Sequence Logos: A Powerful, Yet Simple, Tool

version = 3.03 of hawaii.tex 2003 Feb 12

Mark C. Shaner *†‡ Ian M. Blair*§ and Thomas D. Schneider*¶

Abstract

A powerful technique for analyzing DNA and protein sequences, the sequence logo, is now available to researchers. This method has advantages over the conventional method of creating a consensus. For example, a logo of DNA shows all the bases found in the binding site of a protein, indicates how much each base is conserved, and depicts the location of the conservation with respect to the major and minor grooves of the DNA. It is the intent of this paper to explain sequence logos, to show their usefulness for biological and linguistic research, and to encourage their use by others.

1 What is a Sequence Logo?

The object of a sequence logo (Fig. 1) [1] is to visualize the information contained in a set of DNA, RNA, or protein sequences by examining the order and frequency of the chemical subunits which make up the sequences. The name "sequence logo" comes from the fact that a set of sequences is being represented as a single graphic which contains one or more separate elements (that's the definition of the word "logo" [4]). For example, when an economist wants to show a trend in a market, he creates a graph of the conditions of the market to make the trend apparent with just a quick glance. The sequence logo functions in a similar manner by graphically representing the conservation ("information content") of a set of sequences in a clear, concise and mathematically sound manner. Sequence logos are generated by programs which look at the sequences and analyze them using the information theory developed by Claude Shannon [5, 6, 7]. The process of generating a sequence logo is somewhat similar to that of creating a consensus sequence, but unlike a consensus, subtle features of the data are retained. To understand sequence logos one must first understand the concepts of information and uncertainty.

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

†Frederick High School, 650 Carroll Parkway, Frederick, MD 21701.

‡Present address: University of Maryland, College Park, MD 20742.

§Present address: University of Pittsburgh, Pittsburgh, PA 15213.

¶(301) 846-5581 (-5532 for messages), (301) 846-5598 fax, Internet address: toms@ncifcrf.gov, to whom correspondence should be addressed.

2 Information and Uncertainty

We said that a sequence logo is a graphic representation of the amount of information to be found in a set of DNA, RNA, or protein sequences. But what is information? To understand that, we first must understand the technical meaning of uncertainty. Imagine yourself seated in front of a television screen. We are going to test your psychic powers, except that you haven't told us that you don't have any. Still, we will test your powers by flashing random symbols on the screen to see if you can guess them before we project them. Let's start with the alphabet. You are very uncertain as to which symbol will appear next. After all, there are twenty-six of them, and you aren't psychic. You are only able to guess correctly about once every twenty-six times a new symbol is flashed on the screen. Now, say we throw out all of the letters except "a," "c," "g," and "t". We wouldn't tell you though, as far as you know, there are still twenty-six possibilities. But you're a smart person, so eventually you start to realize that only those four letters are being displayed, and your uncertainty decreases. This makes sense because now you are certain that it will be either an "a," "c," "g," or "t," and not one of the other twenty-two letters. Next, we bring in a new person, and we consistently flash the same letter on the screen, an "a" for instance. Their uncertainty would be zero; they'd know that the symbol was always going to be an "a", and if we started flashing other letters, like "c," "g," and "t" their uncertainty would suddenly increase. So, we see that whenever information is gained, for example when you determined that we stopped using all twenty-six letters, the level of uncertainty decreases. Likewise when information is lost, as in the case where the person no longer knew

