Consensus Sequence Zen

Thomas D. Schneider *

version = 2.33 of zen.tex 2002 Dec 5

runningtitle: Consensus Sequence Zen

Key words: consensus sequence, information theory, sequence logo, sequence walker, binding site, genetic control.

Applied Bioinformatics 1(3) 111-119, 2002. (Schneider, 2002)

*National Cancer Institute at Frederick, 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/

Abstract. Consensus sequences are widely used in molecular biology but they have many flaws. As a result, binding sites of proteins and other molecules are missed during studies of genetic sequences and important biological effects cannot be seen. Information theory provides a mathematically robust way to avoid consensus sequences. Instead of using consensus sequences, sequence conservation can be quantitatively presented in bits of information by using sequence logo graphics to represent the average of a set of sites and sequence walker graphics to represent individual sites.

"All models are wrong but some are useful."

— George E. P. Box (Box, 1979)

How to be sure to make a mistake. Genes are controlled by proteins that bind to specific spots on the DNA sequence. Molecular biologists often represent the patterns at these spots by using a consensus sequence. For example, after aligning some binding sites so that they match each other, one position might contain 70% adenine, 10% cytosine, 10% guanine, and 10% thymine. The consensus is the most frequent base, 'A'. This is the simplest (and possibly the most commonly applied) approach, but there are alternatives (Day & McMorris, 1992). Various kinds of consensus sequence commonly found in the molecular biology literature will be considered here, while the controversy over the use of consensus *trees* used in phylogenetic inference (Barrett *et al.*, 1991; Nelson, 1993; Barrett *et al.*, 1993; de Queiroz, 1993) will not be covered.

The main difficulty with using consensus sequences is that they present distorted pictures of binding sites. In order to locate new binding sites, consensus sequences are compared to various locations in a sequence and the number of matches is tallied. A difficulty arises because a position that is always an ‘A’ in the original set is treated the same as a position that is just 70% A. If we think that a position has A, then when we use this observation to look for additional binding sites, we will find mismatches for 30% of the acceptable sequences. This problem is compounded across the entire binding site, which may be 20 or even 40 bases long (Schneider, 1996; Zheng *et al.*, 1999). For example, a commonly cited consensus sequence is TA$\boxed{\text{TAA}}$T (Lewin, 1997), which represents the -10 region of bacterial promoters originally discovered by David Pribnow (1975). The most prominent bases for the boxed positions are only 49%, 58%, and 54% respectively (Lisser & Margalit, 1993). If one demands that a site have all of the consensus bases, one finds only 14 TATAAT sequences out of 291 sequences in the database. To deal with this, people often count mismatches, but it is not obvious from the simple consensus which bases are allowed to be more variable. Sometimes variations such as allowing C or G are indicated but, again, the degree of allowed variation is lost. It is not surprising then, that consensus sequences frequently fail to identify binding sites or that they predict sites where there are none.

Consensus sequences have other serious problems, many of which are revealed by using information theory to measure the amount of conservation in bits. In a set of aligned binding sites, a DNA position that is always an A stays that way during evolution because the molecule that binds to it always selects A from the four possible bases (Schneider, 2000). Such a selection can be made with a minimum of two yes-no questions: ‘Is it in the set A or T?’ and ‘Is it in the set A or C?’, so the selection takes two bits of information, one to answer each question. Likewise, a position that is either A or T only requires one yes-no question — the other one being ignored — so has one bit of sequence conservation. The late Claude Shannon figured out how to consistently measure the average information when the frequencies are not so simple (Shannon, 1948; Schneider *et al.*, 1986; Schneider, 1995). One can plot the sequence conservation across all positions in the set of aligned binding sites. This continuous quantitative measure often follows a sine wave, reflecting the binding of a protein to one face of helically twisting B-form DNA (Papp *et al.*, 1993; Schneider, 1996; Schneider, 2001). This subtle effect cannot be seen by using consensus sequences.

A Paradox: How can two things be the same but different? Intriguingly, the binding sites for human splice junction donor and acceptor sites have the same consensus sequence for a portion of each site around the junction. Yet, when we measured the sequence conservation in bits we found that the information curves are quite different (Stephens & Schneider, 1992). How could two sites have the same consensus sequence but be different? This conundrum led us to introduce a computer graphic, called a sequence logo, in order to understand the difference (Fig. 1). A logo depicts an average picture of the set of binding sites by a series of stacks of letters. The height of each stack is the sequence conservation (measured in bits of information; the vertical black bar at each junction is 2 bits high) and the heights of the letters show the relative proportions of the bases, sorted so that the more frequent bases are on top. From the logos shown, it is clear that the donor and acceptor sites have different 'emphasis', but this cannot be seen with the consensus sequence CAG|GT, which matches both of them at the junction. The difference in emphasis is important because it shows that there is more information on the intron side of each junction. This allows more freedom during the evolution of the protein-coding exon side, which ⇐Fig 1 is a biologically sensible result. The resemblance between the two junctions suggests that the splice machinery that binds to donors and acceptors have a common ancestor (Stephens & Schneider, 1992).

Walking along the genome. One can depict individual sites using another graphic called a sequence walker, in which the height of a letter above or below zero shows how much that base contributes to the average sequence conservation of the entire collection of sites shown in the logo (Fig. 2) (Schneider, 1997*b*; Schneider, 1997*a*). Instead of counting ⇐Fig 2 matches to a consensus, one sums the information contributions for a given sequence to obtain the information for an individual binding site. (The Shannon information measure is unique in that it is the only measure that allows addition for statistically independent components (Shannon, 1948). Generally, binding site bases are independent (Stephens & Schneider, 1992).)

