New Approaches in Mathematical Biology: Information Theory and Molecular Machines

Thomas D. Schneider National Cancer Institute,
Frederick Cancer Research and Development Center,
Laboratory of Mathematical 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/

October 13, 2000

Abstract

My research uses classical information theory to study genetic systems. Information theory, founded by Claude Shannon in the 1940's, has had an enormous impact on communications engineering and computer sciences. Shannon found a way to measure information. We use this measure to precisely characterize the sequence conservation at nucleic-acid binding sites. The resulting methods completely replace the use of "consensus sequences", and therefore provide better models for molecular biologists. An excess of sequence conservation at bacteriophage T7 promoters and at F plasmid IncD repeats led us to predict the existence of proteins that bind the DNA. In another application of information theory, the wonderful fidelity of telephone communications and compact disk (CD) music can be traced directly to Shannon's channel capacity theorem. When rederived for molecular biology, this theorem explains the surprising precision of many molecular events. Through connections with the Second Law of Thermodynamics and Maxwell's Demon, this approach also has implications for the development of technology at the molecular level. [1]

The theory of molecular machines describes molecular interactions by using the mathematics of information theory [2, 3]. For convenience, I have divided the theory into three levels, which are characterized by these topics:

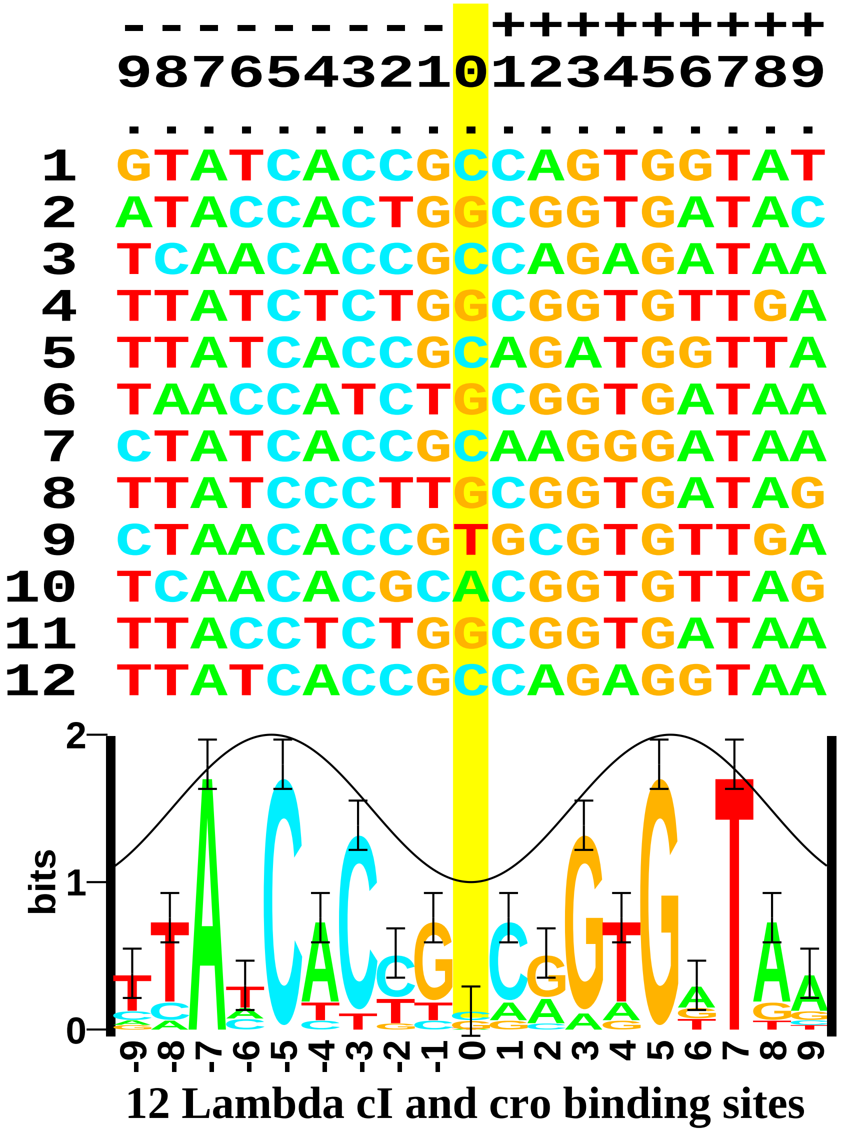
- Level 0. Sequence Logos: patterns in genetic sequences.
- Level 1. Machine Capacity: energetics of macromolecules.
- Level 2. The Second Law: Maxwell's Demon and the limits of computers.

