Information Content of Individual Genetic Sequences

Thomas D. Schneider [*] [†]
with an appendix by John Spouge [‡]

version = 2.20 of ri.tex 1999 December 23
Journal of Theoretical Biology, 189 (4): 427-441, 1997.
http://www-lecb.ncifcrf.gov/~toms/paper/ri/

Running title: Information of Individual Sequences

Key words: DNA-protein interactions, information theory, sequence logo, walker

Subject Category: *Proteins, Nucleic acids and other biologically important macromolecules: molecular structure.*

[*]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/

[†]To whom correspondence should be sent.

[‡]National Library of Medicine, Bethesda, MD 20894, email: spouge@ncbi.nlm.nih.gov

Related genetic sequences having a common function can be described by Shannon's information measure and depicted graphically by a sequence logo. Though useful for many purposes, sequence logos only show the average sequence conservation, and inferring the conservation for individual sequences is difficult. This limitation is overcome by the individual information (R_i) technique described here. The method begins by generating a weight matrix from the frequencies of each nucleotide or amino acid at each position of the aligned sequences. This matrix is then applied to the sequences themselves to determine the sequence conservation of each individual sequence. The matrix is unique because the average of these assignments is the total sequence conservation, and there is only one way to construct such a matrix. For binding sites on polynucleotides, the weight matrix has a natural cutoff that distinguishes functional sequences from other sequences. R_i values are on an absolute scale measured in bits of information so the conservation of different biological functions can be compared to one another. The matrix can be used to rank-order the sequences, to search for new sequences, to compare sequences to other quantitative data such as binding energy or distance between binding sites, to distinguish mutations from polymorphisms, to design sequences of a given strength, and to detect errors in databases. The R_i method has been used to identify previously undescribed but experimentally verified DNA binding sites. The individual information distribution was determined for *E. coli* ribosome binding sites, bacterial Fis binding sites, and human donor and acceptor splice junctions, among others. The distributions demonstrate clearly that the consensus sequence is highly unusual, and hence is a poor method to describe naturally occurring binding sites.

Introduction

A flood of sequence data is appearing in the nucleotide sequence databases. To analyze these data, mathematical methods and computer algorithms are needed that are simple, logical and self-consistent. A mathematics that fits these requirements and also connects directly to the physics underlying molecular binding interactions was created by Shannon with the introduction of information theory (Shannon, 1948; Pierce, 1980; Sloane & Wyner, 1993). Information theory has been successfully used to quantify the sequence conservation in nucleotide and protein sequences (Schneider *et al.*, 1986; Schneider & Stormo, 1989; Eiglmeier *et al.*, 1989; Penotti, 1990; Penotti, 1991; Schneider & Stephens, 1990; Herman & Schneider, 1992; Gutell *et al.*, 1992; Stephens & Schneider, 1992; Papp *et al.*, 1993; Schneider, 1996; Schneider, 1993; Pietrokovski, 1996; Blom *et al.*, 1996). The sequence conservation is given by the average number of bits needed to define a set of aligned sequences. Although this average is useful for understanding the structure of DNA/protein interactions, it does not allow investigation of individual sequences.

This paper describes how the information content of individual sequences can be determined. The method allows direct comparison between the information of particular binding sites to that of other binding sites on the same sequence, to distances between features of the sequence, and to their measured binding energies. It can also be used to search for and to design new binding sites.

Individual information also lends itself to quantitative visualization of complex genetic structures. Previously, only the average picture of a set of binding sites could be depicted graphically by using the sequence logo technique (Schneider & Stephens, 1990). The individual information method described here is the basis of a new graphic method that shows the information contributed by individual bases in a binding site (Schneider, 1997).

With these tools information theory now provides a common framework for investigating many aspects of genetic sequences.

Theory

Individual information of binding sites

The information contained in a set of binding sites can be computed by summing the information content across the base positions of the binding sites (Schneider *et al.*, 1986). But information is an average (Shannon, 1948; Pierce, 1980; Sloane & Wyner, 1993; Schneider, 1995), which suggests that it should be possible to express the average by adding together the information contents of complete individual sequences and then dividing by the number of sequences. This can be done by first creating a weight matrix (Stormo *et al.*, 1982; Schneider *et al.*, 1984; Staden, 1984; Stormo *et al.*, 1986; Stormo, 1990) that assigns an information content to each individual binding site sequence. The matrix is defined so that the average of these values over the entire set of sites is the average information content, as shown below.

The individual information weight matrix is:

$$\begin{aligned} R_{iw}(b,l) &= 2 - (-\log_2 f(b,l) + e(n(l))) \\ &= E(H_{n(l)}) + \log_2 f(b,l) \qquad \text{(bits per base)} \end{aligned} \tag{1}$$

where $f(b,l)$ is the frequency of each base b at position l in the aligned binding site sequences and $e(n(l))$ is a sample size correction factor for the n sequences at position l used to create $f(b,l)$ (Schneider *et al.*, 1986; Penotti, 1990). To simplify the notation, the factor $e(n(l))$ was separated from $\log_2 f(b,l)$ and joined to '2' to create $E(H_{n(l)})$. The reason for writing the double negative will be explained later. Following Shannon's convention, R_{iw} stands for "Rate of information transmission, Individual Weight". Bits per base is a rate like bits per second, especially if we consider the average binding rate in bases per second.

In a set of sequences we represent the j^{th} sequence by a matrix $s(b,l,j)$ that contains only 0's and 1's. For example, the sequence 5′ CAGGTCTGCA 3′ is represented as shown in Fig. 1A. Likewise, an $R_{iw}(b,l)$ matrix for human donor splice junctions is shown in Fig. 1B. ⇐Fig 1

The individual information of a sequence is the dot product between the sequence and the weight matrix:

$$R_i(j) = \sum_l \sum_{b=a}^{t} s(b,l,j) R_{iw}(b,l) \qquad \text{(bits per site).} \tag{2}$$

For the donor splicing weight matrix given in the figure, the sequence 5′ CAGGTCTGCA 3′ is assigned $0.58+1.25+1.64+1.99+1.98+(-3.68)+(-1.59)+1.71+(-0.51)+0.05 = 3.42$ bits per site. Essentially, each base of the sequence "picks out" a particular entry from a column of the $R_{iw}(b,l)$ matrix, and these weights are added together to produce the total R_i.

The average information of the n individual sequences that were used to create the frequency matrix $f(b,l)$ is the expectation (*i.e.* mean) of R_i:

$$E(R_i) = \frac{1}{n} \sum_{j=1}^{n} R_i(j). \tag{3}$$

Now substitute equation (1) into (2) and then substitute equation (2) into (3). By using the definition of the frequency matrix:

$$f(b,l) = \frac{1}{n} \sum_{j=1}^{n} s(b,l,j) \tag{4}$$

and since the frequencies sum to 1:

$$\sum_{b=a}^{t} f(b,l) = 1 \tag{5}$$

some manipulation gives:

$$E(R_i) \;\; = \sum_l \left(E(H_{n(l)}) - \left(- \sum_{b=a}^{t} f(b,l) \log_2 f(b,l) \right) \right). \tag{6}$$

The right hand side is exactly the definition of $R_{sequence}$ (Schneider *et al.*, 1986). This demonstrates that the average of individual information contents is the average information content of the sites. There is only one function that has this property, as shown in Appendix 1.

Relationship between individual information and the roots of information theory: surprisal of bases

By expressing formula (6) as a subtraction, we emphasize that information is a state function defined as a difference of uncertainties (Shannon, 1948; Tribus & McIrvine, 1971; Schneider *et al.*, 1986; Penotti, 1990; Penotti, 1991; Schneider, 1991*a*; Schneider, 1991*b*;

Schneider, 1994). The individual information method is consistent with early work on information theory. Selecting one symbol from a set of M symbols, requires no more than $\log_2 M$ binary decisions (Shannon, 1948). Rearranging the formula gives:

$$\log_2 M = -\log_2 P \tag{7}$$

where $P = 1/M$ is the probability of the equally likely symbols. In general the symbols are not equally likely, as is the case for frequencies of bases in binding sites. To handle this, Tribus (Tribus, 1961) proposed the concept of "surprisal", h_i as the negative logarithm of a symbol's probability in the midst of a stream of symbols:

$$h_i = -\log_2 p_i \tag{8}$$

where p_i is the i^{th} symbol's probability so that (8) is an extension of the form given in equation (7). The advantage of using this definition becomes clear when we consider the average surprisal for the entire stream of symbols. To find this, take the individual surprisals and weight them by their occurrence, p_i, and find the total:

$$H = \sum_i p_i h_i = -\sum_i p_i \log_2 p_i \qquad \text{(bits per symbol).} \tag{9}$$

This is the Shannon uncertainty measure, so H is an average of surprisals (Schneider, 1995).

