}$$

[4]The table is the output of the program CalHnb.
http://www.lecb.ncifcrf.gov/~toms/delila/calhnb.html

That is, the uncertainty of the pattern is increased because there is only a small sample. Substituting equations (17) and (18) into (5) gives equation (6). H_g could also be corrected but the correction is negligible if H_g is calculated from a large sample of the organism's sequence.

The curve for $E(H_{nb})$ as a function of the number of example sites, n, (Fig. 5) has several important general properties. As the number of example sites increases, $E(H_{nb})$ approaches H_g (= 2 bits/base in the figures) since the error $e(n)$ becomes smaller. As the number of examples drops, $E(H_{nb})$ also drops (the error increases), until at only one example $E(H_{nb})$ is zero. With only one example, the uncertainty of what the sequence is, $Hs(L)$, is also zero. At this point, $R_{sequence}$ is forced to zero (from equation 6): one cannot measure an information content from only one example.

The sampling error correction results in an interesting effect. If $R_{sequence}$ could be measured for an infinite number of *Hin*cII sites (this would look something like Fig. 1a), the two *peaks* would be 2 bits/base. When the correction is made for a small sample, the peaks are less than 2 bits/base (Figs. 1b and 1c). This appears odd if we *know exactly* what *Hin*cII recognizes. However, given only six examples, we would not be so sure what the "real" pattern is. The sampling error correction prevents us from assuming that we have more knowledge than can be obtained from the sequences alone. That is, the value $e(n)$ represents our uncertainty of the pattern, owing to a small sample size. In the extreme case of one sequence, we have no knowledge of what the pattern at the site is, even though we see a sequence. Because of the correction, $R_{sequence}$ will be underestimated at truly conserved positions when only a few sites are known. $R_{sequence}$ for six *Hin*cII sites in Fig. 1c is estimated to be 8 bits even though we "know" (by looking at more than six examples) that *Hin*cII recognizes 10 bits.

(d) *Variance of the Correction Factor*

$E(H_{nb})$ is the mean of the noisy estimate H_{nb}. The variance (calculated exactly) can be shown to be:

$$Var(H_{nb}) = \left(\sum_{\text{all nb}} P_{nb}(H_{nb})^2 \right) - E(H_{nb})^2. \tag{19}$$

This can be used to estimate the standard deviation of $R_{sequence}$ owing to sampling error. If a site is r bases wide then the standard deviation is $\sqrt{rVar(H_{nb})}$.

References

1. T. D. Schneider. Information content of individual genetic sequences. *J. Theor. Biol.*, 189(4):427–441, 1997. http://www.lecb.ncifcrf.gov/~toms/paper/ri/.

2. G. D. Stormo. Information Content and Free Energy in DNA-Protein Interactions. *J. Theor. Biol.*, 195:135–137, 1998.

3. T. D. Schneider. Measuring molecular information. *J. Theor. Biol.*, 201:87–92, 1999. http://www.lecb.ncifcrf.gov/~toms/paper/ridebate/.

REFERENCES

Abramson, N. (1963). *Information Theory and Coding*, McGraw-Hill Book Co., New York.

An, G., Bendiak, D.S., Mamelak, L.A. and Friesen, J.D. (1981). *Nucl. Acid Res.* **9**, 4163-4172.

An, G. and Friesen, J.D. (1980). *Gene.* **12**, 33-39.

Bachmann, B.J. and Low, K.B. (1980). *Microb. Rev.* **44**, 1-56.

Backendorf, C., Brandsma, J.A., Kartasova, T. and van de Putte, P. (1983). *Nucl. Acids Res.* **11**, 5795-5810.

Basharin, G.P. (1959). *Theory Probability Appl.*, **4**, 333-336.

Beckwith, J.R. (1978). *The Operon*, eds. Miller, J.H. and Reznikoff, W.S., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 11-30.

Bennett, G.N., Schweingruber, M.E., Brown, K.D., Squires, C. and Yanofsky, C. (1976). *Proc. Natl. Acad. Sci., U. S. A.* **73**, 2351-2355.

Bennett, G.N. and Yanofsky, C. (1978). *J. Mol. Biol.* **121**, 179-192.

Bertrand, K., Squires, C. and Yanofsky, C. (1976). *J. Mol. Biol.* **103**, 319-337.

Bogosian, G., Bertrand, K. and Somerville, R. (1981). *J. Mol. Biol.* **149**, 821-825.

Bogosian, G. and Somerville, R. (1983). *Mol. Gen. Genet.* **191**, 51-58.

Brandsma, J.A., Bosch, D., Backendorf, C. and van de Putte, P. (1983). *Nature (London).* **305**, 243-245.

Breiman, L. (1969). *Probability and Stochastic Processes, with a View Toward Applications*, Houghton Mifflin Co., Boston.

Brent, R. and Ptashne, M. (1981). *Proc. Natl. Acad. Sci., U. S. A.* **78**, 4204-4208.

Burton, Z., Burgess, R.R., Lin, J., Moore, D., Holder, S. and Gross, C.A. (1981). *Nucl. Acid Res.* **9**, 2889-2903.

Campbell, J. (1982). *Grammatical Man: Information, Entropy, Language, and Life*, Simon and Schuster, New York.

