A Brief Review of Molecular Information Theory

Thomas D. Schneider*

National Institutes of Health, National Cancer Institute at Frederick, P. O. Box B, Frederick, MD 21702-1201

Abstract

The idea that we could build molecular communications systems can be advanced by investigating how actual molecules from living organisms function. Information theory provides tools for such an investigation. This review describes how we can compute the average information in the DNA binding sites of any genetic control protein and how this can be extended to analyze its individual sites. A formula equivalent to Claude Shannon's channel capacity can be applied to molecular systems and used to compute the efficiency of protein binding. This efficiency is often 70% and a brief explanation for that is given. The results imply that biological systems have evolved to function at channel capacity, which means that we should be able to build molecular communications that are just as robust as our macroscopic ones.

Keywords: information theory, sequence logo, sequence walker, channel capacity, molecular efficiency, nanotechnology

1. An approach to constructing molecular communications

The fundamental step in communications, including communications at the molecular level, is the accurate reception of a signal. As is well known in communications engineering fields, the mathematical foundation for obtaining good reception was developed by Claude Shannon in 1948 and 1949 [1, 2, 3]. How can we apply these ideas to the construction of molecular communications? One approach is to first find out how biomolecules interact with each other and how they set their states. With some changes in perspective from conventional biochemistry, the states and patterns of molecules can be measured by using information theory and the field of study can be called molecular information theory [4, 5, 6].

So how can we apply Shannon's theory to molecules found in living organisms? A good starting point is to study the interactions between DNA and proteins that control the expression of DNA. Controlling a gene requires that the protein bind to specific points on the nucleic acid to either prevent or activate expression. There are only four nucleotide bases in DNA, named A, C, G and T, so the DNA to which a protein binds can be represented by the pattern of these letters. A protein is a fi-

*Corresponding author

Email address: schneidt@mail.nih.gov (Thomas D. Schneider)

URL: http://alum.mit.edu/www/toms/ (Thomas D. Schneider)

[1]version = 1.34 of brmit.tex 2011 Feb 26

[2]Nano Communication Networks, 1: 173-180, 2010. doi 10.1016/j.nancom.2010.09.002 http://alum.mit.edu/www/toms/papers/brmit/

nite molecule, which means that it can contact only a segment of a DNA at one time, typically 10 to 20 base pairs. The pattern that one protein searches for and subsequently binds is not always the same string of the four bases; the patterns vary. Some of this variation simply doesn't matter to the biological system, while other variations can be used by the protein to do different jobs.

2. Sequence logos show binding site information

We can use information theory to measure how much pattern is in a set of binding sites [13]. As an example, consider the Fis protein. In a starving bacterial cell there are below 100 molecules of Fis, but when the cell encounters nutrients, the numbers increase to over 50,000 molecules [14] and the Fis molecules then control many genes in the cell [7]. Fig. 1 shows several experimentally proven Fis sites from the front of the Fis gene itself. When there is not much Fis in the cell, the Fis gene is active, making more Fis molecules. Eventually those molecules occupy the sites at the Fis gene and shut the synthesis of Fis down as a negative feedback loop. How does Fis find these locations on the genome? From inspecting the figure, it is clear that the sequences are quite different, but the central region (around zero) has many A and T bases, while position −7 is almost always a G and +7 is almost always a C. The logo shows that different parts of the site are conserved by different amounts.

How can we characterize the sites precisely? We know that before Fis has bound to the DNA it can see all four possible bases at one position. So, following Shannon, we can say that the protein is *uncertain* as to what base it will see and that uncertainty can be measured as $\log_2 4 = 2$ bits [15, 16].

In contrast, once the protein has bound to a Fis site, the uncertainty of what it is touching in different cases is lower. It is nearly just one base in positions −7 and +7 and so there the uncertainty is near $\log_2 1 = 0$ bits. However, that is only an approximation since there *are* other bases at those positions. So the uncertainty is not zero. Fortunately Shannon worked out how to compute the uncertainty from the frequency of symbols [1]:

$$H(l) = -\sum_b f_{b,l} \log_2 f_{b,l} \quad \text{(bits per symbol)} \tag{1}$$

where $f_{b,l}$ are the frequencies of the bases $b \in \{A, C, G, T\}$ at a position l in the sequence alignment.

There are several caveats to note at this point. First, since we don't have infinite numbers of sequences, as Shannon's theory would require, we substituted the frequencies of bases for the probabilities. This requires making a correction for small sample size [13]. Another point is that the uncertainty is sometimes called the 'Shannon entropy', but if we say that then there can be confusion later when we have to discuss the actual entropy changes in the molecular binding process. Finally, it is important to be clear that the uncertainty given by equation (1) is *not* the information, as we discuss below.

Before binding to a site, the Fis molecule is somewhere on the DNA and sees 2 bits of uncertainty. After binding, it has lower uncertainty, $H(l)$. Shannon realized that the receiver of a message will get less information because of noise in the signal, and he showed that the information R received is reduced from the transmitted uncertainty $H(x)$:

$$R = H(x) - H_y(x) \quad \text{(bits per symbol)} \tag{2}$$

```
   -                    +
   1--------- +++++++++1
   09876543210123456789 0
   ....................  bits
 1 tttgccgattatttacgcaaa 12.2
 2 tttgcgtaaataatcggcaaa 12.2
 3 tgagtgtaaattttagtcact 11.8
 4 agtgactaaaatttacactca 11.8
 5 tatgagtaattatcgcaccac  9.0
 6 gtggtgcgataattactcata  9.0
 7 cacgctcattttaaatgcaat  6.5
 8 attgcatttaaaatgagcgtg  6.5
 9 attggtcaaagtttggcctt 12.2
10 aaaggccaaactttgaccaat 12.2
11 tatggtcatttcttaactcat  8.5
12 atgagttaagaaatgaccata  8.5
```

