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

Thomas D. Schneider *

version = 1.26 of aust2002.tex 2007 Feb 8

Biological Theory: Integrating Development, Evolution, and Cognition, 1 (3): 250–260 (2006)

A brief personal history is given about how information theory can be applied to binding sites of genetic control molecules on nucleic acids. The primary example used is ribosome binding sites in *Escherichia coli*. Once the sites are aligned, the information needed to describe the sites can be computed using Claude Shannon's method. This is displayed by a computer graphic called a sequence logo. The logo represents an average binding site, and the mathematics easily allows one to determine the components of this average. That is, given a set of binding sites, the information for individual binding sites can also be computed. One can go further and predict the information of sites that are not in the original data set. Information theory also allows one to model the flexibility of ribosome binding sites, and this led us to a simple model for ribosome translational initiation in which the molecular components fit together only when the ribosome is at a good ribosome binding site. Since information theory is general, the same mathematics applies to human splice junctions, where we can predict the effect of sequence changes that cause human genetic diseases and cancer. The second example given is the Pribnow 'box' which, when viewed by the information theory method, reveals a mechanism for initiation of both transcription and DNA replication. Replication, transcription, splicing, and translation into protein represent the central dogma, so these examples show how molecular information theory is contributing to our knowledge of basic biology.

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

In 1978 I went to graduate school with the explicit intention of finding a mathematics that describes living things. Living things are too beautiful for them not to be described by mathematics. In the University of Colorado, it is the practice for students to do rotations in various labs before setting down in one lab. At the same time we all took a 'core' course covering the fields of the Department of Molecular, Cellular and Developmental Biology. In his lectures, Larry Gold presented translational initiation which he was elegantly dissecting by using the powerful r_{II} genetics of bacteriophage T4.

By 1961 the r_{II} region had been used in remarkable genetic experiments by Francis Crick, Sydney Brenner and their colleagues to prove that the genetic code is read in steps of three [Crick *et al.*, 1961]. David Pribnow, who had identified the -10 region of bacterial promoters while at Harvard [Pribnow, 1975], was hard at work with Sid Shinedling and others in Larry Gold's lab sequencing the old T4 mutations and showing in molecular detail what had only been inferred by elegant genetics before. Many cute molecular puzzles were revealed about translational initiation [Singer *et al.*, 1981, Shinedling *et al.*, 1987]. Larry presented to my class the Shine and Dalgarno (SD) region which is about 10 bases in front of the initiation codon [Shine & Dalgarno, 1974]. The SD is similar to Pribnow's 'box' as both are about 10 bases upstream of the initiation point of translation and transcription, respectively. It was known that the $3'$ end of the 16S rRNA, which forms the main skeleton of the 30S subunit of the ribosome, bound to the SD. The initiation codon is the first codon translated and it is usually AUG but sometimes GUG, rarely UUG and perhaps one CUG in *E. coli*. The SD is a pattern in the mRNA and so one challenge was to characterize the pattern since it is not always a perfect complementary match to the 16S rRNA. However, the problem that intrigued me was to look for other SD-like patterns around the initiation codon. I never found anything, but it launched my career.

Working on this problem meant that we had to gather sequences since GenBank—the international repository of genetic sequences at the National Library of Medicine in Bethesda, Maryland—did not exist yet. So we typed the sequences of known *E. coli* genes into the computer.

I immediately realized that I had a problem. If I typed only those parts that I was interested in—the regions just around the ribosome binding site (RBS)—and later decided that I wanted a bigger or different region, then I would have lots of detailed and tricky editing to do. With only 4 letters, DNA is hard to read and errors would abound! So we decided to enter entire published sequences. But this led to another problem: how to extract just the sequences I needed for a particular problem? From this need was born Delila—DEoxyribonucleic acid LIbrary LAnguage. Delila is a small computer language specifically developed for extracting a set of sequences from a library of sequences [Schneider *et al.*, 1982, Schneider *et al.*, 1984, Schneider, 2002*a*]. With this tool in hand, we could investigate ribosome binding sites and, of course, work on many other problems.

By this time Gary Stormo and Jeff Haemer had joined the effort. Jeff, a brilliant geneticist, slowly transformed himself into a computer scientist. From Jeff I learned the powerful Unix idea of building good tools that each do one job well. For example, Jeff's elegant translation of the atchange program into Perl made atchange into a generally useful automation tool [Schneider, 2002*b*]. Together Jeff and I realized that the output of Delila should have the same format as its input. This allowed us to gather all *E. coli* DNA sequences into one database and then to extract just those sequences that represented mRNA. Then we used Delila a second time to extract the regions around ribosome binding sites, thus guaranteeing that our analysis was only with sequences that the ribosome could come in contact with.

Jon McCabe and Stephen O'Haire of the computer science department wrote a searching program, and Gary Stormo set out to determine 'rules' (regular expressions) for finding the Shine-Dalgarno. He discovered that no single set of rules would work [Stormo *et al.*, 1982*b*]. This lesson is still not understood by most molecular biologists today! The lesson is: don't use rules and don't use consensus sequences. Consensus sequences are a model of the binding site usually created by taking the most frequent base at each position of the site. At that time, it was not clear how to replace consensus sequences or rules with better models of binding sites. In this paper I will briefly describe how neural nets and later information theory elegantly filled that niche.