Sequence walkers can be stepped along the sequence (hence the name) to discover positions that match a particular model, and one can predict whether or not a sequence change will destroy the site and cause a genetic disease (Rogan *et al.*, 1998). In the case shown, splicing is normally accomplished using a 12.7 bit acceptor at position 5154. Nearby, however, is an 8.9 bit 'cryptic' acceptor that is not used apparently because the strongest site in any local region normally wins the competition for splice factors. An A to G mutation at 5153 destroys the normal site, making it 4.5 bits while simultaneously raising the cryptic site to 16.5 bits. This results in a single base frame shift, the loss of the protein, and Hunter disease. Cases like this are difficult to understand using consensus sequences because sites are affected by all of their parts and quantitative differences are missed. Using information theory and sequence walkers we have interpreted about 100 mutations in two human workdays (the computer time is only a few seconds).

Statistical effects of making a consensus. The overall strength of a binding site is found by summing the individual bit contributions. A distribution of these strengths is roughly Gaussian and shows that most natural binding sites have much less information than the consensus sequence (Schneider, 1997*a*). The strict consensus (where only the most frequent base is used) is the strongest possible binding site and is on the far high end of the distribution. For example, only one in 270 acceptor sites matches the strict consensus. For this reason it is generally inappropriate to say that one has a consensus binding site at such-and-such a position on a sequence.

As mentioned earlier, using consensus sequences to find binding sites by counting mismatches can lead to errors. How does this compare to the information theory approach? If matches to the consensus are assigned to have 1 unit and mismatches 0 units, then the total count is an integer. In contrast, the information theory weights are $2 + \log_2(\text{base frequency}) + (\text{a small-sample correction})$, which includes the real numbers. Summing the information theory weights gives continuous results, while counting mismatches gives blocky results that will often be off the mark. The commonly used 'percent identity' between two sequences, such as proteins, is flawed for the same reasons.

Sometimes counting matches or mismatches can give results opposite to the information measure weights so that a base in a site could have a mismatch to the consensus and yet that base could contribute positive information. For example, for a position that has 60% A, 30% T, 5% G, and 5% C the consensus base is A by two-fold, and yet a T in an individual binding site would contribute $2+\log_2 0.30 = 0.26$ bits. Only by noting the total distribution can we learn that the T contributes positively to the information. A related effect that is hidden by a consensus is that the diversity of the less frequent bases affects the total sequence conservation. For example, a position with 70% A, 30% T, 0% G, and 0% C has 1.12 bits of conservation, but a position with 70% A and 10% for each of C, G and T has only 0.64 bits. The consensus for both cases, A, does not distinguish between these.

When there are very few sequences, statistical artifacts crop up. Even if there's no information in the set, it can look like there is. For example, if one has only 6 random sequences, one will frequently observe positions that have 50% or more of one base. If, as is commonly done, one uses 50% as the cutoff for writing the consensus base, then one can get the false impression that there is pretty good sequence conservation. In the example shown in Fig. 3, 25 of 41 positions would be identified as 'conserved' even though the sequences were randomly generated! In general, 26 ± 3 of the 41 positions in 6 randomly generated sequences would be marked as the consensus. ⇐Fig 3

Missing the trees in the data forest. As a result of counting mismatches to a consensus it is possible to entirely miss a binding site. One of the most striking examples is a Fis site in the *tgt/sec* promoter of *E. coli* (Fig. 4). We carefully collected 60 sequences shown by DNase I footprinting to be bound by Fis, used information theory to maximize the information content of the alignment (Schneider & Mastronarde, 1996) and produced a model of how Fis binds (Hengen *et al.*, 1997). The model did a good job of predicting where Fis binds in the original footprint regions. However the model also predicted a site in *tgt/sec* that was not noted by the original authors. Surprisingly, four pieces of data support ⇐Fig 4 the existence of a Fis site centered at position −73 relative to the start of transcription (Slany & Kersten, 1992):

1. Fis often induces DNase I hypersensitive phosphates; these are seen between bases −53 and −50, corresponding to the Fis site at position −58. In addition there are hypersensitive positions between −80 and −78, which correspond to the site shown by the walker at −73.

2. The $5'$ end of the original DNase I footprints was not well determined, but could have extended to cover the Fis site at −73.

3. Gel mobility shift assays showed two band shifts when the entire region was used, one band shift when the region $5'$ to the conveniently (!) placed *Cla*I site was removed (which would eliminate the proposed Fis site) and none when both were removed.

4. Assays with these same nested DNAs showed that both sites activate transcription.

The authors were aware that there was a second binding site, but placed its location somewhere in the 69 bases upstream of the *Cla*I site. Why did they miss the site? Two positions (indicated by arrows in the figure) did not match the 'accepted' consensus sequence (Hübner & Arber, 1989). The consensus method gave these positions far more weight than was appropriate. The information for the site at −73 is 10.8 bits, which is 2 bits more than the average. To determine if there is really a site there, we performed a gel shift experiment using a DNA containing only the proposed Fis site at −73 and showed that the sequence is indeed bound by Fis (Hengen *et al.*, 1997). Because the consensus sequence failed to predict a site that had been documented experimentally, that site could not be seen, and to the scientists it did not exist (Kuhn, 1970).