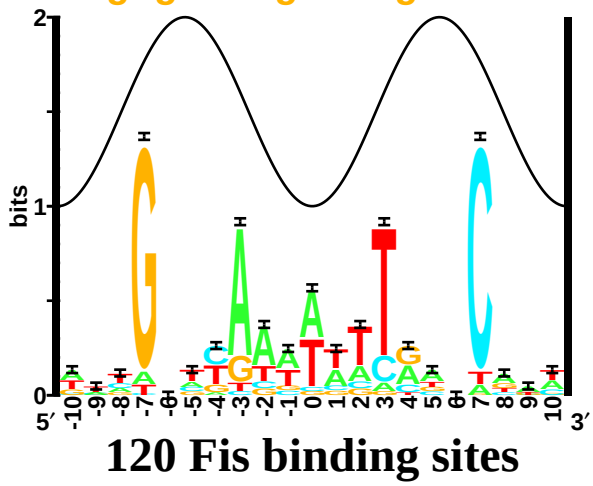

Figure 1: Aligned listing (top) and sequence logo (bottom) for DNA binding sites of the Fis protein from the bacterium *Escherichia coli*. The bar of numbers ('numbar') on the top is to be read vertically and it shows the range from −10 to +10 for positions across the site. Below the numbar are 6 Fis sites and their complementary sequences. Both are given since Fis binds as a dimer. On the right is the individual information for each sequence. The sequence logo on the bottom shows the sequence conservation in the complete data set which consists of 60 Fis sites and their complements [7]. The height of each letter is proportional to the frequency of that base at that position and the letters are sorted. The height of the entire stack of letters is the information, measured in bits. The possible variation of the height from small sample effects is shown by error bars. The peak of the sine wave shows where the major groove of DNA faces the protein [8, 9, 10, 11]. It can be used to infer some aspects of how the protein contacts the DNA [12].

He said "The conditional entropy $H_y(x)$ will, for convenience, be called the equivocation. It measures the average ambiguity of the received signal." [1].

Likewise, $H(l)$ is the ambiguity 'observed' by the DNA binding protein once it has bound to a site, so the information in the binding site is the uncertainty before binding less that after binding:

$$R_{sequence}(l) = 2 - H(l) \quad \text{(bits per base)} \qquad (3)$$

To show this graphically, we can create a sequence logo [17], as shown on the bottom of Fig. 1. We plot $R_{sequence}(l)$ across the binding site and use that to vary the heights of stacks of letters representing the relative abundances of bases at each position. Sequence logos are now widely used in molecular biology to present patterns in DNA, RNA or protein.

Variation of one part of a binding site is generally independent of other parts [18], so the information values at all of the positions in a binding site can be added together to find the total information in a binding site,

$$R_{sequence} = \sum_l R_{sequence}(l) \quad \text{(bits per site).} \qquad (4)$$

This is the 'area' under the logo, found by adding the heights of the letter stacks together.

3. Evolution of information

The significance of $R_{sequence}$ was found by comparing it to another measure of information. In many cases (but not Fis) the number of binding sites on the genome is known. So the problem facing the DNA binding protein is to locate a number of sites, γ, from the entire genome of size G. In information theory terms, the uncertainty before being bound to one of the sites is $\log_2 G$, while after being bound it reduces to $\log_2 \gamma$. So, as with the computation of the information in the binding sites, the information required to find the binding sites is

$$\begin{aligned} R_{frequency} &= \log_2 G - \log_2 \gamma \\ &= -\log_2 \gamma / G \quad \text{(bits per site).} \end{aligned} \qquad (5)$$

Natural binding sites have $R_{sequence}$ close to $R_{frequency}$ [13]. In other words, the information found in binding sites is just sufficient to locate the binding sites on the genome. Since the genome size and number of sites are more or less fixed by the environment, the information at the binding sites, $R_{sequence}$, has to evolve towards that required, $R_{frequency}$, and a computer model called Ev has verified this prediction [19]. A Java version that you can run on your own computer is now available at http://alum.mit.edu/www/toms/papers/ev/. Using this Evj, which was written by Paul C. Anagnostopoulos, one can watch the sequence logo evolve from scratch.

4. Sequence walkers show individual information of binding sites

The significance of $R_{sequence}$ is that it reflects how the genetic control system evolves to meet the demands of the environment as represented by $R_{frequency}$. If we inspect how $R_{sequence}$ is computed from equations (1), (3) and (4), we can see that it relies on the probability-weighted sum in equation (1). That is, $R_{sequence}$ is an average of the function $-\log_2 f_{b,l}$. But what does this average represent? It turns out that it can be expressed as an average of the information of individual sequences by adding together weights of the form $R_i(b,l) = 2 + \log_2 f_{b,l}$ for a sequence

[20, 21, 22]. The computed values for each binding site sequence in Fig. 1 are on the right side of the figure.

Like a sequence logo, the information of a single binding site can be presented graphically [25]. For example, Fig. 2 shows a portion of the Fis promoter region using sequence walkers. The figure reveals the marvelous but previously unsuspected complexity of the genetic control system.

As with Fis in Fig. 2, by using sequence walkers we find a similar complexity of potential promoters and ribosome binding sites between known genes [26]. The method produces a wealth of experimental predictions and can also be used to understand human diseases. Many genetic changes alter a single base at a time (single nucleotide polymorphism, SNP) and about 15% of those that cause genetic disease affect RNA splicing. The SNP changes can be analyzed using individual information and sequence walkers to predict whether the change is likely to be the cause of a disease [27]. Likewise, more than 50% of tumors in some types of cancer have mutations in the DNA binding p53 protein, so it is important for understanding cancer biology. We built an information theory model of how p53 binds to DNA, predicted that 16 genes should be under p53 control and then demonstrated experimentally that 15 of the 16 are indeed under p53 control [28]. Since 11 of the 16 sites had not been identified by other methods, information theory provides another tool for understanding the genetics of diseases.