The recognition process can be modeled by the change an individual recognition "finger" sees when it goes from non-specific binding (the *before* state) to specific binding (the *after* state) (Schneider, 1991*a*; Schneider, 1994). In the *before* state the average surprisal is 2 bits since there are 4 bases, while afterwards it will depend on the frequency of the bases $f(b, l)$ in the binding sites. The decrease in surprisal is:

$$R_{iw}(b, l) = 2 - (-\log_2 f(b, l)) \qquad \text{(bits per base).} \tag{10}$$

This is equation (1) except for the sampling correction. The 2 in equation (10) represents the 2 bits of uncertainty that a recognizer has before it binds to a binding site. Alternatively, the uncertainty associated with binding anywhere on a particular genome ($H_g \leq 2$ (Schneider *et al.*, 1986)) could be used. However, since the recognizer does not make physical contact with the nucleic acid bases in the *before* state the composition of the genome should not matter, so the value 2 seems more appropriate (Schneider, 1991*a*; Schneider, 1991*b*; Schneider, 1994).

Since the individual information is the sum of $R_{iw}(b, l)$ across a binding site, it is the total surprisal decrease from the viewpoint of a particular recognizer binding to a particular sequence. This model allows a recognizer to have different responses to different sequences. Different recognizers have different surprisals for the same sequence because they have different molecular recognition surfaces.

Properties of the individual information distribution

The $R_{iw}(b,l)$ matrix can be applied to each sequence that was used to generate the $R_{iw}(b,l)$ itself. A histogram of the number of sites with a given information versus the information displays the R_i distribution (Fig. 2). The expectation of this distribution is by definition $R_{sequence}$, the total sequence conservation represented by the area under a sequence logo (Schneider & Stephens, 1990). ⇐Fig 2

According to equation (1), by picking out the most frequent base at each position of the weight matrix, the consensus sequence is assigned the largest R_i value, so the upper bound of the R_i distribution is at the consensus. Likewise, choosing the least frequent base at each position gives the lower bound of the distribution, at the "anti-consensus". Since R_i is the sum of a number of small components, its distribution tends to be Gaussian, as dictated by the central limit theorem (Breiman, 1969), assuming that there is only one class of recognizer.

Variance of R_i

Analogous to the mean of the R_i distribution is the spread or variance of the R_i distribution, given by

$$var(R_i) = \frac{1}{n-1}\sum_{j=1}^{n}\left(R_i(j) - E(R_i)\right)^2. \tag{11}$$

For ease of calculation, this may be rewritten as:

$$var(R_i) = \frac{1}{n-1}\sum_{j=1}^{n} R_i(j)^2 - R^2_{sequence}. \tag{12}$$

The standard deviation of the distribution is:

$$\sigma_{R_i} = \sqrt{var(R_i)} \qquad \text{(bits per site).} \tag{13}$$

This number measures how variable the binding sites are.

Standard error of the mean

By using the R_i distribution, we can determine the standard deviation of the mean ($R_{sequence}$), which is known as the standard error of the mean (SEM). The SEM can be determined directly from the standard deviation of the R_i distribution (σ_{R_i}) by

$$SEM = \frac{\sigma_{R_i}}{\sqrt{n}} \tag{14}$$

where n is the number of examples (Taylor, 1982). When many complete sites are available, one can determine the variation of $R_{sequence}$ directly from the individual information distribution. When there are few sequences, the variation of $R_{sequence}$ can also be estimated by

Monte Carlo approximation (Stephens & Schneider, 1992). The SEM plays an important role in molecular information theory, as it allows one to determine quantitatively how much $R_{sequence}$ differs from $R_{frequency}$ (Stephens & Schneider, 1992) or from a multiple of $R_{frequency}$ (Herman & Schneider, 1992).

Individual information variance at each position in a binding site

$R_{iw}(b,l)$ may also be used to determine the variance at *each* position l in the binding site. First we define the individual information at each position l of each sequence j:

$$R_i(l,j) = \sum_{b=a}^{t} s(b,l,j) R_{iw}(b,l). \tag{15}$$

Since the mean at each position is:

$$R_{sequence}(l) = \frac{1}{n}\sum_{j=1}^{n} R_{iw}(l,j) \tag{16}$$

the variance is

$$\begin{aligned} var(R_i(l)) &= \frac{1}{n-1}\sum_{j=1}^{n}\left(R_{iw}(l,j) - R_{sequence}(l)\right)^2 \\ &= \frac{1}{n-1}\sum_{j=1}^{n}\left(R_{iw}(l,j)\right)^2 - R_{sequence}(l)^2. \end{aligned} \tag{17}$$

The standard deviation is:

$$\sigma_{R_i(l)} = \sqrt{var(R_i(l))} \tag{18}$$

Finally, the standard deviation of the mean is the variation of $R_{sequence}(l)$ at each position in the site:

$$SEM(l) = \frac{\sigma_{R_i(l)}}{\sqrt{n}}. \tag{19}$$

These measures may have practical application for producing error bars in the sequence logo display (Schneider & Stephens, 1990) and for testing the hypothesis that positions are independent by calculating individual covariance.

Thermodynamics and individual information

In the case of a molecule binding to a nucleic acid, the zero coordinate on the R_i distribution can be understood from a thermodynamic viewpoint. So far, by avoiding the concept of energy when studying pure sequences, we have avoided making assumptions about the relationship between information and energy. That relationship is not a proportionality, it is the inequality

$$k_{\mathrm{B}} T \ln 2 \leq -q/R \qquad \text{(joules per bit)}, \tag{20}$$

where k_B is Boltzmann's constant, T is the absolute temperature, $-q > 0$ is the heat dissipation to the surroundings (*i.e.* "energy") and R is the information gain. This is an alternative form of the Second Law of Thermodynamics (Schneider, 1991*b*; Schneider, 1994). Because of the inequality, it is not possible to make statements about absolute binding energies given only sequence data, since the latter are purely informational.

Consider a binding site that has a negative evaluation by an $R_{iw}(b,l)$ matrix:

$$R_i < 0. \tag{21}$$

Since Boltzmann's constant k_B, temperature T (under most circumstances, (Waldram, 1985; Atkins, 1984)) and the natural logarithm of 2 are all positive, $k_\mathrm{B}T\ln 2 > 0$. We can therefore multiply both sides of (21) by $k_\mathrm{B}T\ln 2$ and switch sides to obtain

$$0 > R_i k_\mathrm{B} T \ln 2. \tag{22}$$

If binding by only one species of recognizer is responsible for the observed sequence conservation, so that the situations at T7 promoters (Schneider *et al.*, 1986; Schneider & Stormo, 1989) and F *incD* regions (Herman & Schneider, 1992) are excluded, then $R = R_i$ in equation (20). Multiplying both sides of equation (20) by the negative valued R_i gives:

$$R_i k_\mathrm{B} T \ln 2 \geq -q. \tag{23}$$

Transitive combination of equations (22) and (23) and rearranging gives

$$q > 0. \tag{24}$$

Since $q < 0$ corresponds to heat flow out, equation (24) means that heat must flow from the surrounding heat bath *into* the small region of the recognizer/nucleic acid system to make it stay together when the individual information is negative (equation (21)). This is equivalent to pressing on a spring to get the two ends closer together. As soon as one lets go, the energy flows out and the two come apart again. In molecular terms, examples of this are two positive charges or a steric hindrance that would have to be overcome to get the two molecules together. In contrast, the heat flows outward when $R_i > 0$. This increases the entropy of the molecule and the surrounding heat bath and so is (usually) favored. Therefore positive R_i values correspond to binding sites.

Searches using individual information

New sequences can be evaluated and searched for by applying the $R_{iw}(b,l)$ matrix to sequences other than those from which it was derived. Since the numerical value assigned to each position in a sequence by an $R_{iw}(b,l)$ matrix is in bits, the evaluations can be directly compared to the average measures $R_{sequence}$ and $R_{frequency}$ (Schneider *et al.*, 1986).

If a particular base does not appear in the data set used to create the frequency matrix $f(b,l)$, then $f(b,l) = 0$ and so $R_{iw}(b,l) = -\infty$ at that position (see equation (1)). Since there are no known examples of a functioning site containing the base b at position l, there is a high degree of surprisal there. This cannot happen if the matrix is only used to analyze the sequences that were used to make up the matrix itself because the infinite positions are never selected. (Also, when using the dot product method, the fact that $\lim_{f\to 0} f \log f = 0$ ensures that the infinite quantities are suppressed.) Search programs can handle this situation by replacing $-\infty$ with a large negative value. Alternatively, the search may be relaxed by using a less severe penalty (Staden, 1984). The Ri program therefore allows substitution with $1/(n+t)$, with the condition that $t \geq 0$. For example, using $t = 1$ suggests that the missing base would be found if just one more binding site sequence were obtained. However, the "law of succession of Laplace" states that given n trials in which there were k results of one kind, the best estimate for the probability in another trial is $(k+1)/(n+2)$ (Feller, 1968; Papoulis, 1990). In the present case, we need the probability of the absence of a particular base when searching for *another* binding site, so $k = 0$ and the best estimate is $1/(n+2)$. For this reason we set $t = 2$ for most purposes.

