1 Some Lessons for Molecular Biology from Information Theory

Thomas D. Schneider

National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Experimental and Computational Biology, P. O. Box B, Frederick, MD 21702-1201. toms@ncifcrf.gov, http://www.lecb.ncifcrf.gov/~toms/

version = 1.23 of lessons2000.tex 2003 April 4

This paper is a chapter in a book, a festscrift in honour of Prof. J. N. Kapur on “Entropy Measures, Maximum Entropy and Emerging Applications” which has been published by Springer [1]. Dr. Karmeshu, (Professor School of Computer and Systems sciences, Jawaharlal Nehru University New Delhi-110067) is the editor. This book is in the series ”Studies in Fuzziness and Soft Computing”, edited by Prof. Janusz Kacprzyk.

Abstract. Applying information theory to molecular biology problems clarifies many issues. The topics addressed are: how there can be precision in molecular interactions, how much pattern is stored in the DNA for genetic control systems, and the roles of theory violations, instrumentation, and models in science.

This paper is a short review of a few of the lessons I've learned from applying Shannon's information theory to molecular biology. Since there are so many distinct results, I call this emerging field ‘molecular information theory’. Many of the references and figures can be found at my web site [2], along with an earlier review [3] and a primer on information theory [4].

1.1 Precision in Biology

Information theory was first described by Claude Shannon in 1948 [5]. It sets out a mathematical way to measure the choices made in a system. Although Shannon concentrated on communications, the mathematics applies equally well to other fields [6]. In particular, all of the theorems apply in biology because the same constraints occur in biology as in communication. For example, if I call you on the phone and it is a bad connection, I may say ‘let me call you back’. Then I hang up. I may even complain to the phone company who then rips out the bad wires. So the process of *killing the phone line* is equivalent to *selecting against a specific phenotype* in biology.

A second example is the copying of a key. In biology that's called ‘replication’, and sometimes there are ‘mutations’. We go to a hardware store and have a key copied, but we get home only to find that it doesn't fit the door.

When we return to the person who copied it, they throw the key away (kill it) and start fresh.

This kind of selection does not occur in straight physics. It turns out that the requirement of being able to make distinct selections is critical to Shannon's channel capacity theorem [7]. Shannon defined the channel capacity, C (bits per second) as the maximum rate that information can be sent through a communications channel in the presence of thermal noise. The theorem has two parts. The first part says that if the data rate one would like to send at, R, is greater than C, one will fail. At most C bits per second will get through. The second part is surprising. It says that as long as R is less than *or equal to* C the error rate may be made as low as one desires. The way that Shannon envisioned attaining this result was by encoding the message before transmission and decoding it afterwards. Encoding methods have been explored in the ensuing 50 years [8,9], and their successful application is responsible for the accuracy of our solar-system spanning communications systems.

To construct the channel capacity theorem, Shannon assigned each message to a point in a high dimensional space. Suppose that we have a volt meter that can be connected by a cable to a battery with a switch. The switch has two states, on and off, and so we can send 1 bit of information. In geometrical terms, we can record the state (voltage) as one of two points on a line, such as $X = 0$ and $X = 1$. Suppose now that we send two pulses, X and Y. This allows for 4 possibilities, 00, 01, 10 and 11 and these form a square on a plane. If we send 100 pulses, then any particular sequence will be a point in a 100 dimensional space (hyperspace).

If I send you a message, I first encode it as a string of 1s and 0s and then send it down the wire. But the wire is hot and this disturbs the signal [10,11]. So instead of X volts you would receive $X \pm \sigma_X$, a variation around X. There would be a different variation for Y: $Y \pm \sigma_Y$. σ_X and σ_Y are independent because thermal noise does not correlate over time. Because they are the sum of many random molecular impacts, for 100 pulses the σs would have a Gaussian distribution if they were plotted on one axis. But because they are independent, and the geometrical representation of independence is a right angle, this represents 100 different directions in the high dimensional space. There is no particular direction in the high dimensional space that is favored by the noise, so it turns out that the original message will come to the receiver somewhere on a sphere around the original point [7,12,3].

