Measuring Molecular Information

Thomas D. Schneider *

version = 2.32 of ridebate.tex 1999 November 3
Journal of Theoretical Biology, 201: 87-92, 1999

Although information theory was developed more than 50 years ago (Shannon, 1948; Shannon, 1949), it is widely accepted (Gappmair, 1999), and a complete compendium of Claude Shannon's works was recently published (Sloane & Wyner, 1993). The application of information theory to understanding binding sites of proteins on DNA or RNA was published more than 10 years ago (Schneider *et al.*, 1986), and since then it has been profitably used to study many genetic systems (see http://www.lecb.ncifcrf.gov/~toms/). Shannon measured information as an average property of signals passing through a communications channel, so a natural extension is to understand the information contributed by individual symbols. The same extension can be applied to the study of binding sites as an "individual information theory" (Schneider, 1997*a*; Schneider, 1997*b*) and this has also been successfully used to understand a variety of genetic and medically relevant systems (Hengen *et al.*, 1997; Rogan *et al.*, 1998; Allikmets *et al.*, 1998; Kahn *et al.*, 1998; Shultzaberger & Schneider, 1999; Zheng *et al.*, 1999). Dr. Stormo subsequently published a letter in this journal promoting an alternative to the Shannon approach and

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

pointing out some consequences of that choice (Stormo, 1998). In this letter I will address other consequences and interpretations of the two approaches. However, before addressing the deep and difficult issues that Dr. Stormo has raised, which we have been discussing for more than 15 years, I would like to make some small factual corrections.

First, the Staden method (Staden, 1984) *is* discussed in my *J. Theor. Biol.* paper (Schneider, 1997*a*). Staden's method has no cutoff, while the individual information (R_i) method has a natural one and although they are similar, no one derived the R_i formula from Staden's approach. I did not derive the R_i method from Staden; it is a natural extension of information theory inspired by Tribus (Tribus, 1961). The connection between the information contributed by individual binding sites (as represented by the sequence walker computer graphics (Schneider, 1997*b*)) and their ensemble average (as represented by the sequence logo computer graphics (Schneider & Stephens, 1990)) is not obvious from the Staden approach, nor is the relationship to energy (Schneider, 1991*b*).

Second, in his letter (Stormo, 1998) Dr. Stormo implied that I "claim an inequality relationship with the enthalpy of binding". My papers do not claim any relationship with enthalpy; indeed I have not published the word "enthalpy" before now. While it is possible for q in the Second Law $dS \geq dq/T$ to refer to enthalpy (the increase in entropy of the surroundings of a system), the more appropriate measure for molecular machines is the total dissipation, and this corresponds to the free energy. (In this letter I use the terms energy and free energy synonymously.) At this point it would appear that we finally agree, but information is not energy as will be discussed in section II below.

I. What Does Dr. Stormo's I_{seq} Measure?

1. I_{seq} is not a state function. I_{seq} is a relative entropy that is not a distance measure because it is asymmetric and does not follow the triangle inequality (Cover & Thomas, 1991). So why isn't I_{seq} a state function? The previous argument used a simple 3 state case (Schneider, 1991*b*). A more general argument is to consider a series of N states that form a closed loop. Let $N-1$ of the steps between these states be made independently so that

$$\sum_{k=1}^{N-1} I_k \geq 0 \tag{1}$$

since "information" is (supposedly) additive for each independent event, and each step gives a zero or positive value. Irrespective of whether or not the last step is independent, $I_N \geq 0$ because the function is nonnegative. Therefore the sum around a loop is nonnegative:

$$\sum_{k=1}^{N} I_k \geq 0. \tag{2}$$

The only condition where $\sum I_k = 0$ is where no step had a change. By making many excursions to different composition regions of the genome, a recognizer would gain an arbitrarily large (and variable) information by Dr. Stormo's measure. In contrast, free energy and entropy are state functions (*i.e.*, functions of the current state of a system and not its history) and so their integration around a closed loop is always zero and prior history can be ignored. I_{seq} therefore cannot be used to compute energy as Dr. Stormo claims.