Fortunately Andrzej Ehrenfeucht, a computer scientist, suggested to us to try a simple neural network—the perceptron. This worked beautifully, and, using Delila and other programs that he built, Gary was able to find weight matrices that separated the known 124 ribosome binding sites from the 78,000 possible ribosome binding sites in our mRNA library [Stormo *et al.*, 1982*a*]. To my knowledge, this was the first use of a neural net in what became the field of bioinformatics, though we didn't call it that at the time.

So the perceptron replaced rules and we could identify ribosome binding sites. For example, translational fusions of *lacZ* to the *uncB* gene had the odd property of giving a high signal if the fusion was early in the gene, but after a certain point further downstream the signal dropped. I thought that there might be an internal ribosome binding site, and found one by using the perceptron. We then confirmed this experimentally [Matten *et al.*, 1998].

Two lessons came from results like this. First, we could learn from the physicists the idea of doing both theory and experiments. At that time, and to a good extent this is still true today, most molecular biology is entirely experimental. However, theory allows one to guide experiments and to identify anomalies. Physicists have come to accept the two approaches, and to appreciate the tension between them that spurs further work. This has yet to occur in molecular biology. If one makes a prediction in a submitted paper, one may get the complaint from a reviewer that it should be tested by experiment before publication. If an experiment is done, then the complaint is that one doesn't need theory! Yet when theory and experiment go hand-in-hand, we often discover things that go unnoticed by others [Schneider *et al.*, 1986, Schneider & Stormo, 1989, Papp *et al.*, 1993, Lyakhov *et al.*, 2001, Schneider, 2001]. The second lesson is that one should be careful not to look under the lamppost all the time. Everybody 'knows' that ribosome binding sites are at gene starts, but they could be in other places too. If one builds search tools that are too rigid, the others won't be found. This is quite common these days with 'gene finding' programs that do not identify alternative splice junctions. We frequently find good splice junctions in places that 'they shouldn't be' and sometimes we can demonstrate that these cryptic sites have interesting effects which can explain genetic diseases [Rogan *et al.*, 1998].

Being able to find binding sites did not help me to understand what the sites are like. I wanted to see more than just strings of letters, as shown in Fig. 1; I wanted to get an intuitive feeling for ⇐Fig 1
their characteristics. Although one can easily see the ATG at the initiation codon in the figure at positions 0 to 2, the SD—in the region of −9—is difficult to pick out unambiguously. We had realized by this time that one could count the number of each base at each position l in the sites, and these could be normalized to give the frequencies of bases at each position. I presented my work about the frequencies of bases around ribosome binding sites to Andrzej Ehrenfeucht's group. Afterwards, when everyone else had left, he asked in his wonderful thick accent "Why don't you

try the information transform?" "What's that?" I asked. He (probably!) wrote:

$$-\sum p \ln p \tag{1}$$

on the blackboard. "What does that mean?" I inquired. "Go look it up!" So, like a good Zen master, he gave me a virtual kick in the pants and launched my career.

Three quarters of a year later I was working on a program and at one point in the code I had access to the number of each base at each position around the ribosome binding sites. I decided to try the 'information transform' and soon recognized that I had to compute information as a difference. In modern terms (which took years to understand and develop!), I had to compute two uncertainties and subtract them to get the information. The first uncertainty is what bases a ribosome sees as it scans the mRNA *before* binding. There are four bases and the ribosome does not know which will be available next as it moves by random Brownian motion along an mRNA before it finds a ribosome binding site where it can start translation into protein. Indeed, it must be prepared for anything. So the ribosome is 'uncertain' by one possibility in four for each base it encounters.

To pick one thing out of two equally likely events takes 1 bit of information. Following earlier work by Hartley, Claude Shannon, father of information theory, argued that information should be additive and so must be based on the logarithm of the number of possibilities [Shannon, 1948, Pierce, 1980, Schneider, 1995]. That is, $\log_2 2 = 1$ bit. It takes one yes-no question and an answer of either 'heads' or 'tails' to specify the state of a coin. Likewise, to pick one base out of the four in DNA takes $\log_2 4 = 2$ bits. For example, if the bases are arranged in a square, then two questions will pick out one of them: 'Is it on the top?' and 'Is it on the right?'

Why did Shannon use the logarithm? Suppose that we have two independent communication channels, one with symbols **h** and **t** (a coin) and the other with **A**, **C**, **G**, and **T** (DNA). Together these channels can send $2 \times 4 = 8$ possible symbol pairs— **hA**, **hC**, **hG**, **hT**, **tA**, **tC**, **tG**, and **tT**. Each symbol pair would carry $\log_2 8 = 3$ bits of information. The information is additive since $\log_2 2 + \log_2 4 = \log_2(2 \times 4) = \log_2 8$.

So before a ribosome binds to a binding site, it sees all four bases and is *uncertain* by $\log_2 4 = 2$ bits. After binding the ribosome sees various frequencies of bases. The initiation codon AUG, GUG, and rarely UUG or CUG, always has a U in the second and a G in the third position. (When DNA is copied—transcribed—into RNA, U replaces T.) There is only one possibility for the second position, so $\log_2 1 = 0$ bits. The information that the ribosome gains is the difference between its uncertainty before (2 bits) and its uncertainty after (0 bits), which is 2 bits.

I cannot overemphasize the important concept that information must always be computed as a difference. This was the way Shannon did it, but the literature is littered with failed attempts to use information theory in molecular biology because authors did not realize this.