Carter, A.D., Morris, C.E. and McAllister, W.T. (1981). *J. Virol.* **37**, 636-642.

Center, M.S., Studier, F.W. and Richardson, C.C. (1970). *Proc. Natl. Acad. Sci., U. S. A.* **65**, 242-248.

Chamberlin, M., McGrath, J. and Waskell, L. (1970). *Nature (London).* **228**, 227-231.

Chamberlin, M. and Ring, J. (1973). *J. Biol. Chem.* **248**, 2235-2244.

Chamberlin, M.J. (1974). *Ann. Rev. Biochem.* **43**, 721-775.

Cole, S.T. (1983). *Mol. Gen. Genet.* **189**, 400-404.

Cossart, P., Katinka, M. and Yaniv, M. (1981). *Nucl. Acid Res.* **9**, 339-347.

Cunin, R., Eckhardt, T., Piette, J., Boyen, A., Pi*eacute*erard, A. and Glansdorff, N. (1983). *Nucl. Acids. Res.* **11**, 5007-5019.

Davidson, E.H., Jacobs, H.T. and Britten, R.J. (1983). *Nature (London).* **301**, 468-470.

Dickson, R.C., Abelson, J., Barnes, W. and Reznikoff, W.S. (1975). *Science.* **187**, 27-35.

Dunn, J.J. and Studier, F.W. (1983). *J. Mol. Biol.* **166**, 477-535.

Dykes, G., Bambara, R., Marians, K., Wu, R. (1975). *Nucl. Acids Res.* **2**, 327-345.

Ebina, Y., Kishi, F., Miki, T., Kagamiyama, H., Nakazawa, T. and Nakazawa, A. (1981). *Gene.* **15**, 119-126.

Fickett, J.W. (1982). *Nucl. Acids Res.* **10**, 5303-5318.

Flashman, S.M. (1978). *Mol. Gen. Genet.* **166**, 61-73.

Gatlin, L.L. (1972). *Information Theory and the Living System*, Columbia University Press, New York.

Gicquel-Sanzey, B. and Cossart, P. (1982). *EMBO J.* **1**, 591-595.

Gilbert, W. and Maxam, A. (1973). *Proc. Natl. Acad. Sci., U. S. A.* **70**, 3581-3584.

Gilbert, W. and M*uumlaut*ller-Hill, B. (1970). *The Lactose Operon*, Beckwith, J.R. and Zipser, D., eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, p. 104.

Goeddel, D.V., Yansura, D.G. and Caruthers, M.H. (1978). *Proc. Natl. Acad. Sci., U. S. A.* **75**, 3578-3582.

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

Gould, S.J. (1977). *Ever Since Darwin*, W.W. Norton & Co., Inc., New York.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981). *Nucl. Acid Res.* **9**, r43-r74.

Gay, N.J. and Walker, J.E. (1981a). *Nucl. Acid Res.* **9**, 2187-2194.

Gay, N.J. and Walker, J.E. (1981b). *Nucl. Acid Res.* **9**, 3919-3926.

Greene, P.J., Gupta, M., Boyer, H.W., Brown, W.E. and Rosenberg, J.M. (1981). *J. Biol. Chem.* **256**, 2143-2153.

Gunsalus, R.P. and Yanofsky, C. (1980). *Proc. Natl. Acad. Sci., U. S. A.* **77**, 7117-7121.

Harel, D. (1980). *Comm. ACM.* **23**, 379-389.

Hawley, D.K. and McClure, W.R. (1983). *Nucl. Acids Res.* **11**, 2237-2255.

Hesselbach, B.A. and Nakada, D. (1977a). *J. Virol.* **24**, 736-745.

Hesselbach, B.A. and Nakada, D. (1977b). *J. Virol.* **24**, 746-760.

Heyneker, H.L., Shine, J., Goodman, H.M., Boyer, H.W., Rosenberg, J.,

Dickerson, R.E., Narang, S.A., Itakura, K., Lin, S. and Riggs, A.D. (1976). *Nature (London).* **263**, 748-752.

Hochschild, A., Irwin, N. and Ptashne, M. (1983). *Cell.* **32**, 319-325.

Horii, T., Ogawa, T. and Ogawa H. (1981). *Cell.* **23**, 689-697.

Humayun, Z., Jeffrey, A. and Ptashne, M. (1977a). *J. Mol. Biol.* **112**, 265-277.

Humayun, Z., Kleid, D. and Ptashne, M. (1977b). *Nucl. Acids Res.* **4**, 1595-1607.

Inouye, M., Arnheim, N. and Sternglanz, R. (1973). *J. Biol. Chem.* **248**, 7247-7252.

Jaurin, B. and Grundström, T. (1981). *Proc. Natl. Acad. Sci., U. S. A.* **78**, 4897-4901.

Jaurin, B., Grundström, T., Edlund, T. and Normark, S. (1981). *Nature (London).* **290**, 221-225.

Jensen, H.B. and Pryme, I.F. (1974). *Biochem. Biophys. Res. Com.* **59**, 1117-1123.

Jensen, K. and Wirth, N. (1978). *Pascal User Manual and Report*, Second Edition, Springer-Verlag, New York.

