Brief Review of Molecular Information Theory

Peter K. Rogan, Cytognomix Inc. and Western University, London Ontario Canada
info@cytognomix.com

Abstract

Shannon information theory can be used to quantify overall sequence conservation among sets of related sequences. Variation in nucleic acid sequences recognized by proteins can be comprehensively modeled with information weight matrices that permit each member sequence to be rank-ordered according to its respective individual information contents. These rankings can be used to compute the affinities of recognition sites by proteins and to predict the effects of nucleotide substitutions in the sequences of these sites. The distribution of information across a set of protein-binding sites in DNA is related to the pattern of intermolecular contacts that stabilize the protein-nucleic acid complex (i.e., the corresponding helical structure of double-stranded DNA).

1. Theory

Shannon and Weaver (1949) developed their theory of information in order to understand the transmission of electronic signals and model the communication system. Gatlin (1972) first described its extension to biology. Information theory is an obvious tool to use in looking for patterns in DNA and protein sequences (Schneider, 1995). Information theory has been applied to the analysis of DNA and protein sequences in several ways: (1) by analyzing sequence complexity from the Shannon–Weaver indices of smaller DNA windows contained in a long sequence; (2) by comparing homologous sites in a set of aligned sequences by means of their information content; and (3) by examining the pattern of information content of a sequence divided into successively longer words (symbols) consisting of a single base, base pairs, base triplets, and so forth.

Some of the most useful applications of molecular information theory have come from studies of binding sites (typically protein-recognition sites) in DNA or RNA recognized by the same macromolecule, which typically contain similar but nonidentical sequences. Because average information measures the choices made by the system, the theory can comprehensively model the range of sequence variation present in nucleic sequences that are recognized by individual proteins or multisubunit complexes.

Treating a discrete information source (i.e., telegraphy or DNA sequences) as a Markov process, Shannon defined entropy (H) to measure how much information is generated by such a process. The information source generates a series of symbols belonging to an alphabet with size J (e.g., 26 English letters or 4 nucleotides). If symbols are generated according to a known probability distribution p, the entropy function $H(p_1, p_2, ¼, p_J)$ can be evaluated. The units of H are in bits, where one bit is the amount of information necessary to select one of two possible states or choices. In this section, we describe several important concepts regarding the use of entropy in genomic sequence analysis.

1.1. Entropy

Entropy is a measure of the average uncertainty of symbols or outcomes. Given a random variable X with a set of possible symbols or outcomes $A_X = \{a_1, a_2, ¼, a_J\}$, having probabilities $\{p_1, p_2, ¼, p_J\}$, with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{x \in A_X} P(x) = 1$, the Shannon *entropy* of X is defined by

$$H(X) = \sum_{x \in A_X} P(x) \log_2 \frac{1}{P(x)} \quad (1)$$

Two important properties of the entropy function are: (1) $H(X) \geq 0$ with equality for one x, $P(x) = 1$ and (2) Entropy is maximized if $P(x)$ follows the uniform distribution. Here the *uncertainty* or *surprisal*, $h(x)$, of an outcome (x) is defined by

$$h(x) = \log_2 \frac{1}{P(x)} \text{ (bits)} \quad (2)$$

For example, given a DNA sequence, we say each position corresponds to a random variable X with values $A_X = \{A, C, G, T\}$, having probabilities $\{p_a, p_c, p_g, p_t\}$, with $P(x = A) = p_a$, $P(x = C) = p_c$, and so forth. Suppose the probability distribution $P(x)$ at a position of DNA sequence is $P(x = A) = 1/2$; $P(x = C) = 1/4$; $P(x = G) = 1/8$; $P(x = T) = 1/8$.

The uncertainties (surprisals) in this case are $h(A) = 1$, $h(C) = 2$, $h(G) = h(T) = 3$ (bits). The entropy is the average of the uncertainties
$H(X) = E[h(x)] = 1/2(1) + 1/4(2) + 1/8(3) + 1/8(3) = 1.75$ bits. In a study of genomic DNA sequences, Schmitt and Herzel (1997) found that genomic DNA sequences are closer to completely random sequences than to written text, suggesting that higher-order interdependencies between neighboring or adjacent sequence positions make little contributions to the block entropy.