Sometimes the uncertainty *after* is not zero and so the information is lower. This corresponds to noise in a communications channel, and Shannon called it the equivocation. It represents sequence variations that the ribosome does not care about. If a DNA binding protein accepted two possible bases, T or C in its binding sites, the uncertainty after would be 1 bit and the information at that position would be $2 - 1 = 1$ bit. In the extreme, if the ribosome doesn't care about a position, as when it is outside the binding site, then all four bases are allowed and the uncertainty after is 2 bits. So the information is $2 - 2 = 0$ bits.

A more complicated example is the first base of the initiation codon, which has the frequencies: A: 3551, C: 1, G: 298, and T: 50. How can the uncertainty be computed? Shannon recognized that the *average* is the important quantity to compute:

$$\text{uncertainty} = H = -\sum_{i=1}^{M} p_i \log_2 p_i \quad \text{bits/symbol.} \tag{2}$$

See my information theory primer for how this can be derived intuitively [Schneider, 1995]. p_i is the probability of the i^{th} symbol out of M possible symbols. For ribosomes we know the frequencies of bases $b \in \{A,C,G,T\}$ of each position (l), which we can write as $f(b,l)$. The frequencies are an estimate of the probability of the bases, so plugging this into (2) gives:

$$H_{after}(l) = -\sum_{b\in\{A,C,G,T\}} f(b,l)\log_2 f(b,l) \quad \text{bits/base.} \tag{3}$$

Frequencies are only an *estimate* of the probabilities and a correction (not shown) must be made to account for this, especially when there are few sequences [Schneider *et al.*, 1986].

Before binding, for simplicity, we will assume all bases are equally likely. This is true for *E. coli*, but see reference [Schneider, 1999] for a discussion. For $M = 4$ equally likely bases, equation (2) collapses to

$$H_{before} = \log_2 4 = 2 \quad \text{bits/base.} \tag{4}$$

Showing that this is indeed the case is a worthy exercise for the reader.

The information at l is the decrease of uncertainty that the ribosome experiences:

$$R_{sequence}(l) = H_{before} - H_{after}(l) \quad \text{bits/base.} \tag{5}$$

Following Shannon, R stands for the rate of information transmission, bits per base in this case. The perceptive reader will notice that the uncertainty (equation (2)) corresponds to the entropy and that the information represents a decrease of the entropy. The relationship between entropy, uncertainty and information has been discussed in reference [Schneider, 1991*b*], but that fascinating topic is beyond the scope of this paper.

Fig. 2 shows the information curve for ribosome binding sites in *E. coli*. Note that the initiation codon shows up as a peak at positions 0, 1 and 2. Since information is additive for independent systems, and since the positions of ribosomes are independent by our measurement of correlations [Stephens & Schneider, 1992], one can compute the total information as: ⇐Fig 2

$$R_{sequence} = \sum_{l} R_{sequence}(l) \quad \text{bits/site.} \tag{6}$$

The total information is a nice additive measure of sequence conservation for biology. The implications of this important number are beyond the scope of this paper. Briefly, however, one can use the size of the genome and the number of sites to predict how much information is needed to find the binding sites. This is often close to $R_{sequence}$ [Schneider *et al.*, 1986, Schneider & Stormo, 1989, Herman & Schneider, 1992, Schneider, 2000].

Ten years after starting this work, in 1990, Mike Stephens (a high school student at that time) and I invented a way to show the patterns [Schneider & Stephens, 1990]. Fig. 3 shows the sequence logo for the curve of Fig. 2. The logo consists of stacks of letters representing the DNA bases. The height of each stack is the information in bits. The height of each letter is proportional to the ⇐Fig 3 frequency of the corresponding base, and the bases are sorted to put the most frequent one on top. With sequence logos, one can finally see the patterns in binding sites.

How can we see what individual binding sites look like? Again, the approach begins with Shannon's uncertainty equation, (2), which we can rewrite as

$$H = \sum_{i=1}^{M} p_i(-\log_2 p_i) \quad \text{bits/symbol.} \tag{7}$$

From this viewpoint, the uncertainty can be seen as the average of the function

$$u_i = -\log_2 p_i \quad \text{bits/symbol.} \tag{8}$$

This quantity was recognized by Tribus in 1961 and called the surprisal [Tribus, 1961]. With this in mind, we can look at the sequence logo (Fig. 3) and recognize that it is representing the *average* of many ribosome binding sites.

We know that the 'area' under the logo, $R_{sequence}$, is the average sequence conservation. Suppose that we could assign to each ribosome binding site an individual information, so that the average of these is $R_{sequence}$. It turns out that this is easy [Schneider, 1997*a*, Schneider & Rogan, 1999]. The state change is from being anywhere on the sequence to being at a specific location, so we compute the difference between the average *before* surprisal (the uncertainty) and *after* surprisal:

$$R_i(b,l) = 2 - (-\log_2 f(b,l)) \quad \text{bits/base.} \tag{9}$$

This forms a matrix of 4 by l numbers, as shown in Fig. 4. A specific sequence will pick out one number at each of the l positions [Stormo *et al.*, 1982*a*, Schneider, 1997*a*]. Add these together to get the individual information of the sequence, R_i. It can be shown that the average of these over all of the input sequences is indeed the total $R_{sequence}$. John Spouge proved that formula (9) is unique; there is no other way to compute the individual information [Schneider, 1997*a*]. ⇐Fig 4