Sampling problems and assumptions

It is not possible to determine the information content from a single sequence alone. One reason is that the actual contacts could be anywhere within the sequence, and some positions could be absolutely required (2 bits) while others are completely ignored (0 bits). Without further data, these cannot be distinguished. Another reason is that when frequencies are substituted for probabilities, the information measure becomes biased, and so a small sample correction must be applied (Schneider *et al.*, 1986). When there is only one sequence the bias is so large that the information content calculated at every position is zero. Yet this paper presents a method for evaluating the sequences of individual binding sites, which may at first appear to be impossible. It is possible because the method is performed in two steps: creating a weight matrix and then evaluating the binding sites with that matrix. There is no contradiction because the individual sites are always evaluated by a model created from a large collection of sequences.

If parts of the sequences are unknown, then the average of the individual information contents generally will not equal the $R_{sequence}$ as calculated from the frequencies of bases at each position because individual sequences can be strongly affected by missing data. Missing sequences do not affect the overall frequencies much, so $R_{sequence}$ hardly changes. For this reason calculation of $R_{sequence}$ should still be done by the original frequencies method (Schneider *et al.*, 1986), and individual information values taken from partial sequence data should be interpreted cautiously.

The individual information method depends on an aligned set of sequences. While multiple alignment is a difficult problem in general, for most binding sites gaps are not required to make good alignments because protein binding sites are generally small objects with little flexibility observed along the sequence. We have recently shown that it is possible to perform rapid gap-free multiple alignment based on information theory (Schneider & Mastronarde, 1996). A general theory for individual information with gaps is not available, although the uncertainty introduced by gaps has been considered (Schneider *et al.*, 1986) and hidden Markov models may provide the basis for a solution (Krogh *et al.*, 1994).

The model described here assumes that positions along the site are statistically independent from one another. Fortunately, in the cases which have enough sequence samples to be tested, binding sites show almost complete statistical independence. For example, at most 2% of the information in human splice donor sites is in correlations, and none was observed for acceptors (Stephens & Schneider, 1992). This is also supported by the success of one-layer neural net training (the perceptron) (Stormo *et al.*, 1982; Nakata *et al.*, 1985; Brunak *et al.*, 1990*a*; Brunak *et al.*, 1990*b*; O'Neill, 1991; Horton & Kanehisa, 1992; Bisant & Maizel, 1995). Single layer neural networks depend on additivity, and hence their success demonstrates a good degree of positional independence. Furthermore, the closeness of $R_{sequence}$ and $R_{frequency}$ also supports independence in a number of cases (Schneider *et al.*, 1986). In cases that do not show independence, it should be possible to extend the individual information method to account for bases correlated to their neighbors, or even longer relationships (Stephens & Schneider, 1992; Gutell *et al.*, 1992). However, to do this or to apply it to protein patterns requires many more sequences to avoid the severe effects of small sample size with a large alphabet (Schneider *et al.*, 1986).

Multiple recognizers in a genetic region can affect information theory based models. This problem breaks down into two parts. First, when two or more recognizers have binding sites that are always in the same register with respect to each other, the sequence conservation is higher than expected from the size of the genome and the number of binding sites (Schneider *et al.*, 1986; Herman & Schneider, 1992). If a thorough information analysis has been done, the situation is easy to detect and in such cases it is unwise to use the individual information matrix because it does not represent a single entity. Second, when nearby sites are not in the same register, the sequence conservation of one site is blurred out in the alignment of the other site. For example, there is no hint of a promoter near the *Escherichia coli* CRP binding sites (Schneider *et al.*, 1986; Papp *et al.*, 1993).

Results and Discussion

Information of individual sequences

The first step in individual information analysis of nucleotide binding sites is to gather a number of example sites and to align them using information content as a criterion for good alignment (Schneider *et al.*, 1982; Schneider & Mastronarde, 1996). After computation of the average information content of the binding sites ($R_{sequence}$) (Schneider *et al.*, 1986) and generation of a sequence logo graphic to inspect the average sequence conservation (Schneider & Stephens, 1990; Schneider, 1996), the aligned sequences are used to generate a model of the binding sites that is called the $R_{iw}(b, l)$ matrix (equation (1)). Because this weight matrix is created from many sequences, it can give statistically significant evaluations of individual sequences, including those used to create the matrix itself. Surprisingly, only one simple criterion is needed to completely determine the weight matrix: it must give individual evaluations to a set of binding sites such that the average of the evaluations is $R_{sequence}$.

Single binding site conservation distributions

The individual conservation distributions for ribosome, donor and acceptor sites are shown in Fig. 3. The majority of the individual information values are above zero (99%, 98%, and 97%, respectively, Fig. 3). This confirms the idea that zero has special significance on the distribution (Fig. 2). A particular sequence might have some parts rated negatively, and other parts rated positively such that the total R_i is zero. These sequences have at best no binding energy according to equation (23), so $R_i = 0$ classifies sequences into sites and non-sites. Shannon's channel capacity theorem shows that this can be a sharp demarkation (Schneider, 1991*a*). ⇐Fig 3

Although the distributions are approximately Gaussian, they cannot be exactly Gaussian because the smallest values are truncated at zero. There is also a softer limit at the high end because of the consensus sequence, so the distribution is contained much like a binomial but for practical purposes may be treated as Gaussian.

In rare cases the calculated R_i value is less than zero. This may occur for various reasons. (1) Site sequences may contain sequence or database errors. (2) The R_i is often underestimated when only part of a site's sequence is available. (3) When a limited number of sequences are available to define the distribution, the error for any individual sequence may be appreciable. (4) There may be correlations between parts of the site that are not properly accounted for, although for ribosomes and splice sites these are minimal effects (Stormo *et al.*, 1982; Schneider, 1991*a*; Stephens & Schneider, 1992). (5) There may be several kinds of recognizer sites in the data set, an example of which is the new class of splice junctions discussed below.

Identification of distinct classes of sites

Hall and Padgett (1994) have observed a new class of splice junctions. Using the acceptor site model developed for Fig. 3, the human acceptor sites in the CMP intron G (GenBank accession M55682, coordinate 396) and P120 intron F (GenBank accession M33132, coordinate 7205) are rated as -3.5 and -3.7 bits respectively. This shows that if the binding sites for several different recognizers have been lumped together, the individual information may be used to help identify the different classes. With enough sequences, multiple classes of sites might be detected by a bimodal or multimodal distribution.

Detecting database errors

Like neural networks that have been used to detect errors in a sequence database (Brunak *et al.*, 1990*a*; Brunak *et al.*, 1990*b*), negative individual information values have been used to detect errors in data sets for splice junctions, ribosome binding sites and other binding sites. For example, a search of GenBank (72.0 6/15/92) for entries with "Homo sapiens" in the source line and "exon" in the features gave 4873 entries. The ends of exons were extracted (Schneider *et al.*, 1982) and analyzed with the $R_{iw}(b, l)$ for donor sites from Fig. 1. Of the 6405 exon ends in the 3756 entries that really had exon features, many were not donor sites because many exons end at the poly A site (unfortunately donor and acceptor sites are not explicitly recorded in the database). A large number of exon ends with large negative R_i values were expected (1438 were found), but 842 entries were discovered that had *all* negative values. An example is the locus HUMEMPB42 (accession M60298) which turned out to be a spliced transcript (Korsgren & Cohen, 1991). Although portions of the introns are known (figure 2 of that paper) they were not reported to GenBank, only the abutted exons were. (After the error was reported to GenBank, the entry was corrected.)

Effect of adding new sites to a binding site model

When new sites are added to an individual information model, the evaluation of both the old and new sites changes. Generally this has only a small effect on the old sites once the model has been reasonably well established (Fig. 4). In contrast, new sites almost always ⇐Fig 4 increase in value as underrepresented bases become more appropriately represented. On occasion, addition of one site will significantly increase the value of an old site because the new site contains a second example of a base that previously only appeared in the old site.