5. Information and energy

Having determined practical measures of information in biological systems, the question arises, how is this information related to the binding energy? Surprisingly, the answer comes from two apparently different directions [29].

The first approach is from the Second Law of Thermodynamics expressed as the Clausius inequality [30, 31, 32]:

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

Here S is the total entropy of a system and it has units of joules per kelvin since Q is heat and T is the absolute temperature. The Boltzmann-Gibbs entropy of a physical system is

$$S \equiv -k_{\mathrm{B}} \sum_{i=1}^{\Omega} p_i \ln p_i \quad \left(\frac{\text{joules}}{\mathrm{K} \cdot \text{microstate}}\right) \tag{7}$$

in which k_{B} is Boltzmann's constant, p_i is the probability of a microstate and Ω is the number of microstates. Within binding sites for a protein there are only four possible states, represented by the individual bases, and these therefore correspond to the 'microstates' of the system. So we can equate the estimated probabilities of equation (1) with the probabilities of equation (7). This allows us to find the relationship between the uncertainty and the entropy:

$$S = k_{\mathrm{B}} \ln(2) H \tag{8}$$

Applying this to the Second Law, one can show that when the temperature is constant (as it is before and after a molecule functions since it equilibrates with the surroundings in a few picoseconds)

$$\mathcal{E}_{min} = k_{\mathrm{B}} T \ln(2) \quad \text{(joules per bit)} \tag{9}$$

where $\mathcal{E}_{min}$ is the minimum energy that must be dissipated out of a system for that system

