Use of Information Theory in Molecular Biology

Thomas D. Schneider
Laboratory of Mathematical Biology
National Cancer Institute
Frederick Cancer Research and Development Center
P. O. Box B, Frederick, MD 21702-1201*

Abstract

Applying Shannon's information theory to examples from molecular biology reveals a new world for exploration by physicists interested in the limits of computation.

1 Introduction

What are the limits to computation? One way to approach this question is to investigate the properties of molecular systems that have already been in existence for millions of years. Biochemical systems perform many subtle chemical reactions. Genes are turned on or off under the control of many physical stimuli, such as oxygen, light, and temperature. These control systems function directly at the molecular level by blocking or enhancing the copying of DNA into RNA or the RNA into protein. Because DNA and RNA are long strings of only 4 letters, specific patterns are required to position the controlling machinery in the right places. We have been investigating these patterns using the tools provided by Shannon's information theory [1, 2, 3]. The approach has been fruitful. Because Shannon's measure corresponds directly to the entropy measure of physics when one is discussing molecular systems [4], we have also forged a direct connection between the information processing of molecular systems and the physics of those systems.

2 Physics of Computation at a Molecular Binding Site

It is not possible to make sensible measures of molecular phenomena without knowing exactly what one is doing and why. A careful definition of "information" is therefore absolutely necessary to avoid many difficulties and confusions in the field of the physics of computation. There are a number of these pitfalls.

Suppose that I am sending you a stream of characters from an alphabet of 4 possible symbols: A, C, G and T. *Before* you receive each character, you are uncertain as to which character you will get. Shannon defined *uncertainty* as:

$$H = -\sum_{i=1}^{M} p_i \log_2 p_i \quad \text{(bits per symbol)} \qquad (1)$$

where p_i is the probability of the i$^{\text{th}}$ symbol out of M symbols [1, 2, 3, 5]. This equation simplifies to $H_{eq} = \log_2 M$ when the symbols are equally likely. H is a state function, meaning that it depends entirely on the state of a system at a certain time or under specified conditions.

To avoid the first pitfall, we must realize that uncertainty is rarely the same as information. Rather, the information you gain is the amount that your uncertainty decreases as a result of receiving the characters. Following Shannon, we define the information to be the drop in uncertainty from *before* receiving to *after* reception:

$$R = H_{before} - H_{after} \quad \text{(bits per symbol).} \qquad (2)$$

Thus information is always a measure of the change of state of a system. Consider the situation *after* you have received some characters. If you were completely successful, you would know exactly what each character was. In this case $H_{after} = 0$, so $R = H_{before}$. Unfortunately people often assume this is true all the time. But consider the case where the communication is noisy, so that you can't tell A from T or G from C. In this case $H_{before} = 2$ bits per symbol, $H_{after} = 1$ bit per symbol so $R = 1$ bit. Because of the noise, only part of the information gets through. This makes it

*(301) 846-5581 (-5532 for messages), (301) 846-5598 fax, Internet address: toms@ncifcrf.gov.

clear that neither uncertainty is the same as the information you receive.

To see how this is important, let's take a mental ride along with a genetic control molecule. In our bodies there are odd little machines called spliceosomes. After DNA is copied into RNA in the nucleus of the cell, the spliceosomes remove pieces of the RNA and splice the ends back together. Nobody really knows why they do this, but they do it for almost every gene in the body. The RNA then moves out to the cytoplasm where it is translated into proteins. The translational machinery steps 3 letters ("bases") after each amino acid has been inserted into a growing protein. If there is even one base missing, the reading frame is disrupted and the protein is made incorrectly. Since proteins are made by reading from spliced RNA, it is important that the splicing be precise.