Using this method, we can represent individual binding sites with a computer graphic called a sequence walker (Fig. 5). These walkers correspond to the 10 sequences in Fig. 1. Unlike the logo, a walker consists of only one letter per position, because it is an evaluation of a single sequence by an individual information weight matrix. The height of each letter in a walker gives the information weight of the base according to equation (9). Positive values represent good binding ($\Delta G < 0$) while negative values represent bases that are not favored ($\Delta G > 0$) [Schneider, 1997*a*, Schneider, 1991*b*]. ⇐Fig 5

With the advent of the sequence logo, individual information, and sequence walker techniques we can finally avoid using neural networks. The advantage is that there is no training process to compute the information, and one can build a model directly from sequences known to bind. In neural net training one needs examples of sequences that do not bind to the recognizer and, generally, good data are not available. Often people will assume that there are no sites near to the known ones, which experience has shown us is a bad assumption because there are often important sites near by [Schneider, 1997*b*, Hengen *et al.*, 1997], or worse, they make up data for training! With information theory we can gain a theoretical understanding of the data.

From the logo (Fig. 3) we can immediately see that the SD is not very big. It is only a small lump to the left of the initiation codon. The SD does not show up well in the walkers either (Fig. 5). Since we aligned the sequences by the initiation codon, the SD are not well aligned and their patterns are spread out, making the picture of the SD blurred. The reason is that in different genes the $3'$ end of 16S rRNA binds at different distances from the initiation codon. We have recently shown that this variable distance can be nicely accounted for by using information theory [Shultzaberger *et al.*, 2001].

To dissect the ribosome binding sites into their SD and initiation region (IR) parts, we need to align the SD region. An extremely clean way to do this is to maximize the information content. The method is simple [Schneider & Mastronarde, 1996]. The SD regions are first isolated away from the IR by embedding them in random sequences. Then the SD sequences are shuffled back and forth while the total information content $R_{sequence}$ is computed. With a few tricks, such as making a look-up table for computing $-f(b,l)\log_2 f(b,l)$ because there are a finite number of frequencies, this method is very fast. Fortunately binding sites are tight enough that we can avoid introducing gaps, which would make the alignment problem explode exponentially in the number of sequences and number of allowed gaps. To our delight this multiple alignment process converged nicely. The left side of Fig. 6 shows the sequence logo for the aligned SD [Shultzaberger *et al.*, 2001]. The pattern that appears matches the $3'$ end of the 16S rRNA. This is remarkable because we did not use the 16S sequence to do the alignment. The correlation is a strong confirmation that the SD exists and is bound by the 16S. Thus, for the first time, we were able to create an unbiased picture of what the SD 'looks like'. ⇐Fig 6

The right side of Fig. 6 is the initiation region where the first tRNA delivers the N-formylmethionine to initiate translation. The middle of the figure shows the relative distribution of distances between the SD and the IR. How can we take this into account when computing the individual information?

Using information theory, the solution is, again, quite simple. We have a distribution of distances produced during the multiple alignment process. This forms the probability distribution shown in Fig. 6. The uncertainty of any probability distribution can be computed from equation (2). Therefore, the surprisal for each individual distance can be computed from equation (8). The total information for a single ribosome binding site can be computed by adding the individual information of the SD and IR and subtracting the spacing surprisal. With this parameter-free approach, we were able to model the majority of ribosome binding sites in *E. coli* [Shultzaberger *et al.*, 2001].

How can we see what one site looks like with this flexible model? Fig. 7 shows examples of flexible sequence walkers. The model is searched across a sequence, with all SD-IR distances allowed and the ones with the highest information content are displayed. In most cases the SD shows up as a distinct lump of information at various distances from the initiation codon.

The SD lump is about 10 or 11 bases away from the initiation codon, which suggested to us a simple model for translational initiation [Shultzaberger *et al.*, 2001]. Since 11 bases is a single twist of double helical RNA, the idea is that the double helix formed by the SD and the mRNA and the interaction of the initiation fMet-tRNA$_f^{Met}$ with the first codon may form a single structure that nestles onto the surface of the 30S subunit. The sinusoidal shape of the logo suggested that the SD helix is bound on one side, as we had observed for DNA-protein interactions [Papp *et al.*, 1993, Schneider, 2001]. Since the 30S and 50S ribosomal subunits are compact objects [Nissen *et al.*, 2000, Ban *et al.*, 2000, Wimberly *et al.*, 2000], we proposed that the recognition of the SD might occur by the double helix fitting into a slot on the 30S subunit surface. If the SD in the mRNA does not match the 16S $3'$ end well, then the helix would not fit into the slot. When there is a good fit, all the parts come together compactly and this would be the initiation configuration. Three-dimensional X-ray crystals with and without the mRNA were obtained by Harry

Noller's laboratory [Yusupova *et al.*, 2001]. They observed that the SD helix is indeed enclosed in a cleft, with the N terminus of protein S8 pointing into the major groove. These results account for the sequence logo and support the idea that initiation occurs by via a compact bound state.

I'd like to end this essay by returning to the Pribnow 'box', which resembles the SD in that it is also 10 bases upstream of the point of initiation, but for transcription instead of translation. The sequence that David Pribnow observed is often called a TATAAT consensus [Lewin, 1997] since those are the most frequent bases. But the logo reveals something very different (Fig. 8) ⇐Fig 8 [Schneider, 2001]. What's going on here? The logo shows that three of the middle bases are far less conserved than conventionally understood. The highly conserved T on the right side at position -7 is in the region opened by RNA polymerases during transcriptional initiation (shown by the solid and dashed boxes). The bases to the left of position -9 are outside the opened region. We propose that after sigma factor binding, the initiation is accomplished by swinging the T at position -7 out of the DNA. This 'base flipping' has been observed in X-ray crystal structures of protein-DNA complexes [Roberts, 1995, Roberts & Cheng, 1998] and we have made similar observations with sequence logos in several other systems which are known to open DNA [Schneider, 2001].