Correlating binding site conservation with another binding site or a distance

The "exon definition" model for splicing proposes that the acceptor site is bound first and that the spliceosome then scans downstream *across the exon* to locate the next donor site (Robberson *et al.*, 1990; Talerico & Berget, 1990; Niwa *et al.*, 1992). A weak donor might be compensated by a strong acceptor, or the strength of the donor might be related to the distance from the acceptor, so it is important to check whether there are relationships between the donor and acceptor conservation and the exon and intron lengths. Human donor and acceptor splice sites were collected across complete introns and exons, and the individual information of each donor site was plotted against the corresponding acceptor individual information. The R_i site conservations were also plotted against neighboring intron and exon lengths and the total intron-exon interval surrounding each site. No strong correlations were observed (data not shown). A similar lack of correlations between individual splice junctions and each other or with distances between sites across the intron was first noted by F. E. Penotti (Penotti, 1991). This implies that each human binding site evolves independently to match the spliceosome's molecular surface. Thus R_i can play a role in quantitative analysis of genetic structures.

Correlations of conservation within a single binding site

Not only can correlations between whole sites be made, but also correlations between parts of sites can be investigated. A previous analysis of splice junctions suggested that each comes in two parts (Stephens & Schneider, 1992). To see whether this has an effect on the conservation of these parts, the left half of all donor sites (positions -3 to $+1$ of Fig. 1) was correlated to the right half of the same sites ($+2$ to $+6$) giving $r = -0.37$, and the left half of the acceptor (-25 to -4) was correlated to the right half (-2 to $+2$) giving $r = -0.12$. In each case only a weak negative correlation was observed (data not shown), as expected from the requirement for the whole site to have positive R_i.

Strong binding sites are not always natural binding sites

Probabilities computed from individual information distributions are curious because sequences with evaluations significantly higher than the mean have low probabilities of being real sites, as can be seen in the distributions (Fig. 3). Strong sites are less likely to appear in the set of natural sites. Evidently the sites evolve to what is required for their function rather than to become the strongest binder. That is, the average of the distribution ($R_{sequence}$) evolves to match the information needed to locate the set of sites in the genome ($R_{frequency}$) (Schneider *et al.*, 1986; Schneider, 1988; Schneider, 1994).

Consensus sequences are abnormal binding sites

Many authors have proposed methods for searching for binding sites in nucleic-acid sequences. The "consensus sequence" is widely used by practicing molecular biologists (Day & McMorris, 1992; Prestridge & Stormo, 1993) even though it destroys subtle distinctions in the frequencies of bases in a set of binding sites. This is because choosing the most frequent base at a position is mathematically equivalent to forcing one frequency to 1.0 and all others to 0.0. A glance at some sequence logos (Stephens & Schneider, 1992; Papp *et al.*, 1993) demonstrates that in many binding sites the observed frequencies lie between 0.0 and 1.0 and are not simple fractions such as 1/2 or 1/3.

A consensus sequence is a model of the binding sites. However, to many authors the idea of a consensus sequence has become synonymous with the actual binding sites (Mount *et al.*, 1992; Toledano *et al.*, 1994; Cui *et al.*, 1995). Thus, for example, it is said that "the splice site machinery searches a region of the precursor RNA for a consensus $5'$ splice site" (Robberson *et al.*, 1990) or "The splice points are marked by consensus sequences that act as signals for the splicing process" (Seidel *et al.*, 1992). The simplest consensus sequence is found by selecting the most frequent base at each position, and therefore by equation (1) gives the largest value obtainable from the R_{iw} matrix. As a result, the consensus sequence lies at the high end of the R_i distribution (Fig. 2). The histograms for ribosomes and splice junctions (Fig. 3) show that most binding sites are *not* the consensus.

For *E. coli* ribosomes, the individual information distribution over the base range -21 to $+18$ is characterized reasonably well as a Gaussian distribution having a mean and standard deviation of 8.68 ± 3.42 bits (Fig. 3). The consensus is at 23.98 bits, which is $Z = 4.48$ standard deviations from the mean, so the probability of finding such a sequence in wild type *E. coli* ribosome binding sites is $p < 3.8 \times 10^{-6}$. No single site (of 1055) was the consensus. Since there are only about 4300 genes (GenBank accession U00096), chances are slim that even one consensus sequence exists in the natural population.

For the compact human donor sites, over the range -3 to $+6$, the mean and standard deviation are 7.93 ± 3.22 bits with the consensus at 13.13 bits, giving a Z of 1.6, and a probability of 0.05. In the set of 1799 sites, only 5 (0.3%) were the consensus. Even with such a compact binding site, the consensus is not representative of the whole set.

Acceptor sites, with the range -25 to $+2$, are much more flexible, allowing for a larger consensus. Their mean and standard deviation are 9.44 ± 4.57 bits with the consensus at 21.68 bits, giving a Z of 2.7, and a probability of $p < 3.7 \times 10^{-3}$; none were in the set of 1744 sites.

Thus the consensus, rather than being typical, is improbable. If the consensus were the pattern being searched for by a recognizer molecule, as suggested by the statements quoted above, most sites would not be found. One cannot rescue the consensus method by allowing discrete variations such as "A or G" (Day & McMorris, 1992) since this still distorts the frequency data. For these reasons, consensus sequences are extremely poor models for binding sites.

Comparison with other quantitative methods

Individual information, although independently derived as described above, is related to several other methods that use a matrix. However, important distinctions exist. Information is the only measure that allows one to consistently add together "scores" from each position in a binding site (Shannon, 1948), so other proposed search methods (Mulligan *et al.*, 1984; Shapiro & Senapathy, 1987; Goodrich *et al.*, 1990; Gribskov *et al.*, 1990; Bucher, 1990; Quandt *et al.*, 1995) will give inconsistent results. The logarithm of probabilities was proposed as a useful information measure because it allows addition of the components, assuming their independence (Shannon, 1948). Likewise, various authors have used the natural logarithm of the base frequencies to create a weight matrix (Staden, 1984; Berg & von Hippel, 1987; Bucher, 1990; Rice *et al.*, 1992), but a logarithm alone is not sufficient to identify sites; some kind of cutoff is required, and usually it is chosen arbitrarily. For example, because Staden's method does not add the factor of 2 bits in equation (1), all scores are negative with strong ones closest to zero and so it is not clear where to place a cutoff. Furthermore, all weights at positions with equiprobable bases would be assigned $\ln(0.25)$ so the scale shifts depending on the width of the frequency matrix, and one cannot compare sites for different recognizers to each other. Using a consensus to express a weight matrix evaluation as a percentage of a maximum (Goodrich *et al.*, 1990; Bucher, 1990; Quandt *et al.*, 1995) also prevents comparison between recognizers. Staden's measure also lacks a correction for small sample size. Because these sequence evaluation methods lack an absolute scale of measure, they cannot be used to create a graphic display of binding sites, such as the walker (Schneider, 1997), that is consistent for different recognizers. With natural logarithms the units of the score are "nits", which have to be divided by $\ln 2 = 0.693\ldots$ to be directly comparable to the "bits" used in modern computing and communications systems (Schneider, 1995).

The log-odds method, a derivative of the information theory approach (Schneider, 1984; Schneider *et al.*, 1986; Stormo, 1990), does put different kinds of sites on a common scale in bits. However, the average of the log-odds distribution is not the Shannon information content and does not produce a state function (Schneider, 1991*b*; Schneider, 1994). It therefore cannot be related to standard definitions of entropy and energy, which are state functions. Further, the log-odds computation of the average can give values larger than 2 bits (Stormo, 1990) even though there are only 4 possible bases. This is because the log-odds method measures the information an observer gains rather than the information gained by the molecular system (Schneider, 1991*b*). The states of an external observer are not relevant to molecular interactions, so this computation is not appropriate for the goal of modeling molecular interactions.

The individual information method avoids these various problems by giving an average, consistent with information theory, that allows one to compare different recognizer's sites to each other on an absolute scale given in bits.

The relationship between individual information and "discrimination energy"

The statistical mechanical approach to the analysis of binding sequences assumes that the ratio of the frequencies of bases is related to the energy by a Boltzmann function (Berg & von Hippel, 1987; Berg & von Hippel, 1988*a*; Berg, 1988; Berg & von Hippel, 1988*b*; Stormo, 1990; Penotti, 1990; Penotti, 1991). Strictly following this approach leads to a serious difficulty. At bacteriophage T7 promoters only half of the 35 bit pattern surrounding the transcriptional start is required for transcriptional initiation (Schneider *et al.*, 1986; Schneider, 1988; Schneider & Stormo, 1989). If the observed patterns actually represent energy dissipations, then 35 bits worth of energy is dissipated by the T7 polymerase when it binds. Yet, experiments show that the polymerase only requires 18 ± 2 bits of sequence pattern (Schneider & Stormo, 1989). Since energy must be dissipated to the surroundings to be useful for molecular binding, what happened to the "undissipated" energy? How can there be "discrimination energy" that is not dissipated by polymerase binding? This difficulty can be avoided by referring only to the information in the sequence patterns: half of the 35 bit pattern is used by the polymerase, and the other half is presumably used by a different recognizer when it binds. The difficulty with the statistical mechanical approach stems from an assumption that energy is equivalent to information. The Second Law of Thermodynamics shows that information and energy are related, but by the inequality in equation (20).