Joachimiak, A., Kelley, R.L., Gunsalus, R.P., Yanofsky, C. and Sigler, P.B. (1983). *Proc. Natl. Acad. Sci., U. S. A.* **80**, 668-672.

Johnson, A., Meyer, B.J. and Ptashne, M. (1978). *Proc. Natl. Acad. Sci., U. S. A.* **75**, 1783-1787.

Johnson, A.D., Meyer, B.J. and Ptashne, M. (1979). *Proc. Natl. Acad. Sci., U. S. A.* **76**, 5061-5065.

Johnson, A.D., Poteete, A.R., Lauer, G., Sauer, R.T., Ackers, G.K. and Ptashne, M. (1981). *Nature (London).* **294**, 217-223.

Johnson, D.I. and Somerville, R.L. (1983). *J. Bact.* **155**, 49-55.

Kalnins, A., Otto, K., R*uumlaut*ther, U. and M*uumlaut*ller-Hill, B. (1983). *EMBO J.* **2**, 593-597.

Kanazawa, H., Mabuchi, K., Kayano, T., Tamura, F. and Futai, M. (1981). *Biochem. and Biophys. Res. Comm.* **100**, 219-225.

Kenyon, C.J., Brent, R., Ptashne, M. and Walker, G.C. (1982). *J. Mol. Biol.* **160**, 445-457.

Kerr, C. and Sadowski, P.D. (1975). *Virol.* **65**, 281-285.

Kleid, D., Humayun, Z., Jeffrey, A. and Ptashne, M. (1976). *Proc. Natl. Acad. Sci., U. S. A.* **73**, 293-297.

Kleppe, G., Jensen, H.B. and Pryme, I.F. (1977). *Eur. J. Biochem.* **76**, 317-326.

Kr*uumlaut*ger, D.H. and Schroeder, C. (1981). *Microb. Rev.* **45**, 9-51.

Lin, S. and Riggs, A.D. (1975). *Cell.* **4**, 107-111.

Lipman, D.J. and Maizel, J. (1982). *Nucl. Acids Res.* **10**, 2723-2739.

Little, J.W., Mount, D.W. and Yanisch-Perron, C.R. (1981). *Proc. Natl. Acad. Sci., U. S. A.* **78**, 4199-4203.

Little, J.W. and Mount, D.W. (1982). *Cell.* **29**, 11-22.

Little, J.W. (1983). *J. Mol. Biol.* **167**, 791-808.

Maas, W.K. and Clark, A.J. (1964). *J. Mol. Biol.* **8**, 365-370.

Maas, W.K., Maas, R., Wiame, J.M. and Glansdorff, N. (1964). *J. Mol. Biol.* **8**, 359-364.

Mackie, G.A. (1981). *J. Biol. Chem.* **256**, 8177-8182.

Maniatis, T., Ptashne, M., Backman, K., Kleid, D., Flashman, S., Jeffrey, A. and Maurer, R. (1975). *Cell.* **5**, 109-113.

Markham, B.E., Little, J.W. and Mount, D.W. (1981). *Nucl. Acid Res.* **9**, 4149-4161.

Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., Fisher, R.G., Takeda, Y. (1983). *Trends Biochem. Sci.* **8**, 25-29.

McAllister, W.T. and Wu, H. (1978). *Proc. Natl. Acad. Sci., U. S. A.* **75**, 804-808.

McAllister, W.T., Morris, C., Rosenberg, A.H. and Studier, F.W. (1981). *J. Mol. Biol.* **153**, 527-544.

Meyer, B.J., Kleid, D.G. and Ptashne, M. (1975). *Proc. Natl. Acad. Sci., U. S. A.* **72**, 4785-4789.

Meyer, B.J., Maurer, R. and Ptashne, M. (1980). *J. Mol. Biol.* **139**, 163-194.

Miki, T., Ebina, Y., Kishi, F. and Nakazawa, A. (1981). *Nucl. Acid Res.* **9**, 529-543.

Miller, G.A. (1955). *Information Theory in Psychology*,

Quastler, H., ed., Free Press, Glencoe Illinois, pp. 95-100.

Miyazaki, J., Ryo, Y., Fujisawa, H. and Minagawa, T. (1978). *Virol.* **89**, 327-329.

Morlon, J., Lloubes, R., Chartier, M., Bonicel, J. and Lazdunski, C. (1983). *EMBO J.* **2**, 787-789.

Mulligan, M.E., Hawley, D.K., Entriken, R., McClure, W.R. (1984). *Nucl. Acid Res.* **12**, 789-800.

M*uumlaut*ller-Hill, B., Gronenborn, B., Kania, J., Schlotmann, M. and Beyreuther, K. (1977). *Nucleic Acid-Protein Recognition*, Vogel, H.J., ed., Academic Press, New York, p. 219.

Nei, M. and Li, W. (1979). *Proc. Natl. Acad. Sci., U. S. A.* **76**, 5269-5273.

Neidhardt, F.C., Vaughn, V., Phillips, T.A. and Bloch, P.L. (1983). *Microb. Rev.* **47**, 231-284.

Newman, A.K., Rubin, R.A., Kim, S. and Modrich, P. (1981). *J. Biol. Chem.* **256**, 2131-2139.