These observations led us to perform experiments and the results indicate that bacteriophage P1 probably uses base flipping to initiate its DNA replication [Lyakhov *et al.*, 2001]. It is likely that base flipping is a general mechanism used to open DNA to initiate both RNA transcription and DNA replication, as predicted by Rich Roberts [Roberts, 1995]. This discovery was possible only because sequence logos give such a clear picture of binding sites.

This paper is a brief introduction to the field I call Molecular Information Theory. I have mentioned only a few of the results. Notably missing is our work with human splice junctions, which has led to a form of medical diagnosis [Rogan *et al.*, 1998]. What does the future hold? Shannon not only worked out how to measure information, but he also derived an equation for the maximum information that can be transmitted over a channel. This channel capacity can be linked to fundamental thermodynamics and molecular biology [Schneider, 1991*a*, Schneider, 1991*b*] and from this connection many new discoveries are coming.

Acknowledgments. I thank Ryan Shultzaberger for creating Fig. 6; Krishnamachari Annangarachari, Brent Jewett, Jerry Chandler, Ryan Shultzaberger, and Jim Ellis for comments on the manuscript.

References

[Ban *et al.*, 2000] Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science,* **289**, 905–920.

[Blattner *et al.*, 1997] Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453–1474.

[Crick *et al.*, 1961] Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature,* **192**, 1227–1232.

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

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

[Lewin, 1997] Lewin, B. (1997). *Genes VI.* Oxford University Press, Oxford.

[Lyakhov *et al.*, 2001] Lyakhov, I. G., Hengen, P. N., Rubens, D. & Schneider, T. D. (2001). The P1 Phage Replication Protein RepA Contacts an Otherwise Inaccessible Thymine N3 Proton by DNA Distortion or Base Flipping. *Nucleic Acids Res.* **29** (23), 4892–4900. http://www.ccrnp.ncifcrf.gov/~toms/paper/repan3/.

[Matten *et al.*, 1998] Matten, S. R., Schneider, T. D., Ringquist, S. & Brusilow, W. S. A. (1998). Identification of an intragenic ribosome binding site that affects expression of the *uncB* gene of the *Escherichia coli* proton-translocating ATPase (*unc*) operon. *J. Bacteriol.* **180**, 3940–3945.

[Nissen *et al.*, 2000] Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science,* **289**, 920–930.

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

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

[Pribnow, 1975] Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA,* **72**, 784–788.

[Roberts, 1995] Roberts, R. J. (1995). On base flipping. *Cell,* **82**, 9–12.

[Roberts & Cheng, 1998] Roberts, R. J. & Cheng, X. (1998). Base flipping. *Annu Rev Biochem,* **67**, 181–198.

[Rogan *et al.*, 1998] Rogan, P. K., Faux, B. M. & Schneider, T. D. (1998). Information analysis of human splice site mutations. *Human Mutation,* **12**, 153–171. Erratum in: Hum Mutat 1999;13(1):82. http://www.ccrnp.ncifcrf.gov/~toms/paper/rfs/.

[Rudd, 2000] Rudd, K. E. (2000). EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.* **28**, 60–64.

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

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

[Schneider, 1995] Schneider, T. D. (1995). *Information Theory Primer*. http://www.ccrnp.ncifcrf.gov/~toms/paper/primer/.

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

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

[Schneider, 1999] Schneider, T. D. (1999). Measuring molecular information. *J. Theor. Biol.* **201**, 87–92. http://www.ccrnp.ncifcrf.gov/~toms/paper/ridebate/.

[Schneider, 2000] Schneider, T. D. (2000). Evolution of biological information. *Nucleic Acids Res.* **28** (14), 2794–2799. http://www.ccrnp.ncifcrf.gov/~toms/paper/ev/.

[Schneider, 2001] Schneider, T. D. (2001). Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res.* **29** (23), 4881–4891. http://www.ccrnp.ncifcrf.gov/~toms/paper/baseflip/.

[Schneider, 2002*a*] Schneider, T. D. (2002*a*). Introduction to Delila Instructions. http://www.ccrnp.ncifcrf.gov/~toms/delilainstructions.html.

[Schneider, 2002*b*] Schneider, T. D. (2002*b*). The atchange Program. http://www.ccrnp.ncifcrf.gov/~toms/atchange.html.

[Schneider & Mastronarde, 1996] Schneider, T. D. & Mastronarde, D. (1996). Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Applied Mathematics,* **71**, 259–268. http://www.ccrnp.ncifcrf.gov/~toms/paper/malign.

[Schneider & Rogan, 1999] Schneider, T. D. & Rogan, P. K. (1999). Computational analysis of nucleic acid information defines binding sites, United States Patent 5867402.

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

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

[Schneider *et al.*, 1986] Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431. http://www.ccrnp.ncifcrf.gov/~toms/paper/schneider1986/.

[Schneider *et al.*, 1982] Schneider, T. D., Stormo, G. D., Haemer, J. S. & Gold, L. (1982). A design for computer nucleic-acid sequence storage, retrieval and manipulation. *Nucleic Acids Res.* **10**, 3013–3024.