Associated with the idea of "discrimination energy" is a parameter called λ that defines the relationship between sequence information and measured binding energies. λ could be a function of the position in the binding site since the information could be closer or further from its ideal maximum given by equation (20). That is, some binding positions could dissipate more energy than absolutely necessary to specify a bit while other positions could dissipate just exactly the minimum amount. (An entire binding site should not be able to beat the Second Law, but it would be interesting to look for parts of a binding site that do so by "coming along for the ride" as negative weights within functional sites. Several potential candidates are shown by the upside-down bases of the walker positioned on the functional site at base 180 in the middle sequence of Figure 1 in (Schneider, 1997). Confirmation would require experimental studies of the binding energetics of these positions.)

The discrimination energy method compares the frequency of a base at a position in a binding site to the frequency of the consensus base at the same position (Berg & von Hippel, 1988*b*; Stormo & Hartzell, 1989). However, the discrimination energy can easily be calculated from the $R_{iw}(b,l)$ matrix. Let $R_{iw}(\text{consensus},l)$ be the evaluation of the consensus base at position l, where "consensus" is the most frequent base. Then, equation (1) gives:

$$R_{iw}(\text{consensus},l) = E(H_{n(l)}) + \log_2(f(\text{consensus},l)). \tag{25}$$

The discrimination energy measure (DE) is:

$$\begin{aligned} DE(b,l)/\ln 2 &= R_{iw}(\text{consensus},l) - R_{iw}(b,l) &(26)\\ &= \log_2(f(\text{consensus},l)/f(b,l)) &(27)\end{aligned}$$

where the factor of $\ln 2$ converts the nits in the original definition of DE into bits for direct comparison to the R_i values. When this DE matrix is used to evaluate a sequence, we sum equation (27) over the sequence (using equation (2)), so the result is:

$$DE(sequence)/\ln 2 = R_i(consensus) - R_i(sequence). \tag{28}$$

Thus the R_i method can produce the DE result, but the DE method is relative to a standard sequence, usually the consensus, and therefore the scale changes between different binding sites while R_i does not. From the second law and the results for ribosome binding sites and splice junctions it appears that $R_i = 0$ corresponds to a cutoff for functional sites, a feature that the DE method lacks. Furthermore a larger "discrimination energy" corresponds to *worse* binding (Berg & von Hippel, 1988*b*), which is counter-intuitive. Finally, DE uses the consensus, which is an extreme binding site (Fig. 2). However, as Berg and von Hippel note (Berg & von Hippel, 1988*b*), one can use a different reference sequence to obtain similar results, but this would shift the scale. In contrast, the R_i method compares the sites to a string of equiprobable bases, which accurately represents the non-specific binding of the recognizer to the nucleic acid. Using this reference state, the individual information measure is directly comparable to the energy dissipated by the molecular machine during its operation (Schneider, 1991*a*; Schneider, 1991*b*).

Individual information compared to training methods

With ideal data sets the individual information search method probably cannot give results as good as artificial intelligence methods such as the perceptron (Stormo *et al.*, 1982), neural nets (Nakata *et al.*, 1985; Brunak *et al.*, 1990*a*; Brunak *et al.*, 1990*b*; O'Neill, 1991; Horton & Kanehisa, 1992; Bisant & Maizel, 1995), categorical discrimination (Iida, 1987) or hidden Markov models (Krogh *et al.*, 1994) because those methods have the advantage of training on sequences that are not sites.

In practice, however, extensive experimental analysis is needed to avoid contamination of the negative training set by functional sites. In contrast, the information theory method does not require such sites or any cyclic training. As soon as a few experimentally proven sample sites are available, the R_{iw} matrix can be constructed by using an equation. The difficult task of collecting large sets of experimentally demonstrated non-functional sequences is avoided, so there is no concern that one may have contaminated the non-functional examples with real sites (Horton & Kanehisa, 1992). This lack of training is particularly advantageous for identifying errors for any kind of binding site recorded in a sequence database (Schneider, 1997).

Conclusion

The individual information method is simple, but has many useful properties. By this method, individual binding sites can be compared directly to the overall information content, $R_{sequence}$, since by definition $R_{sequence}$ is the average of R_i over the sites. This also allows direct comparison to the predicted average information content given the size of the genome and number of sites ($R_{frequency}$) (Schneider *et al.*, 1986). It shows that there is a relationship between the evolution of specific genetic control points and the overall control mechanism in the cell. Individual sequence conservation is measured in standard units, bits, that are easy to manipulate (Schneider, 1995) and allow a wide variety of biological systems to be compared to each other. Because R_i calculations make no assumptions about binding energies, the relationship between energy and information can be investigated experimentally. Applications of the method include the graphical display and engineering of entire genetic control systems (Schneider, 1997) and dissection of binding sites to reveal new kinds of genetic control systems (Hengen *et al.*, in preparation).

Materials and Methods

Programs

Programs of the Delila system (Schneider *et al.*, 1982; Schneider *et al.*, 1984; Schneider *et al.*, 1986; Schneider & Stephens, 1990) were used to collect and analyze the sites. The Ri program (version 2.37) generates a $R_{iw}(b,l)$ matrix and correlates individual sites with quantitative data. The Scan program (version 2.88) uses the weight matrix to perform searches (Hengen *et al.*, 1997). It reports the evaluation of each sequence position j in three ways: as the individual information ($R_i(j)$), as the standard deviation from the wild type distribution mean ($Z(j) = (R_i(j) - R_{sequence})/\sigma_{R_i}$) and as the one tailed probability ($p(j)$, computed from $Z(j)$ assuming a normal distribution). The DNAplot program (version 3.40) graphs the results in PostScript (Hengen *et al.*, 1997). Histograms (Fig. 3) were generated by the GenHis program (version 1.73) written by G. Stormo, and displayed in PostScript by the GenPic program (version 2.20). X-Y plots and correlation coefficient computations (Fig. 4) were performed by the Xyplo program (version 8.63). See http://www-lmmb.ncifcrf.gov/~toms/ for further information about the programs.

Sequences

Ribosome binding sites were from Kenn Rudd's EcoSeq5 database (Rudd & Schneider, 1992). Human donor and acceptor splice sites were those described in (Stephens & Schneider, 1992). Fis sites are described in (Hengen *et al.*, 1997).

Acknowledgments

I thank Greg Alvord for his help with the statistics, John S. Garavelli for pointing out the mathematics behind the Laplace method, David Lipman for access to GenBank, John Spouge, John S. Garavelli, Kenn Rudd, Paul N. Hengen, Elaine Bucheimer, Denise Rubens, Hugo Martinez, Dong Xu, Stacy Bartram, Kenn Rudd, and especially Pete Rogan for useful conversations and comments on the manuscript, Stacy Bartram for writing DNAscan, and the Frederick Biomedical Supercomputing Center for computer resource support.

APPENDIX 1 Proof that R_i is the only function whose average is $R_{sequence}$

John Spouge, National Library of Medicine, Bethesda, MD 20894

First, from equation (6) and $E(R_i) = R_{sequence}$,

$$\sum_l E(H_{n(l)}) - R_{sequence} = \sum_{l=1}^{L} j(b,l) \tag{29}$$

where $j(b,l) = -\sum_{b=A}^{T} f(b,l)\log_2 f(b,l)$ and $1 \ldots L$ is the range of positions l in the site. This Appendix asserts that if the function h in

$$j(b,l) = -\sum_{b=A}^{T} f(b,l)h[f(b,l)] \tag{30}$$

also satisfies (29) for all base frequencies at position l, then the problem's constraints imply that $h(p) = \log_2 p$. Hence R_i is the only function whose average is $R_{sequence}$.

We reduce the problem further by adding a new position to the range. Initially the site runs from $1 \ldots L$ and

$$\sum_{l=1}^{L} E(H_{n(l)}) - Rs_1 = \sum_{l=1}^{L} j(l). \tag{31}$$

With an extended range $0 \ldots L$,

$$\sum_{l=0}^{L} E(H_{n(l)}) - Rs_2 = \sum_{l=0}^{L} j(l) \tag{32}$$

so

$$Rs_2 - Rs_1 = E(H_{n(0)}) - j(0). \tag{33}$$

Thus the problem is reduced to considering a single (arbitrary) position.

Now we only need to show that at position 0 if

$$j(0) = -\sum_{b=A}^{T} f(b,0)\log_2 f(b,0) = -\sum_{b=A}^{T} f(b,0)h[f(b,0)] \tag{34}$$

for all frequency vectors $f(b,0)$, then $h(p) = \log_2 p$, where $p = f(b,0)$.