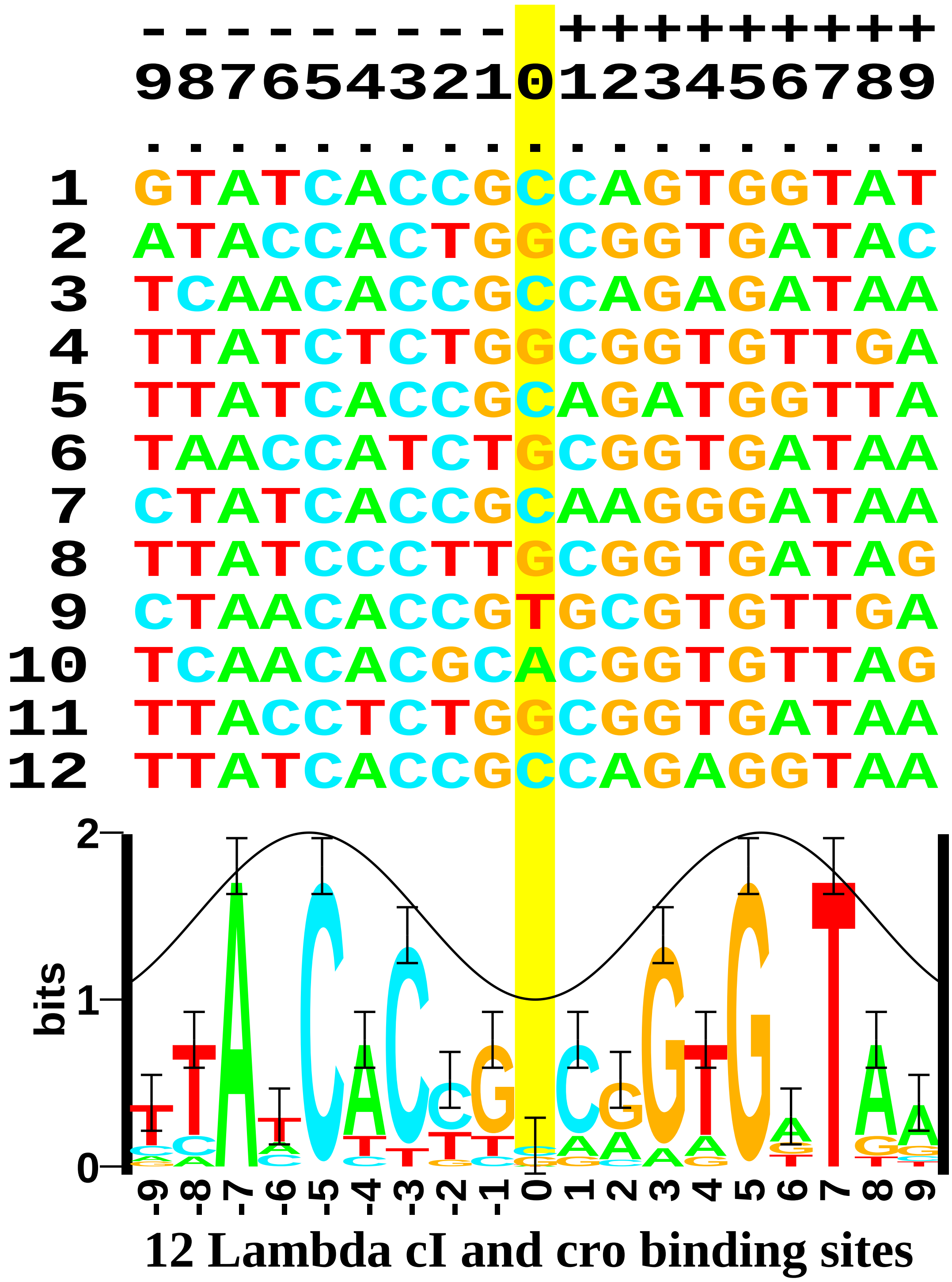

Figure 1: Some aligned sequences and their sequence logo.
At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [2]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [3].

that the symbol would always be an “a,” uncertainty increases. So how do you measure information and uncertainty?