The entropy (average uncertainty), H, is 2 bits if each of the four bases is equally probable (uniform distribution) before the site is decoded. The information content (IC) is a measure of a reduction in average uncertainty after the binding site is decoded. $IC(X) = H_{\text{before}} - H_{\text{after}} = \log_2|A_x| - H(X$, provided the background probability distribution P(before) is uniform (Schneider, 1997a). If the background distribution is not uniform, the Kullback–Leibler distance (relative entropy) can be used (Stormo, 2000). The information content calculation needs to be corrected for the fact that a finite number of sequences were used to estimate the information content of the ideal binding site, resulting in the corrected IC, R_{sequence} (Schneider *et al.*, 1986). This measures the decrease in uncertainty before versus after the macromolecule is bound to a set of target sequences. Positions within a binding site with high information are conserved between binding sites, whereas low-information content positions exhibit greater variability. The R_{sequence} 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 conserved features of the site.

1.2. Relative entropy

For two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet, the *relative entropy* (also known as the Kullback–Leibler divergence or KL-distance) is defined by

$$D_{KL}(P||Q) = \sum_{x \in A_X} P(x) \log \frac{P(x)}{Q(x)} \tag{3}$$

Note that the relative entropy is not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$; and although it is sometimes called the KL-distance, it is not strictly a distance (Koski, 2001; Lin, 1991). Relative entropy is an important statistic for finding unusual motifs/patterns in genomic sequences (Durbin *et al.*, 1998; Lawrence *et al.*, 1993; Bailey and Elkan, 1994; Hertz and Stormo, 1999; Liu *et al.*, 2002).

1.3. $R_{sequence}$ versus $R_{frequency}$

The fact that proteins can find their required binding sites among a huge excess of nonsites (Lin and Riggs, 1975; von Hippel, 1979) indicates that more information is required to identify an infrequent site than a common binding site in the same genome. The amount of information required for these sites to be distinguished from all sites in the genome, $R_{frequency}$, is derived independently from the size and frequency of sites in the genome. $R_{frequency}$, like $R_{sequence}$, is expressed in bits per site. $R_{sequence}$ cannot be less than the information needed to find sites in the genome. With few exceptions, it has been found that $R_{sequence}$ and $R_{frequency}$ are similar (Schneider *et al.*, 1986). This empirical relationship is strongly constrained by the fact that all DNA-binding proteins operating on the genome are encoded in the genome itself (Kim *et al.*, 2003).

1.4. Molecular machines

Molecular machines are characterized by stable interactions between distinct components, for example, the binding of a recognizer protein to a specific genomic sequence. The behavior of a molecular machine can be described with information theory. The properties of molecular machine theory may be depicted on multiple levels: on one level, sequence logos, which describe interactions between the molecules (see Figure 1), are equivalent to transmission of information by the recognizer as a set of binary decisions; on another level, the information capacity of the machine, which represents the maximum number of binary decisions (or bits) that can be made for the amount of energy dissipated by the binding event; and finally, the relationship between information content and the energy cost of performing molecular operations (Schneider, 1991; Schneider, 1994). The molecular machine capacity is derived from Shannon's channel capacity (Shannon, 1949). The error rate of the machine can be specified to be as low as necessary to ensure the survival of the organism, so long as the molecular machine capacity is not exceeded. Entropy decreases as the machine makes choices, which corresponds to an increase in information.