A more critical example is in the hMSH2 gene, which is associated with familial non-polyposis colon cancer (Rogan & Schneider, 1995). A 'T' to 'C' transition occurred at position -5 of an acceptor site and this change was proposed to be the cause of the disease (Fishel *et al.*, 1993). Inspection of the logo in Fig. 1 shows that the consensus at position -5 (base zero is just to the left of the vertical bar, the first base on the intron side) is a T, but that close to half of the bases in the polypyrimidine tract are C. When the transition is made, the individual information changes by only 0.2 bits, which is not significantly different. A study of 20 normal people found that only 2 had this change (Leach *et al.*, 1993), so the change is a polymorphism unrelated to the disease.

Why did this potential 'misdiagnosis' happen? We suppose that T was taken to be the consensus sequence. Given this, one would interpret *any* change from that consensus to be detrimental. In this case the consensus sequence was so rigid that it could not handle a subtle change and a site disappeared from the scientist's view even though it was still functional. As DNA sequencing technologies become widely available to doctors, this situation will come up repeatedly. Serious malpractice suits could occur as a result of using the consensus model.

Flipping the light on to see an unseen world. Two recently published examples demonstrate some of the interesting biology that one can miss by using a consensus sequence. The first example is the RepA binding site (Fig. 5). The sine wave over the logo ⇐Fig 5 represents the twist of B-form DNA. The crests of the wave represent the protein facing a major groove and the troughs represent the protein facing a minor groove. Experimental data indicate that the major groove sides of Gs at positions 1 and 12 (black dots) are facing the protein, while the major groove sides of those base pairs at positions 6 and 8 (open circles) are facing away from the protein (Papp *et al.*, 1993). Hydroxyl radical and ethylation interference data also support this assignment and indicate that RepA binds to only one face of the DNA (Papp & Chattoraj, 1994).

RepA and other DNA binding proteins show sequence conservation up to 2 bits where they contact the major groove and only 1 bit where they face a minor groove (Papp *et al.*, 1993; Schneider, 2001). The upper bound of 2 bits is achievable because all 4 bases can be distinguished using contacts in the major groove (Seeman *et al.*, 1976). In contrast, the minor groove of B-form DNA is essentially symmetrical and can only provide up to 1 bit of sequence conservation.

Intriguingly, as seen in Fig. 5, RepA violates this rule at positions $+7$ and $+8$ where the protein faces a minor groove (Papp *et al.*, 1993). The violation implies that the DNA is not B-form. To understand this anomaly, we substituted a variety of chemically modified base pairs at position $+7$ (and its complement $+7'$) and found that the N3 proton on the thymine is responsible for contacting RepA through the minor groove (Lyakhov *et al.*, 2001). Since the N3 proton is normally sequestered in the center of the DNA helix, the DNA must indeed be distorted, as predicted from the sequence logo. Furthermore, the acceptable contact points for hydrogen bonding vary by several angstroms more than an H-bond could withstand in a rigid structure, suggesting that the base may rotate towards the minor groove for binding to occur. In other words, the T at $+7$ may be 'flipping' out of the DNA.

Base flipping was discovered by Rich Roberts in the co-crystal of the *Hha*I methyltransferase (Roberts, 1995; Roberts & Cheng, 1998). This solved a puzzle of how that enzyme functions, since the chemistry of methylation requires attack from above or below the plane of the base. Such an attack is not possible inside the DNA helix. The *Hha*I methyltransferase solves the problem by flipping the base out of the helix and into a pocket of the enzyme. Other DNA modification proteins also flip bases (Cheng *et al.*, 1993; Klimašauskas *et al.*, 1994; Verdine, 1994; Reinisch *et al.*, 1995).

Why would RepA be flipping a base? RepA is used by the bacteriophage P1 plasmid for DNA replication (Abeles, 1986; Abeles *et al.*, 1989). DNA replication requires that the helix be opened before synthesis can begin. The first step of this process would be the binding of RepA to the DNA. A very simple second step would be the flipping of a base out of the DNA, since DNA 'breathing' occurs naturally on a millisecond scale (Guéron *et al.*, 1987; Leroy *et al.*, 1988). If the thymine at +7 flips, is captured, and then held out of the DNA helix by RepA, weakened stacking could allow the remainder of the DNA to be more easily opened by a DNA helicase.

Sequence logos of other DNA replication protein binding sites have similar anomalies (Schneider, 2001), suggesting that base flipping may be a general mechanism for the second step of DNA replication.

How is this related to consensus sequences? The consensus sequence for RepA sites can be determined by reading the top letters of the sequence logo (Fig. 5) because the letters are sorted so that the most frequent base is on top. One finds: $5'$ ATGTGTGCTGGAGGGAAA $3'$. By viewing the binding site through the restrictive glasses of a consensus sequence, the unusual base becomes indistinguishable from the other bases!

A second example is the TATAAT sites mentioned earlier, for which the sequence logo is shown in Fig. 6. The logo shows that there is much lower sequence conservation in positions -10, -9 and -8 than in positions -12, -11 and -7, but the low region is significantly above background since the error bars are so small. (Contrast these tiny error bars to the ones in Fig. 3 and Fig. 5, where there are fewer sites.) The DNA region opened by RNA polymerase straddles the gap, leaving a highly conserved T at -7 near the $5'$ edge of the opened region. We propose that -7 is the first base opened during RNA transcriptional initiation, and a reasonably large body of experimental evidence supports this hypothesis (Schneider, 2001). As with RepA sites, the unusual thymine is obscured if one uses the consensus sequence. ⇐Fig 6