3 The Magnificent Bit

The most common unit for measuring information and uncertainty is called the bit. A bit is the amount of information necessary to choose between two equally probable choices. To demonstrate this, let's play a game of twenty questions. Say that I put 1,048,576 identical boxes in a straight line and put a blue ball into one of them at random. I'll let you have twenty yes-or-no questions to find the ball, and if you do, I'll give you the ball (it's a special ball). Are you going to play? Of course you are, it's a *really special* ball. All of the boxes look identical, so what are you going to do? You can hope you get lucky, and guess randomly at twenty boxes, but then you'd only win that really special ball about once every 52,429 times you play. However, there is a better way! And best of all, you'll always win! Can you guess the best method? Well, here it is: the best questions to ask are the ones which eliminate the largest possible set of choices NO MATTER WHAT THE ANSWER IS. In other words, whether I say “yes” or “no”, you can eliminate the same number of choices. That number is one-half of the total number of choices. Therefore, your first question would be “Is the ball in the first half of the line of boxes?” My answer would be “Yes,” and now you only have 524,288 boxes left to chose from, and you still have 19 questions. Although this seems like a lot of boxes to choose from, notice that 1,048,576 (the number you started with) is equal to 2^{20}. So if you keep dividing the set of boxes in half, your twentieth question will determine which of the two remaining boxes contains that really special blue ball. The answer to “Where is the blue ball?” is given by 20 bits of information.

This method of searching is by no means a new concept (computer programmers call it a binary search and have been using it for quite a long time), and it has been used in biology before, but only recently has someone used it for studying binding sites [8, 9, 10, 11, 12, 13]. It takes two bits of information to determine which of four equiprobable DNA bases occurs at a certain location. The first 1 bit decision divides the set in half, leaving only 2 choices. The second 1 bit decision determines which base is at the current location.

4 The Nitty Gritty Bit

Now that you have a basic understanding of information theory, we'll give you a simplified mathematical derivation of the formulas.[1]

First, we've seen that to make a choice between two equiprobable symbols requires one bit of information, while four symbols require two bits. Going one step further, you should see that eight symbols require three bits to pick one of them. From this we see the relationship:

$$2^b = M \tag{1}$$

or

$$b = \log_2 M \tag{2}$$

where b is the number of bits required to determine which of M different symbols is the current one. This formula assumes that all of the symbols are equally likely, but what if they are not?

First, we rearrange equation (2). Since $(M^{-1})^{-1} = M$, we can substitute into (2) yielding the equation $b = \log_2[(M^{-1})^{-1}]$. By pulling out the exponent (according to the rule that $\log_a b^n = n \log_a b$) and by using the rule that $M^{-1} = \frac{1}{M}$ we obtain the equation:

$$b = -\log_2 \left(\frac{1}{M}\right). \tag{3}$$

For M equally probable symbols, $\frac{1}{M}$ is the probability of each symbol P, so:

$$b = -\log_2(P). \tag{4}$$

While information theory [5, 6, 7] is based on probabilities, sequence data can only be used to generate frequencies, so we have to use them instead, and we have to be a little bit careful since they differ. *Probability* is the chance that something (in this case a particular symbol) will occur at any specific location in an infinite set; but *frequency* is the number of times something occurs in a finite set divided by the number of elements in the set (our sequence). That is, probability is from an entire population, while frequency is from a sample of that population. So, the larger the sample one is working with, the closer the frequency will be to the probability, but the frequency will only be an approximation of the probability. Now, suppose that we have M different symbols (like a, b, c, etc.) and that each has a *different* probability P_i where i

[1] This section is necessary for a thorough understanding of how sequence logos work, but is not necessary for understanding their function. However, if your mathematical background includes the calculus, we would recommend that you wade through all of this.

denotes which symbol is being referred to. From these, we create a sequence that is N symbols long (for example, the sequence aabcba has $N = 6$ and $M = 3$). When there is a large sample size, the frequency of a symbol approximates its probability, so we substitute frequency for probability in equation (4). Because of this substitution, it is important to use a correction for small sample sizes [8], but we won't discuss that here. Finally, we define probability so that the sum of the probabilities is 1 (100%). In other words no symbols which are not part of the set will appear.

When each symbol appears, you are surprised to see it. If a relatively common symbol appears, you are not very surprised, and if a very rare symbol appears, then you are extremely surprised to see it. Tribus [14] quantified this "surprisal" and defined it as:

$$u_i = -\log_2(P_i). \tag{5}$$

Thus, when the probability of a symbol is low, its surprisal is high and when the probability is high, then its surprisal is low. Notice the resemblance of equation (5) equation to (4).