The spliceosome looks for two patterns in the RNA. The one nearer to the start of the RNA is called a donor site, the one nearer to the end is called the acceptor site. Spliceosomes bind to both of these and remove the region in between (called the "intron") leaving the rest for translation (the "exon"). Here are some examples of donor sites:

```
                                      +
                     -------- +++++++++1
                     8765432101234567890

                     ...................
HUMA1ACMB    458   1 TGAGGCAGGTAATCCATGA
HUMA1AR1     814   2 CTTTCACGGTAAGGTAGCA
HUMA1AR2     679   3 GACATAAGGTGATTTCCAG
HUMA1AR2    2241   4 TCTCCAAGGTGAGGTCACC
HUMA1ATP    2002   5 AGTGAATCGTAAGTATGCC
HUMA1ATP    7962   6 CTTTAAAGGTAAGGTTGCT
HUMA1ATP    9683   7 GACAGAAGGTGATTCCCCA
HUMA1ATP   11087   8 TCTCCAAGGTGAGATCACC
HUMA1GLY2   1723   9 TGGACCGGGTGAGTGCCTG
HUMA1GLY2   2281  10 CAGACCCGGTGAGAGCCCC
HUMA1GLY2   2570  11 CAGATACGGTGAGGGCCAG
HUMA1GLY2   3382  12 TTTCTATGGTAGGCATGCT
HUMA1GLY2   3636  13 GGAAAAAGGTAAACGCAAG
HUMA2PIG1   1105  14 CAGGGAGGGTAGCCCTCTC
HUMA2PIG2    551  15 GCTCCGTGGTGAGCTGGTG
HUMA2PIG2    695  16 GCCGGCAGGTACTGGGGAG
HUMA2PIG2    878  17 GCAACCAGGTACAACCAGG
HUMA2PIG2   1390  18 GGCACTAGGTACCCTGGCA
HUMA2PIG3    280  19 GCAGAAAGGTAGGCGCTGA
HUMA2PIG3    630  20 CTTCCAGGGTGCGCTCCTC
```

HUMA1ACMB is the name assigned to some of the sequence in the human alpha-1-antichymotrypsin gene. At position 458 of this gene there is a donor site. We assign this the coordinate zero, and show bases −8 through +10 (the coordinates are written vertically).

Notice that position +1 is always a 'T'. *Before* the spliceosome has found any donor sites, it faces all 4 bases, so it's uncertainty is 2 bits per base. *After* the spliceosome has found a donor site, it has an uncertainty of 0 bits at position +1, so the information there is the difference, 2 bits. Now look at position +2. Here there are 11 A's and 9 G's, so the uncertainty *after* binding is about 1 bit. Position 1 contains about $2 - 1 = 1$ bits of sequence conservation (as a biologist would call it).

If we call H_{after} the information (and thus fall into a pit), then we would incorrectly say that position +1 has *no* information [6, 7]. But clearly there must be information there for the spliceosome to use. Further, positions outside the binding site should not contain information for the spliceosome. In those regions all 4 bases appear whether or not we are at a binding site, so $H_{before} = H_{after} = 2$ bits. If we called this the information, we would have to say that there was more information for the spliceosome outside the binding sites, which is absurd. By taking information as a state function, the *decrease of uncertainty*, we avoid this pitfall and get the correct answer that there are $R = 0$ bits of information outside the binding site.

We used this method to study the patterns at donor and acceptor sites [8]. In our study we had nearly 1800 donor and acceptor sites, so the statistical analysis was very accurate. We calculated the information R at every position across all the sites and then displayed the results by the method shown in Fig. 1. These "sequence logos" consist of several stacks of letters. The height of the stack is R for that position. The vertical bar is 2 bits high. The letters are drawn at a size proportional to their frequency in the original data set, and they are sorted so that the most frequent one is on top. Sequence logos are useful for studying binding sites because we can get a sense for what is important and what is not just by looking at them. The list of donor sites given above is almost indigestible, but the donor site logo instantly shows patterns in almost 100 times as much data. The vertical bar is between −1 and 0, so the high conservation at positions 0 and +1 is obvious, as is the partial conservation at +2. In this figure the region between the two vertical bars is removed by the spliceosome.