Figure 1. Examples of sequence logos. Models of human (a) 108 079 acceptor and (b) 111 772 donor splice sites derived from both strands of the human genome reference sequence (April, 2003) are shown. A sequence logo visually represents the sequence conservation among a common set of recognition sites, with the height of each nucleotide stack corresponding to the average information content at that position. The height of each nucleotide is proportional to its frequency. Sampling error is indicated by error bar at the top of each stack. The zero coordinate represents the intronic position immediately adjacent to the splice junction. The average information contents ($R_{sequence}$) of the acceptor and donor sites are respectively, 9.8118 ± 0.0001 bits/site and 8.12140 ± 0.00009 bits/site

$$\varepsilon_{min} = \kappa_B T \ln(2) \leq \frac{-q}{IC} \text{ (joules per bit)}$$

The second law of thermodynamics can be expressed by the equation $dS \geq dQ/T$. The equation states that for a given increment of heat dQ entering a volume at some temperature T, the entropy will increase dS at least by dQ/T. If we relate entropy to Shannon's uncertainty, we can rewrite the second law in the following form:
where k_B is Boltzman constant and q is the heat. This equation states that there is a minimum amount of heat energy that must be dissipated (negative q) by a molecular machine in order for it to gain IC = 1 bit of information.

1.5. Individual information

The information contained in a set of binding sites is an average of the individual contributions of each of the sites (Shannon, 1948; Pierce, 1980; Sloane and Wyner, 1993; Schneider, 1995). The information content to each individual binding-site sequence can be determined by a weight matrix so that the average of these values over the entire set of sites is the average information content (Schneider, 1997a).

The individual information weight matrix is

$$R_{iw}(b,l) = 2.0 - (-\log_2(f(b,l) + e(n(l)))) \text{ (bits per base)} \quad (5)$$

$$R_i(j) = \sum_l \sum_{b=a}^{t} s(b,j,j) R_{iw}(b,l) \text{ (bits per site)} \quad (6)$$

in which $f(b,l)$ is the frequency of each base b at position l in the binding-site sequences; $e(n(l))$ is a correction of $f(b,l)$ for the finite sample size (n sequences at position l) (Schneider *et al.*, 1986). The jth sequence of a set of binding sites is represented by a matrix $s(b,l,j)$, which contains 1's in cells from base b at position l of a binding site and zeros at all other matrix locations. The individual information of a binding site sequence is the dot product between the sequence and the weight matrix:

2. Applications

2.1. Displaying sequence conservation

Sequence logos, which display information about both consensus and nonconsensus nucleotides, are visual representations of the information found in a binding site (an example is shown in

Figure 1). This is the information that the decoder (i.e., a binding protein) uses to evaluate potential sites in order to recognize actual sites. The calculation of sequence logos uses the assumption that each site is evaluated independently, that is, that there is no correlation between a change in nucleotide at one position with another position, which is reasonable for most genomic sequences (Schmitt and Herzel, 1997). An advantage of the information approach is that the sequence conservation can be interpreted quantitatively. R_{sequence}, which is the total area under the sequence logo and measures the average information in a set of binding site sequences, is related to the specific binding interaction between the recognizer and the site. R_{sequence} is an additive measure of sequence conservation; thus, it is feasible to quantitatively compare the relative contributions of different portions within the same binding site.

Structural features of the protein–DNA complex can be inferred from sequence logos. When positions with high information content are separated by a single helical turn (10.4 bp), this suggests that the protein makes contacts across the same face of the double helix. Sequence conservation in the major groove can range anywhere between 0 and 2 bits depending on the strength of contacts involved, and usually correlates with the highest information content positions (Papp *et al.*, 1993). Minor groove contacts of B-form DNA allow both orientations of each kind of base pair so that rotations about the dyad axis cannot easily be distinguished; hence, a single bit is the information content in native B-form DNA (Schneider, 1996). Higher levels of conservation for bases within the minor groove indicate that these positions are accessed protein distortion of the helix, that is, bending accompanied by base-pair opening and flipping (Schneider, 2001).