Just say no! We can express a consensus sequence in bits and so quantify the effect of making one. Each unique base (A, C, G or T) of a consensus counts as 2 bits. When two variations such as C or G are allowed (*e.g.* Fig. 3) we count 1 bit; there is, of course, no small-sample correction. 3 bases would count as $2 - \log_2 3$ (Schneider *et al.*, 1986), and 4 bases (N) would be zero. Plotting the total information of each sequence logo shown in this paper against this 'consensus information' we obtain Fig. 7 (summarized in Table 1). The figure shows that the consensus is always larger than the information content, even drastically so. However, the consensus could be tweaked (in individual cases) by arbitrarily playing around with the rules (Day & McMorris, 1992) to reduce the number of bits. But all the tweaking in the world will never give the proper weights because the frequencies are always rounded to obtain a consensus sequence. The examples in this paper show that when faced with the prospect of using a consensus sequence, we should 'just say no'. ⇐Fig 7 ⇐Table 1

Models and Illusions. One sometimes reads about how a particular DNA sequence has a consensus sequence at such-and-such a position (Robberson *et al.*, 1990). Thus using consensus sequences has led these biologists into a philosophical trap: confounding the model of reality (the consensus sequence) with reality (the binding sites). Even the original title of our paper on sequence logos reflects our initial confusion on this issue (Schneider & Stephens, 1990). Logos and walkers let us see more deeply into the genetic structure, revealing the details of sites and how mutations work. But no matter how sophisticated we are in depicting the patterns at binding sites, all we have are models. Logos and walkers are clearly better than consensus sequences (and can replace them completely), but they are still only representations of the universe 'out there' (Box, 1979). It is surprising, then, that scientists forget this and treat the consensus as reality. The effect was understood more than 30 years ago by Thomas Kuhn: once a paradigm is formed it occludes other ways of thinking and molds the way scientists perceive the world (Kuhn, 1970). Yet a consensus can no more be 'in' a DNA sequence than the meaning of these words is on the page. These words are interpretations in your mind; the page only has some disconnected black squiggles. One way to see this is to consider the perennial myth of a face on Mars that appears in American tabloid magazines. Whether or not there is a face on Mars, most of us have seen faces in clouds. Are there really faces there? Evidently not. Experiments with sheep and monkeys have identified neurons that become excited when a face is presented in the visual field (Kendrick & Baldwin, 1987). So faces in clouds, words, and consensus sequences are all constructs in our brains. Stranger still, the words may not be there when you perceive them since neural impulses take 300 milliseconds to travel from your eye to your brain (Rager & Singer, 1998), where, after another 80 milliseconds, they are finally perceived (Eagleman & Sejnowski, 2000). All that we see, hear, feel, smell, touch, and taste is delayed, so the entire perceived world is a model in our minds. Optical illusions remind us, and Zen masters understood, that everything is illusion (Purves *et al.*, 2002).

Acknowledgments. I thank Ilya Lyakhov, Becky Chasan, Peter K. Rogan, Zehua Chen, Krishnamachari Annangarachari, and Jeff Haemer for comments on the manuscript.

References

Abeles, AL. 1986. P1 plasmid replication. Purification and DNA-binding activity of the replication protein RepA. *J. Biol. Chem.* 261, 3548–3555.

Abeles, AL, Reaves, LD & Austin, SJ. 1989. Protein-DNA interactions in regulation of P1 plasmid replication. *J. Bacteriol.* 171, 43–52.

Barrett, M, Donoghue, MJ & Sober, E. 1991. Against Consensus. *Syst. Zool.* 40, 486–493.

Barrett, M, Donoghue, MJ & Sober, E. 1993. Crusade? A Reply to Nelson. *Syst. Biol.* 42, 216–217.

Basharin, GP. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probability Appl.* 4 (3), 333–336.

Box, GEP. 1979. Robustness is the strategy of scientific model building. In *Robustness in Statistics*, (Launer, RL & Wilkinson, GN, eds), pp. 201–236, Academic Press, New York, NY. 'ALL MODELS ARE WRONG BUT SOME ARE USEFUL' is a subheader in the chapter on page 202.

Cheng, X, Kumar, S, Posfai, J, Pflugrath, JW & Roberts, RJ. 1993. Crystal structure of the HhaI DNA methyltransferase complexed with S-Adenosyl-L-Methionine. *Cell,* 74, 299–307.

Day, WHE & McMorris, FR. 1992. Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res.* 20, 1093–1099.

de Queiroz, A. 1993. For Consensus (Sometimes). *Syst. Biol.* 42, 368–372.

Eagleman, DM & Sejnowski, TJ. 2000. Motion integration and postdiction in visual awareness. *Science,* 287, 2036–2038.

Fishel, R, Lescoe, MK, Rao, MRS, Copeland, NG, Jenkins, NA, Garber, J, Kane, M & Kolodner, R. 1993. The human mutator gene homolog *MSH2* and its association with hereditary nonpolyposis colon cancer. *Cell,* 75, 1027–1038.

Guéron, M, Kochoyan, M & Leroy, JL. 1987. A single mode of DNA base-pair opening drives imino proton exchange. *Nature,* 328, 89–92.

Hengen, PN, Bartram, SL, Stewart, LE & Schneider, TD. 1997. Information analysis of Fis binding sites. *Nucleic Acids Res.* 25 (24), 4994–5002. http://www.lecb.ncifcrf.gov/~toms/paper/fisinfo/.

Hübner, P & Arber, W. 1989. Mutational analysis of a prokaryotic recombinational enhancer element with two functions. *EMBO J.* 8, 577–585.

Kendrick, KM & Baldwin, BA. 1987. Cells in temporal cortex of conscious sheep can respond preferentially to the sight of faces. *Science,* 236, 448–450.