If we define $g(p) = p\log_2 p - ph(p)$ then showing that $g(p) = 0$ for all p would finish the proof, since then $h(p) = \log_2 p$ except possibly at $p = 0$, but the latter value can be ignored because of the multiplication by $p = f(b,0)$ in (34).

We shall insist that $h(p)$ should be continuous, so the same is true of $g(p)$. Moreover, equation (34) gives

$$\sum_{b=A}^{T} g(f(b,0)) = 0 \tag{35}$$

for all frequency vectors $f(b,0)$.

Because a new base β (Piccirilli *et al.*, 1990) with frequency zero (*i.e.* $f(\beta, 0) = 0$ always) should not affect equation (35), $g(f(\beta, 0)) = g(0) = 0$. The frequency vector

$$f(b, 0) = \{1, 0, 0, 0\} \tag{36}$$

in equation (35) gives us that $g(1) = 0$. This shows that the two ends of the distribution are the same. Substituting the two frequency vectors $p = \{p + q, 1 - p - q, 0, 0\}$ and $p = \{p, q, 1 - p - q, 0\}$ into equation (35) gives

$$g(p + q) + g(1 - p - q) + 2g(0) = g(p) + g(q) + g(1 - p - 1) + g(0) \tag{37}$$

so

$$g(p) + g(q) = g(p + q) \tag{38}$$

$g(p)$ is therefore continuous and linear with $g(0) = g(1) = 0$, so for any p, $g(p) = 0$ must follow as originally desired. This last step can be justified as follows (*e.g.* Melzak 1976, page 325).

Rewriting integer multiplications as repeated additions gives

$$\begin{aligned} g(mn^{-1}) &= g(n^{-1} + n^{-1} + \ldots + n^{-1}) \\ &= g(n^{-1}) + g(n^{-1}) + \ldots + g(n^{-1}) \\ &= mg(n^{-1}) \end{aligned} \tag{39}$$

Setting $m = n$ shows that $g(1) = ng(n^{-1})$. Solving for this last for $g(n^{-1})$ and back-substituting into equation (39) gives $g(mn^{-1}) = mn^{-1}g(1)$. Thus $g(p) = pg(1)$ for any rational $p = mn^{-1}$. Since any frequency p is arbitrarily close to a rational number and $g(p)$ is continuous, $g(p) = pg(1) = 0$ for any p.

A) the j^{th} sequence matrix, $s(b, l, j)$

base b	position l									
	C	A	G	G	T	C	T	G	C	A
	-3	-2	-1	0	1	2	3	4	5	6
A	0	1	0	0	0	0	0	0	0	1
C	1	0	0	0	0	1	0	0	1	0
G	0	0	1	1	0	0	0	1	0	0
T	0	0	0	0	1	0	1	0	0	0

B) individual information weight matrix, $R_{iw}(b, l)$

base b	position l									
	-3	-2	-1	0	1	2	3	4	5	6
A	$+0.42$	$\boxed{+1.25}$	-1.41	$-\infty$	-5.81	$+1.12$	$+1.51$	-1.81	-0.68	$\boxed{+0.05}$
C	$\boxed{+0.58}$	-0.78	-2.40	-7.81	-5.49	$\boxed{-3.68}$	-1.56	-2.24	$\boxed{-0.51}$	-0.17
G	-0.58	-1.04	$\boxed{+1.64}$	$\boxed{+1.99}$	-6.23	$+0.72$	-1.06	$\boxed{+1.71}$	-0.32	$+0.44$
T	-1.02	-0.87	-1.67	-5.81	$\boxed{+1.98}$	-3.38	$\boxed{-1.59}$	-2.21	$+0.90$	-0.49
	C	A	G	G	T	C	T	G	C	A

Figure 1: Matrix representation of a sequence and a sequence recognizer.
A) The sequence 5′ CAGGTCTGCA 3′ represented in matrix format. There is only one "1" in each column, marking the base at that position. The remainder of the column is filled with "0"s. B) The individual weight matrix for human donor splice junctions derived from data given in (Stephens & Schneider, 1992). The weights of the matrix in B that are selected by the sequence in A are enclosed by boxes.

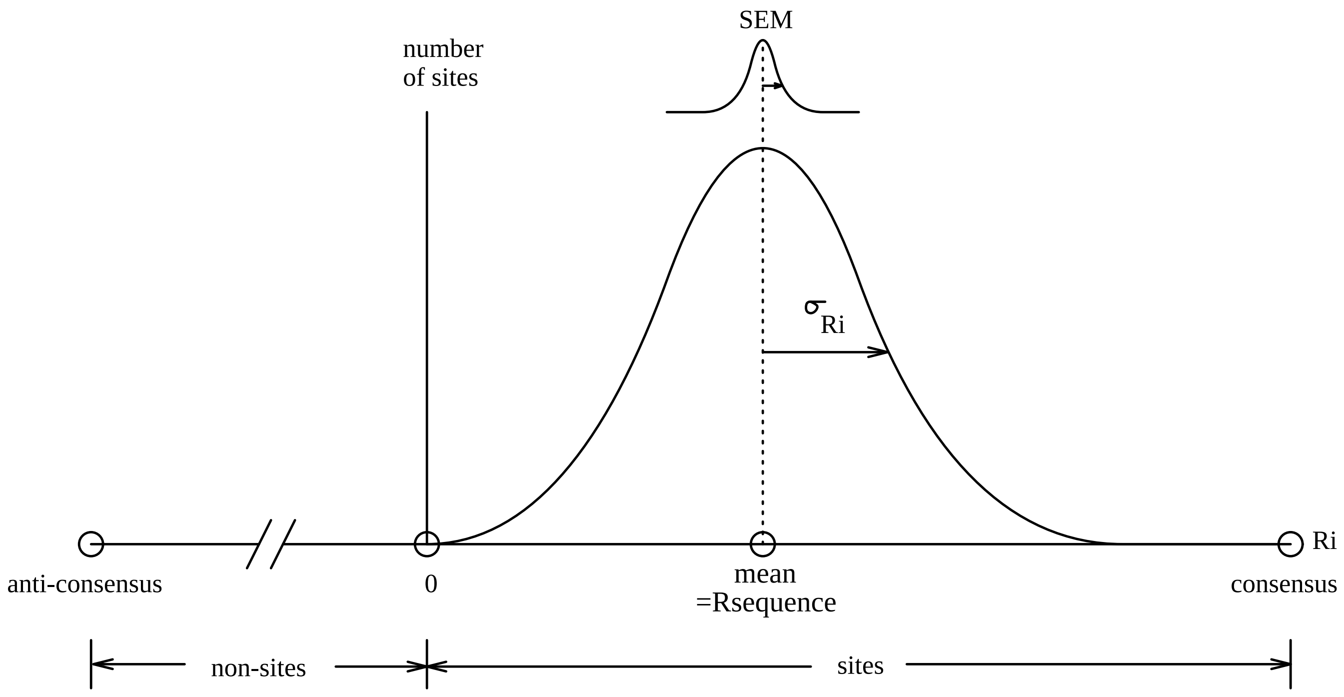

Figure 2: Important landmarks on the individual information (R_i) scale.
The abscissa is the R_i scale in bits, while the ordinate is the number of sites. The R_i distribution is approximately Gaussian. By definition, the mean of the distribution is $R_{sequence}$. The standard deviation of the distribution is σ_{R_i}. The standard deviation of $R_{sequence}$ is the standard error of the mean, SEM. The consensus is the highest possible sequence evaluation by the $R_{iw}(b,l)$ matrix; the anti-consensus is the lowest. For binding sites, sequences with $R_i = 0$ separate sites ($R_i > 0$) from non-sites ($R_i < 0$).