2.2. Visualizing individual binding-site information

Because sequence logos display the average information content in a set of binding sites, they may not accurately convey protein–DNA interactions with individual DNA sequences, especially at highly variable positions within a binding site. The walker method (Schneider, 1997b) graphically depicts the nucleotide conservation of a known or suspected site compared to other valid binding sites defined by the individual information weight matrix (Schneider, 1997a). Walkers apply to a single sequence (rather than a set of binding sites); only a single letter is visualized at each position of the binding site (Figure 2). The height of the letter, which is in units of bits, represents the contribution of that nucleotide at each position in the binding site by the information weight matrix, $R_{iw}(b,l)$. Evaluation of the R_i value at each position in a genomic DNA sequence is equivalent to moving the walker along that sequence. Walkers are displayed for sequences with positive R_i values, since these are more likely to be valid binding sites (see equation 4 and discussion below). Sequence walkers facilitate visualization and interpretation of the structures and strengths of binding sites in complex genetic intervals and can be used to understand the effects of sequence changes (see below), and engineer overlapping or novel binding sites.

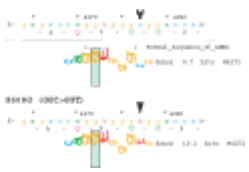

Figure 2. Examples of sequence walkers. A synonymous C > T substitution at codon 608 activates a cryptic donor splice site in exon 11, the *LMNA* gene in patients with Hutchinson–Gifford progeria (Eriksson *et al.*, 2003). The walker, shown below the sequence, indicates a preexisting 8.7-bit cryptic site that is strengthened by the mutation to 10.2 bits (≥ 2.8-fold). The height and orientation of each nucleotide correspond to contribution that nucleotide makes to the overall information content in the binding site. The green rectangle indicates the location of valid splice site ($R_i > 0$) and delineates the scale displayed; the lower and

upper limits shown are, respectively, –4 bits and +2 bits. Sequence coordinates are from GenBank accession L12401 (4277 C > T)

2.3. Mutation and polymorphism analysis

Because the relationship between information and energy can be used to predict the effects of natural sequence variation at these sites, phenotypes can be predicted from corresponding changes in the individual information contents (R_i, in bits) of the natural and variant DNA binding sites (Rogan *et al.*, 1998; *see* Normal DNA Sequence Variations in Humans). For splice site variants, mutations have lower R_i values than the corresponding natural sites, with null alleles having values at or below zero bits (equation 4; Kodolitsch *et al.*, 1999). The decreased R_i values of mutated splice sites indicate that such sites are either not recognized or are bound with lower affinity, usually resulting in an untranslatable mRNA. Decreases in R_i are more moderate for partially functional (or leaky) mutations that reduce but do not abolish splice site recognition and have been associated with milder phenotypes (Rogan *et al.*, 1998). The minimum change in binding affinity for leaky mutations is $\geq 2^{DRi}$ lower fold than cognate wild-type sites. Mutations that activate cryptic splicing may decrease the R_i value of the natural site, increase the strength of the cryptic site, or concomitantly affect the strengths of both types of sites (see Figure 2). Nondeleterious changes do not alter the R_i value of splice sites significantly (Rogan and Schneider, 1995). Increases in R_i indicate stronger interactions between protein and cognate binding sites.

2.4. Information evolution

How do genetic systems gain information in a binding site of genomic DNA sequence by evolutionary processes? Schneider (2000) proposed an answer to this question. Given a binding site for an artificial protein, his simulation begins with zero information and, as in naturally occurring genetic systems, the information measured in the fully evolved binding sites (R_{sequence}) is close to that needed to locate the sites in the genome ($R_{\text{frequency}}$).

2.5. Model refinement

Information models based on small numbers of proven binding sites may fail to detect valid binding sites and tend to predict R_i inaccurately. Iterative selection of functional binding sites has been used to optimize (Lund *et al.*, 2000) and to introduce bias (Shultzaberger and Schneider, 1999) into the frequencies of each nucleotide in computing the information theory–based weight matrices of binding sites. Significant differences between information weight matrices have been determined from their respective evolutionary distance metrics (e.g., see Shultzaberger *et al.*, 2001). The effects of model refinements can be monitored by comparing the genome scan results for pairs of successive information weight matrices based on additional binding sites (Gadiraju *et al.*, 2003). Other potential applications include the determination of the locations of overlapping binding sites recognized by different proteins and comparisons of binding sites detected with information models of orthologous proteins from different species.