What Shannon recognized is that these little noise spheres have very sharply defined edges. This is an effect of the high dimensionality: in traversing from the center of the sphere to the surface there are so many ways to go that essentially everything is on the surface [13,14,12]. If one packs the message spheres together so that they don't touch (with some error because they are still somewhat fuzzy) then one can get the channel capacity. The positions in hyperspace that we choose for the messages is the *encoding.* If we were to allow the spheres to intersect (by encoding in a poor way) then the receiver wouldn't be able to distinguish overlapping messages. The crucial point is that we must choose non-overlapping spheres. This only matters in human and animal communications systems where failure can mean death. It does not happen to rocks on the moon because there is no consequence for 'failure' in that case. So Shannon's channel capacity theorem only applies when there is a living creature associated with the system. From this I conclude that Shannon is a biologist and that his theorem is about biology.

The capacity theorem can be constructed for biological molecules that interact or have different states [12]. This means that these molecular machines are capable of making precise choices. Indeed, biologists know of many amazingly specific interactions; the theorem shows that not only is this possible but that **biological systems can evolve to have as few errors as necessary for survival**.

1.2 The Address is the Message

Keys select one lock in a set of locks and so are capable (with a little motive force from us) of making a 'choice'. The base 2 logarithm of the number of choices is the number of bits. (More details about information theory are described in a *Primer* [4].)

In a similar way, there are many proteins that locate and stick to specific spots on the genome. These proteins turn on and off genes and perform many other functions. When one collects the DNA sequences from these spots, which are typically 10 to 20 base pairs long, one finds that they are not all exactly the same. Using Shannon's methods, we can calculate the amount of information in the binding sites, and I call this $R_{sequence}$ because it is a rate of information measured in units of bits per site as computed from the sequences [15]. (See figure 1.1 for the details of this computation.)

For example, in our cells the DNA is copied to RNA and then big chunks of the RNA are cut out. This splicing operation depends on patterns at the two ends of the segment that gets removed. One of the end spots is called the donor and the other is called the acceptor. Let's focus on the acceptor because the story there is simple (what's happening at the donor is beyond the scope of this paper). Acceptor sites can be described by about 9.4 bits of information on the average [16]. Why is it that number?

A way to answer this is to see how the information is used. In this case there are acceptor sites with a frequency of roughly one every 812 positions along the RNA, on average. So the splicing machinery has to pick one spot from 812 spots, or $\log_2 812 = 9.7$ bits; this is called $R_{frequency}$ (bits per site). So **the amount of pattern at a binding site ($R_{sequence}$) is just enough for it to be found in the genome ($R_{frequency}$)**. Also, notice that we are using the fact that the capacity theorem says that it is possible for the sites to be distinguished from the rest of the genome.

1. The number of bases $b \in \{a, c, g, t\}$ at each position l in a set of aligned binding sites is called $n(b, l)$. The total number of sequences at a given position is
$$n(l) = \sum_{b=a}^{t} n(b, l), \tag{1.1}$$
where the sum is over all 4 bases b. In Fig. 1.2, the range of l is from -9 to $+9$ bases and $n(l) = 12$ for all positions. Many times data will be missing, in which case $n(l)$ will vary with position l.
2. The frequency of bases at each position is then computed as
$$f(b, l) = \frac{n(b, l)}{n(l)}. \tag{1.2}$$
3. Shannon's uncertainty [4,5] is estimated from
$$H = -\sum_{i=1}^{M} f_i \log_2 f_i + e(n) \qquad \text{(bits/symbol)} \tag{1.3}$$
for M symbols, where $e(n)$ is a correction for replacing the probability of the i^{th} symbol with a frequency f_i, which leads to a small-sample bias when the number of samples n is small [15].
4. Protein-DNA interactions are modeled at two thermodynamic states, *before* and *after* binding [3]. Before a protein binds DNA, all four bases are possible, so f_i is the frequency of each base in the genome, about 0.25, and equation 1.3 reduces to:
$$H_{before} \cong 2 \qquad \text{(bits/base)}. \tag{1.4}$$
After binding, the uncertainty is computed by equation 1.3 for each position l across the set of aligned binding sites, using equation 1.2 and $f_i = f(b, l)$:
$$H_{after} = H(l) = -\sum_{b=a}^{t} f(b, l) \log_2 f(b, l) + e(n(l)) \qquad \text{(bits/base)} \tag{1.5}$$
5. The information at each position is the *decrease in uncertainty* from *before* to *after* binding:
$$R_{sequence}(l) = H_{before} - H_{after} \qquad \text{(bits/base)} \tag{1.6}$$
R stands for a 'rate', in this case information gain in bits *per base*.
6. If the positions in a binding site are independent (which is generally true, but can be tested [16]) then the total information at the binding sites is the sum of the information over all positions:
$$R_{sequence} = \sum R_{sequence}(l) \qquad \text{(bits/site)}. \tag{1.7}$$

Fig. 1.1. Method of computing information content at protein binding sites ($R_{sequence}$) from DNA sequences.