Klimašauskas, S, Kumar, S, Roberts, RJ & Cheng, X. 1994. HhaI methyltransferase flips its target base out of the DNA helix. *Cell,* 76, 357–369.

Kuhn, TS. 1970. *The Structure of Scientific Revolutions.* second edition, The University of Chicago Press, Chicago.

Leach, FS, Nicolaides, NC, Papadopoulos, N, Liu, B, Jen, J, Parsons, R, Peltomäki, P, Sistonen, P, Aaltonen, LA, Nyström-Lahti, M, Guan, XY, Zhang, J, Meltzer, PS, Yu, JW, Kao, FT, Chen, DJ, Cerosaletti, KM, Fournier, REK, Todd, S, Lewis, T, Leach, RJ, Naylor, SL, Weissenbach, J, Mecklin, JP, Järvinen, H, Petersen, GM, Hamilton, SR, Green, J, Jass, J, Watson, P, Lynch, HT, Trent, JM, de la Chapelle, A, Kinzler, KW & Vogelstein, B. 1993. Mutations of a *mutS* homolog in hereditary nonpolyposis colorectal cancer. *Cell,* 75, 1215–1225.

Leroy, JL, Kochoyan, M, Huynh-Dinh, T & Guéron, M. 1988. Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *J. Mol. Biol.* 200, 223–238.

Lewin, B. 1997. *Genes VI.* Oxford University Press, Oxford.

Lisser, S & Margalit, H. 1993. Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res.* 21, 1507–1516.

Lyakhov, IG, Hengen, PN, Rubens, D & Schneider, TD. 2001. The P1 Phage Replication Protein RepA Contacts an Otherwise Inaccessible Thymine N3 Proton by DNA Distortion or Base Flipping. *Nucl. Acid Res.* 29 (23), 4892–4900. http://www.lecb.ncifcrf.gov/~toms/paper/repan3/.

Miller, GA. 1955. Note on the bias of information estimates. In *Information Theory in Psychology*, (Quastler, H, ed.), pp. 95–100, Free Press, Glencoe, IL.

Nelson, G. 1993. Why Crusade against Consensus? A Reply to Barrett, Donoghue, and Sober. *Syst. Biol.* 42, 215–216.

Papp, PP & Chattoraj, DK. 1994. Missing-base and ethylation interference footprinting of P1 plasmid replication initiator. *Nucleic Acids Res.* 22, 152–157.

Papp, PP, Chattoraj, DK & Schneider, TD. 1993. Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.* 233, 219–230.

Pribnow, D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA,* 72, 784–788.

Purves, D, Lotto, RB & Nundy, S. 2002. Why we see what we do. *Amer. Sci.* 90 (3), 236–243. http://www.americanscientist.org/articles/02articles/Purves.html.

Rager, G & Singer, W. 1998. The response of cat visual cortex to flicker stimuli of variable frequency. *Eur. J. Neurosci.* 10, 1856–1877.

Reinisch, KM, Chen, L, Verdine, GL & Lipscomb, WN. 1995. The crystal structure of

HaeIII methyltransferase covalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. *Cell,* 82, 143–153.

Robberson, BL, Cote, GJ & Berget, SM. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* 10, 84–94.

Roberts, RJ. 1995. On base flipping. *Cell,* 82, 9–12.

Roberts, RJ & Cheng, X. 1998. Base flipping. *Annu Rev Biochem,* 67, 181–198.

Rogan, PK, Faux, BM & Schneider, TD. 1998. Information analysis of human splice site mutations. *Human Mutation,* 12, 153–171. http://www.lecb.ncifcrf.gov/˜toms/paper/rfs/.

Rogan, PK & 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. http://www.lecb.ncifcrf.gov/˜toms/paper/colonsplice/.

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. *Meth. Enzym.* 274, 445–455. http://www.lecb.ncifcrf.gov/˜toms/paper/oxyr/.

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

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

Schneider, TD. 2000. Evolution of biological information. *Nucleic Acids Res.* 28 (14), 2794–2799. http://www.lecb.ncifcrf.gov/~toms/paper/ev/.

Schneider, TD. 2001. Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucl. Acid Res.* 29 (23), 4881–4891. http://www.lecb.ncifcrf.gov/~toms/paper/baseflip/.

Schneider, TD. 2002. Consensus Sequence Zen. *Applied Bioinformatics,* 1 (3), 111–119. http://www.lecb.ncifcrf.gov/~toms/papers/zen/.

Schneider, TD & 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, TD & Stephens, RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100. http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/.

Schneider, TD, Stormo, GD, Gold, L & Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188, 415–431. http://www.lecb.ncifcrf.gov/˜toms/paper/schneider1986/.

Seeman, NC, Rosenberg, JM & Rich, A. 1976. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA,* 73, 804–808.

Shannon, CE. 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.

Slany, RK & Kersten, H. 1992. The promoter of the *tgt/sec* operon in *Escherichia coli* is preceded by an upstream activation sequence that contains a high affinity FIS binding site. *Nucleic Acids Res.* 20, 4193–4198.

Stephens, RM & Schneider, TD. 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/.

Verdine, GL. 1994. The flip side of DNA methylation. *Cell,* 76, 197–200.

Zheng, M, Doan, B, Schneider, TD & Storz, G. 1999. OxyR and SoxRS regulation of *fur*. *J. Bacteriol.* 181, 4639–4643. http://www.lecb.ncifcrf.gov/˜toms/paper/oxyrfur/.