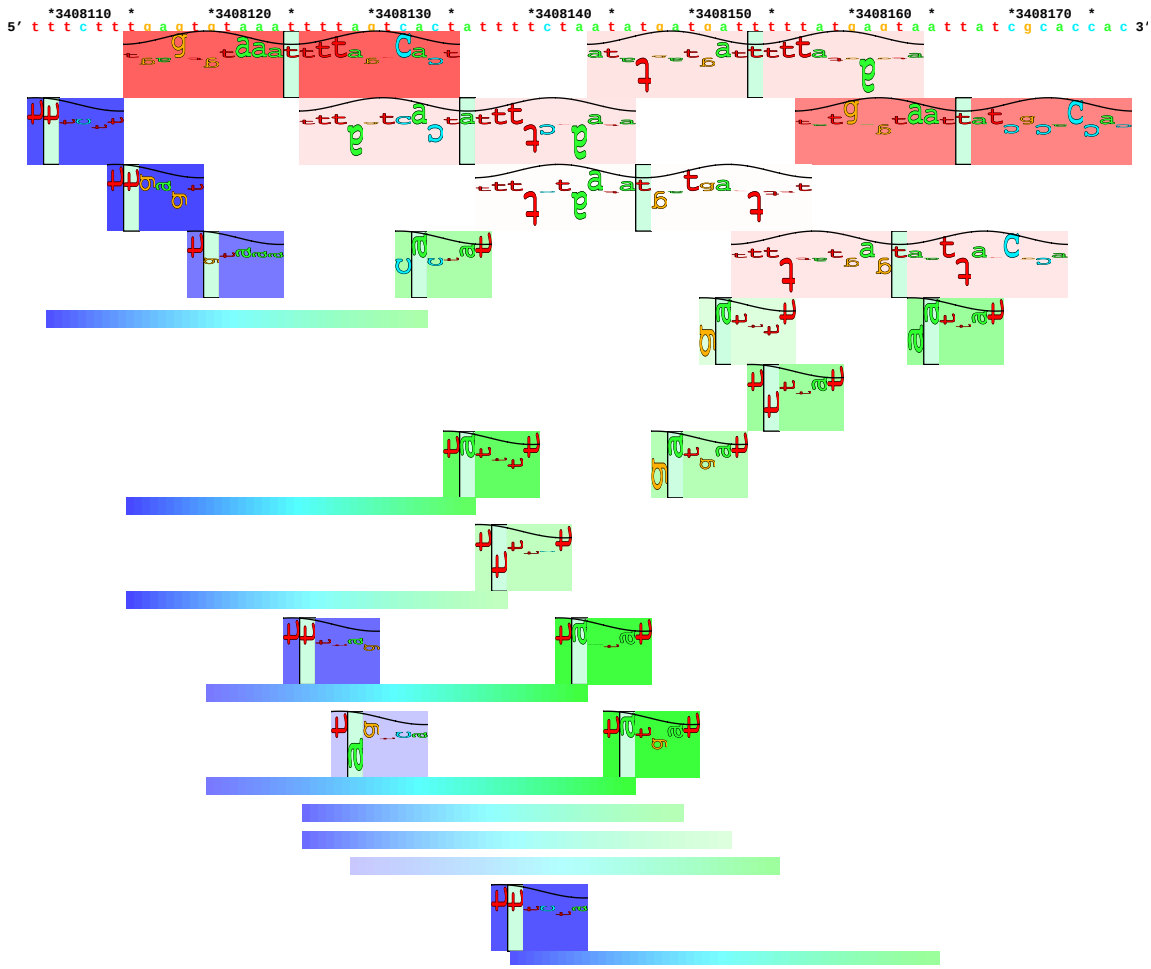

Figure 2: Sequence walkers show a map of DNA binding sites of the Fis protein on part of the Fis promoter on the genome of the bacterium *Escherichia coli* [7]. The DNA sequence from GenBank entry Accession NC_000913.2 is given on the top with every 5th and 10th base marked with an asterisk (*) and every 10th base numbered. A sequence walker, much like a sequence logo, is a set of adjacent letters with varying heights, but there is only one letter per position. A walker represents the location of where a protein binds. As in Fig. 1, sine waves show how the protein is oriented on the DNA. For example, directly below the DNA sequence are red and pink rectangles (called 'petals') each with a green bar in the center that marks the location of the binding site (zero coordinate of the logo in Fig. 1). Sequence walkers for Fis sites are shown superimposed on these red rectangles. The letters of each walker correspond to the DNA sequence directly above. The height of a letter is how well that base is conserved in the original data set of binding sites, measured in bits. The scale runs from 2 bits at the top of each rectangle to −3 bits at the bottom. Preferred bases go upwards while bases that are bad for binding go downwards. The total information for each site is shown by the saturation of the color so a red rectangle means that the site has more information than a pink or white one. A total of six Fis sites (two red, three pink and one white) are predicted to be in this piece of DNA. The two red sites are numbers 3 and 5 of Fig. 1. The bottom of this map also shows a set of promoters for making RNA from this region. The σ^{70} promoters shown have two parts, a '−35', shown in blue rectangles on the left and a '−10' region, shown by green rectangles on the right [23]. These parts are connected together by a variable-length horizontal bar below the rectangles [24]. The color of the connecting bar shades from one side to the other to help identify the corresponding sequence walkers at their zero coordinates, indicated by the green bars. Because Fis binds as a dimer, the binding sites are symmetrical and the sequence walker letters are vertical. Since promoters are asymmetrical, the letters in these sequence walkers are turned sideways so that reading 'down' the sequence of letters indicates the direction that transcription will take place. The map shows that since the gene for Fis is 'downstream' (to the right) of this region, Fis protein binds to its own promoter to control its own synthesis in a negative feedback loop.

to gain one bit of information [29]. This is another form of the Second Law of Thermodynamics [33], and so it provides a hard bound on what a system can do.

The other approach comes from Shannon's channel capacity equation. Shannon's equation describes the maximum bits that can be sent through a communications channel given the bandwidth, the power dissipated at the receiver and the thermal noise at the receiver. An equivalent equation was developed for molecules [34], in which the 'machine capacity' is the maximum bits that a molecular machine can select amongst given that it has a number of independently acting parts d_{space} (equivalent to bandwidth) and that it dissipates energy P_y (equivalent to power for a single molecular binding or selection event) in the presence of thermal noise N_y:

$$C_y = d_{space} \log_2 \left(\frac{P_y}{N_y} + 1 \right) \quad \text{(bits per selection)} \tag{10}$$

This equation relates bits to energy so one can define the actual joules used (*i.e.* dissipated) per bit gained as:

$$\mathcal{E} \equiv \frac{P_y}{C_y} \quad \text{(joules per bit).} \tag{11}$$

Taking the limit of $\mathcal{E}$ as P_y goes to zero gives equation (9) again [29].

$\mathcal{E}_{min}$ is not only a version of the Second Law of Thermodynamics (at constant temperature) but can also be used as an ideal conversion factor between energy and information. So we are now in a position to compare the energy dissipated when a DNA binding protein binds to DNA to the information it gains in doing so. The best example to use for this is the protein EcoRI.

Found in strains of the bacterium *Escherichia coli*, EcoRI acts as a molecular defense system. When a virus attacks a bacterium, the virus injects its DNA into the bacterial cell and the DNA contains signals that cause it to take over the cell metabolism to generate new virus particles. However, if the cell has EcoRI, the EcoRI scans along the viral DNA until it finds the sequence GAATTC, which it then cuts between the G and first A. This happens on both strands of the DNA since GAATTC on one side of the DNA reads the same in the opposite direction on the complementary strand. With both strands cut, the viral DNA falls apart and the virus is stopped. But why don't the bacteria destroy their own DNA? Because they have another enzyme that protects the GAATTC sequences by putting a methyl group on the second A [35]. Since the virus doesn't have that protection, they are attacked.

A sequence logo for EcoRI looks just like GAATTC since there is essentially no variation to the pattern that EcoRI binds. So to define the first position, the G, requires a choice of 1 in 4 or 2 bits. There are 6 positions so the total information to define GAATTC is $6 \times 2 = 12$ bits.

How much energy is dissipated when EcoRI binds? This has been carefully measured [36] and using $\mathcal{E}_{min}$ we find that it is equivalent to 17.3 bits. That is, for the energy dissipation that occurs during binding, EcoRI could make $R_{energy} = 17.3$ bits of choices but from the DNA sequence we see that it only does $R_{sequence} = 12$ bits. It is inefficient, and we can measure its efficiency compared to the ideal as $R_{sequence}/R_{energy} = 12/17.3 = 69\%$ [37]. Many other genetic systems are around 70% efficient (in preparation). What causes this effect?

An answer to why DNA binding proteins are 70% efficient comes from the machine capac-

ity equation. A sketch of the result is given here and details can be found in the original paper [37]. $\mathcal{E}_{min}$ from equation (9) is the best a molecule can do to make decisions by using energy since it gives the minimum joules dissipated to get one bit. In contrast, $\mathcal{E}$ from equation (11) is what a molecule actually does. So we can define a theoretical efficiency as

$$\epsilon_t \equiv \frac{\mathcal{E}_{min}}{\mathcal{E}} \tag{12}$$

Substituting into equation (12) equations (9), (11) and (10), and then using $N_y = d_{space} k_{\mathrm{B}} T$ [29] we find the elegant form

$$\epsilon_t = \frac{\ln\left(\frac{P_y}{N_y} + 1\right)}{\frac{P_y}{N_y}}. \tag{13}$$

It can be shown that this is an upper bound on the molecular efficiency since it comes from the channel capacity, which is an upper bound. The efficiency is unitless and has the range from 0 to 1. Intriguingly, if $P_y = N_y$, then $\epsilon_t = \ln 2 = 0.69$. So we can understand the occurrence of 70% efficiencies if we understand why P_y would equal N_y.

What would make the energy dissipated from a molecule (P_y) exactly equal the thermal noise flowing through the molecule at the same time (N_y)? This is almost a koan in its simplicity, but a distinct answer can be found by looking at the origins of the machine/channel capacity. (Readers who wish a challenge may stop at this point, read references [34] and [29] but not [37] and see if they can find a solution. It took the author 6 years.)