2.7. Genome-wide analyses

Information weight matrices of binding sites can be developed directly from validated sets of binding sites extracted from genome sequences provided that the locations of sequence features are accurately annotated. As this is not always the case, we built a genome-wide human splice junction database by initially extracting the coordinates and sequences of donor and acceptor

splice regions from both strands of the human genome reference sequence (Rogan *et al.*, 2003). After redundant sites were eliminated, the splice site database consisted of 170 144 acceptor and 170 450 donor sites. The information weight matrix was recomputed after each of iteration and scanning of the resultant set of sites. Successive models iteratively utilized the modified matrix by excluding sites with negative R_i values. After eight cycles of refinement, the final models were then defined by 108 079 unique acceptor sites and 111 772 donor sites (sequence logos of model sites are shown in Figure 1).

The average information contents of the acceptor and donor sites are respectively, 9.8 bits/site and 8.1 bits/site. These values are similar to those previously reported by Schneider and Stephens (1990), that is, 9.35 bits for acceptors and 7.92 bits for donors, which were based on about 65-fold fewer splice sites. Individual splice site strengths in the genome have an approximately Gaussian distribution.

3. Prospects for information theory–based analyses of genomic sequences

As the functions for regulatory and expressed nucleic acid sequences are elucidated, it is becoming evident that multiple protein components catalyze these processes. Modeling such molecular machines by information theory will require the development of procedures that account for cooperative and interdependent binding events between two or more recognizers. Frameworks for building multipartite information models will therefore have to incorporate corrections for overlapping sites and mutual information.

There are opportunities to enhance currently available genomic applications by scaling currently available software for information theory analyses (Gadiraju *et al.*, 2003) to investigate genome annotation. For example, changes in IC contents due to mutation may be of assistance in prioritizing single nucleotide polymorphisms for functional analyses. It is also possible that comparative genomic analyses of binding sites with orthologous DNA recognition domains from multiple species may reveal possible identities of functionally analogous regulatory sequences in these systems.

Medical genetic applications of information theory–based binding-site models are a promising avenue to improve diagnosis and prognosis of disease-causing mutations. Accurate models will be required for use of information theory in a clinical setting. To calibrate individual information measures of protein-nucleic acid binding with the thermodynamic properties of these complexes will require more comprehensive models, that is, based on larger numbers of binding sites spanning a wide range of binding affinities.

Further reading

Chen X, Li M, Ma B and Tromp J (2002) DNACompress: fast and effective DNA sequence compression. Bioinformatics, **18**, 1696–1698.

Cover TM and Thomas JA (1991) *Elements of Information Theory*, John Wiley & Sons: New York.

Milosavljevic A and Jurka J (1993) Discovering simple DNA sequences by the algorithmic

significant method. Computer Applications in the Biosciences: CABIOS, **9**, 407–411.
Schneider TD and Mastronarde DN (1996) Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. Discrete Applied Mathematics, **71**, 259–268.
Stephens RM and Schneider TD (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. Journal of Molecular Biology, **228**, 1124–1136.