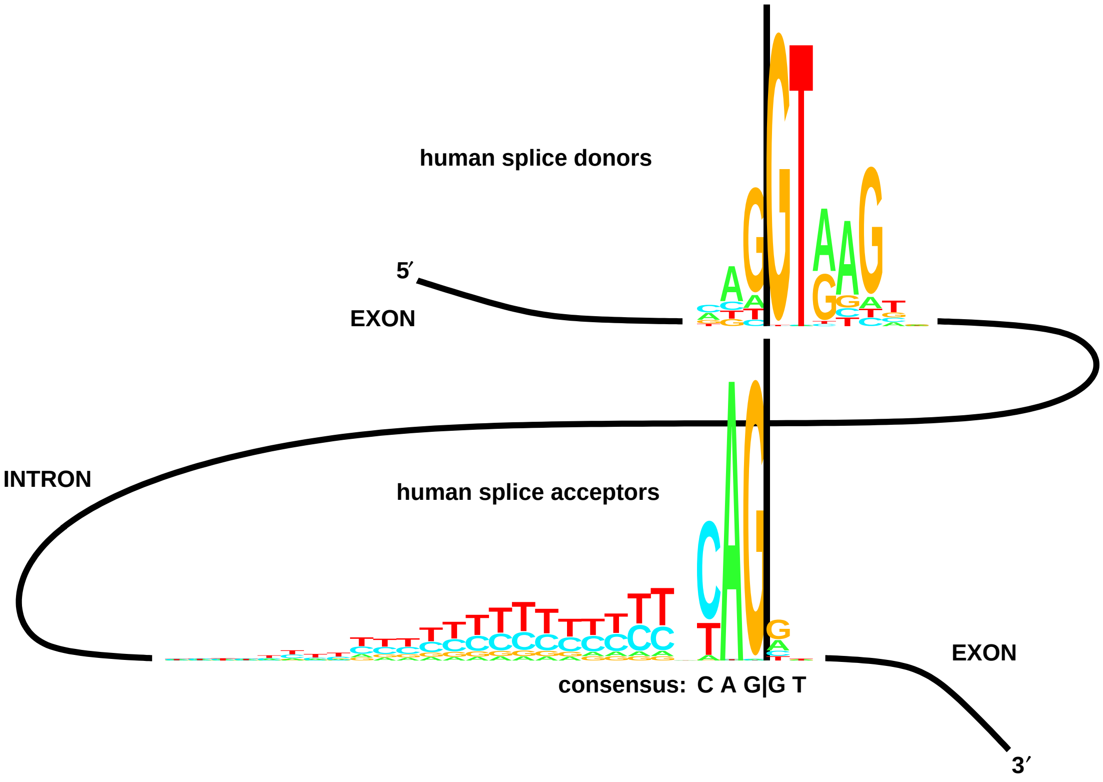

Figure 1: Sequence logos (Schneider & Stephens, 1990) for human donor and acceptor splice junctions (Stephens & Schneider, 1992) compared to the consensus sequence for both sites. Source: Adapted from (Stephens & Schneider, 1992).

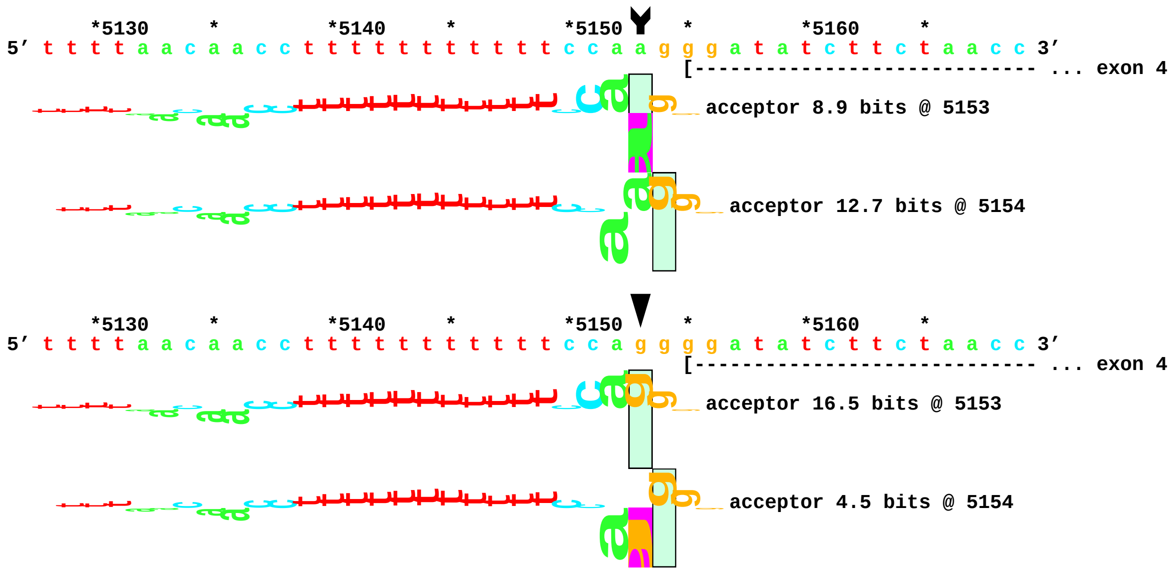

Figure 2: Sequence walkers (Schneider, 1997*b*) for a human acceptor site in the iduronidase synthetase gene and a mutation (indicated by an arrow). On the top sequence, the normal end of exon 4 is shown by a bracket and dashed line. The vertical rectangle on a sequence walker is the 'zero base' used to identify the location of the walker. The vertical rectangles also indicate a scale from -3 to $+2$ bits. A 12.7 bit acceptor at 5154 directs splicing to the correct location. Source: Adapted from (Rogan *et al.*, 1998).