2. I_{seq} **is not an information theory measure.** Shannon's uncertainty

$$H = -\sum_i P_i \log_2 P_i \qquad \text{(bits per symbol)} \tag{3}$$

is related to the physical entropy if the probabilities correspond to the microstates of the system, so that $S = k_{\mathrm{B}} \ln 2 H$ (Schneider, 1991*b*). (H is often incorrectly called an entropy; see Tribus & McIrvine, 1971, for an amusing story about why.)

The theory I work with differs from that of Dr. Stormo in that it uses a definition of information that is path independent. A molecular machine — including not only genetic recognizers but also rhodopsin, myosin, *etc.* — dissipates energy into its surroundings as it makes choices (Schneider, 1991*a*). The information R (a rate of information, following Shannon's original notation) is a decrease in uncertainty:

$$R = H_{before} - H_{after} = -\Delta H \qquad \text{(bits per operation).} \tag{4}$$

For protein binding on a nucleic acid, the *before* state is the recognizer unbound or nonspecifically bound and the *after* state is it being specifically bound (Schneider, 1994). By using the state function H, the measure R is path independent. As a direct consequence Shannon's information can be compared to the measured energy change of such processes because energy changes are also path independent.

The formula ascribed to by Dr. Stormo is

$$\begin{aligned} I_{seq}(l) &= \sum_b f(b,l) \log_2 \frac{f(b,l)}{p(b)} \\ &= \left(-\sum_b f(b,l) \log_2 p(b)\right) - \left(-\sum_b f(b,l) \log_2 f(b,l)\right), \end{aligned} \tag{5}$$

where supposedly $p(b)$ is the probability of base b in the genome, and $f(b,l)$ is the frequency of base b at a position l in a binding site. Writing I_{seq} in the second form shows that it is a difference, but not of state functions since the first part mixes two states: $p(b)$ represents the unbound state and $f(b,l)$ represents the bound state.

Note that R can be computed from the genomic uncertainty

$$H_{before} = H_{genomic} = H_g = -\sum_b p(b) \log_2 p(b) \tag{6}$$

in which case, contrary to Dr. Stormo's claim, it *does* cancel the 'background': the information of regions outside a binding site will fluctuate around zero in a sequence logo (Schneider *et al.*, 1986; Schneider & Stephens, 1990). Therefore equation (6) *can* account for skewed genome composition. However, this may be fundamentally incorrect as there is no physical contact between the recognizer and the nucleic acid bases in this state.

In other words there are three possible formulas:

$$\begin{aligned} H_{before} &= 2 && (7) \\ H_{before} &= H_g && (8) \\ \text{“}H_{before}\text{”} &= -\sum_b f(b,l) \log_2 p(b) && (9) \end{aligned}$$

Formula (7) would be the strict molecular machine view in which contact is not made before binding (Schneider, 1991*a*; Schneider, 1994), so that the uncertainty is $\log_2 4 = 2$ bits. This raises the issue of how it is known to be 4 bases. However, the situation is equivalent to determining the channel capacity and therefore follows Shannon in that sense. Modification of bases, for example by methylation or glycosylation, does not increase the information capacity of DNA beyond 2 bits per base since the modifications depend on the sequence itself, for example in the methylation of adenine by Dam methylase at 5′ GATC 3′. However, increasing the number of symbols everywhere by adding new bases would increase the information, as has been done experimentally (Piccirilli *et al.*, 1990).

In formula (8) H_g can be used to cancel the 'background' around a binding site due to genomic composition skew (Schneider *et al.*, 1986), but this is dangerous because we don't know what causes the skew. For example, it could be caused by a nucleosome binding pattern everywhere in the genome and therefore real information is there. This leaves us with the difficult or unresolvable technical problem to separate and identify the information of other binding sites in such genomes. A similar difficult situation is to use purely theoretical means to distinguish ribosome binding site patterns from the downstream codon biases that occur with 3 base repetition. Aside from the toeprint experiment (Hartz *et al.*, 1988) one doesn't know exactly where the 3′ edge of the ribosome is (Rudd & Schneider, 1992), and it is not clear that complicated subtraction or extraction schemes would provide fair models close to the initiation codon since translation or protein chains may be different when they are just starting as compared to later on. Experimental approaches to determine the patterns, such as SELEX, are also presently inadequate (Schneider, 1996; Shultzaberger & Schneider, 1999).

Formula (9) is not a true Shannon uncertainty of the form $-\sum p \log p$, and is not a state function.