6. Coding theory explains molecular efficiency

In Shannon's model of communications, a series of D independent voltage pulses sent over a wire is represented by a point in a D dimensional space [2]. Although the pulses are initially distinct values, thermal noise distorts them by the time they reach the receiver. Essentially each pulse undergoes a drunkard's walk, which means there will be a Gaussian variation around each signal pulse. The noise on the pulses is also independent and a combination of independent Gaussian distributions forms a sphere in the high dimensional space [2, 34, 38],

Starting from the center of a sphere in high dimensional space (the initial pure message) and moving away from that point because of thermal noise, there are so many ways to go that most of the received points will be on the surface of the sphere [39, 40, 34]. One may think of the sphere as similar to a well-defined ping-pong ball—the continuous thermal noise is almost certain to put the received point on the surface. So for a transmitted point (the sphere center) the received point will almost certainly be somewhere on the surface of a sphere whose radius is determined by the thermal noise.

Each message is a point but what is received is a point on a sphere around the message point. The receiver's job is to determine the original message from the received point, which means to find the nearest possible transmitted message point. This is possible as long as the spheres do not intersect. An arrangement of message points that avoids sphere intersections is called a coding, and finding the nearest sphere center is called 'decoding'. The total space in which spheres can reside is determined by the power and it is also spherical but larger than the thermal noise spheres. Shan-

non noted that the maximum possible number of distinct messages M can be determined by dividing the volume of the large sphere by the volume of a smaller sphere. Then $\log_2 M$ in time t is the number of bits per second, the channel capacity.

Shannon proved an important theorem about the capacity. First, one cannot transmit data at a rate greater than the channel capacity C; at most C bits per seconds will get through. Conversely, so long as one transmits at a rate R less than or *equal* to the channel capacity, one may have as few errors as desired. This result is possible by appropriate coding. After more than 60 years to develop codes near Shannon's limit and the creation of computer chips to compute the algorithms for those codes we now have mobile cell phones that take advantage of this theorem. The result explains why our communications are so good.

An equivalent model can be built for molecules [34]. In this model the individual pulses are replaced by independently moving parts of the molecule. These parts are called 'pins' because they are analogous to the independently moving pins in a lock. Thermal noise impacts on the pins from all directions causing them to gain and lose energy either as potential energy or as a transient velocity. The series of random impacts means that the distribution of velocities for each pin is Gaussian. As with the communications model, the set of independent Gaussian distributions are represented by a sphere in a high dimensional space. A theorem equivalent to the channel capacity theorem says that so long as they do not exceed the machine capacity, living molecular machines may make as few errors as necessary for survival.

With several more concepts we can determine (if not fully understand) why molecular efficiencies are near 70%. The model for molecular machines is of spheres in a high dimensional space. A sphere represents the possible thermal vibrations of a molecule residing in one state. For example, the EcoRI protein bound to GAATTC would be a state. Another state would be EcoRI bound incorrectly to a similar sequence, say TAATTC. If these two states were not distinct for EcoRI, then EcoRI would be able to bind to TAATTC and cut there. This would be a disaster for the bacterium because the TAATTC sequences are not protected by methyl groups, the genome would be chopped up, and the cell would die. Obviously binding to GAATTC and to TAATTC should be different physically so to go from one to the other requires changes in the positions of atoms. In the high dimensional coding space, this is a vector between two spheres.

When EcoRI has a lot of energy it slides along the DNA. At every position it can move closer to the DNA [41] and if it is at a GAATTC sequence, the molecular surfaces will match and EcoRI will bind, dissipating energy to stick there [42, 43]. So before binding, the state of EcoRI is in a large 'before' sphere and after binding it is in a smaller 'after' sphere. If EcoRI were not to discriminate between sequences, it would end up in a sphere at the center of the large sphere. This sphere, called the 'degenerate sphere' represents binding to any sequence without discrimination. That would be fatal. So EcoRI must bind to a sphere representing GAATTC that is some distance away from the central degenerate sphere, the 'forward' sphere.

To get away from the central degenerate sphere to the forward GAATTC sphere requires dissipating energy. EcoRI must move outward by at least the radius of the degenerate sphere. As soon as the molecule has gone that far, it can decode into the GAATTC sphere. That is, the center of motion of the molecule can switch from one to the other state by a

rearrangement of the atoms. Once the rearrangement has occurred, the molecule will 'orbit' around the new state. This continuous motion around a central state is the equivalent of 'decoding' in a communications system. (See the appendix of [37] for more details of this model.)

Getting from the degenerate sphere center to the radius requires expending energy P_y that is at least equal to N_y (which determines the thermal noise radius) and so $P_y > N_y$. Plugging this into equation (13) gives an efficiency of no more than 69%, as we observe. While this result makes mathematical sense, unfortunately it is not intuitive.

In any case, we can draw several important lessons from it. First, the result implies that the molecule EcoRI is functioning at machine/channel capacity because to get this result it must be operating on the efficiency curve. To some people this is a surprising result—how could a tiny molecule approach this ultimate limit?—but to a biologist it should not be surprising. If there is an optimal solution, so long as there are no barriers in the way, we expect an evolutionary process to find the solution.

All the codes people have devised for communications systems potentially have applications in biology—we 'just' have to look at the biological data to see where they might apply. The efficiency result is telling us the reverse too. All the sophisticated mechanisms of biology must be based on codes—we 'just' have to learn what those codes are so that we can apply them to our own technologies. This result is highly encouraging since it means that we too could build molecules that have error rates as low as we may desire.