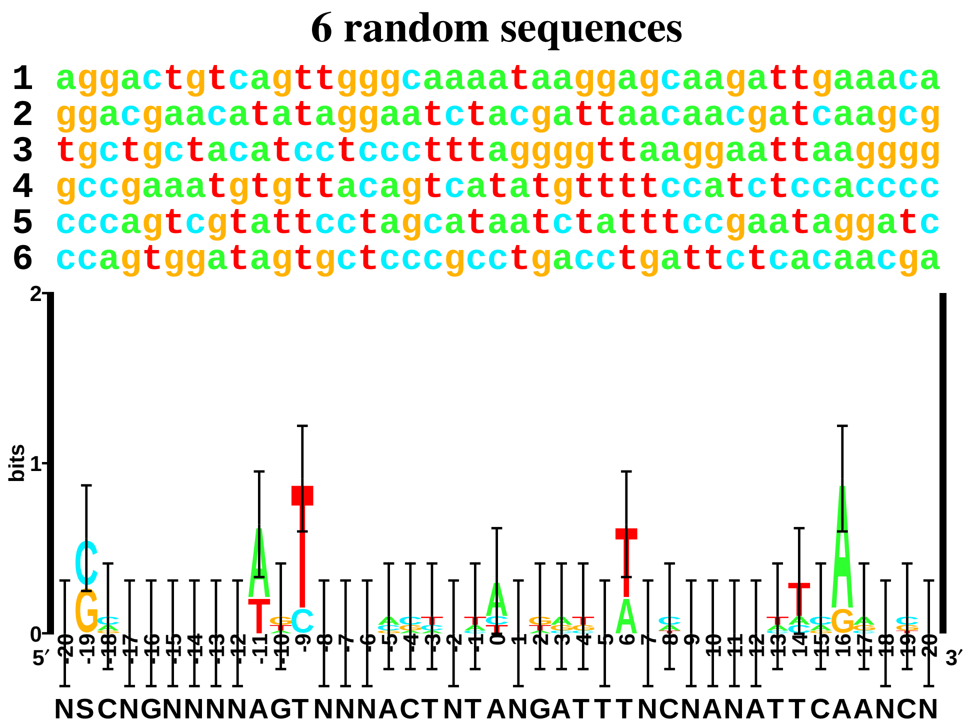

Figure 3: Sequence logo for random sequences.

Error bars, shown by I beams, indicate one standard deviation of the stack height. Note that a small-sample correction (Schneider *et al.*, 1986) suppresses the stack height so that a position such as -19, which is 50% C and 50% G, is lower than 1 bit. The correction is needed to counter a statistical bias that causes an apparent information to appear when one substitutes frequencies for probabilities in Shannon's equation (Schneider *et al.*, 1986; Miller, 1955; Basharin, 1959). The same effect makes one tend to see patterns where there are none. The consensus sequence on the bottom was chosen from positions that have 50% or more of one base. S is the two-letter code for C or G.

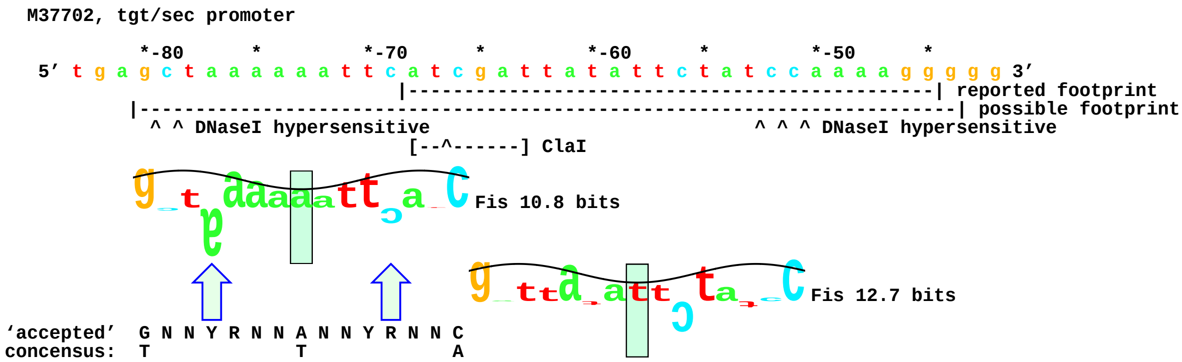

Figure 4: Region upstream of the *tgt/sec* promoter of *E. coli* analyzed by Fis sequence walkers. The information for each Fis site was computed from models that are 21 bases wide (-10 to $+10$) but only the range -7 to $+7$ is shown by walkers. The sine waves represent major (peaks) and minor (valley) grooves faced by the Fis protein. Source: Adapted from (Schneider, 1997*b*).

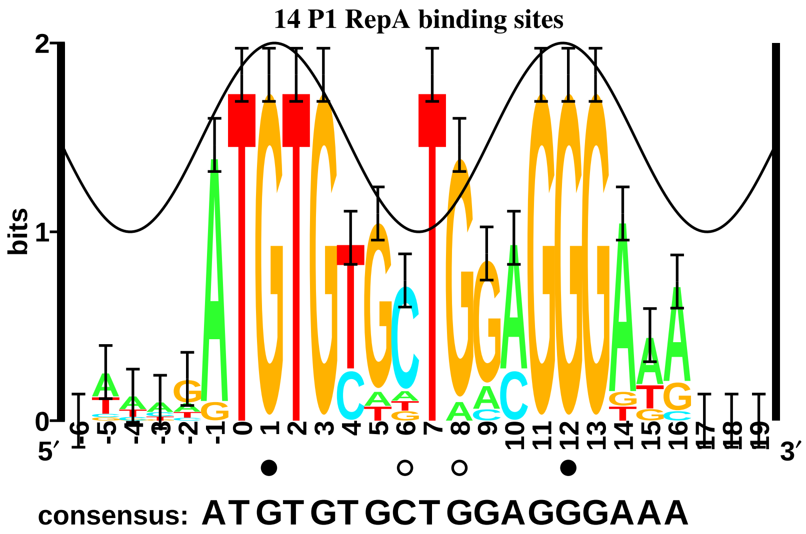

Figure 5: Sequence logo for RepA binding sites.