Now we ask: what is the total amount of sequence conservation at each of these sites? As far as we could tell, there is no strong correlation between the base at any position in the sites and any of its neighbors. Because they are independent, we can add the infor-

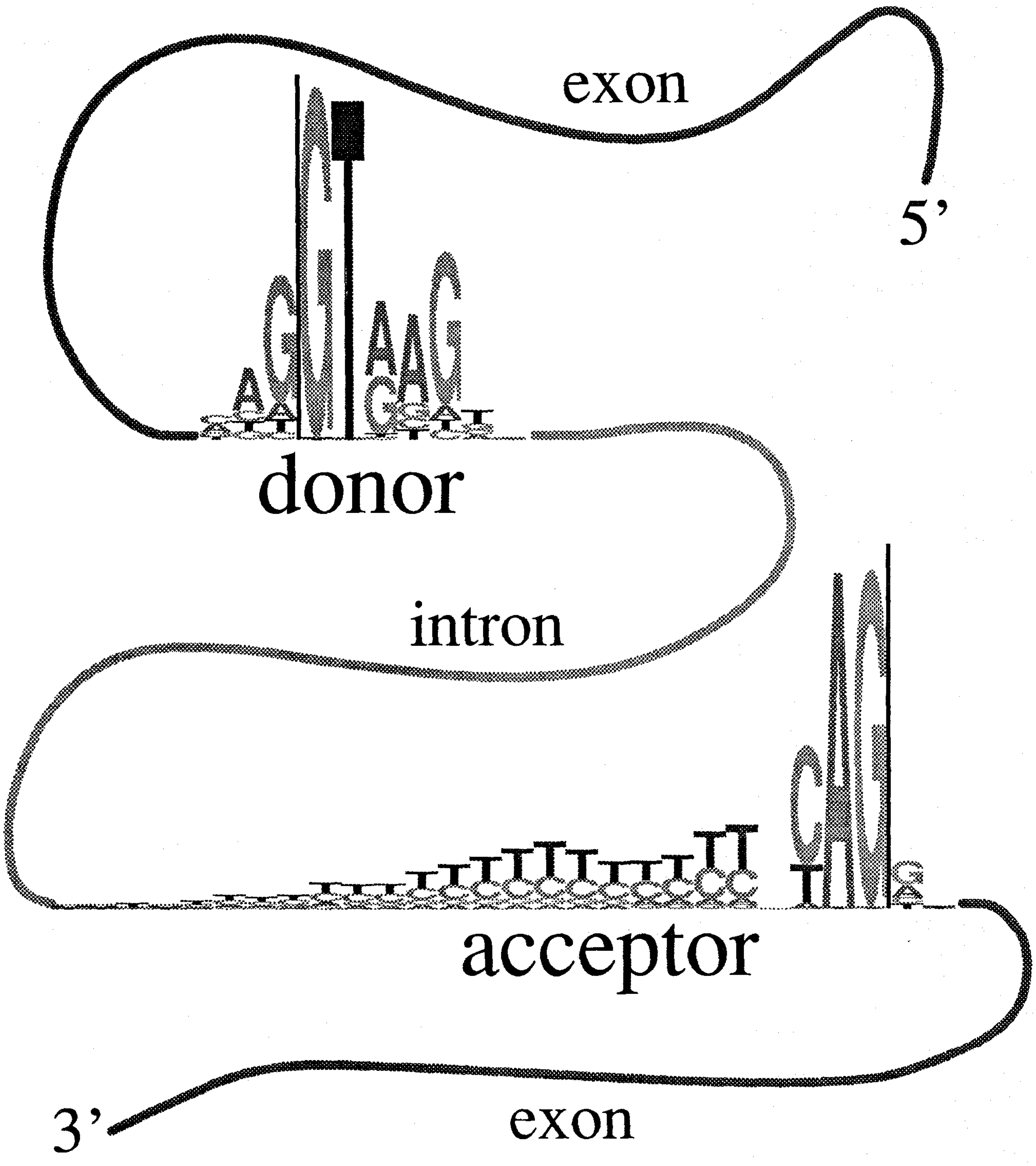

Fig. 1. Sequence logos for human donor and acceptor sites.