Thus formulas (7) or (8) appear reasonable but (9) is not and does not match the physics discussed in Section II.

3. I_{seq} can violate the channel capacity theorem. Shannon's channel capacity theorem provides an upper bound on the information that can be transmitted (Shannon, 1948; Shannon, 1949). It has been used to explain the observed precision of molecular systems (Schneider, 1991*a*; Schneider, 1994). Because I_{seq} can give indefinitely large values, it could be used to transmit more information than the channel capacity of a communications system, in violation of the theorem. Dr. Stormo gives an example where more than 2 "bits" per base are obtained from the string GGGG even though it never takes more than 2 bits to choose one object in four.

When discussing the computation for GGGG, Dr. Stormo does not give a justification for having more than 2 bits/base other than having $R_{sequence}$ (the average information at a set of binding sites) equal $R_{frequency}$ (the information needed to locate the binding sites on the genome). There are now a number of clear cases where $R_{sequence}$ does not equal $R_{frequency}$ for good biological reasons (Schneider *et al.*, 1986; Schneider & Stormo, 1989; Herman & Schneider, 1992; Rudd & Schneider, 1992; Stephens & Schneider, 1992), so forcing one's formula to make them equal means that one could miss important biological phenomena.

4. Interpreting I_{seq} as a macroscopic measure made by an observer. I understand that it is not Dr. Stormo's intent to model the observational process, but it is worthwhile understanding the implications of this possible interpretation. Formulas like I_{seq} directly compare two probability distributions, and because they always have positive values they can be interpreted as measuring the state change of an observer who doesn't forget. If this is the case, then they are not an appropriate measure for single molecules, which do forget where, or even whether, they were previously bound.

5. I_{seq} can measure prejudice. I_{seq}-like functions may be a way of measuring prejudice of an observer. They will give an indefinitely large response when some initial probabilities are small but later turn out to be large ($f(b, l) \gg p(b)$ in equation 5). That is, the more prejudiced the observer is, the more surprised they can be. This has a curious consequence. If there are 2 possible initial states and an observer believes that one of them is highly likely, then when the states change later the observer can gain more than 1 "bit" of information, even though a 2 state system cannot contain more than 1 bit of information since it takes only $\log_2 2 = 1$ yes-no question to completely identify one of the two items. The more prejudiced the person is about the initial state, the more that they 'learn', and they somehow learn more than it is possible to know! This violation of the channel capacity shows that it is not appropriate to assign the units "bits" to this measure.

6. I_{seq} as a global free energy measure. Dr. Stormo (private communication) indicates that I_{seq} is intended to "compare two different situations, the protein occurring equally at all possible positions and its equilibrium distribution." In other words, Dr. Stormo proposes it as a measure of the *macroscopic* binding reaction. By this interpretation, I_{seq} does not measure the state change of a single molecule, so it cannot be used to determine the average energy change *a single molecule* experiences in the transition between being non-specifically bound to the genome and being bound at the binding sites.

The choices made by a single protein cannot be sensitive to the macroscopic chemical equilibrium. For example, the local binding interaction between a single *Eco*RI molecule and the base A cannot be sensitive to the number of A molecules elsewhere on a DNA. The *Eco*RI molecule can only react with the bases it is close to.

II. What is the inequality that Dr. Stormo disputes?

The inequality is a version of the Second Law of Thermodynamics, given in a previous *J. Theor. Biol.* paper (Schneider, 1991*b*). The relationship derived from both the Second Law and (surprisingly!) from Shannon's channel capacity equation is:

$$\mathcal{E}_{min} = k_{\mathrm{B}} T \ln 2 \leq \frac{-q}{R} \quad \text{(joules per bit)} \tag{10}$$

where k_{B} is Boltzmann's constant, T is the absolute temperature and $\ln 2$ is a constant that gives units of bits. Positive q is defined as heat put into the system. The formula shows that to gain one bit of information (set $R = 1$) at least $k_{\mathrm{B}} T \ln 2$ joules must be dissipated ($-q$) to the surroundings. The Second Law forbids a smaller amount but allows a larger amount.