Communications at the molecular level will be limited by the machine/channel capacity. One can build a hybrid model for molecular machines that receive parallel signals and decode them both in space and in time as described in the appendix of [34]. Likewise, we can envision building single molecules that use coding systems to reliably transmit molecular states to the outside world. An example of such a device is the patent pending Medusa™ sequencer [44]. This molecule consists of a DNA or RNA polymerase that reads along a DNA or RNA and produces a series of light pulses that represent the DNA sequence. It is possible to code the output spectrum with a Hamming-like code so that damage to the Medusa™ sequencer will be reported, in which case its signals can be ignored.

7. Information theory in biology

We have used molecular information theory to investigate many biological systems across the 'central dogma' and beyond:

- DNA replication initiation by bacteriophage P1 RepA and other proteins [8, 9, 10, 11]
- transcription factors [13, 45, 7, 46, 47, 48, 28]
- RNA polymerases including those in T7 and related phages [13, 49, 50, 51, 24]
- splice junctions [52] and mutations in splice junctions causing human disease [53, 27, 54, 55], including the cancer-causing xeroderma pigmentosum [56, 57, 58, 59, 60]
- RNA folding [61]
- ribosome binding sites [13, 62, 63, 64, 65, 23]
- protein structure [66]
- evolution and phylogeny [67, 68, 19, 50, 51]
- vision and muscle [34, 29, 4, 37]

These papers only represent ones from my lab; hundreds of others have been published by many other groups, but a fair review of them would be far beyond the scope of this paper. Why is information theory so applicable to biology? Shannon embedded into his mathematics a single idea that does not appear in physics or thermodynamics: that states must not intersect. As discussed above, in communications we demand that messages not be confused, so the spheres in high dimensional space must not touch. Because it is a high dimensional space, this is possible, and Shannon took advantage of it to create the channel capacity formula. There is nothing in physics or thermodynamics that demands this separation. Likewise, molecular states in biological systems must also be distinct—for survival, and Shannon's theorem and Darwinian evolution guarantee that a solution can be found. This inevitably leads biological systems that make choices between two or more states to have 70% efficiency. Claude Shannon probably never realized that his work on the mathematics of communications was about biology [38].

8. Acknowledgments

I thank Don Court, Amar Klar, Ryan Shultzaberger, Rose Chiango and Carrie Paterson for reading and commenting on the manuscript. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

[1] C. E. Shannon, A Mathematical Theory of Communication, Bell System Tech. J. 27 (1948) 379–423, 623–656. http://tinyurl.com/Shannon1948.

[2] C. E. Shannon, Communication in the Presence of Noise, Proc. IRE 37 (1949) 10–21.

[3] C. E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, 1949.

[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 (1994) 1–18. http://alum.mit.edu/www/toms/papers/nano2/.

[5] T. D. Schneider, Some lessons for molecular biology from information theory, in: Karmeshu (Ed.), 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, Springer-Verlag, New York, 2003, pp. 229–237. http://alum.mit.edu/www/toms/papers/lessons2003/.

[6] T. D. Schneider, Twenty Years of Delila and Molecular Information Theory: The Altenberg-Austin Workshop in Theoretical Biology Biological Information, Beyond Metaphor: Causality, Explanation, and Unification Altenberg, Austria, 11-14 July 2002, Biol. Theory 1 (2006) 250–260. http://alum.mit.edu/www/toms/papers/schneider2006/.

[7] P. N. Hengen, S. L. Bartram, L. E. Stewart, T. D. Schneider, Information analysis of Fis binding sites, Nucleic Acids Res. 25 (1997) 4994–5002. http://alum.mit.edu/www/toms/papers/fisinfo/.

[8] P. P. Papp, D. K. Chattoraj, T. D. Schneider, Information analysis of sequences that bind the replication initiator RepA, J. Mol. Biol. 233 (1993) 219–230.

[9] D. K. Chattoraj, T. D. Schneider, Replication control of plasmid P1 and its host chromosome: the common ground, Prog. Nucl. Acid Res. Mol. Biol. 57 (1997) 145–186.

[10] T. D. Schneider, Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation, Nucleic Acids Res. 29 (2001) 4881–4891. http://alum.mit.edu/www/toms/papers/baseflip/.

[11] I. G. Lyakhov, P. N. Hengen, D. Rubens, T. D. Schneider, The P1 Phage Replication Protein RepA Contacts an Otherwise Inaccessible Thymine N3 Proton by DNA Distortion or Base Flipping, Nucleic Acids Res. 29 (2001) 4892–4900. http://alum.mit.edu/www/toms/papers/repan3/.

[12] T. D. Schneider, Reading of DNA sequence logos: Prediction of major groove binding by

information theory, Meth. Enzym. 274 (1996) 445–455. http://alum.mit.edu/www/toms/papers/oxyr/.

[13] T. D. Schneider, G. D. Stormo, L. Gold, A. Ehrenfeucht, Information content of binding sites on nucleotide sequences, J. Mol. Biol. 188 (1986) 415–431. http://alum.mit.edu/www/toms/papers/schneider1986/.

[14] C. A. Ball, R. Osuna, K. C. Ferguson, R. C. Johnson, Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*, J. Bacteriol. 174 (1992) 8043–8056.

[15] J. R. Pierce, An Introduction to Information Theory: Symbols, Signals and Noise, Dover Publications, Inc., NY, 2nd edition, 1980.

[16] T. D. Schneider, Information theory primer, Published on the web at http://alum.mit.edu/www/toms/papers/primer/ (2010).

[17] T. D. Schneider, R. M. Stephens, Sequence logos: A new way to display consensus sequences, Nucleic Acids Res. 18 (1990) 6097–6100. http://alum.mit.edu/www/toms/papers/logopaper/.

[18] P. V. Benos, M. L. Bulyk, G. D. Stormo, Additivity in protein-DNA interactions: how good an approximation is it?, Nucleic Acids Res. 30 (2002) 4442–4451.

[19] T. D. Schneider, Evolution of biological information, Nucleic Acids Res. 28 (2000) 2794–2799. http://alum.mit.edu/www/toms/papers/ev/.