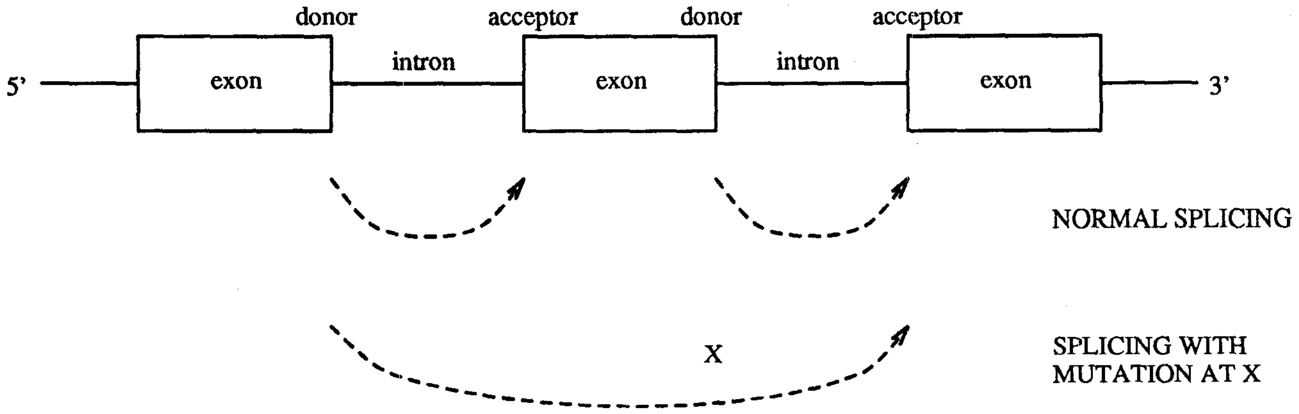

Fig. 2. The phenomenon of exon skipping.

mation from all the positions together. Donor sites have roughly 8 bits of information, while acceptors have about 9 bits.

These numbers are meaningless without something to compare them to. Fortunately, we can determine another number for comparison. To do this, we maintain the same *before-after* state function difference we have been using all along, but view it from a different angle. *Before* binding to either site, the spliceosome is somewhere on the RNA. Its *positional uncertainty* is easy to calculate. Assuming that the spliceosome is equally likely to be at all positions, this is $H_{before} = \log_2$ (number of positions before binding). *After* binding, the spliceosome is at any one of the donor or acceptor sites. It still is "uncertain" as to which donor or acceptor site it is at, but that doesn't matter: having found the site, it's job is to start splicing! Therefore, after having found a site its uncertainty is: $H_{after} = \log_2$ (number of sites after binding). The decrease in the positional uncertainty is the information needed to find the sites, just as in equation (2).

We calculated this for splicing and found that the spliceosome needs about 9 bits to locate its sites. So the acceptors have enough information, but the donors are too low by 1 bit. With the amount of data we have, this is a highly significant result. In studies of other genetic control systems, we have often found that the sequence conservation is close to the amount predicted [9, 10]. Essentially, *the addressing information stored in the RNA sequence is just sufficient for the addresses to be located in the RNA.* So the acceptor sites fit the theory, but donors are anomalous. As in regular physics, we have two options: to reject the theory or to explain the anomaly within the theory. A simple explanation was suggested by R. Michael Stephens. Mike proposed that the spliceosome binds to the acceptor sites first. Having done this *the problem has been (literally!) cut in half*—so the donor bits need 1 bit less pattern to be found, as observed.