No matter what symbol you receive, your uncertainty *before* you receive it is the same because a symbol you don't have can't influence your uncertainty. So uncertainty is the *average* surprisal for all of the N symbols. Uncertainty, defined this way, has several properties which are desirable for information theory [5]. So if we sum the surprisals for each symbol received and divide by N (the total number of symbols), we obtain the average surprisal:

$$\frac{u_1 + u_2 + u_3 + \ldots + u_N}{N}. \tag{6}$$

Using summation notation this can be expressed as:

$$\frac{\sum_{i=1}^{M} N_i u_i}{N} \tag{7}$$

where, once again, M is the number of symbol types, N is the total number of symbols in the sequence, N_i is the number of times the i^{th} symbol appears within the entire sequence, and u_i is the surprisal for that symbol. This summation is equivalent to expression (6) because we have regrouped the surprisals. Instead of grouping them in the order in which they occur, we have grouped them by the symbol which they are related to. Consider the string xxyyzyyzzxxy. The first expression (6) would represent this as:

$$\frac{u_x+u_x+u_y+u_y+u_z+u_y+u_y+u_z+u_z+u_x+u_x+u_y}{12} \tag{8}$$

where the second expression (7) would represent it as:

$$\frac{4 \times u_x + 5 \times u_y + 3 \times u_z}{12}. \tag{9}$$

Now, by bringing the denominator inside the summation of (7), we get:

$$\sum_{i=1}^{M} \frac{N_i}{N} u_i. \tag{10}$$

Since the frequency of the i^{th} symbol occurring (F_i) is found by dividing the number of times the particular symbol occurs (N_i) by the number of symbols (N), F_i can be substituted for $\frac{N_i}{N}$ giving:

$$\sum_{i=1}^{M} F_i u_i. \tag{11}$$

Also since F_i gets closer to P_i as N gets larger, (mathematically that's $P_i = \lim_{N\to\infty} F_i$), and since the number of sequences one could be dealing with might be pretty high, (after all, life has existed for billions of years, and will continue, we hope, to flourish) P_i can be substituted for F_i giving:

$$\sum_{i=1}^{M} P_i u_i. \tag{12}$$

Finally, by substituting for u_i from equation (5), we obtain Shannon's famous general formula for uncertainty:

$$\boxed{H = -\sum_{i=1}^{M} P_i \log_2 P_i \qquad \text{(bits per symbol).}} \tag{13}$$

5 Example

Suppose the symbols "a," "c," "g," and "t" are the four symbols a machine uses to generate a twelve letter sequence "gattttctcttt". So far, we know that $N = 12$, $M = 4$, $N_a = 1$, $N_c = 2$, $N_g = 1$, and $N_t = 8$. We find that the frequencies are $F_a = \frac{1}{12}$, $F_c = \frac{2}{12}$, $F_g = \frac{1}{12}$, and $F_t = \frac{8}{12}$. Now, let's say that the frequencies are always the same no matter how many sequences the machine creates. In other words, if the set was infinite, then the frequency of each letter would equal its probability, and this makes $P_i = F_i$. So, $u_a = -\log_2(0.08) = 3.58$ bits. Similarly, $u_c = 2.58$, $u_g = 3.58$, and $u_t = 0.58$ bits. Using equation (13), and substituting in the values for P_a, P_c, P_g, and P_t, we obtain the following:

$$\begin{aligned} H = -[\quad & \frac{1}{12} \times \log_2(\frac{1}{12}) \ + \ \frac{2}{12} \times \log_2(\frac{2}{12}) \\ + \quad & \frac{1}{12} \times \log_2(\frac{1}{12}) \ + \ \frac{8}{12} \times \log_2(\frac{8}{12})] \end{aligned} \tag{14}$$

so $H = 1.42$ bits. This follows with our earlier discussion, as we find that it requires an average of 1.42 bits to determine which symbol is located at any particular position in the sequence. (Note - For equiprobable choices, equation (2) can be used and is much simpler, we just wanted to show you how the more general formula works. As an exercise, you can show that if all the P_i are equal, equation (13) reduces to the same form as (2).)