Error bars indicate standard deviations of the entire stack height. Source: Adapted from (Schneider, 2001).

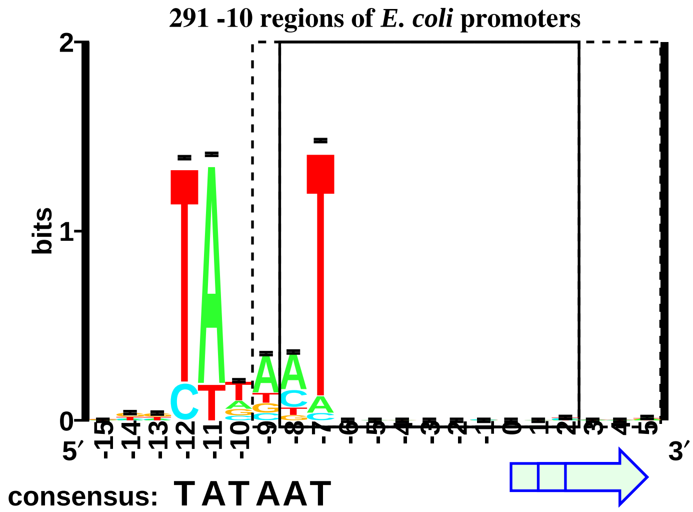

Figure 6: Sequence logo for the -10 region of *E. coli* promoters.

The promoters were from the Lisser-Margalit database (Lisser & Margalit, 1993). The dashed and solid boxes show the regions opened by the polymerase, while the arrow shows the start points of transcription. Source: Adapted from (Schneider, 2001).

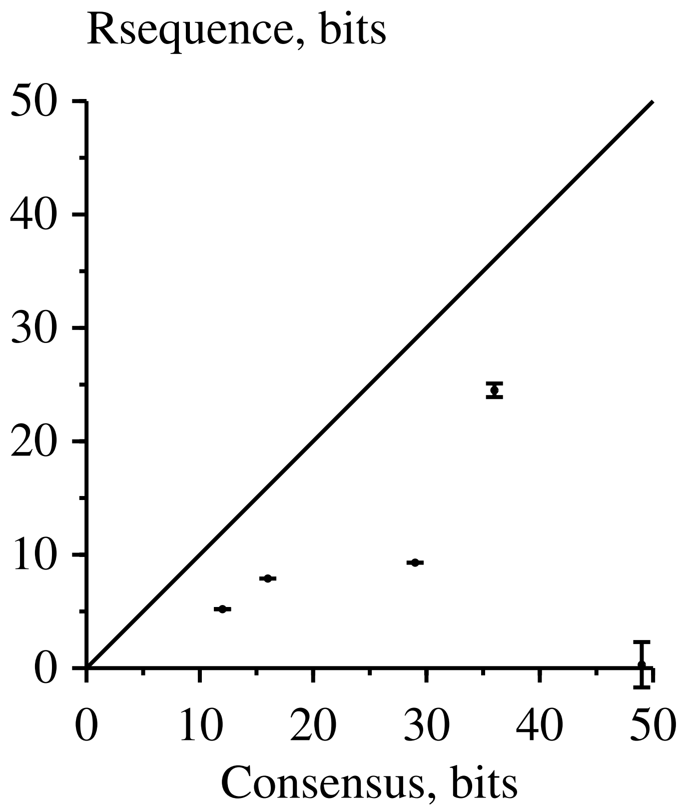

Figure 7: Consensus *versus* $R_{sequence}$.

The information for the 5 sequence logos in figures 1, 3, 5, and 6 was graphed by comparing the information content ($R_{sequence}$) to the information content of the corresponding consensus sequence. $R_{sequence}$ is the average information in a set of binding sites. It is also the summed area under the sequence logo. The line at 45° represents equality between the two measures. The data are summarized in Table 1.

Name	Rs	SD	CON	Consensus	From	To
	bits	bits	bits			
donor	7.9	0.0	16	NAGGTAAGTN	-3	+6
acceptor	9.3	0.0	29	NNNNNNNNNNNTTTTTTTTTYTNCAGGN	-25	+2
random	0.3	2.0	49	NSCNGNNNNAGTNNNACTNTANGATTTNCNANATTCAANCN	-20	+20
repa	24.5	0.6	36	ATGTGTGCTGGAGGGAAA	-1	+16
-10	5.2	0.0	12	TATAAT	-12	-7

Table 1: $R_{sequence}$ and consensus information for sequence logos.

Rs is $R_{sequence}$; SD is the standard deviation of Rs (to one decimal place); CON is the consensus information. The range of the site used is shown in columns From and To. The lowest frequency for using a base in each consensus was 0.4 and the consensus was computed using the program consensus (version 1.16, http://www.lecb.ncifcrf.gov/˜toms/delila/consensus.html).