This prediction is born out by two kinds of experiment. The first were direct measurements with spliceosomes which demonstrate that the acceptor sites are bound first. The second is the observation of a phenomenon called "exon skipping". Genes have many introns and exons. Consider the situation shown in Fig. 2, in which there are two introns interspersed between three exons. The introns are normally removed, joining the exons. In a simple model, we would say that each donor joins to the next acceptor. But if a mutation is made which destroys the donor for the second intron (marked by an X on the figure), *the entire middle exon is skipped!* How could the loss of a donor to the right of the first intron affect splicing there? A solution to this puzzle is the "exon definition" model [11, 12], which proposes that the spliceosome first binds to the acceptor and then searches downstream ("to the right") for the next donor. This defines the exon region. Then two spliceosomes get together to chop out the intron between them. (What happens on the ends of the RNA is current research.) If the donor site doesn't exist, the entire exon is missed (never "defined") so its adjacent neighbors are joined together without it. This model fits the information theory prediction that the acceptor is bound first.

3 More Pitfalls in the Molecular Physics of Computation

For the spliceosome to bind to the donor or acceptor sites, the molecules must dissipate energy into the surrounding solution. If they did not do this they would not stick to the sites. Thus gaining information (as defined above) requires the dissipation of energy away from the system. At the same time, the positional entropy of the spliceosome on the RNA has to be reduced. This is "paid for" by the dissipation of energy away from the spliceosome/RNA complex.

In the literature of computational physics we see statements such as "throwing away information requires dissipation" [13]. This is exactly the reverse of what the spliceosome and many other genetic systems are known to be doing! They *gain* information by dissipating heat to the surroundings!

The difficulty appears to stem from the idea that an erasing operation is somehow different from storing information in general. In general, we store a pattern of 0's and 1's in a memory device. In erasure, we store a pattern of pure 0's in the device. That is, *erasure is merely a specialized form of recording.* When the spliceosome changes from the *before* state to the *after* state it selects certain positions of the RNA from all the other possibilities. Thus this little "molecular machine" makes a simple choice. Having made this choice, there is some information stored in the location of the spliceosome. If we send a pulse of heat into the solution, the spliceosome can fall off the RNA and lose the information. Thus information loss is associated with heat *absorption.*

Erasure is also a two step operation, with energy first going into the device and then flowing back out as heat. Suppose I have some coins on a table. Each coin can store 1 bit of information. Now suppose I put energy *into* the coins by flipping them. Once they are flipping, they have lost information! Now I allow some of that energy to drain out, and each coin goes to one or the other face (which I perhaps have chosen). The coins have regained information by losing energy. A system must lose energy (dissipate) for *that system* to gain information. If we don't allow the energy to dissipate away, the coin keeps flipping and never comes to store information. In summary, information is lost when you put energy *into* a system, not when you extract energy from it. When energy dissipates away from a system, that system can *gain* information. (It might not gain information. If I randomly dump coins on a table, their energy dissipates but as a whole they have not gained information.)

Despite prominent statements to the contrary [14], communication, computation and measurements are essentially the same in that in each case choices are being made by a device. Shannon's theory is not about the channel, it is about what happens at the receiver. The receiver must be prepared in a state which will accept any symbol from the channel. Like bowling pins ready to be toppled, a flipping coin, or a finger hovering over a keyboard, the receiver must be prepared in a high energy state. Selection of one of the symbols coincides with dissipation of the energy. Likewise, in computation, an AND gate must be prepared for either result at its output. The selection dissipates energy. (For how this applies to electrical circuits, see the discussion in the book by Mead and Conway [15], chapter 9.) Finally, a measuring device tells us one of several possible states of the system being measured, so it too makes selections. We can now add to these cases the molecular machines, such as the spliceosome.