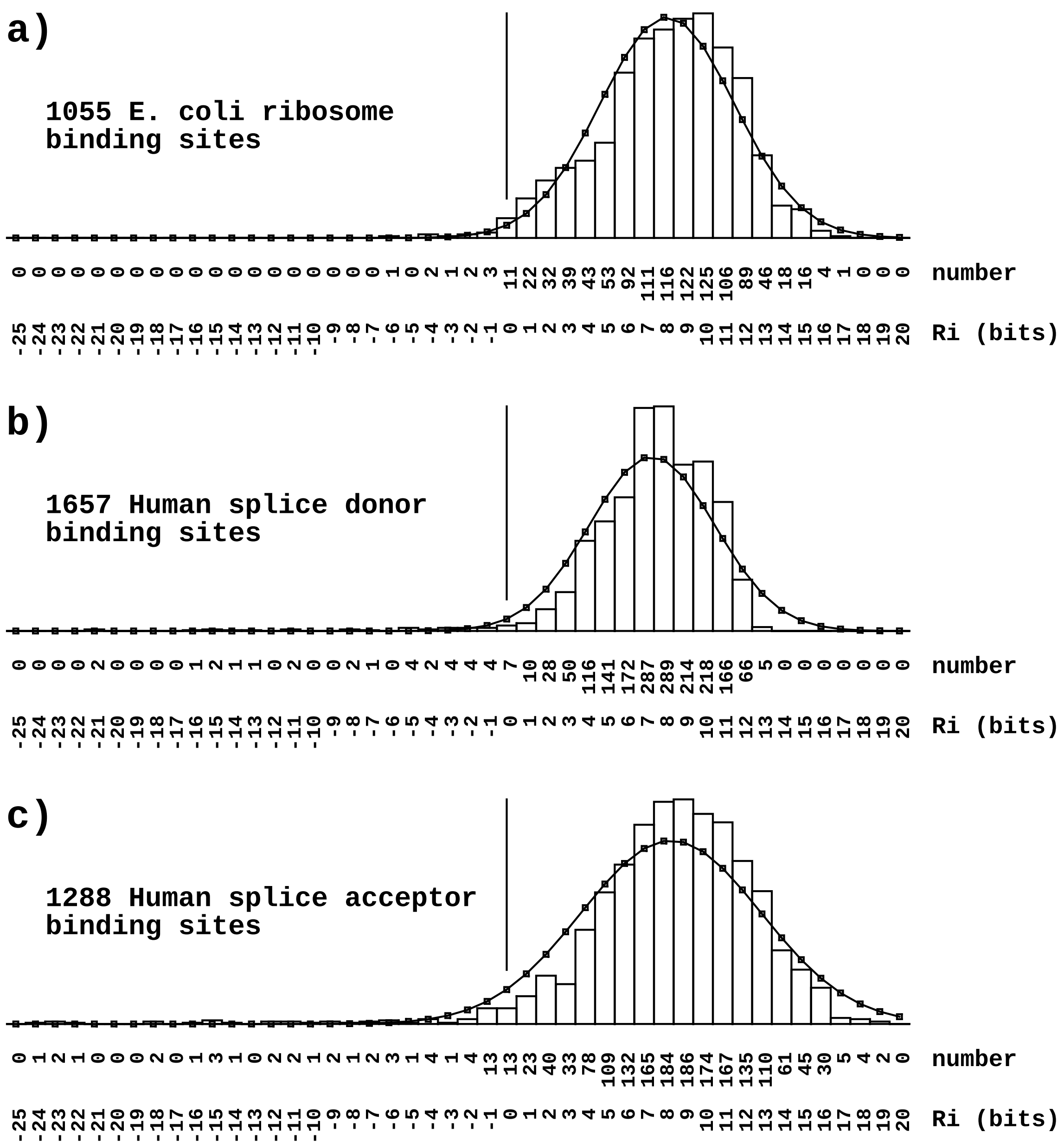

Figure 3: Individual information distribution for binding sites
Partially sequenced sites were eliminated from the distributions shown.
A. Individual information distribution for 1055 *E. coli* ribosome binding sites (Rudd & Schneider, 1992). The mean and standard deviation of the R_i values were fitted by a Gaussian distribution. All sites included the complete range from base -21 to $+18$ of the binding site. A vertical bar marks $R_i = 0$.
B. Sites from a collection of 1799 human donor binding sites (Stephens & Schneider, 1992). Only 1657 sites that included the complete range from -3 to $+6$ were included.
C. Sites from a collection of 1744 human acceptor binding sites (Stephens & Schneider, 1992). Only 1288 sites that included the complete range from -25 to $+2$ were included.

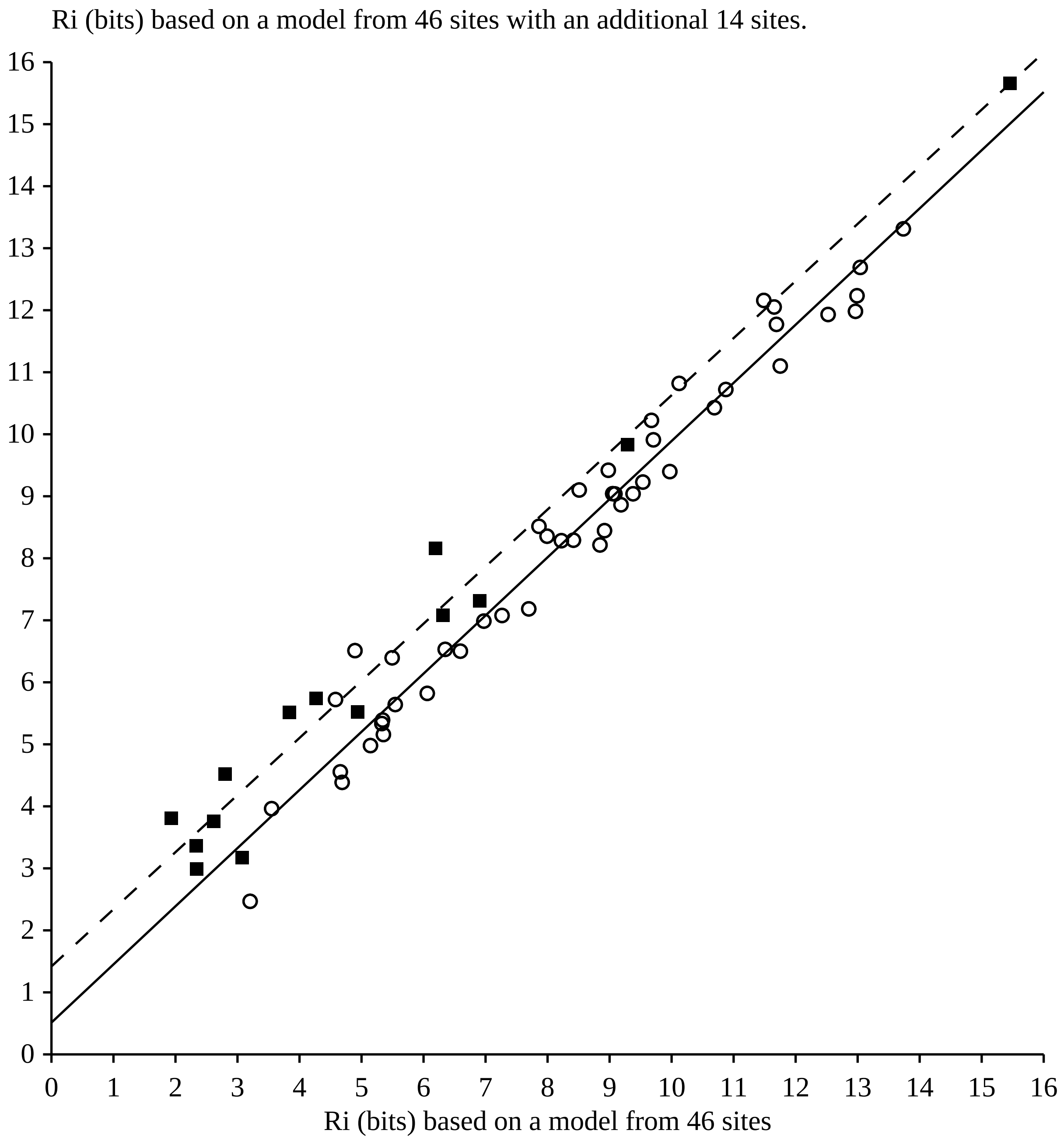

Figure 4: Effect of adding new sites to a binding site model.
14 new sites (Green *et al.*, 1996; Pan *et al.*, 1996; Falconi *et al.*, 1996) were added to a previous set of 46 Fis binding sites (Hengen *et al.*, 1997). ○, the 46 sites evaluated before and after addition of the new sites to the model. Linear regression through these points ($r = 0.982$) is shown by the solid line. ■, the 14 new sites evaluated before and after addition of (the same) new sites to the model. Linear regression through these points ($r = 0.986$) is shown by the dashed line.

References

Atkins, P. W. (1984). *The Second Law.* W. H. Freeman and Co., N. Y.

Berg, O. G. (1988). Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition. *J. Biomol. Struct. Dyn.* **6**, 275–297.

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

Berg, O. G. & von Hippel, P. H. (1988*a*). Selection of DNA binding sites by regulatory proteins. *Trend. Bioch. Sci.* **13**, 207–211.

Berg, O. G. & von Hippel, P. H. (1988*b*). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* **200**, 709–723.

Bisant, D. & Maizel, J. (1995). Identification of ribosome binding sites in *Escherichia coli* using neural network models. *Nucl. Acids Res.* **23**, 1632–1639.

Blom, N., Hansen, J., Blaas, D. & Brunak, S. (1996). Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Science,* **5**, 2203–2216.