Nussinov, R. (1984). *Nucl. Acid Res.* **12**, 1749-1763.

Oakley, J.L. and Coleman, J.E. (1977). *Proc. Natl. Acad. Sci., U. S. A.* **74**, 4266-4270.

Oakley, J.L., Strothkamp, R.E., Sarris, A.H. and Coleman, J.E. (1979). *Biochem.* **18**, 528-537.

Oppenheim, D.S., Bennett, G.N. and Yanofsky, C. (1980). *J. Mol. Biol.* **144**, 133-142.

Owen, J.E., Schultz, D.W., Taylor, A. and Smith, G.R. (1983). *J. Mol. Biol.* **165**, 229-248.

Pabo, C.O., Krovatin, W., Jeffrey, A. and Sauer, R.T. (1982). *Nature (London).* **298**, 441-443.

Pabo, C.O. and Sauer, R.T. (1984). *Ann. Rev. Biochem.* **53**, 293-321.

Pierce, J.R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise*, Second Edition, Dover Publications Inc., New York.

Pingoud, A. (1985). *Eur. J. Biochem.* **147**, 105-109.

Pribnow, D. (1979). *Biological Regulation and Development*,

Goldberger, R.F., ed., Plenum Press, New York, Vol. 1, pp. 219-277.

Ptashne, M., Backman, K., Humayun, M.Z., Jeffrey, A., Maurer, R., Meyer, B., and Sauer, R.T. (1976). *Science.* **194**, 156-161.

Ptashne, M., Jeffrey, A., Johnson, A.D., Maurer, R., Meyer, B.J., Pabo, C.O., Roberts, T.M. and Sauer, R.T. (1980). *Cell.* **19**, 1-11.

Putney, S.D., Mele*acute*ndez, D.L. and Schimmel, P.R. (1981). *J. Biol. Chem.* **256**, 205-211.

Reznikoff, W.S., Winter, R.B. and Hurley, C.K. (1974). *Proc. Natl. Acad. Sci., U. S. A.* **71**, 2314-2318.

Riggs, A.D., Suzuki, H. and Bourgeois, S. (1970). *J. Mol. Biol.* **48**, 67-83.

Roberts, R.J. (1983). *Nucl. Acids Res.* **11**, r135-r167.

Sadler, J.R., Sasmor, H. and Betz, J.L. (1983a). *Proc. Natl. Acad. Sci., U. S. A.* **80**, 6785-6789.

Sadler, J.R., Waterman, M.S. and Smith, T.F. (1983b). *Nucl. Acids Res.* **11**, 2221-2231.

Sadowski, P.D. and Kerr, C. (1970). *J. Virol.* **6**, 149-155.

Sampson, J.R. (1976). *Adaptive Information Processing*, Springer-Verlag, New York.

Sancar, A., Sancar, G.B., Rupp, W.D., Little, J.W. and Mount, D.W. (1982a). *Nature (London).* **298**, 96-98.

Sancar, G.B., Sancar, A., Little, J.W. and Rupp, W.D. (1982b). *Cell.* **28**, 523-530.

Schneider, T.D., Stormo, G.D., Haemer, J.S. and Gold, L. (1982). *Nucl. Acids Res.* **10**, 3013-3024.

Schneider, T.D., Stormo, G.D., Yarus, M.A. and Gold, L. (1984). *Nucl. Acids Res.* **12**, 129-140.

Schneider, T.D. (1984). Ph.D. thesis, University of Colorado.

Shalloway, D., Kleinberger, T. and Livingston, D.M. (1980). *Cell.* **20**, 411-422.

Shannon, C.E. (1948). *Bell System Tech. J.* **27**, 379-423, 623-656.

Shannon, C.E. (1951). *Bell System Tech. J.* **30**, 50-64.

Shannon, C.E. and Weaver, W. (1949). *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.

Shine, J., Dalgarno, L. (1974). *Proc. Natl. Acad. Sci., U. S. A.* **71**, 1342-1346.

Silberstein, S., Inouye, M. and Studier, F.W. (1975). *J. Mol. Biol.* **96**, 1-11.

Singh, J. (1966). *Great Ideas in Information Theory, Language and Cybernetics*, Dover Publications, Inc., New York.

Singleton, C.K., Roeder, W.D., Bogosian, G., Somerville, R.L. and Weith, H.L. (1980). *Nucl. Acids Res.* **8**, 1551-1560.

Smith, H.O. (1979). *Science.* **205**, 455-462.

Stormo, G.D., Schneider, T.D. and Gold, L.M. (1982a). *Nucl. Acid Res.* **10**, 2971-2996.

Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982b). *Nucl. Acids Res.* **10**, 2997-3011.

Stormo, G.D. (1985). "Translational Initiation" in *Maximizing Gene Expression*, Gold, L. and Reznikoff, W., eds., Benjamin/Cummings Publishing Co., Inc., in press.

Studier, F.W. (1969). *Virol.* **39**, 562-574.

Studier, F.W. (1972). *Science.* **176**, 367-376.

Summers, W.C. and Siegel, R.B. (1970). *Nature (London).* **228**, 1160-1162.