[Schneider *et al.*, 1984] Schneider, T. D., Stormo, G. D., Yarus, M. A. & Gold, L. (1984). Delila system tools. *Nucleic Acids Res.* **12**, 129–140.

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

[Shine & Dalgarno, 1974] Shine, J. & Dalgarno, L. (1974). The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA,* **71**, 1342–1346.

[Shinedling *et al.*, 1987] Shinedling, S., Gayle, M., Pribnow, D. & Gold, L. (1987). Mutations affecting translation of the bacteriophage T4 *rIIB* gene cloned in *Escherichia coli*. *Mol Gen Genet,* **207**, 224–232.

[Shultzaberger *et al.*, 2001] Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E. & Schneider, T. D. (2001). Anatomy of *Escherichia coli* Ribosome Binding Sites. *J. Mol. Biol.* **313**, 215–228. http://www.ccrnp.ncifcrf.gov/~toms/paper/flexrbs/.

[Singer *et al.*, 1981] Singer, B. S., Gold, L., Shinedling, S. T., Colkitt, M., Hunter, L. R., Pribnow, D. & Nelson, M. A. (1981). Analysis *in vivo* of translational mutants of the rIIB cistron of bacteriophage T4. *J. Mol. Biol.* **149**, 405–432.

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

[Stormo *et al.*, 1982*a*] Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. (1982*a*). Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10**, 2997–3011.

[Stormo *et al.*, 1982*b*] Stormo, G. D., Schneider, T. D. & Gold, L. M. (1982*b*). Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10**, 2971–2996.

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

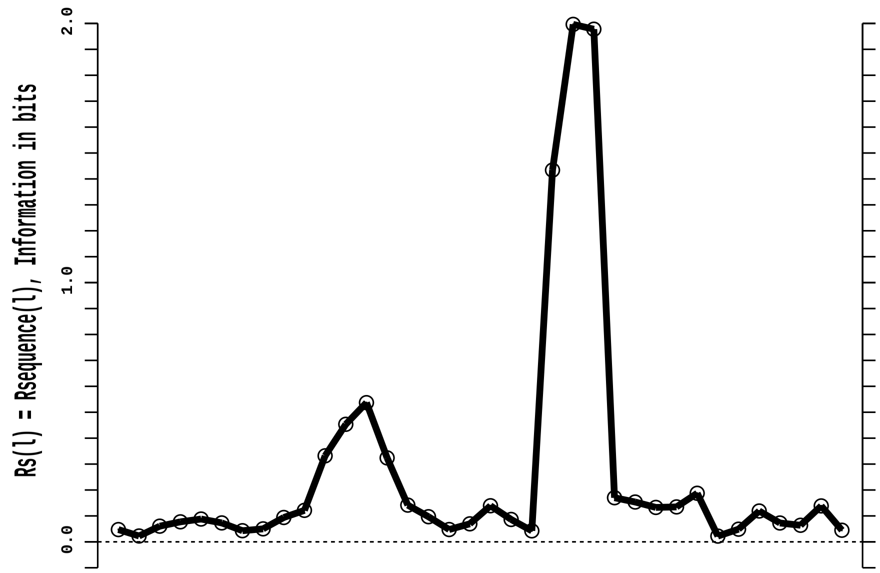
[Wimberly *et al.*, 2000] Wimberly, B. T., Brondersen, D. E., Clemons Jr., W. M., Morgan-Warren, R. J., Carter, A. P., Vonrheln, C., Hartsch, T. & Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. *Nature,* **407**, 327–339.

[Yusupova *et al.*, 2001] Yusupova, G. Z., Yusupov, M. M., Cate, J. H. & Noller, H. F. (2001). The path of messenger RNA through the ribosome. *Cell,* **106**, 233–241.

```
                     ------------                   +++++
                     221111111111--------- +++++++++11111
                     109876543210987654321012345678901234
                     ....................................
U00096  3734 +  1 gcacgagtactggaaaactaaatgaaactctacaat
U00096  8238 +  2 tgtttaaagagaaatactatcatgacggacaaattg
U00096 12163 +  3 atatatagtggagacgtttagatgggtaaaataatt
U00096 14168 +  4 tctaggggcaatttaaaaaagatggctaagcaagat
U00096 17489 +  5 cacctgaaagagaaataaaaagtgaaacatctgcat
U00096 22391 +  6 aaatacggaaccgagaatctgatgagtgactataaa
U00096 25826 +  7 taaatataagagcaaacctgcatgtctgaatctgta
U00096 29651 +  8 gaatattctctggagggtgttttgattaagtcagcg
U00096 30817 +  9 gtaatcaggagtaaaagagccatgccaaaacgtaca
U00096 49823 + 10 tttttttatcgggaaatctcaatgatcagtctgatt
```

Figure 1: Some proven ribosome binding sites.