1.3 Breaking the Rules

Within 5 days of discovering that $R_{sequence} \approx R_{frequency}$ for a number of genetic systems I found an apparent exception [15]. The virus T7 infects the bacterium *Escherichia coli* and replaces the host RNA polymerase with its own. These T7 polymerases bind to sites that have about $R_{sequence} = 35.4$ bits of information on the average. If we compute how much information is needed to locate the sites, it is only $R_{frequency} = 16.5$ bits. So there is twice as much information at the sites as is needed to find them.

The idea that $R_{sequence} \approx R_{frequency}$ is the first hypothesis of molecular information theory. As in physics if we are building a theory and we find a violation we have two choices: junk the theory or recognize that we have discovered a new phenomenon.

One possibility would be that the T7 polymerase really uses all the information at its binding sites. I tested this idea at the lab bench by making many variations of the promoters and then seeing how much information is left among those that still function strongly. The result was 18 ± 2 bits [17], which is reasonably close to $R_{frequency}$. So the polymerase does not use all of the information available to it in the DNA!

An analogy, due to Matt Yarus, is that if we have a town with 1000 houses we should expect to see $\log_{10} 1000 = 3$ digits on each house so that the mail can be delivered. (The analogy as is does not match the biology perfectly, but one can change it to match [3].) Suppose we came across a town and we count 1000 houses but each house has 6 digits on it. A simple explanation is that there are two delivery systems that do not share digits with each other.

In biological terms, this means that there could be another protein binding at T7 promoters. We are looking for it in the lab.

Some years after making this discovery, I asked one of my students, Nate Herman, to analyze the repeat sequences in a replicating ring of DNA called the F plasmid that makes bacteria male. (Yes, they grow little pilli ...) He did the analysis but did not do the binding sites I wanted because we were both ignorant of F biology at that time. Nate found that the *incD* repeats contain 60 bits of information but only 20 bits would be needed to find the sites. The implication is that three proteins bind there. Surprisingly, when we looked in the literature we found that an experiment had already been done that shows three proteins bind to that DNA [18,19]! It seems that **we can predict the minimum number of proteins that bind to DNA**.

1.4 Waves in DNA Patterns

If one calculates the information in many binding sites an interesting pattern emerges [20]: the information often comes in two peaks. The peaks are about 10 base pairs apart, which is the distance over which the DNA helix twists once. DNA has two grooves, a wide one and a narrow one, called the major