[20] T. D. Schneider, Information content of individual genetic sequences, J. Theor. Biol. 189 (1997) 427–441. http://alum.mit.edu/www/toms/papers/ri/.

[21] T. D. Schneider, P. K. Rogan, Computational analysis of nucleic acid information defines binding sites, United States Patent 5867402, 1999. http://alum.mit.edu/www/toms/patent/walker/.

[22] I. Erill, M. C. O'Neill, A reexamination of information theory-based methods for DNA-binding site identification, BMC Bioinformatics 10 (2009) 57.

[23] R. K. Shultzaberger, R. E. Bucheimer, K. E. Rudd, T. D. Schneider, Anatomy of *Escherichia coli* Ribosome Binding Sites, J. Mol. Biol. 313 (2001) 215–228. http://alum.mit.edu/www/toms/papers/flexrbs/.

[24] R. K. Shultzaberger, Z. Chen, K. A. Lewis, T. D. Schneider, Anatomy of *Escherichia coli* σ^{70} promoters, Nucleic Acids Res. 35 (2007) 771–788. http://alum.mit.edu/www/toms/papers/flexprom/.

[25] T. D. Schneider, Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences, Nucleic Acids Res. 25 (1997) 4408–4415. http://alum.mit.edu/www/toms/papers/walker/, erratum: NAR 26(4): 1135, 1998.

[26] M. R. Hemm, B. J. Paul, T. D. Schneider, G. Storz, K. E. Rudd, Small membrane proteins found by comparative genomics and ribosome binding site models, Mol. Microbiol. 70 (2008) 1487–1501. http://alum.mit.edu/www/toms/papers/smallproteins/.

[27] P. K. Rogan, B. M. Faux, T. D. Schneider, Information analysis of human splice site mutations, Human Mutation 12 (1998) 153–171. Erratum in: Hum Mutat 1999;13(1):82. http://alum.mit.edu/www/toms/papers/rfs/.

[28] I. Lyakhov, K. Annangarachari, T. D. Schneider, Discovery of Novel Tumor Suppressor p53 Response Elements Using Information Theory, Nucleic Acids Res. 36 (2008) 3828–3833. http://alum.mit.edu/www/toms/papers/p53/.

[29] T. D. Schneider, Theory of molecular machines. II. Energy dissipation from molecular machines, J. Theor. Biol. 148 (1991) 125–137. http://alum.mit.edu/www/toms/papers/edmm/.

[30] G. W. Castellan, Physical Chemistry, Addison-Wesley Publishing Company, Reading, Mass., second edition, 1971.

[31] R. C. Weast, M. J. Astle, W. H. Beyer, CRC Handbook of Chemistry and Physics, CRC Press, Inc., Boca Raton, Florida, 1988.

[32] P. W. Atkins, The Second Law, W. H. Freeman and Co., N. Y., 1984.

[33] E. T. Jaynes, The evolution of Carnot's principle, in: G. J. Erickson, C. R. Smith (Eds.), Maximum-Entropy and Bayesian Methods in Science and Engineering, volume 1, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988, pp. 267–281. http://bayes.wustl.edu/etj/articles/ccarnot.ps.gz http://bayes.wustl.edu/etj/articles/ccarnot.pdf.

[34] T. D. Schneider, Theory of molecular machines. I. Channel capacity of molecular machines, J. Theor. Biol. 148 (1991) 83–123. http://alum.mit.edu/www/toms/papers/ccmm/.

[35] R. A. Rubin, P. Modrich, *Eco*RI methylase. Physical and catalytic properties of the homogeneous enzyme, J. Biol. Chem. 252 (1977) 7265–7272.

[36] G. M. Clore, A. M. Gronenborn, R. W. Davies, Theoretical aspects of specific and non-specific equilibrium binding of proteins to DNA as studied by the nitrocellulose filter binding assay:

Co-operative and non-co-operative binding to a one-dimensional lattice, J. Mol. Biol. 155 (1982) 447–466.

[37] T. D. Schneider, 70% efficiency of bistate molecular machines explained by information theory, high dimensional geometry and evolutionary convergence, Nucleic Acids Res. 38 (2010) 5995–6006. Doi:10.1093/nar/gkq389, http://alum.mit.edu/www/toms/papers/emmgeo/.

[38] T. D. Schneider, Claude Shannon: Biologist, IEEE Engineering in Medicine and Biology Magazine 25 (2006) 30–33. http://alum.mit.edu/www/toms/papers/shannonbiologist/.

[39] L. Brillouin, Science and Information Theory, Academic Press, Inc., New York, second edition, 1962.

[40] H. B. Callen, Thermodynamics and an Introduction to Thermostatistics, John Wiley & Sons, Ltd., N. Y., second edition, 1985.

[41] I. T. Weber, T. A. Steitz, A model for the non-specific binding of catabolite gene activator protein to DNA, Nucleic Acids Res. 12 (1984) 8475–8487.

[42] I. T. Weber, T. A. Steitz, Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity, Proc. Natl. Acad. Sci. USA 81 (1984) 3973–3977.

[43] Y. Kim, J. C. Grable, R. Love, P. J. Greene, J. M. Rosenberg, Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing, Science 249 (1990) 1307–1309.

[44] T. D. Schneider, I. Lyakhov, D. Needle, PROBE FOR NUCLEIC ACID SEQUENCING AND METHODS OF USE, 2010. US patent claims allowed; European patent number 1960550 granted on 2010 September 15. US patent number 7,871,777 granted on 2011 January 18. http://alum.mit.edu/www/toms/patent/medusa/.

[45] R. K. Shultzaberger, T. D. Schneider, Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX, Nucleic Acids Res. 27 (1999) 882–887. http://alum.mit.edu/www/toms/papers/lrp/.

[46] P. N. Hengen, I. G. Lyakhov, L. E. Stewart, T. D. Schneider, Molecular flip-flops formed by overlapping Fis sites, Nucleic Acids Res. 31 (2003) 6663–6673.