The first 10 experimentally proven ('verified') ribosome binding sites in the EcoGene 12 dataset [Rudd, 2000] are shown aligned by the initiation codon, which covers positions 0 to 2. The sequences are written 5′ on the left to 3′ on the right and translation is to the right. The sequences come from the complete *E. coli* genome, GenBank Accession U00096 [Blattner *et al.*, 1997]. These particular example gene sequences are oriented clockwise (+) on the genome, but about half of all genes have the other orientation. Above the sequences are coordinate positions, l, written vertically. Color coding (or shading) helps one to see patterns.



l	a	c	g	t	Rs(l)
-21	175	115	96	183	0.05
-20	186	120	117	146	0.02
-19	201	121	88	159	0.06
-18	209	109	87	164	0.08
-17	215	99	90	165	0.09
-16	207	112	87	163	0.07
-15	189	105	109	166	0.04
-14	204	96	143	126	0.05
-13	229	84	138	118	0.09
-12	215	74	183	97	0.12
-11	198	48	269	54	0.33
-10	180	32	308	49	0.45
-9	189	16	313	51	0.54
-8	235	40	232	62	0.32
-7	210	57	189	113	0.14
-6	216	70	132	151	0.10
-5	196	97	118	158	0.05
-4	214	108	96	151	0.07
-3	255	100	127	87	0.14
-2	163	162	63	181	0.09
-1	162	177	86	144	0.04
0	509	0	48	12	1.43
1	0	0	0	569	2.00
2	0	0	568	1	1.98
3	251	71	159	88	0.17
4	171	238	84	76	0.15
5	184	106	64	215	0.13
6	238	128	139	64	0.13
7	262	111	59	137	0.19
8	173	104	132	160	0.02
9	210	126	126	107	0.05
10	194	131	57	187	0.12
11	192	100	93	184	0.07
12	207	128	148	86	0.06
13	211	111	59	188	0.14
14	158	142	86	183	0.04

Figure 2: Information curve for verified ribosome binding sites.
The sites are described in [Rudd, 2000, Shultzaberger *et al.*, 2001]. The table under the curve gives and the position l, the number of a, c, g and t at each position, and the information $Rs(l)$. $Rs(l)$ stands for $R_{sequence}(l)$. The curve corresponds to the data from Fig. 1.

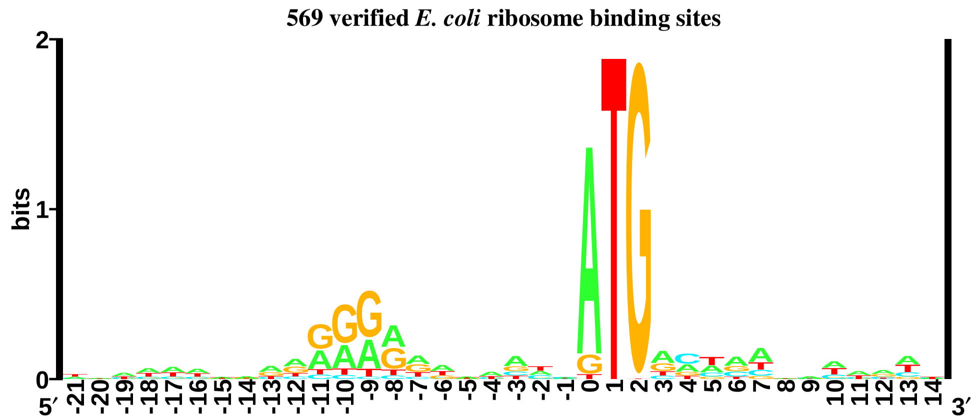


Figure 3: Sequence logo for ribosome binding sites.
The logo corresponds to the curve in Fig. 2 and the data from Fig. 1.

base b	position l		
	0	1	2
A	$\boxed{+1.84}$	-7.16	-7.16
C	-7.16	-7.16	-7.16
G	-1.57	-7.16	$\boxed{+1.99}$
T	-3.57	$\boxed{+2.00}$	-7.16

Figure 4: Initiation codon information weight matrix, $R_i(b,l)$.

The weights for the sequence 5′ ATG 3′ are boxed. The value -7.16 represents positions where that base was not observed. Since $f(b,l) = 0$ at these positions, equation (9) shows that such weights could be set to $-\infty$, but since there is only a finite sample of sequences, an estimate based on the probability of observing that base is substituted [Schneider, 1997*a*]. This prevents the model from being overly reactive to new data.