This paper is a brief guide to papers presented elsewhere. See ftp://ftp.ncifcrf.gov/pub/delila/cover.ps for a list of references and http://www.lecb.ncifcrf.gov/~toms/paper/nano2 [4] for a review. Other information is available on the world wide web at http://www.lecb.ncifcrf.gov/~toms/. Discussions of these topics are held on the internet newsgroup bionet.info-theory.

1 Level 0. Sequence Logos: patterns in genetic sequences.

Genetic expression is usually controlled by proteins and other macromolecules ("recognizers") that bind to specific sequences on DNA or

[1] This paper, version = 1.16 of trieste1996.tex 2000 Oct 13, was published in [1]. I presented it at the Informatics session of the Trieste Conference on Chemical Evolution, IV: Physics of the Origin and Evolution of Life, Cyril Ponnamperuma Memorial. Trieste, Italy, September 4-8, 1995.



```
    --------- +++++++++
    9876543210123456789
    ...................
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

Figure 1: Example of 12 DNA sequences and their corresponding sequence logo.
There are 6 binding sites, and both proteins are dimers so both the sequence (odd rows) and its complementary sequence (even rows) were used for the analysis. This makes the resulting logo have more data at each position and it also makes the logo symmetrical. Error bars show the expected variation of the stack heights. The cosine wave represents the major (crest) and minor (trough) grooves of DNA facing the protein. This can be used to predict the face of the DNA bound by the protein [5].

RNA. Molecular biologists often characterize these sequences by a “consensus sequence” in which the most frequent base is chosen for every position of the binding site. Because the frequency information is lost, this method destroys subtle patterns in the data. How can we model binding sites without losing data? Fig. 1 shows the DNA sequences that the cI and cro proteins from bacteriophage λ bind to. Below these is shown a “sequence logo” [6]. Consider position -7 in the sequences. This is always an A in each of the 12 binding sites, so it is represented as a tall A in the logo. Position -8 has mostly T’s, 2 C’s and an A; this is represented in the logo as a stack of letters. The height of each letter is drawn proportional to its frequency and the letters are sorted so that the most frequent one is on top. The entire height of the stack is the sequence conservation at that position, measured in bits of information. A “bit” is the choice between two equally likely possibilities. There are 4 bases in DNA, and these can be arranged in a square:

```
A   C

G   T
```

To pick one of the 4 it suffices to answer only two yes-no questions: “is it on top?” and “is it on the left?”. Thus the scale for the sequence logo runs from 0 to 2 bits. When the frequencies of the bases are not exactly 0, 50 or 100 percent, a more sophisticated calculation must be made. The uncertainty is a function of the frequency $f(b,l)$ of each base b at position l:

$$H(L) = -\sum_{b=A}^{T} f(b,l)\log_2 f(b,l) + e(n(l)) \tag{1}$$

where $e(n(l))$ is a correction for the small sample size n at position l. The information content (or sequence conservation) is then:

$$R_{sequence}(L) = 2 - H(L). \tag{2}$$

The reasoning behind this formula is described in a primer on information theory that can be obtained from
ftp://ftp.ncifcrf.gov/pub/delila/primer.ps.

The sequence logo shows not only the original frequencies of the bases, but also shows the conservation at each position in the binding sites. Because it is a graphic, one can immediately see the pattern at the binding sites. In contrast to the sequence logo, one can be fooled by the distortions of a consensus sequence in which, for example, one cannot distinguish 100% A from 75% A.

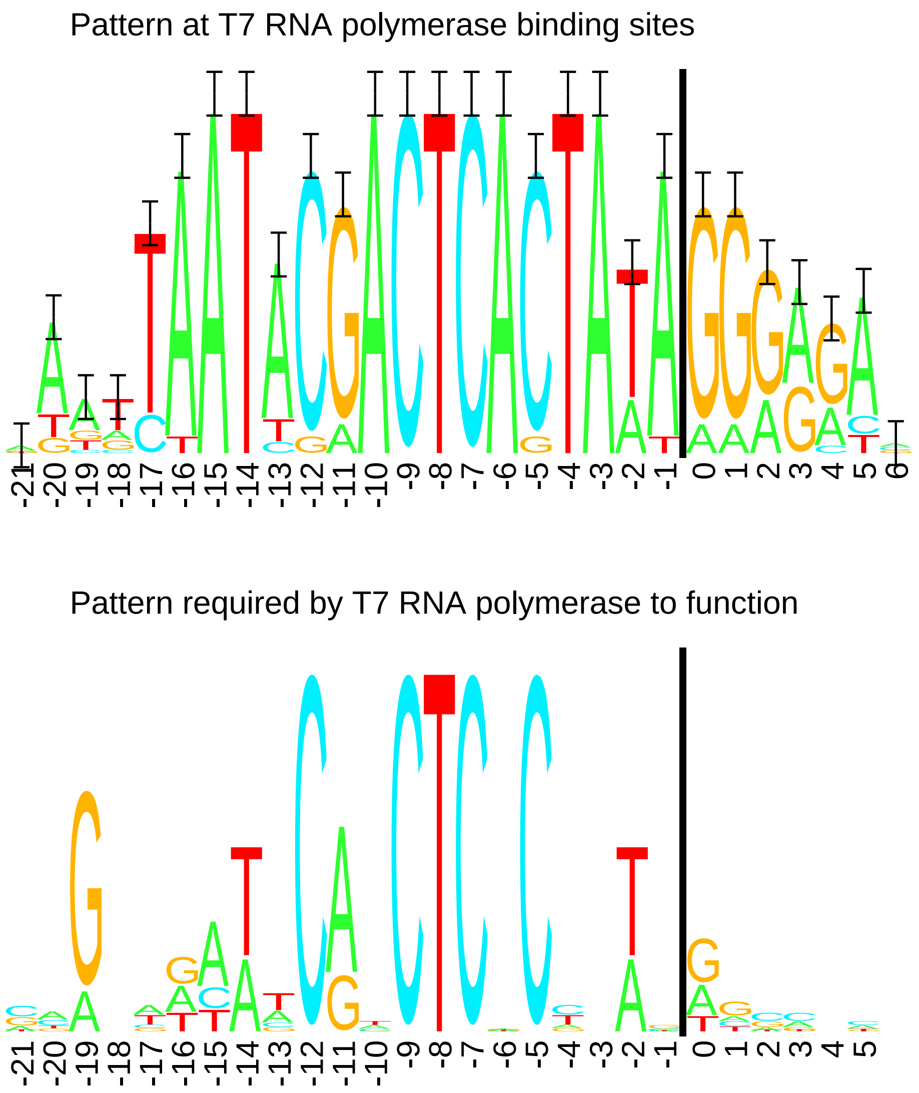
An important reason that we measure the sequence conservation using bits of information is that bits are additive. One can get the total sequence conservation in the binding site simply by adding together the heights of the sequence logo stacks:

$$R_{sequence} = \sum_{l} R_{sequence}(L). \tag{3}$$

This single number alone does not teach us anything, so we use an entirely different perspective to approach the problem of how a recognizer finds its binding sites. The recognizer must select the binding sites from all possible sequences in the genetic material, so we can calculate how many bits of choices it makes by determining the size of the genetic material G and the number of binding sites γ. Before the sites have been located, the initial number of bits of choice is $\log_2 G$, while after the set of sites have been found there remain $\log_2 \gamma$ choices that have not been made. So the decrease in uncertainty measures the number of choices made:

$$\log_2 G - \log_2 \gamma = \log_2 \frac{G}{\gamma} = -\log_2 \frac{\gamma}{G} = R_{frequency}. \tag{4}$$

The name $R_{frequency}$ was chosen because $\frac{\gamma}{G}$ is the frequency of the sites. This number is often close to the value of $R_{sequence}$, which means that the information content in binding site patterns is just sufficient for the sites to be found in the genome [7].


Pattern at T7 RNA polymerase binding sites

Pattern required by T7 RNA polymerase to function

Figure 2: Sequence logos for T7 promoters.
The vertical bars are 2 bits high. Transcription starts at base 0 and proceeds to the right.

Matt Yarus suggested a simple analogy that makes this clear. If we have a town with 1000 houses, how many digits must we put on each house to be sure the mail is delivered correctly? The answer is 3 digits since the houses can be numbered 000 through 999. So there is a relationship between the size of the town (size of genetic material and number of sites) and the digits on the mail box (pattern at the binding sites).

A surprising exception appears in the case of bacteriophage T7 promoters (Fig. 2 top), where $R_{sequence} = 35.4 \pm 2.0$ bits per site but $R_{frequency} = 16.5$ bits per site. There is a $R_{sequence}/R_{frequency} = 2.1 \pm 0.1$ fold excess of sequence conservation. Either the theory is wrong or we are learning something new. In the town analogy, there are 1000 houses, but each house has 6 digits on it. One explanation is that there are two independent mail delivery systems that could not agree on a common address system. The biological explanation is that there are two proteins binding at these patterns.