6 From Uncertainty to Information

Now that we have defined uncertainty and given it a unit of measure, we shall put information in terms of uncertainty. Let us examine our boxes with the blue ball again, but for simplicity, let's say there are just eight boxes. At the start, before the ball's location is known, there is an uncertainty of 3 bits ($\log_2 8 = 3$) [equations (2) and (13)]. After the ball has been located, there is no uncertainty as to where the ball is. After all, you are staring right at it. Since your uncertainty has gone from 3 bits to 0 bits, you have GAINED 3 bits of information.

Now suppose we play the guessing game, but that I refuse to tell you the answer to the last yes-no question you ask. *Before* playing the game your uncertainty is 3 bits, but *after* the game is over, you still don't know which box the blue ball is in. Because you are still uncertain by 1 bit, you learned from me only $3 - 1 = 2$ bits. In other words, the uncertainty (H) *before* the input is received minus the uncertainty *after* it is received is the information you gained (R):

$$H_{before} - H_{after} = R. \tag{15}$$

7 Tying This to Sequence Logos

Understanding the concepts of information and uncertainty is crucial to understanding how a sequence logo is designed and what it shows. On a logo, the horizontal axis represents the position of the base, and the vertical axis represents the amount of information (in bits) which that position holds. So, if there is a letter "T" that is two bits tall at position 7 on a logo, this tells us that whatever the number of sequences analyzed, all, or close to all, had a "T" at that position.

The sequence logo also stacks the letters at each position in order of importance. In other words, the most common letter at a position will be placed at the top of the stack, while the least common letters will be placed at the bottom. So, the letters on top are the equivalent of the consensus sequence.

While the height of the stack is the information content at that position, the height of each letter in the stack corresponds to the frequency of the letter at that position. Take position +3 of the "Lambda cI and cro" logo (Fig. 1). Here, the predominant base is guanine, but there is a case of adenine. So, the "G" is both taller, and on top of the "A" meaning that guanine is more common than cytosine, adenine, and thymine at position +3.

Looking now at column -9, you will see the same letters there—"gattttctcttt"—as we used in the example for calculating equation (14). That was the calculation of H_{after}, which is the uncertainty seen *after* the sites are found. Before the sites are found the protein is not in contact with the DNA and all 4 bases are possible. So the uncertainty H_{before} is 2 bits. Using equation (15), the information at position -9 is $2 - 1.42 = 0.58$ bits. A small-sample correction [8] reduces this to the 0.38 bits high you see in Fig. 1.

This method is ideal for analysis of binding sites on both DNA and mRNA, as well as for analyzing proteins. In a consensus sequence, the base at each position is merely the most common one appearing at that location. This suggests that each base of the binding site is of equal importance. However, the more highly conserved bases are usually the most important for binding, and if a consensus was a good model, then all of the bases across the binding site would be of equal importance and their corresponding letters on the logo would be of equal heights. With the exception of certain restriction enzymes, logos almost invariably show that this is not the case, because they display varying conservation at different points in binding sites. In the sequence logo of Lambda cI and cro binding sites, the difference in importance of each position can easily be seen in the ups and downs of the logo. (It is curious that the conservation alternates between high and low values, but this is not true for other binding sites, so whether it is significant to the biology of these sites is unknown.)

Also in the logo, you will notice that there are error bars on the top of each stack of letters. These bars,