Swift, G., McCarthy, B.J. and Heffron, F. (1981). *Mol. Gen. Genet.* **181**, 441-447.

Tribus, M. and McIrvine, E.C. (1971). *Sci. Am.* **225** (Sept.), 179-188.

Uhlin, B.E., Völkert, M.R., Clark, A.J., Sancar, A. and Rupp, W.D. (1982). *Mol. Gen. Genet.* **185**, 251-254.

van den Elzen, P.J.M., Maat, J., Walters, H.H.B., Veltkamp, E. and Nijkamp, H.J.J. (1982). *Nucl. Acids Res.* **10**, 1913-1928.

Völker, T.A., Gafner, J., Bickle, T.A. and Showe, M.K. (1982). *J. Mol. Biol.* **161**, 479-489.

von Hippel, P.H. (1979). *Biological Regulation and Development*, Goldberger, R.F., ed., Plenum Press, New York, Vol. 1, pp. 279-347.

Warner, J.R. (1979). *DIGRAF: Device Independent Graphics from FORTRAN, User's Guide Version 2.0,* Graphics Development Group, University Computing Center, University of Colorado, Boulder.

Weaver, W. (1949). *Sci. Am.* **181**, 11-15.

Wiberg, J.S. and Karam, J.D. (1983). in *Bacteriophage T4*, Mathews, C.K., Kutter, E.M., Mosig, G., and Berget, P.B., eds., American Society for Microbiology, pp. 193-201.

Winter, R.B. and von Hippel, P.H. (1981). *Biochem.* **20**, 6948-6960.

Yokota, T., Sugisaki, H., Takanami, M. and Kaziro, Y. (1980). *Gene.* **12**, 25-31.

Young, I.G., Rogers, B.L., Campbell, H.D., Jaworoski, A. and Shaw, D.C. (1981). *Eur. J. Biochem.* **116**, 165-170.

Zavriev, S.K. and Shemyakin, M.F. (1982). *Nucl. Acids Res.* **10**, 1635-1652.

Zolg, J.W. and H*aumlaut*nggi, U.J. (1981). *Nucl. Acid Res.* **9**, 697-710.

Zurawski, G., Gunsalus, R.P., Brown, K.D. and Yanofsky, C. (1981). *J. Mol. Biol.* **145**, 47-73.

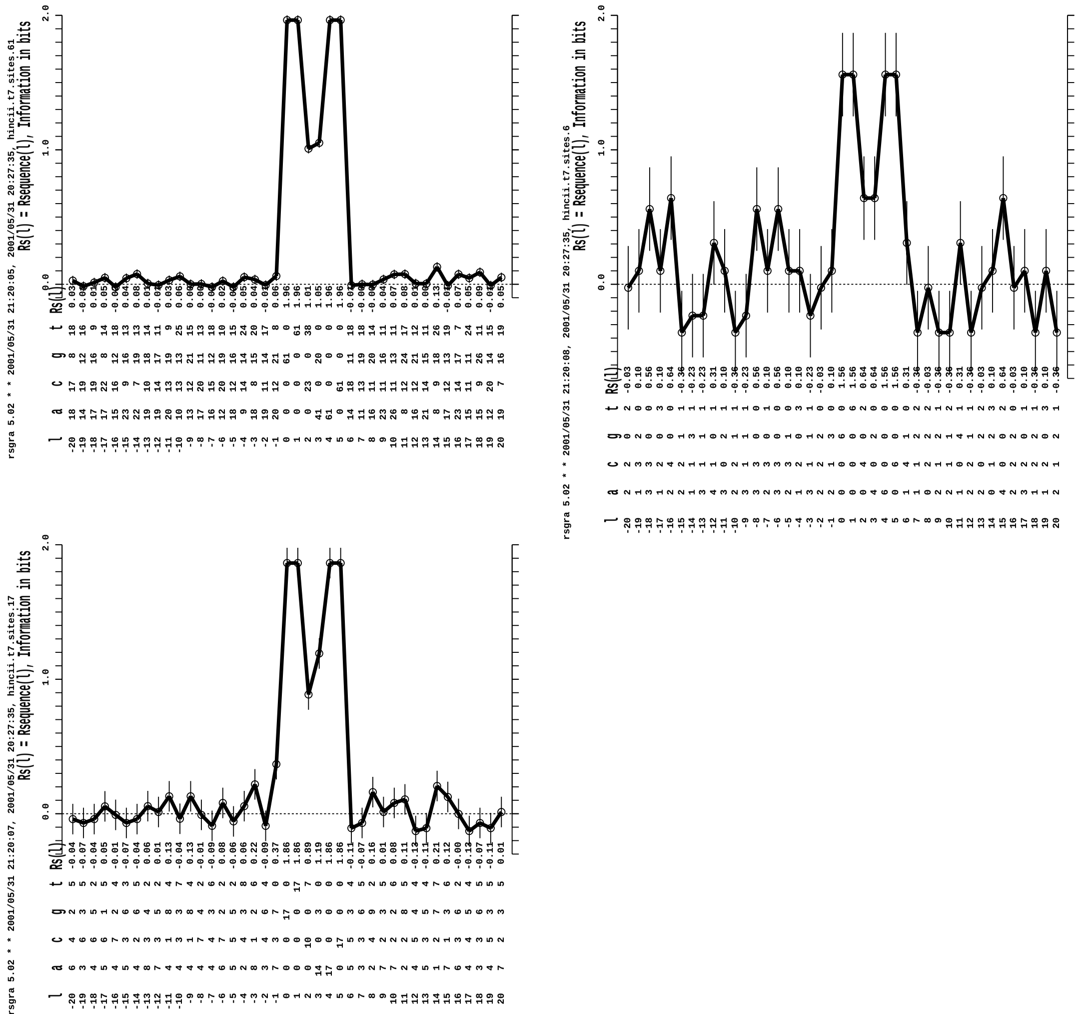
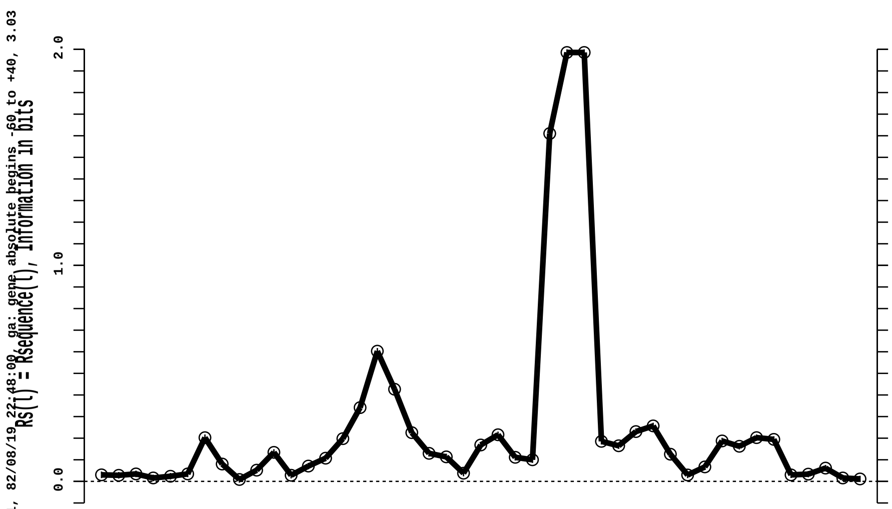