[47] R. K. Shultzaberger, L. R. Roberts, I. G. Lyakhov, I. A. Sidorov, A. G. Stephen, R. J. Fisher, T. D. Schneider, Correlation between binding rate constants and individual information of *E. coli* Fis binding sites, Nucleic Acids Res. 35 (2007) 5275–5283. http://alum.mit.edu/www/toms/papers/fisbc/ http://dx.doi.org/10.1093/nar/gkm471.

[48] Z. Chen, K. A. Lewis, R. K. Shultzaberger, I. G. Lyakhov, M. Zheng, B. Doan, G. Storz, T. D. Schneider, Discovery of Fur binding site clusters in *Escherichia coli* by information theory models, Nucleic Acids Res. 35 (2007) 6762–6777. http://alum.mit.edu/www/toms/papers/fur/.

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

[50] Z. Chen, T. D. Schneider, Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases, Nucleic Acids Res. 33 (2005) 6172–6187. http://alum.mit.edu/www/toms/papers/t7like/.

[51] Z. Chen, T. D. Schneider, Comparative analysis of tandem T7-like promoter containing regions in enterobacterial genomes reveals a novel group of genetic islands, Nucleic Acids Res. 34 (2006) 1133–1147. http://alum.mit.edu/www/toms/papers/t7island/.

[52] R. M. Stephens, T. D. Schneider, Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites, J. Mol. Biol. 228 (1992) 1124–1136. http://alum.mit.edu/www/toms/papers/splice/.

[53] P. K. Rogan, T. D. Schneider, Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites, Human Mutation 6 (1995) 74–76. http://alum.mit.edu/www/toms/papers/colonsplice/.

[54] R. Allikmets, W. W. Wasserman, A. Hutchinson, P. Smallwood, J. Nathans, P. K. Rogan, T. D. Schneider, M. Dean, Organization of the ABCR gene: analysis of promoter and splice junction sequences, Gene 215 (1998) 111–122. http://alum.mit.edu/www/toms/papers/abcr/.

[55] S. R. Svojanovsky, T. D. Schneider, P. K. Rogan, Redundant designations of BRCA1 intron 11 splicing mutation; c. 4216-2A>G; IVS11-2A>G; L78833, 37698, A>G, Human Mutation 16 (2000) 264. http://www3.interscience.wiley.com/cgi-bin/abstract/73001161/START.

[56] 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, K. H. Kraemer, Xeroderma Pigmentosum Group

C splice mutation associated with mutism and hypoglycinemia - A new syndrome?, J. Investigative Dermatology 111 (1998) 791–796.

[57] S. Emmert, T. D. Schneider, S. G. Khan, K. H. Kraemer, The human XPG gene: Gene architecture, alternative splicing and single nucleotide polymorphisms, Nucleic Acids Res. 29 (2001) 1443–1452.

[58] S. G. Khan, V. Muniz-Medina, T. Shahlavi, C. C. Baker, H. Inui, T. Ueda, S. Emmert, T. D. Schneider, K. H. Kraemer, The human *XPC* DNA repair gene: arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice acceptor site on alternative splicing and function, Nucleic Acids Res. 30 (2002) 3624–3631.

[59] S. G. Khan, A. Metin, E. Gozukara, H. Inui, T. Shahlavi, V. Muniz-Medina, C. C. Baker, T. Ueda, J. R. Aiken, T. D. Schneider, K. H. Kraemer, Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk, Hum Mol Genet 13 (2004) 343–352. http://hmg.oxfordjournals.org/cgi/content/full/13/3/343.

[60] H. Inui, K. S. Oh, C. Nadem, T. Ueda, S. G. Khan, A. Metin, E. Gozukara, S. Emmert, H. Slor, D. B. Busch, C. C. Baker, J. J. Digiovanna, D. Tamura, C. S. Seitz, A. Gratchev, W. H. Wu, K. Y. Chung, H. J. Chung, E. Azizi, R. Woodgate, T. D. Schneider, K. H. Kraemer, Xeroderma Pigmentosum-Variant Patients from America, Europe, and Asia, J Invest Dermatol 128 (2008) 2055–2068. http://alum.mit.edu/www/toms/papers/xpv/.

[61] E. Bindewald, T. D. Schneider, B. A. Shapiro, CorreLogo: An online server for 3D sequence logos of RNA and DNA alignments, Nucleic Acids Res. 34 (2006) w405–w411. http://alum.mit.edu/www/toms/papers/correlogo/.

[62] G. D. Stormo, T. D. Schneider, L. Gold, Quantitative analysis of the relationship between nucleotide sequence and functional activity, Nucleic Acids Res. 14 (1986) 6661–6679.

[63] D. N. Arvidson, P. Youderian, T. D. Schneider, G. D. Stormo, Automated kinetic assay of β-galactosidase activity, BioTechniques 11 (1991) 733–738.

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

[65] D. Barrick, K. Villanueba, J. Childs, R. Kalil, T. D. Schneider, C. E. Lawrence, L. Gold, G. D. Stormo, Quantitative analysis of ribosome binding sites in *E. coli.*, Nucleic Acids Res. 22 (1994) 1287–1295.

[66] N. Kannan, T. D. Schneider, S. Vishveshwara, Logos for amino acid preferences in different backbone packing density regions of protein structural classes, Acta Crystallographica Section D 56 (2000) 1156–1165.

[67] P. K. Rogan, J. J. Salvo, R. M. Stephens, T. D. Schneider, Visual display of sequence conservation as an aid to taxonomic classification using PCR amplification, in: C. A. Pickover (Ed.), Visualizing Biological Information, World Scientific, Singapore, 1995, pp. 21–32.

[68] P. W. Tooley, J. J. Salvo, T. D. Schneider, P. K. Rogan, Phylogenetic inference based on information theory-based PCR amplification, Journal of Phytopathology 146 (1998) 427–430.