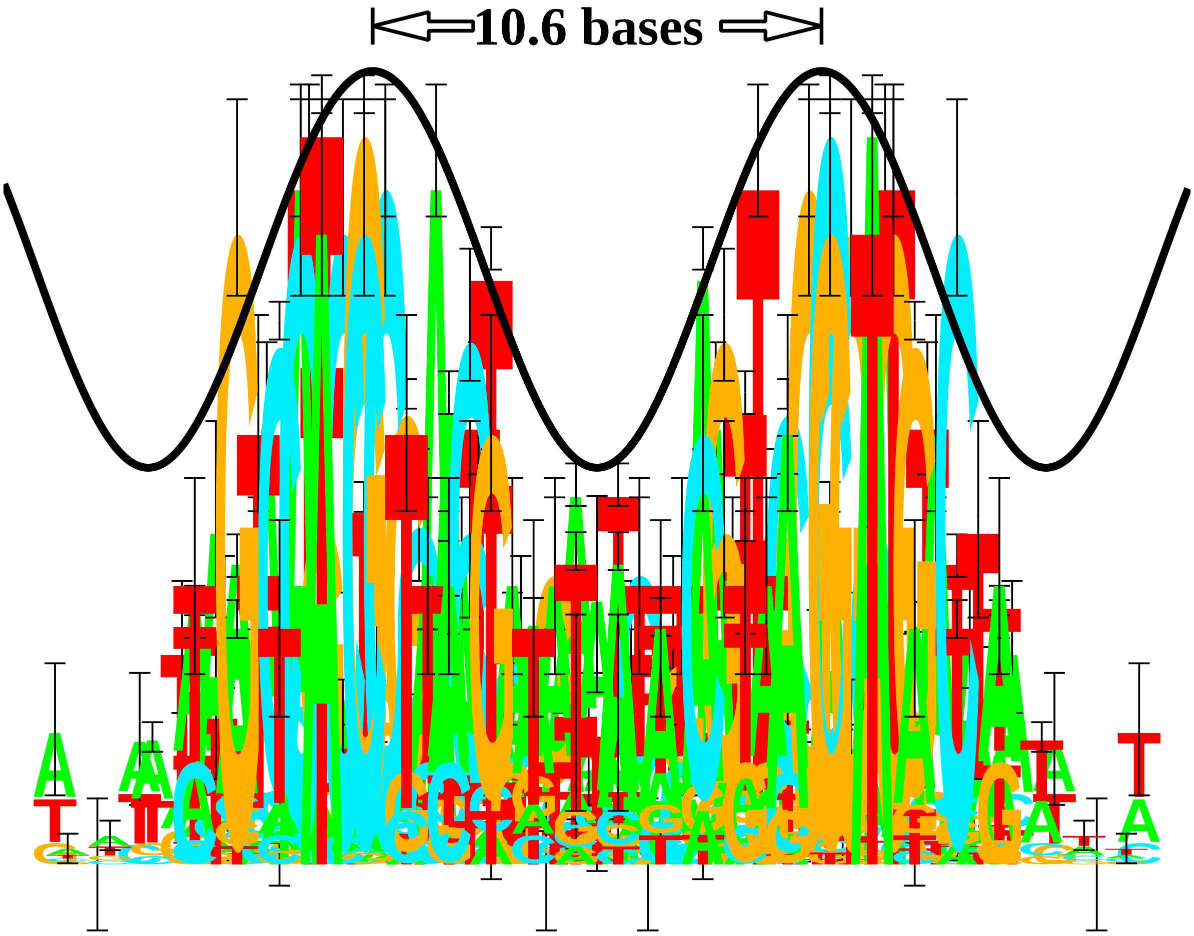

Figure 2: Many sequence logos printed on top of each other.
There were 12 lambda cI and cro binding sites, 8 lambda O binding sites, 58 CRP binding sites, 34 ArgR binding sites, 38 LexA binding sites, 8 TrpR binding sites and 12 AraC binding sites. See [15].

Figure 3: Sequence logo of words in the English language.
Words from the Unix dictionary, /usr/dict/words, were aligned by their first letter. Vowels are darker and consonants are lighter. In color versions of this paper, vowels are red, consonants are green and the letter Y is in blue.