The relationship between the energy dissipated and the information gained is widely known but not clearly recognized. It is:

$$k_{\mathrm{B}}T\ln(2) \leq -q/R \quad \text{joules per bit} \qquad (3)$$

where k_{B} is Boltzmann's constant, T is the absolute temperature, $-q$ is the heat dissipation to the surroundings and R is the information gain. The literature contains many statements about this, but they often miss three fundamentally important points. First, this is an ***equation***, it is not some disconnected mathematical phrase. Second, the equation is associated with ***units***: joules per bit. A bit is the precise choice between two equally likely possibilities. There is nothing vague about this as has been suggested [14]: the factor of ln(2) assures that the result is given in bits. If the factor were ln(10) the units of information would be ***digits***! Finally, *equation (3) is a novel form of the Second Law of Thermodynamics* (see [4] for two derivations). We cannot dismiss it lightly.

The Second Law, as embodied above, says that for a system to gain 1 bit of information, by selecting between two equally likely states, *at least* $k_{\mathrm{B}}T\ln(2)$ joules of energy must be dissipated to the surroundings under isothermal conditions. There is no way to escape this without breaking the Second Law.

So how are we to understand the recent work which proposes that one can do enormous amounts of computation for a small energy dissipation? (For reviews see [14, 13].) Part of the solution was understood by Feynman [16], who pointed out that what really counts is the *output* of the computation. This is where the selections are made. If one has a problem to which the answer is 40 bits, then one must dissipate $40 \times k_{\mathrm{B}}T\ln(2)$ joules to capture the solution. The Second Law dictates that this is unavoidable. We must add to this the dissipation required to set the inputs to the computer. What happens in between ***doesn't matter*** for the simple reason that no commitment to selections need be made except for the inputs and the outputs.

Why is it necessary to commit to dissipation at the input and output? For the input, we must commit ourselves to setting the states so that the computer has distinct values to work on. If they are not set, the computer cannot solve our problem. For the output, the computer must set the display so that humans or other devices can read it later. The computation in between need not take any dissipation. This is related to the interchangeability of Boolean circuits. For example, NOT(A) can be expanded to NOT(NOT(NOT(A))). We can make a circuit more complex by adding lots of extra unnecessary steps but as long as we run the most reduced version, we can claim to have arbitrarily large amounts of computation for low energetic cost because the extra gates don't really cost anything. How much can the circuit be reduced? Boolean algebra sets a lower limit for each task. However, if we use other kinds of logic, steps can be joined together into one single concerted step: the choice of the right "answer" from all other possibilities. This is exactly what a spliceosome does. Many other examples of this "dissipationless computation" can be found in biology, as described elsewhere [17].

4 Internet News Group and Archive

We have formed an internet news group for discussing the use of information theory in biology; people interested in the limits of computation are welcome to join us. The newsgroup is called "bionet.infotheory". A Frequently Asked Questions (FAQ) sheet is available by anonymous ftp from ncifcrf.gov in directory pub/delila. If you have net access but not net news access, you can receive the newsgroup by electronic mail. Subscribe by sending email to biosci@net.bio.net (North or South America or the Pacific Rim) or biosci@daresbury.ac.uk (Europe, Africa, and Central Asia). Questions may be addressed by email to toms@ncifcrf.gov. Programs to create sequence logos may also be found in the archive.

5 Abstracts of Papers

A series of recent papers demonstrate the usefulness of Shannon's information theory for molecular biologists. Rather than describe them in detail, the abstracts from several are given below. Please refer to the original papers for more details. Copies of the papers are available upon request.

6 Level 0 Molecular Machine Theory

Papers [9, 10, 18]: Repressors, polymerases, ribosomes and other macromolecules bind to specific nucleic acid sequences. They can find a binding site only if the sequence has a recognizable pattern. We define a measure of the information ($R_{sequence}$) in the sequence patterns at binding sites. It allows one to investigate how information is distributed across the sites and to compare one site to another. Given the frequency of sites in the genome, one can also calculate the amount of information ($R_{frequency}$) that is required to locate them. Several *Escherichia coli* binding sites were analyzed using these two independent empirical measurements.