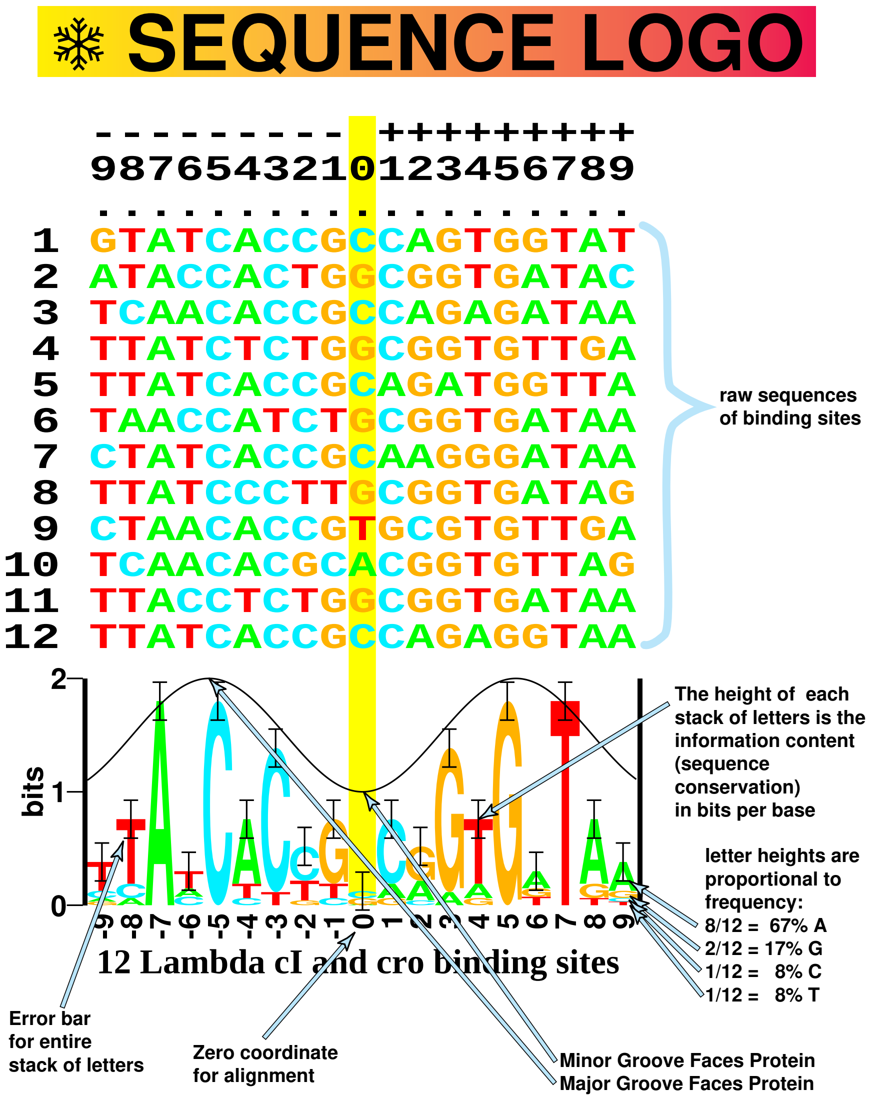

Fig. 1.2. Sequence logo for the 6 sequences (and their complements) bound by both the bacteriophage λ cI repressor and the cro proteins. The sequences are given 5′ to 3′. The method of computing the stack heights is given in Fig. 1.1.

and minor groove respectively. Using experimental data I found that the peaks of information correspond to places where a major groove faces the protein [20]. (See Fig. 1.2 for an example.)

This effect can be explained by inspecting the structure of bases [21]. There are enough asymmetrical chemical moieties in the major groove to allow all four of the bases to be completely distinguished. Thus any base pair from the set AT, TA, CG and GC is distinct from any other pair in the set. But because of symmetry in the minor groove it is difficult or impossible for a protein contact there to tell AT from TA, while CG is indistinguishable from GC. So a protein can pick 1 of the 4 bases when approaching the DNA from the major groove and it can make $\log_2 4 = 2$ bits of choices, but from the minor groove it only make 1 bit of choice because it can distinguish AT from GC but not the orientation ($\log_2 2 = 1$). This shows up in the information curves as a dip that does not go higher than 1 bit where minor grooves face the protein. In contrast, the major groove positions often show sequence conservation near 2 bits.

There is another effect that the information curves show: as one moves across the binding site the curve increases and decreases as a sine wave according to the twist of the DNA. This pretty effect can be explained by understanding how proteins bind DNA and how they evolve [22,23].

Proteins first have to locate the DNA and then they will often skim along it before they find and bind to a specific site. They move around by Brownian motion and also bounce towards and away from the DNA. So during the evolution of the protein it is easiest to develop contacts with the middle of a major groove, because there are many possibilities there. However, given a particular direction of approach to the DNA, contacts more towards the back side (on the opposite "face") would be harder to form and would develop more rarely. So we would expect the DNA accessibility for the major groove to go from 2 bits (when a major groove faces the protein) to zero (when a minor groove faces the protein). The same kind of effect occurs at the same time for the minor groove but the peak is at 1 bit. The sum of these effects is a sine wave from 2 bits for the major groove down to 1 bit for the minor groove, as observed. **The patterns of sequence conservation in DNA follow simple physical principles.**

1.5 On Being Blind

Why weren't the waves noticed before? The sine waves in binding site sequences cannot be seen with a method often used to handle sequences. Most molecular biologists will collect binding sites or other sequences, align them, and then determine the most frequent base at each position. This is called a 'consensus sequence'.

Suppose that a position in a binding site has 70% A, 10% C, 10% G and 10% T. Then if we make a consensus model of this position, we could call it 'A'. This means that when we come to look at new binding sites, 30% of the time we will not recognize the site! If a binding site had 10 positions like this, then we would be wrong $(1 - 0.7^{10}) = 97\%$ of the time! Yet this method is extremely widespread in the molecular biology literature.