[2] We already know about one of them, it is the T7 RNA polymerase. To test this idea, a large number of random DNA sequences were constructed and then ones which still functioned as T7 promoters were selected [8]. If there is another protein, then it would not be binding in this test and so the excess information would disappear. This is indeed what happened (Fig. 2 bottom): the binding sites for T7 promoters alone only have 18 ± 2 bits of information, close to the predicted value of $R_{frequency} = 16.5$ bits per site. The hypothesis that there is a second protein was upheld, but to date we have not identified it experimentally.

Later on we discovered another case in the F plasmid *incD* region where $R_{sequence} = 60.2 \pm 2.6$ bits per site and $R_{frequency} = 19.6$ bits per site so that there is a $R_{sequence}/R_{frequency} = 3.07 \pm 0.13$ fold excess of sequence conservation. Three proteins have been seen to bind to this DNA, and we were able to tentatively identify them [9].

[2] The question comes up as to why the information content for the two λ proteins of Fig. 1 do not give rise to a ratio of 2. The reason is that cI and cro bind to the same location in competition, so they share information. Presumably the two proteins that bind at T7 promoters do so simultaneously and do not use the same molecular contacts.

2 Level 1. Machine Capacity: energetics of macromolecules.

The results described above indicate that we can successfully apply ideas from information theory to molecular interactions. This suggests that other concepts from information theory should also apply. An important concept is that of the channel capacity. A given communications channel, such as a radio signal, will operate over a certain range of frequencies W and the signal will dissipate some power P into the receiver. The receiver must distinguish the signal from thermal noise N it is also receiving. Shannon found that these factors alone define the highest rate of information that can pass across the channel:

$$C = W \log_2 \left(\frac{P}{N} + 1 \right) \qquad \text{(bits per second).} \tag{5}$$

He also proved a remarkable theorem about the channel capacity [10]. If the rate of communication R is greater than the capacity, at most C bits per second will get through. On the other hand if $R \leq C$, *the error rate may be made as small as desired* but not zero. The way to do this is to encode the signal to protect it from noise so that when the signal is decoded, errors can be corrected. Coding is used in compact disks to correct up to 4000 simultaneous bit errors [11], which is why CD music is so clear.

The corresponding ideas can be constructed for molecular interactions in which a molecule ("molecular machine") makes choices from among several possibilities [12, 4]. The corresponding statement of the theorem is that so long as the molecular machine does not exceed the machine capacity, the molecular interactions can have as few errors as necessary for survival of the organism. Of course statements cannot be about "desires" in molecular biology, so the theorem is related to the evolution of the system. This mathematical result explains the observed precision of genetic control systems.

3 Level 2. The Second Law: Maxwell's Demon and the limits of computers

The Second Law of Thermodynamics can be expressed by the equation:

$$dS \geq \frac{dQ}{T}. \tag{6}$$

(See ftp://ftp.ncifcrf.gov/pub/delila/secondlaw.ps for discussion of this equation.) The equation states that for a given amount of heat dQ entering a volume at some temperature T, the entropy will increase dS at least by $\frac{dQ}{T}$.