The two amounts of information are similar for most of the sites we analyzed. In contrast, bacteriophage T7 RNA polymerase binding sites contain about twice as much information as is necessary for recognition by the T7 polymerase, suggesting that a second protein may bind at T7 promoters. The extra information can be accounted for by a strong symmetry element found at the T7 promoters. This element may be an operator. If this model is correct, these promoters and operators do not share much information. The comparisons between $R_{sequence}$ and $R_{frequency}$ suggest that the information at binding sites is just sufficient for the sites to be distinguished from the rest of the genome.

Paper [19]: In our previous analysis of the information at binding sites on nucleic acids, we found that most of the sites examined contain the amount of information expected from their frequency in the genome. The sequences at bacteriophage T7 promoters are an exception, because they are far more conserved (35 bits of informati›n content) than should be necessary to distinguish them from the background of the *Escherichia coli* genome (17 bits). To determine the information actually used by the T7 RNA polymerase, promoters were chemically synthesized with many variations and those that function well in an *in vivo* assay were sequenced. Our analysis shows that the polymerase uses 18 bits of information, so the sequences at phage genomic promoters have significantly more information than the polymerase needs. The excess may represent the binding site of another protein.

Paper [20]: The 12 *incD* repeats in the F plasmid each contain about 60 bits of information, which is three times the amount of conservation that a single protein would need to distinguish the repeats from the rest of the *Escherichia coli* genome. This is the first reported discovery of a case of threefold excess information and it implies that at least three proteins bind independently to the repeats. In support of this observation, other workers have shown that three polypeptides bind to this region, but only one, SopB, is known to bind independently of other factors. Identification of the other two proteins should help us to understand the mechanism of plasmid partitioning during cell division.

Paper [8]: An information analysis of the 5′ (donor) and 3′ (acceptor) sequences spanning the ends of nearly 1800 human introns has provided evidence for structural features of splice sites that bear upon spliceosome evolution and function:

(1) 82% of the sequence information (*i.e.* sequence conservation) at donor junctions and 97% of the sequence information at acceptor junctions is confined to the introns, allowing codon choices throughout exons to be largely unrestricted. The distribution of information at intron-exon junctions is also described in detail and compared with footprints.

(2) Acceptor sites are found to possess enough information to be located in the transcribed portion of the human genome, whereas donor sites possess about one bit less than the information needed to locate them independently. This difference suggests that acceptor sites are located first in humans and, having been located, reduce by a factor of two the number of alternative sites available as donors. Direct experimental evidence exists to support this conclusion.

(3) The sequences of donor and acceptor splice sites exhibit a striking similarity. This suggests that the

two junctions derive from a common ancestor and that during evolution the information of both sites shifted onto the intron. If so, the protein and RNA components which are found in contemporary spliceosomes, and which are responsible for recognizing donor and acceptor sequences, should also be related. This conclusion is supported by the common structures found in different parts of the spliceosome.

7 Level 1 Molecular Machine Theory

Paper [21]: Like macroscopic machines, molecular-sized machines are limited by their material components, their design, and their use of power. One of these limits is the maximum number of states that a machine can choose from. The logarithm to the base 2 of the number of states is defined to be the number of bits of information that the machine could "gain" during its operation. The maximum possible information gain is a function of the energy that a molecular machine dissipates into the surrounding medium (P_y), the thermal noise energy which disturbs the machine (N_y) and the number of independently moving parts involved in the operation (d_{space}): $C_y = d_{space} \log_2(\frac{P_y+N_y}{N_y})$ bits per operation. This "machine capacity" is closely related to Shannon's channel capacity for communications systems.

An important theorem that Shannon proved for communication channels also applies to molecular machines. With regard to molecular machines, the theorem states that if the amount of information which a machine gains is less than or equal to C_y, then the error rate (frequency of failure) can be made arbitrarily small by using a sufficiently complex coding of the molecular machine's operation. Thus, the capacity of a molecular machine is sharply limited by the dissipation and the thermal noise, but the machine failure rate can be reduced to whatever low level may be required for the organism to survive.