A coin is a useful example for understanding this. A coin can carry one bit of information, since it has 2 states and $\log_2(2) = 1$ bit. Consider a coin flipping in the air or bouncing around in a box. In such a condition it has no particular state and so its uncertainty is 1 bit. To 'store' information in the coin, it must come to rest on one or the other face. This requires that the energy in the coin be allowed to flow out to the surrounding environment. The point here is that the initial energy of the coin can have different values relative to the final value. The Second Law tells us that there is a minimum energy that must be dissipated per bit ($k_{\mathrm{B}}T \ln 2$ joules), but there can be extra dissipation that is merely wasted because under all conditions no more than 1 bit can be stored in the coin. With even a small inefficiency, the relationship between energy dissipated and information gained will be an inequality, contrary to Dr. Stormo's claim (see Tribus & McIrvine, 1971).

A coin is also a good analogy for the situation of a protein binding to DNA. Before specific binding, the protein/DNA complex has high energy, while after binding at specific DNA sites it has lower energy. The excess energy must be dissipated to the surroundings for the molecule to stick, since if the energy were not dissipated the molecule would move on. As with the coin, there can be an excess dissipation so there is no *a priori* relationship between energy and information aside from the Second Law bound.

If, in attempting to model binding energetics, $p(b)$ and $f(b, l)$ are to represent the time-average of various bases bound by the protein, then the non-equivalence of energy and information means that it is not correct to assume that these are the same as the base frequencies observed in the genome and in binding sites, respectively, since those correspond to information. In this case, these probabilities are not yet experimentally accessible and the measure Dr. Stormo proposes cannot be made.

On the other hand, these probabilities are usually presented as estimatable from observed base frequencies, in which case Dr. Stormo is working entirely on the information side of the energy/information equation (10) to compute his "specific free energy of binding". In this interpretation, I_{seq} cannot be a measure of energy. Because of the Second

Law inequality, the *only* way to know what the real energy is, is to go and make direct measurements of it.

Acknowledgments

I thank Lakshmanan Iyer for comments and for useful discussions leading to equation (2), and Elaine Bucheimer, John S. Garavelli, Denise Rubens, Peter K. Rogan, John Spouge, Bruce Shapiro, and Ryan Shultzaberger for comments on the manuscript.

References

Allikmets, R., Wasserman, W. W., Hutchinson, A., Smallwood, P., Nathans, J., Rogan, P. K., Schneider, T. D. & Dean, M. (1998). Organization of the ABCR gene: analysis of promoter and splice junction sequences. *Gene,* **215**, 111–122. http://www.lecb.ncifcrf.gov/~toms/paper/abcr/.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley & Sons, Inc., N. Y.

Gappmair, W. (1999). Claude E. Shannon: The 50th anniversary of information theory. *IEEE Communications Magazine,* **37** (4), 102–105.

Hartz, D., McPheeters, D. S., Traut, R. & Gold, L. (1988). Extension inhibition analysis of translation initiation complexes. *Meth. Enzym.* **164**, 419–425.

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

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

Kahn, S. G., Levy, H. L., Legerski, R., Quackenbush, E., Reardon, J. T., Emmert, S., Sancar, A., Li, L., Schneider, T. D., Cleaver, J. E. & Kraemer, K. H. (1998). Xeroderma Pigmentosum Group C splice mutation associated with mutism and hypoglycinemia - A new syndrome? *Journal of Investigative Dermatology,* **111**, 791–796.

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

Rogan, P. K., Faux, B. M. & Schneider, T. D. (1998). Information analysis of human splice site mutations. *Human Mutation,* **12**, 153–171. http://www.lecb.ncifcrf.gov/˜toms/paper/rfs/.

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

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

Schneider, T. D. (1991*b*). Theory of molecular machines. II. Energy dissipation from

molecular machines. *J. Theor. Biol.* **148**, 125–137.
http://www.lecb.ncifcrf.gov/˜toms/paper/edmm/.

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

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

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

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

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

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

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

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

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. IRE,* **37**, 10–21.

Shultzaberger, R. K. & Schneider, T. D. (1999). Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res,* **27** (3), 882–887. http://www.lecb.ncifcrf.gov/˜toms/paper/lrp/.

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

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

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

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

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

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

Zheng, M., Doan, B., Schneider, T. D. & Storz, G. (1999). OxyR and SoxRS Regulation of *fur*. *J. Bact,* **181**, 4639–4643.