Breiman, L. (1969). *Probability and Stochastic Processes: With a View Toward Applications.* Houghton Mifflin Company, Boston.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1990*a*). Cleaning up gene databases. *Nature,* **343**, 123.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1990*b*). Neural network detects errors in the assignment of mRNA splice sites. *Nucl. Acids Res.* **18**, 4797–4801.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578.

Cui, Y., Wang, Q., Stormo, G. D. & Calvo, J. M. (1995). A consensus sequence for binding of Lrp to DNA. *J. Bact.* **177**, 4872–4880.

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

Eiglmeier, K., Honoré, N., Iuchi, S., Lin, E. C. C. & Cole, S. T. (1989). Molecular genetic analysis of FNR-dependent promoters. *Mol. Microb.* **3**, 869–878.

Falconi, M., Brandi, A., Teana, A. L., Gualerzi, C. O. & Pon, C. L. (1996). Antagonistic involvement of FIS and H-NS proteins in the transcriptional control of *hns* expression. *Molec. Microb.* **19**, 965–975.

Feller, W. (1968). In *An Introduction to Probability Theory and Its Applications* vol. I, pp. 123–124, John Wiley & Sons, Inc., New York.

Goodrich, J. A., Schwartz, M. L. & McClure, W. R. (1990). Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucl. Acids Res.* **18**, 4993–5000.

Green, J., Anjum, M. F. & Guest, J. R. (1996). The *ndh*-binding protein (Nbp) regulates the *ndh* gene of *Escherichia coli* in response to growth phase and is identical to Fis. *Molec. Microb.* **19**, 1043–1055.

Gribskov, M., Lüthy, R. & Eisenberg, D. (1990). Profile analysis. *Meth. Enzym.* **183**, 146–159.

Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. & Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.* **20**, 5785–5795.

Hall, S. L. & Padgett, R. A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.* **239**, 357–365.

Hengen, P. N., Bartram, S. L., Stewart, L. E. & Schneider, T. D. (1997). Information analysis of Fis binding sites. *Nucl. Acids Res.* **25** (24), 4994–5002. http://www.lecb.ncifcrf.gov/~toms/paper/fisinfo/.

Herman, N. D. & Schneider, T. D. (1992). High information conservation implies that at least three proteins bind independently to F plasmid *incD* repeats. *J. Bact.* **174**, 3558–3560.

Horton, P. B. & Kanehisa, M. (1992). An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucl. Acids Res.* **20**, 4331–4338.

Iida, Y. (1987). DNA sequences and multivariate statistical analysis. Categorical discrimination approach to 5′ splice site signals of mRNA precursors in higher eukaryotes' genes. *CABIOS,* **3**, 93–98.

Korsgren, C. & Cohen, C. M. (1991). Organization of the gene for human erythrocyte membrane protein 4.2: structural similarities with the gene for the a subunit of factor XIII. *Proc. Natl. Acad. Sci. USA,* **88**, 4840–4844.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology, applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.

Melzak, Z. A. (1976). *Mathematical Ideas, Modeling and Applications,* Vol. 2 *of Companion to Concrete Mathematics.* John Wiley & Sons, Inc., New York.

Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucl. Acids Res.* **20**, 4255–4262.

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

Nakata, K., Kanehisa, M. & DeLisi, C. (1985). Prediction of splice junctions in mRNA sequences. *Nucl. Acids Res.* **13**, 5327–5340.

Niwa, M., MacDonald, C. C. & Berget, S. M. (1992). Are vertebrate exons scanned during splice-site selection? *Nature,* **360**, 277–280.

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

Pan, C. Q., Johnson, R. C. & Sigman, D. S. (1996). Identification of new Fis binding sites by DNA scission with Fis-1,10-phenanthroline-copper(I) chimeras. *Biochem.* **35**, 4326–4333.

Papoulis, A. (1990). *Probability & Statistics.* Prentice Hall, Englewood Cliffs, NJ.

Papp, P. P., Chattoraj, D. K. & Schneider, T. D. (1993). Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.* **233**, 219–230.

Penotti, F. E. (1990). Human DNA TATA boxes and transcription initiation sites: a statistical study. *J. Mol. Biol.* **213**, 37–52.

Penotti, F. E. (1991). Human pre-mRNA splicing signals. *J. Theor. Biol.* **150**, 385–420.

Piccirilli, J. A., Krauch, T., Moroney, S. E. & Benner, S. A. (1990). Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature,* **343**, 33–37.

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

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucl. Acids Res.* **24**, 3836–3845.

Prestridge, D. S. & Stormo, G. (1993). SIGNAL SCAN 3.0: new database and program features. *CABIOS,* **9**, 113–115.

Quandt, K., Freche, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* **23**, 4878–4884.

Rice, P. M., Elliston, K. & Gribskov, M. (1992). Identification of simple sites and transcriptional signals. In *Sequence Analysis Primer* pp. 23–43, W. H. Freeman and Co., New York.

Robberson, B. L., Cote, G. J. & Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* **10**, 84–94.

Rudd, K. E. & Schneider, T. D. (1992). Compilation of *E. coli* ribosome binding sites. In *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for* Escherichia coli *and Related Bacteria*, (Miller, J. H., ed.), pp. 17.19–17.45, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schneider, T. D. (1984). *Information content of binding sites on nucleotide sequences* (Ph.D. thesis). University of Colorado.

Schneider, T. D. (1988). Information and entropy of patterns in genetic switches. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, (Erickson, G. J. & Smith, C. R., eds), vol. 2, pp. 147–154, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Schneider, T. D. (1991*a*). Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.* **148**, 83–123. http://www.lecb.ncifcrf.gov/~toms/paper/ccmm/.

Schneider, T. D. (1991*b*). Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.* **148**, 125–137. http://www.lecb.ncifcrf.gov/~toms/paper/edmm/.

Schneider, T. D. (1993). Protein patterns as shown by sequence logos. In *Visual Cues - Practical Data Visualization*, (Keller, P. R. & Keller, M. M., eds), p. 64, IEEE Press, Piscataway, NJ.

Schneider, T. D. (1994). Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology,* **5**, 1–18. http://www.lecb.ncifcrf.gov/~toms/paper/nano2/.

Schneider, T. D. (1995). *Information Theory Primer.* http://www.lecb.ncifcrf.gov/~toms/paper/primer/.

Schneider, T. D. (1996). Reading of DNA sequence logos: prediction of major groove binding by information theory. *Meth. Enzym.* **274**, 445–455. http://www.lecb.ncifcrf.gov/~toms/paper/oxyr/.

Schneider, T. D. (1997). Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucl. Acids Res.* **25**, 4408–4415. http://www.lecb.ncifcrf.gov/~toms/paper/walker/, erratum: NAR 26(4): 1135, 1998.

Schneider, T. D. & Mastronarde, D. (1996). Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Applied Mathematics,* **71**, 259–268. http://www.lecb.ncifcrf.gov/~toms/paper/malign.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100. http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/.

Schneider, T. D. & Stormo, G. D. (1989). Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucl. Acids Res.* **17**, 659–674.

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

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

Schneider, T. D., Stormo, G. D., Yarus, M. A. & Gold, L. (1984). Delila system tools. *Nucl. Acids Res.* **12**, 129–140.

Seidel, H. M., Pompliano, D. L. & Knowles, J. R. (1992). Exons as microgenes? *Science,* **257**, 1489–1490.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423, 623–656. http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

Shapiro, M. B. & Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucl. Acids Res.* **15**, 7155–7174.

Sloane, N. J. A. & Wyner, A. D. (1993). *Claude Elwood Shannon: Collected Papers.* IEEE Press, Piscataway, NJ.

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

Stephens, R. M. & Schneider, T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**, 1124–1136. http://www.lecb.ncifcrf.gov/~toms/paper/splice/.

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

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

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

Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucl. Acids Res.* **10**, 2997–3011.

Talerico, M. & Berget, S. M. (1990). Effect of 5′ splice site mutations on splicing of the preceding intron. *Mol. Cell. Biol.* **10**, 6299–6305.

Taylor, J. R. (1982). *An Introduction to Error Analysis.* University Science Books, Mill Valley, CA.

Toledano, M. B., Kullik, I., Trinh, F., Baird, P. T., Schneider, T. D. & Storz, G. (1994). Redox-dependent shift of OxyR-DNA contacts along an extended DNA binding site: a mechanism for differential promoter selection. *Cell,* **78**, 897–909.

Tribus, M. (1961). *Thermostatics and Thermodynamics.* D. van Nostrand Company, Inc., Princeton, N. J.

Tribus, M. & McIrvine, E. C. (1971). Energy and information. *Sci. Am.* **225** (3), 179–188. (Note: the table of contents in this volume incorrectly lists this as volume **224**).

Waldram, J. R. (1985). *The Theory of Thermodynamics.* Cambridge University Press, Cambridge.