References

Bailey TL and Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings International Conference on Intelligent Systems for Molecular Biology, **2**, 28–36.
Bi C-P, Vyhidal CA, Leeder JS and Rogan PK (2004) A minimization entropy based bipartite algorithm with application to PXR/RXR a binding sites. *Proceedings of the RECOMB2004 Annual Symposium, San Diego*, 453–454.
Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press: Cambridge.
Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robins CM, Moses TY, Berglund P, *et al*. (2003) Recurrent de novo point mutations in lamin A cause Hutchinson-Gifford progeria syndrome. Nature, **423**, 293–298.
Gadiraju S, Vyhlidal CA, Leeder JS and Rogan PK (2003) Genome-wide prediction, display and refinement of binding sites with information theory-based models. BMC Bioinformatics, **4**, 38.
Gatlin LL (1972) *Information Theory and the Living System*, Columbia University Press: New York.
Hengen PN, Bartram SL, Stewart LE and Schneider TD (1997) Information analysis of Fis binding sites. Nucleic Acids Research, **25**, 4994–5002.
Hertz GZ and Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, **15**, 563–577.
Kim JT, Martinez T and Polani D (2003) Bioinformatic principals underlying the information content of transcription factor binding sites. Journal of Theoretical Biology, **220**, 529–544.
Kodolitsch Yv, Pyeritz RE and Rogan PK (1999) Splice site mutations in atherosclerosis candidate genes: relating individual information to phenotype. Circulation, **100**, 693–699.
Koski T (2001) *Hidden Markov Models for Bioinformatics*, Kluwer Academic Publishers.
Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, **262**, 208–214.
Lin J (1991) Divergence measure based on the Shannon entropy. IEEE Transactions on Information Theory, **37**, 145–151.
Lin S and Riggs AD (1975) The general affinity of lac repressor for E. coli DNA: implications for gene regulation in procaryotes and eucaryotes. Cell, **4**, 107–111.
Liu XS, Brutlag DL and Liu JS (2002) An algorithm, for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature Biotechnology, **20**, 835–839.

Lund M, Tange TO, Dyhr-Mikkelsen H, Hansen J and Kjems J (2000) Characterization of human RNA splice signals by iterative functional selection of splice sites. RNA, **6**, 528–544.

Papp PP, Chattoraj DK and Schneider TD (1993) Information analysis of sequences that bind the replication initiator RepA. Journal of Molecular Biology, **233**, 219–230.

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

Rogan PK, Faux BM and Schneider TD (1998) Information analysis of human splice site mutations. Human Mutation, **12**, 153–171.

Rogan PK and Schneider TD (1995) Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. Human Mutation, **6**, 74–76.

Rogan PK, Svojanovsky S and Leeder JS (2003) Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics, **13**, 207–218.

Schmitt AO and Herzel H (1997) Estimating the entropy of DNA sequences. Journal of Theoretical Biology, **188**, 369–377.

Schneider TD (1991) Theory of molecular machines. Journal of Theoretical Biology, **148**, 83–137.

Schneider TD (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. Nanotechnology, **5**, 1–18.

Schneider TD (1995) Information Theory Primer.
http://www.lecb.ncifcrf.gov/~toms/paper/primer/.

Schneider TD (1996) Reading of DNA sequence logos: prediction of major groove binding by information theory. Methods in Enzymology, **274**, 445–455.

Schneider TD (1997a) Information content of individual genetic sequences. Journal of Theoretical Biology, **189**, 427–441.

Schneider TD (1997b) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. Nucleic Acids Research, **25**, 4408–4415.

Schneider TD (2000) Evolution of biological information. Nucleic Acids Research, **28**, 2794–2799.

Schneider TD (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. Nucleic Acids Research, **29**, 4881–4891.

Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Research, **18**, 6097–6100.

Schneider TD, Stormo GD, Gold L and Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. Journal of Molecular Biology, **188**, 415–431.

Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal, **27**, 379–432.

Shannon CE (1949) Communication in the presence of noise. Proceedings of the Institute of Radio Engineers, **37**, 10–21.

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

Shultzaberger RK, Bucheimer RE, Rudd KE and Schneider TD (2001) Anatomy of *Escherichia coli* ribosome binding sites. Journal of Molecular Biology, **313**, 215–228.

Shultzaberger RK and Schneider TD (1999) Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. Nucleic Acids Research, **27**, 882–887.

Sloane NJA and Wyner AD (1993) *Claude Elwood Shannon: Collected Papers*, IEEE Press: Piscataway.

Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics, **16**, 16–23.

Stormo GD and Fields DS (1989) Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences of the United States of America, **88**, 5699–5703.

von Hippel PH (1979). In *Biological Regulation and Development*, Vol. **1**, Goldberger RF (Ed.), Plenum Press: New York, pp. 279–347.