Figure 1: Information content, $R_{sequence}(L)$ in bits/base, at various positions (L) in and around *Hinc*II sites [GT(T/C)(A/G)AC].
The numbers of bases at each position, $n(B, L)$, are given. The sites were obtained starting at the left end of the bacteriophage T7 DNA sequence (Dunn and Studier, 1983) and only one orientation of each site was used. The left-most base in each site (G) was placed at position 0 in each case, and the sequence examined for 20 nucleotides in each direction from this base. The solid lines are the zero without sampling error correction. The dashed lines are the zero when the correction is made. The bars show one standard deviation above or below $R_{sequence}(L)$. They show the variation of the sampling error correction. (a) 61 sites, $R_{sequence} = 10.7 \pm 0.2$ bits; (b) 17 sites, $R_{sequence} = 9.9 \pm 0.7$ bits; (c) 6 sites, $R_{sequence} = 8.3 \pm 2.0$ bits.



l	a	c	g	t	Rs(l)
-26	33	34	25	50	0.03
-25	40	26	29	48	0.03
-24	45	28	25	45	0.03
-23	40	27	31	46	0.02
-22	43	33	24	44	0.02
-21	47	29	25	44	0.03
-20	71	31	16	27	0.20
-19	54	27	21	44	0.08
-18	43	34	28	44	0.01
-17	52	24	29	44	0.05
-16	59	34	14	42	0.13
-15	43	27	30	49	0.03
-14	53	28	22	46	0.07
-13	61	19	39	30	0.11
-12	67	15	44	23	0.20
-11	50	11	73	15	0.34
-10	47	9	88	5	0.60
-9	44	9	82	14	0.43
-8	53	10	60	26	0.23
-7	65	19	36	29	0.13
-6	49	14	36	50	0.11
-5	48	25	30	46	0.04
-4	65	27	15	42	0.17
-3	76	28	20	25	0.22
-2	48	31	16	54	0.11
-1	53	31	17	48	0.10
0	140	0	7	2	1.61
1	0	0	0	149	1.99
2	0	0	149	0	1.99
3	57	13	54	25	0.19
4	48	62	20	19	0.16
5	56	20	13	60	0.23
6	75	35	27	12	0.26
7	67	29	23	30	0.13
8	49	31	26	43	0.03
9	53	20	33	43	0.07
10	41	31	12	65	0.19
11	56	23	16	54	0.16
12	75	26	27	21	0.20
13	63	36	10	40	0.19
14	45	32	25	47	0.03
15	53	28	38	30	0.03
16	41	45	18	45	0.06
17	34	36	29	50	0.02
18	44	28	44	33	0.01