For example, a Fis binding site in the *tgt/sec* promoter was missed even though four pieces of experimental data pointed to the site. Although the site was 2 bits stronger than an average Fis site, it was overlooked because it did not match the consensus used by the authors [24]. We tested the site experimentally and found that it does indeed bind to Fis [25]. Likewise the sine waves were missed before information analysis was done because creating a consensus sequences smashes the delicate sequence conservation in natural binding sites. Surprisingly, in retrospect, information theory provides good "instrumentation" for understanding the biology of DNA sequences.

In addition, information theory has been shown to be quite useful for biomedical applications. My colleague Pete Rogan found a paper that claimed to have identified a T to C change at a splice acceptor site as the cause of colon cancer. Presumably, the reason that the authors thought this is that the most frequent base at that position is a T. Then they apparently forgot that almost 50% of the natural sites have a C, so when they came across the T to C change it was misinterpreted as a mutation. Using information theory we were able to show that this is unlikely [26]. Our prediction was confirmed by experimental work which showed that of 20 normal people, 2 people had the change. If the initial claim had been made in a doctor's office it would have been a misdiagnosis, with legal ramifications. Since that time we have analyzed many splice junctions in a variety of genes and we have found that the information theory approach is powerful [27–30].

Consensus sequences apparently cause some scientists to make a classical scientific error. The first time that promoter binding site sequences were obtained (by David Pribnow) they were aligned. How can one deal with this fuzzy data? One way is to simplify the data by making a model, the consensus sequence. Although biologists are well aware that these frequently fail, they apparently don't recognize that the problem is with the model itself, and as a consequence they will often write that there is a consensus site in such and such a location and that, for example a protein binds to the consensus [31]. That is, they think that the *model* (a consensus sequence) is the same as the *reality* (a binding site). But a model of reality is not reality itself. This problem has a Zen-like quality, since even our perceptions are models of reality. Indeed, it is now thought that our minds are running a controlled hallucination that is continuously matching data coming from our senses, and when there is no input or a mismatch, some rather odd illusions occur [32].

We have developed two models that use information theory to get away from the errors caused by using consensus sequences. The first is a graphic called a sequence logo [33]. (An example is Fig. 1.2.) Sequence logos show an average picture of binding sites. Fortunately the mathematics of information theory also allows one to compute the information for individual binding sites and these models are called sequence walkers [34,24]. Many examples of logos and walkers can be found in the references or at my web site.

Consensus sequences are dangerous to use and should be avoided. Using the best available instrumentation can be critical to science. We should always be aware that we are always working with models because no model fully captures reality.

1.6 Acknowledgments

I thank Karen Lewis, Ilya Lyakhov, Ryan Shultzaberger, Herb Schneider, Denise Rubens, Shu Ouyang and Pete Lemkin for comments on the manuscript.

References

1. T. D. Schneider. Some lessons for molecular biology from information theory. In J. Kacprzyk, editor, *Entropy Measures, Maximum Entropy Principle and Emerging Applications. Special Series on Studies in Fuzziness and Soft Computing. (Festschrift Volume in honour of Professor J.N. Kapour, Jawaharlal Nehru University, India)*, volume 119, pages 229–237, New York, 2003. Springer-Verlag. Errata to the book: the two figures given in the paper are missing from the book!
2. T. D. Schneider, 2000. http://www.lecb.ncifcrf.gov/~toms/.
3. T. D. Schneider. Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology*, 5:1–18, 1994. http://www.lecb.ncifcrf.gov/~toms/paper/nano2/.
4. T. D. Schneider. *Information Theory Primer.* http://www.lecb.ncifcrf.gov/~toms/paper/primer/, 1995.
5. C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948. http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.
6. J. R. Pierce. *An Introduction to Information Theory: Symbols, Signals and Noise.* Dover Publications, Inc., New York, second edition, 1980.
7. C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37:10–21, 1949.
8. W. Gappmair. Claude E. Shannon: The 50th anniversary of information theory. *IEEE Communications Magazine*, 37(4):102–105, April 1999.
9. S. Verdú and Steven W. McLaughlin. *Information Theory: 50 Years of Discovery.* IEEE Press, New York, 1998.
10. H. Nyquist. Thermal agitation of electric charge in conductors. *Physical Review*, 32:110–113, 1928.
11. J. B. Johnson. Thermal agitation of electricity in conductors. *Physical Review*, 32:97–109, 1928.
12. T. D. Schneider. Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.*, 148:83–123, 1991. http://www.lecb.ncifcrf.gov/~toms/paper/ccmm/.