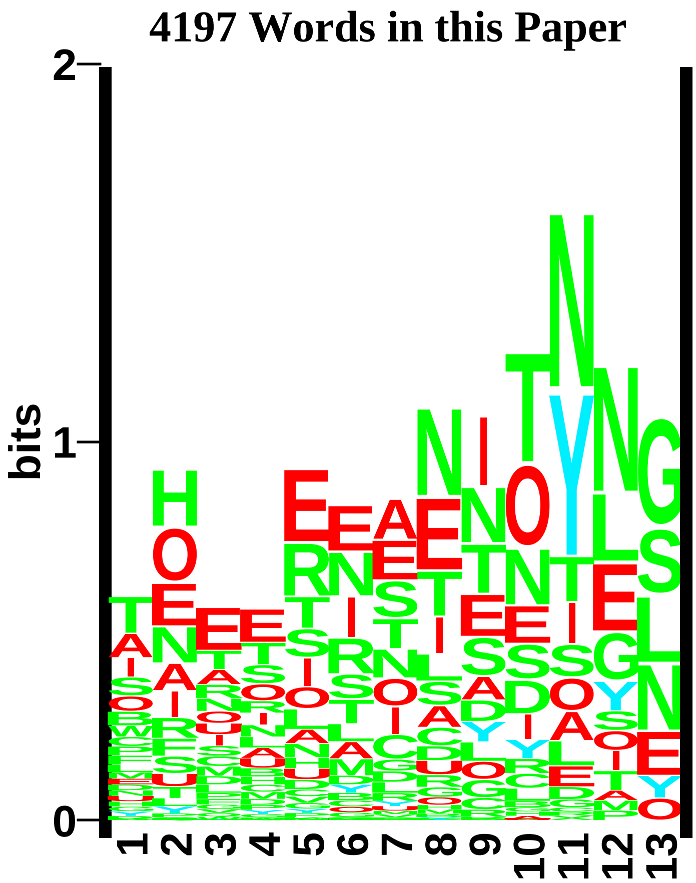

Figure 4: Sequence logo of words in this paper.
Words were aligned by their first letter. In color versions of this paper, vowels are red, consonants are green and the letter Y is in blue.

which look like the letter "I", represent the error that is possible (1 standard deviation) in the value of the *entire* letter stack (the height of the whole stack) due to a limited sample size.

The cosine wave running above the logo represents the major and minor grooves of the DNA helix as seen from one side. The high points of the wave represent the major groove facing the protein, while the low points represent it facing away. Conversely, the low points of the wave represent the minor groove facing the protein, while the high points represent it facing away. This wave is there merely to help you visualize the grooves of the DNA and is not some sort of measurement of how much information should be at the site. However, the height of the wave on the vertical axis is significant. In B-form DNA, two bits of information can be conserved by protein contacts approaching the DNA from the major groove, but only one bit of information can be found by those in the minor groove [15]. The logo demonstrates this effect, since the stacks of letters are not as high where the minor groove faces the protein in the middle. An overlay diagram (Fig. 2) shows many sequence logos printed on top of each other, and as you can see, the height of the letter stacks rarely goes above the wave. Those stacks that do, have error bars which allow for them to pass under the wave.

Now that we've cleared that up, let's take a look at what's happening in that peculiar minor groove. The reason for the depression in the middle of these logos is that if a protein is using contacts to the minor groove, it is difficult or impossible to determine base pair orientation. That is, it is possible to determine an A-T pair, or a G-C pair, but it is not possible to determine whether an A-T pair is oriented as A-T or T-A, and if a G-C pair is G-C or C-G. The structure of the bases and phosphate backbone which make up the DNA are such that the minor groove will only allow a distinction between the two possible base pairs (a one bit decision) but not their orientation. The major groove will not only allow the distinction between A-T or G-C, but also the orientation of the individual base pairs.[2] You may have noticed that the cosine wave is at a height of one bit in the minor groove. This is because, as we said earlier, only one binary question can be answered in the minor groove, so only one bit of information can be obtained there; and in the major groove, the cosine wave has a height of two bits, since two binary questions can be answered there. And now with this last loose end tied up, your sequence logo lesson is complete.