We can relate entropy to the Shannon uncertainty if the probabilities describing the system states are the same for both functions, as is the case for molecular machines [13, 4]. This connection and the constant temperature at which molecular machines operate at allow us to rewrite the Second Law in the following form:

$$\mathcal{E}_{min} = k_{\mathrm{B}} T \ln(2) \leq \frac{-q}{R} \quad \text{(joules per bit)}, \tag{7}$$

where k_{B} is Boltzmann's constant. This indicates that there is a relationship between the information R and the heat q. Remarkably, this same limit can be determined from the channel capacity (equation (5)) and the machine capacity. The interpretation of this equation is straightforward—there is a minimum amount of heat energy that must be dissipated (*negative* q) by a molecular machine in order for it to gain $R = 1$ bit of information.

Maxwell's Demon is a mythical creature who is supposedly able to open and close a tiny door between two containers of gas [14, 15]. By observing the molecules that approach the door and by controlling the opening appropriately, the demon can allow the fast molecules through to one side and the slow ones to the other. Although any molecular biologist would expect the muscles and eyes of the demon to use energy, this is neglected by physicists. Also, they presume that the energy used to open the door can be regained when it is shut if it is attached to a spring. Such a demon could presumably create a macroscopic temperature difference between the two containers, and this could be used to run a heat engine. Apparently, the demon can supply energy merely by choosing between two alternatives. This would violate the Second Law of Thermodynamics.

Equation (7) applies to this problem. The demon always selects molecules in every scenario that he appears. We become duped by the story because the selective process is not explicitly stated as invoking the Second Law. But the Second Law always requires dissipation of heat energy to counterbalance selections made. Thus the demon is no longer a puzzle.

Equation (7) also applies both to molecular machines and to computers, so it sets a limit on computation. It is impossible to get temperatures of absolute zero because that would require infinite energy to remove all the heat energy. At any temperature above absolute zero, a computer must dissipate energy to make choices. As this energy must come from somewhere, we must feed the computer energy so that the computer can dissipate the energy while calculating our answers.

I thank Denise Rubens, Elaine Bucheimer and Paul N. Hengen for comments on the manuscript.

References

[1] T. D. Schneider. New approaches in mathematical biology: Information theory and molecular machines. In Julian Chela-Flores and Francois Raulin, editors, *Chemical Evolution: Physics of the Origin and Evolution of Life*, pages 313–321, Dordrecht, The Netherlands, 1996. Kluwer Academic Publishers.

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

[3] N. J. A. Sloane and A. D. Wyner. *Claude Elwood Shannon: Collected Papers*. IEEE Press, Piscataway, NJ, 1993.

[4] 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/.

[5] 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.

[6] 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/.

[7] 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.

[8] 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.

[9] 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.

[10] C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37:10–21, 1949.

[11] K. A. S. Immink. *Coding Techniques for Digital Recorders.* Prentice-Hall, Inc., N. Y., 1991.

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

[14] J. C. Maxwell. *Theory of Heat.* Longmans, Green and Co., London, 1904.

[15] H. S. Leff and A. F. Rex. *Maxwell's Demon: Entropy, Information, Computing.* Princeton University Press, Princeton, N. J., 1990.