Figure 2: Ribosome binding site information content, determined as for Fig. 1. Position 0 is the first base of the initiation codon.

```
na := n; nc := 0; ng := 0; nt := 0; done := false;
repeat
   (* Calculate equations 11 to 13 here *)
   if nt > 0
   then begin (* ending on a t - do outer loops *)
      if ng > 0
      then begin (* turn g into t *)
         ng := ng - 1;
         nt := nt + 1
      end
      else if nc > 0
      then begin (* turn one c into g,
         and all t to g (note ng = 0 initially) *)
         nc := nc - 1;
         ng := nt + 1;
         nt := 0
      end
      else if na > 0
      then begin (* turn one a into c and
         all g and t to c. (note ng=nc=0 initially) *)
         na := na - 1;
         nc := nt + 1;
         nt := 0
      end
      else done := true (* since nt = n *)
   end
   else begin (* no t - increment innermost loop *)
      if ng > 0
      then begin (* turn g into t *)
         ng := ng - 1;
         nt := nt + 1
      end
      else if nc > 0
      then begin (* turn c into g *)
         nc := nc - 1;
         ng := ng + 1
      end
      else begin (* na > 0; turn a into c *)
         na := na - 1;
         nc := nc + 1
      end
   end
until done;
```

Figure 3: Algorithm corresponding to formula (14).

```
* calhnb 2.25 calculate statistics of hnb
*
* genomic composition:  a = 1,  c = 1,  g = 1,  t = 1
* genomic entropy, hg =   2.00000 bits
*
* n          is the number of sequence examples
* e(hnb)     is the expectation of the entropy hnb calculated from n examples
* ae(hnb)    an approximation of e(hnb) that is calculated
*            more rapidly than e(hnb) for large n
* e diff     e(hnb)-ae(hnb)
* var(hnb)   is the variance of hnb
* avar(hnb)  is the approximate variance of hnb
* std diff   is the difference between the standard deviations
*            (square roots of) var(hnb) and avar(hnb)
* e(n)       hg - e(hnb), the sampling error.
* sd(n)      square root of var(hnb).
*
* units are bits/base, except for the variances which
* are the square of these.
*
*  n    e(hnb)   ae(hnb)    e diff  var(hnb) avar(hnb)  std diff      e(n)     sd(n)
*
   1   0.00000  -0.16404   0.16404   0.00000   4.68308  -2.16404   2.00000   0.00000
   2   0.75000   0.91798  -0.16798   0.18750   1.17077  -0.64901   1.25000   0.43301
   3   1.11090   1.27865  -0.16775   0.18227   0.52034  -0.29441   0.88910   0.42694
   4   1.32399   1.45899  -0.13500   0.15171   0.29269  -0.15151   0.67601   0.38950
   5   1.46291   1.56719  -0.10429   0.12148   0.18732  -0.08427   0.53709   0.34854
   6   1.55923   1.63933  -0.08010   0.09639   0.13009  -0.05021   0.44077   0.31046
   7   1.62900   1.69085  -0.06185   0.07661   0.09557  -0.03237   0.37100   0.27678
   8   1.68129   1.72949  -0.04821   0.06129   0.07317  -0.02294   0.31871   0.24756
   9   1.72155   1.75955  -0.03800   0.04947   0.05782  -0.01802   0.27845   0.22243
  10   1.75328   1.78360  -0.03031   0.04034   0.04683  -0.01555   0.24672   0.20086
  11   1.77879   1.80327  -0.02448   0.03325   0.03870  -0.01439   0.22121   0.18234
  12   1.79966   1.81966  -0.02000   0.02769   0.03252  -0.01392   0.20034   0.16641
  13   1.81699   1.83354  -0.01654   0.02331   0.02771  -0.01379   0.18301   0.15267
  14   1.83159   1.84543  -0.01384   0.01982   0.02389  -0.01380   0.16841   0.14077
  15   1.84403   1.85573  -0.01170   0.01701   0.02081  -0.01384   0.15597   0.13043
  16   1.85475   1.86475  -0.00999   0.01473   0.01829  -0.01387   0.14525   0.12138
  17   1.86408   1.87270  -0.00862   0.01287   0.01620  -0.01385   0.13592   0.11344
  18   1.87227   1.87978  -0.00750   0.01133   0.01445  -0.01379   0.12773   0.10644
  19   1.87952   1.88610  -0.00658   0.01004   0.01297  -0.01367   0.12048   0.10022
  20   1.88598   1.89180  -0.00582   0.00897   0.01171  -0.01352   0.11402   0.09468
  21   1.89177   1.89695  -0.00518   0.00805   0.01062  -0.01333   0.10823   0.08972
  22   1.89699   1.90163  -0.00465   0.00727   0.00968  -0.01311   0.10301   0.08526
  23   1.90172   1.90591  -0.00419   0.00660   0.00885  -0.01287   0.09828   0.08122
  24   1.90604   1.90983  -0.00380   0.00601   0.00813  -0.01262   0.09396   0.07755
  25   1.90998   1.91344  -0.00346   0.00551   0.00749  -0.01235   0.09002   0.07421
  50   1.95594   1.95672  -0.00078   0.00130   0.00187  -0.00726   0.04406   0.03602
  75   1.97081   1.97115  -0.00034   0.00057   0.00083  -0.00501   0.02919   0.02385
 100   1.97817   1.97836  -0.00019   0.00032   0.00047  -0.00381   0.02183   0.01783
 125   1.98257   1.98269  -0.00012   0.00020   0.00030  -0.00308   0.01743   0.01423
 150   1.98549   1.98557  -0.00008   0.00014   0.00021  -0.00258   0.01451   0.01185
 175   1.98757   1.98763  -0.00006   0.00010   0.00015  -0.00222   0.01243   0.01015
 200   1.98913   1.98918  -0.00005   0.00008   0.00012  -0.00195   0.01087   0.00887
```

Figure 4: Statistics of H_{nb} for equiprobable genomic composition.