Other strings can be analyzed by this method [5]. For example, in the logo for an English dictionary (Fig. 3), we can see that the first letter is predominantly a consonant (s, c or p) , the second letter is a vowel, and that the third is again a consonant. Curiously, E trails over a hump for the remainder of the words. For the text of this paper the letter usage is different (Fig. 4). There are so many 'the's (7%) that they show up in the first three positions of the logo. The predominance of N and Y at position 11 is particularly telling: it indicates the prominent use of the words *probability*, *uncertainty* and *information* in this paper.

The sequence logo is a powerful tool for analyzing DNA, RNA, protein sequences and words in a language [1]. It goes far beyond the old consensus method. The logo method of analysis reveals the importance of each position in a sequence, along with the importance of each base occurring at each position. A logo represents the amount of information present using a standard unit of measure, which allows for comparison of different types of sites.

Acknowledgements

This work was sponsored by the National Cancer Institute Student Intern Program, the Frederick County Board of Education and the Frederick Cancer Research and Development Center Laboratory of Mathematical Biology. Computational facilities were furnished by the Biomedical Supercomputing Center. Proofreading and friendly conversation were provided by Denise Rubens, R. Michael Stephens, Paul N. Hengen, and Mort Schultz.

Appendix: Obtaining Delila Files

Delila programs (the ones which create sequence logos) are available on Internet by anonymous ftp from ftp.ncifcrf.gov in the directory pub/delila. In the ftp directory, all files except README are compressed by the UNIX compress command. (So they end with a ".Z"). Don't forget to use the binary transfer mode when you transport them.

The uncompress program can be obtained by anonymous ftp from "uunet.uu.net" in "compress.tar" There is also a "help" file there. For VAX VMS users, it may also be obtained from genbank.bio.net in directory pub/vms as the file "lzdcmp.exe" (Contact Dave Kristofferson, kristoff@genbank.bio.net for more information.)

[2]For a complete description of why this is so, see reference [16].

The files are also available to people on BITNET from Dan Davison (davison@uh.edu) on "geneserver%bchs.uh.edu@CUNYVM" [17] (many thanks to Dan for this service).

For additional assistance contact Tom Schneider (toms@ncifcrf.gov). Future significant upgrades will be announced on the newsgroup bionet.info-theory [18]. If you do obtain any programs, please contact us with comments, so we may improve the archive for you.

This paper was originally published in [19].

References

[1] 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/.

[2] M. Ptashne, K. Backman, M. Z. Humayun, A. Jeffrey, R. Maurer, B. Meyer, and R. T. Sauer. Autoregulation and function of a repressor in bacteriophage lambda. *Science*, 194:156–161, 1976.

[3] M. Ptashne, A. Jeffrey, A. D. Johnson, R. Maurer, B. J. Meyer, C. O. Pabo, T. M. Roberts, and R. T. Sauer. How the λ repressor and cro work. *Cell*, 19:1–11, 1980.

[4] A. H. Soukhanov and K. Ellis, editors. *Webster's II New Riverside University Dictionary.* Houghton Mifflin Co., Boston, MA, 1984.

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

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

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

[8] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986. http://www.lecb.ncifcrf.gov/˜toms/paper/schneider1986/.

[9] 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.

[10] 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.

[11] 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.

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

[15] 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.

[16] N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, 73:804–808, 1976.

[17] D. B. Davison and J. E. Chappelear. The Genbank-server at the University of Houston. *Nucleic Acids Res.*, 18:1571–1572, 1990.

[18] A. Bleasby, P. Griffiths, R. Harper, D. Hines, K. Hoover, D. Kristofferson, S. Marshall, N. O'Reilly, and M. Sundvall. Electronic communications and the new biology. *Nucleic Acids Res.*, 20:4127–4128, 1992.

[19] M. C. Shaner, I. M. Blair, and T. D. Schneider. Sequence logos: A powerful, yet simple, tool. In T. N. Mudge, V. Milutinovic, and L. Hunter, editors, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences, Volume 1: Architecture and Biotechnology Computing*, pages 813–821, Los Alamitos, CA, 1993. IEEE Computer Society Press. http://www.lecb.ncifcrf.gov/˜toms/paper/hawaii/.