13. L. Brillouin. *Science and Information Theory.* Academic Press, Inc., New York, second edition, 1962.
14. H. B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. John Wiley & Sons, Ltd., N. Y., second edition, 1985.
15. T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986. http://www.lecb.ncifcrf.gov/~toms/paper/schneider1986/.
16. R. M. Stephens and T. D. Schneider. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.*, 228:1124–1136, 1992. http://www.lecb.ncifcrf.gov/~toms/paper/splice/.
17. T. D. Schneider and G. D. Stormo. Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucleic Acids Res.*, 17:659–674, 1989.
18. Y. Hayakawa, T. Murotsu, and K. Matsubara. Mini-F protein that binds to a unique region for partition of mini-F plasmid DNA. *J. Bacteriol.*, 163:349–354, 1985.
19. N. D. Herman and T. D. Schneider. High information conservation implies that at least three proteins bind independently to F plasmid *incD* repeats. *J. Bacteriol.*, 174:3558–3560, 1992.
20. P. P. Papp, D. K. Chattoraj, and T. D. Schneider. Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.*, 233:219–230, 1993.
21. N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, 73:804–808, 1976.
22. T. D. Schneider. Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Meth. Enzym.*, 274:445–455, 1996. http://www.lecb.ncifcrf.gov/~toms/paper/oxyr/.
23. T. D. Schneider. Evolution of biological information. *Nucleic Acids Res.*, 28(14):2794–2799, 2000. http://www.lecb.ncifcrf.gov/~toms/paper/ev/.
24. T. D. Schneider. Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.*, 25:4408–4415, 1997. http://www.lecb.ncifcrf.gov/~toms/paper/walker/, erratum: NAR 26(4): 1135, 1998.
25. P. N. Hengen, S. L. Bartram, L. E. Stewart, and T. D. Schneider. Information analysis of Fis binding sites. *Nucleic Acids Res.*, 25(24):4994–5002, 1997. http://www.lecb.ncifcrf.gov/~toms/paper/fisinfo/.
26. P. K. Rogan and T. D. Schneider. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Human Mutation*, 6:74–76, 1995. http://www.lecb.ncifcrf.gov/~toms/paper/colonsplice/.
27. P. K. Rogan, B. M. Faux, and T. D. Schneider. Information analysis of human splice site mutations. *Human Mutation*, 12:153–171, 1998. http://www.lecb.ncifcrf.gov/~toms/paper/rfs/.
28. C. Kannabiran, P. K. Rogan, L. Olmos, S. Basti, G. N. Rao, M. Kaiser-Kupfer, and J. F. Hejtmancik. Autosomal dominant zonular cataract with sutural opacities is associated with a splice mutation in the betaA3/A1-crystallin gene. *Mol Vis*, 4:21, 1998.

29. R. Allikmets, W. W. Wasserman, A. Hutchinson, P. Smallwood, J. Nathans, P. K. Rogan, T. D. Schneider, and M. Dean. Organization of the ABCR gene: analysis of promoter and splice junction sequences. *Gene*, 215:111–122, 1998. http://www.lecb.ncifcrf.gov/~toms/paper/abcr/.
30. S. G. Khan, H. L. Levy, R. Legerski, E. Quackenbush, J. T. Reardon, S. Emmert, A. Sancar, L. Li, T. D. Schneider, J. E. Cleaver, and K. H. Kraemer. Xeroderma Pigmentosum Group C splice mutation associated with mutism and hypoglycinemia - A new syndrome? *J. Investigative Dermatology*, 111:791–796, 1998.
31. C. Speck, C. Weigel, and W. Messer. From footprint to toeprint: a close-up of the DnaA box, the binding site for the bacterial initiator protein DnaA. *Nucleic Acids Res.*, 25:3242–3247, 1997.
32. V. S. Ramachandran and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. William Morrow & Co, New York, 1998.
33. T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990. http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/.
34. T. D. Schneider. Information content of individual genetic sequences. *J. Theor. Biol.*, 189(4):427–441, 1997. http://www.lecb.ncifcrf.gov/~toms/paper/ri/.