8 Level 2 Molecular Machine Theory

Paper [4]: Single molecules perform a variety of tasks in cells, from replicating, controlling and translating the genetic material to sensing the outside environment. These operations all require that specific actions take place. In a sense, each molecule must make tiny decisions. To make a decision, each "molecular machine" must dissipate an energy P_y in the presense of thermal noise N_y. The number of binary decisions that can be made by a machine which has d_{space} independently moving parts is the "machine capacity" $C_y = d_{space} \log_2(\frac{P_y+N_y}{N_y})$. This formula is closely related to Shannon's channel capacity for communications systems, $C = W \log_2(\frac{P+N}{N})$.

This paper shows that the minimum amount of energy that a molecular machine must dissipate in order to gain one bit of information is $\mathcal{E}_{min} = k_{\mathrm{B}}T\ln(2)$ joules per bit. This equation is derived in two distinct ways. The first derivation shows that this equation is a restatement of the Second Law of Thermodynamics. The second derivation begins with the machine capacity formula, which shows that the machine capacity is also related to the Second Law of Thermodynamics.

One of Shannon's theorems for communications channels shows that as long as the channel capacity is not exceeded, the error rate may be made as small as desired by a sufficiently involved coding. This result also applies to the dissipation formula for molecular machines. So there is a *precise* upper bound on the number of choices a molecular machine can make for a given amount of energy loss. This result will be important for the design and construction of molecular computers.

I thank Paul N. Hengen, Stacy L. Bartram and Denise Rubens for their comments on this manuscript.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

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

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

[4] T. D. Schneider. Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.*, 148:125–137, 1991.

[5] R. Bharath. Information theory. *Byte*, 12(14):291–298, December 1987.

[6] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, 220:49–65, 1991.

[7] A. V. Lukashin, J. Engelbrecht, and S. Brunak. Multiple alignment using simulated annealing: branch point definition in human mRNA splicing. *Nucl. Acids Res.*, 20:2511–2516, 1992.

[8] 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.*, in press, 1992.

[9] 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.

[10] T. D. Schneider. Information and entropy of patterns in genetic switches. In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering*, volume 2, pages 147–154, Dordrecht, The Netherlands, 1988. Kluwer Academic Publishers.

[11] B. L. Robberson, G. J. Cote, and S. M. Berget. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.*, 10:84–94, 1990.

[12] M. Talerico and S. M. Berget. Effect of 5′ splice site mutations on splicing of the preceding intron. *Mol. Cell. Biol.*, 10:6299–6305, 1990.

[13] R. Landauer. Information is physical. *Physics Today*, 44(5):23–29, May 1991.

[14] R. Landauer. Dissipation and noise immunity in computation and communication. *Nature*, 335:779–784, 1988.

[15] C. Mead and L. Conway. *Introduction to VLSI Systems*. Addison-Wesley Publishing Company, Reading, Mass., 1980.

[16] R. P. Feynman. Tiny computers obeying quantum mechanical laws. In N. Metropolis, D. M. Kerr, and G. Rota, editors, *New Directions in Physics: The Los Alamos 40th Anniversary Volume*, pages 7–25, Boston, 1987. Academic Press, Inc.

[17] T. D. Schneider. Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology*, 1992. submitted.

[18] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res.*, 18:6097–6100, 1990.

[19] T. D. Schneider and G. D. Stormo. Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucl. Acids Res.*, 17:659–674, 1989.

[20] N. D. Herman and T. D. Schneider. High information conservation implies that at least three proteins bind independently to F plasmid *incD* repeats. *J. Bact.*, 174:3558–3560, 1992.

[21] T. D. Schneider. Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.*, 148:83–123, 1991. (Note: The figures were printed out of order! Fig. 1 is on p. 97).