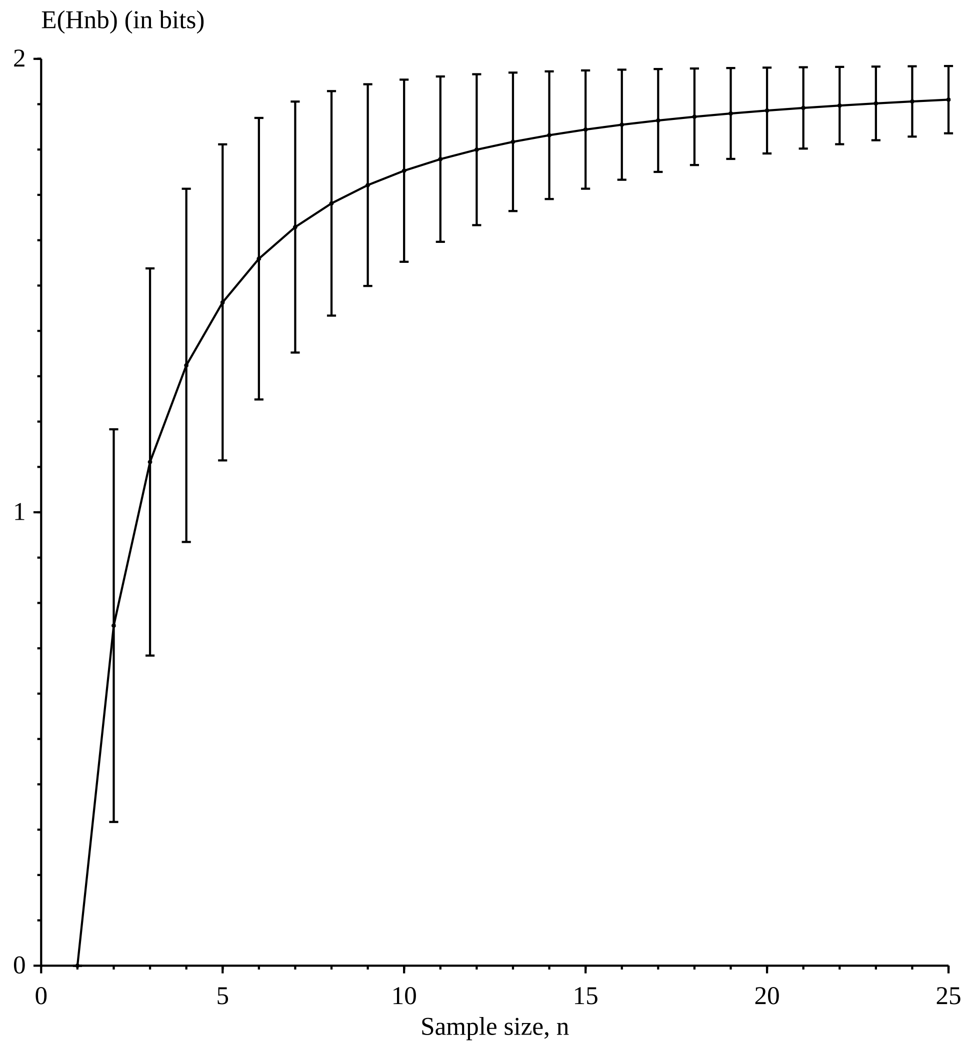


Figure 5: $E(H_{nb})$ vs number of sites, n.

These data are for an equiprobable genomic composition. The curve is less than 1% lower for the composition of *E. coli*. Each bar represents one standard deviation above and below the curve.

Figures

■ **(to be done)**

Figure 6: LexA operator information content, determined as for Figure 1.

■ **(to be done)**

Figure 7: TrpR operator information content, determined as for Figure 1.

■ **(to be done)**

Figure 8: LacI operator information content, determined as for Figure 1.

■ **(to be done)**

Figure 9: ArgR operator information content, determined as for Figure 1.

■ **(to be done)**

Figure 10: λ cI/cro operator information content, determined as for Figure 1.

■ **(to be done)**

Figure 11: T7 promoter information content, determined as for Figure 1.
The center of the symmetry element is marked by a bar and the points of symmetry by dots. The start of transcription at base zero is shown by an arrow.

■ **(to be done)**

Figure 12: T7 promoter symmetry element.
The sequences of the 17 T7 polymerase binding sites are shown. Position zero is presumed to be the start point for transcription (Dunn and Studier, 1983). The position numbers are written vertically. The positions found to be part of the symmetry (Table 2) are shown as capital letters printed in bold face. The GAG's that may be shifted to the left by one base are indicated by an underline.

■ **(to be done)**

Figure 13: T7 symmetry element information content, determined as for Figure 1.
The information content outside the 12 positions of the symmetry element is from the asymmetric promoter sequences.