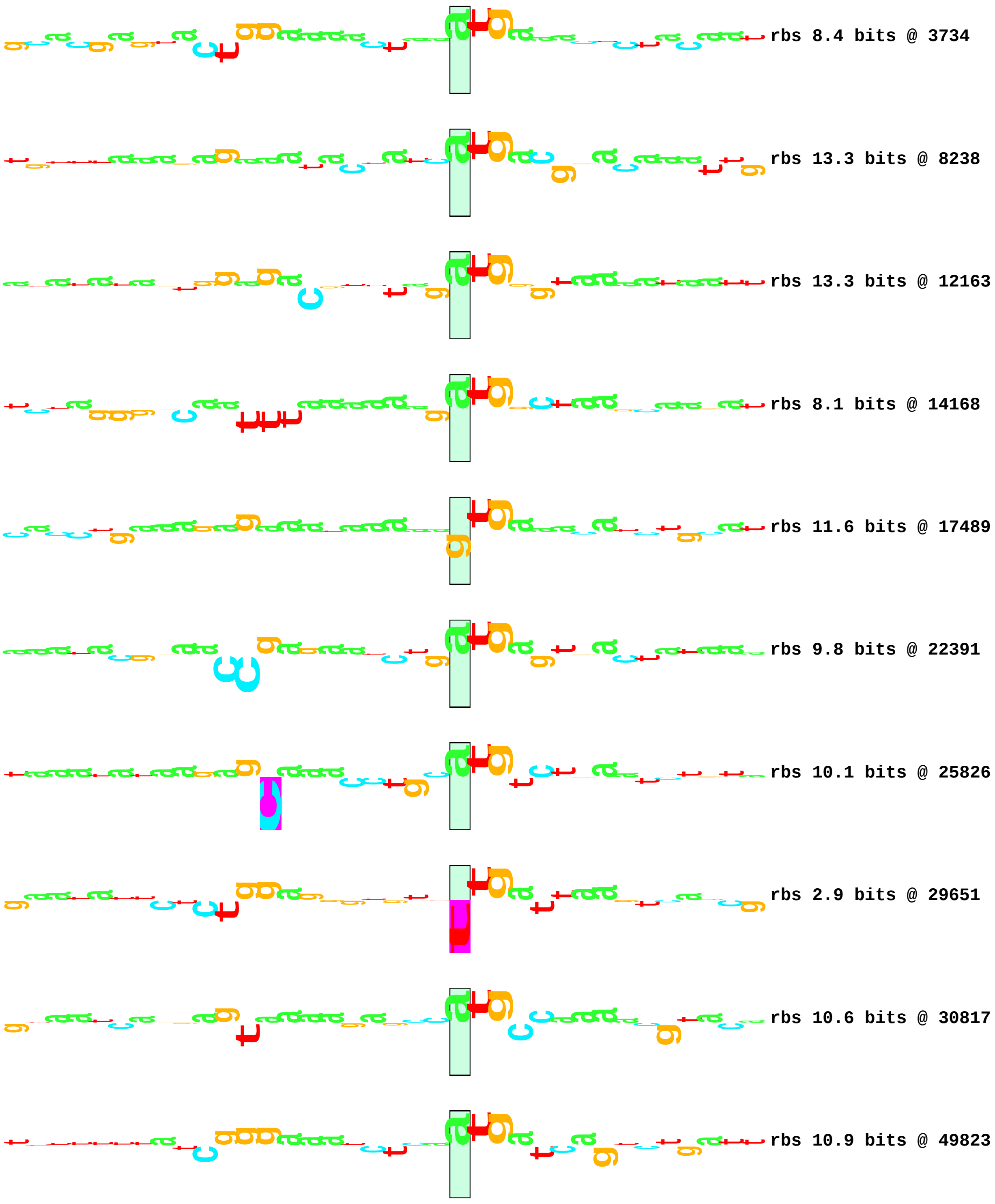


Figure 5: Sequence walkers for individual ribosome binding sites (rbs).
These are the first 10 verified sites used in Fig. 1 evaluated by the individual information model corresponding to Figures 2 and 3. The green (lightly shaded when black and white) box indicates the scale, which runs from -3 to $+2$ bits. A purple (dark shaded when black and white) box indicates that the information is less than -3 bits. The information content of each site is given followed by the coordinates on the *E. coli* genome [Blattner *et al.*, 1997].

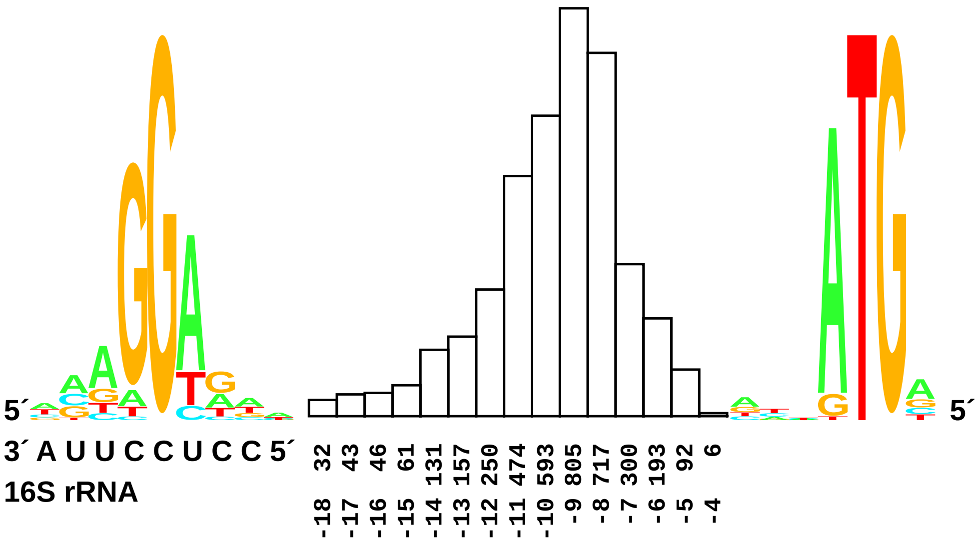


Figure 6: Flexible ribosome binding site model: sequence logos for the SD and IR, and the distance distribution between them.

Note how the SD sequence logo nicely complements the $3'$ end of the 16S rRNA, although the 16S sequence was not used to align the SD. This demonstrates that the SD pattern exists independently of models for 16S binding. The smooth shape of the information curve indicates that not all positions are equally important. Also, this sinusoidal shape is characteristic of interactions in which nucleic acids are recognized while in double helical form [Papp *et al.*, 1993, Schneider, 2001]. The peak of the spacing represents a distance of -9 bases between the peak of the SD and the first base of the initiation codon, with larger distances to the left of the histogram [Shultzaberger *et al.*, 2001]. The number of ribosome binding sites at each spacing is given above the distance numbers.

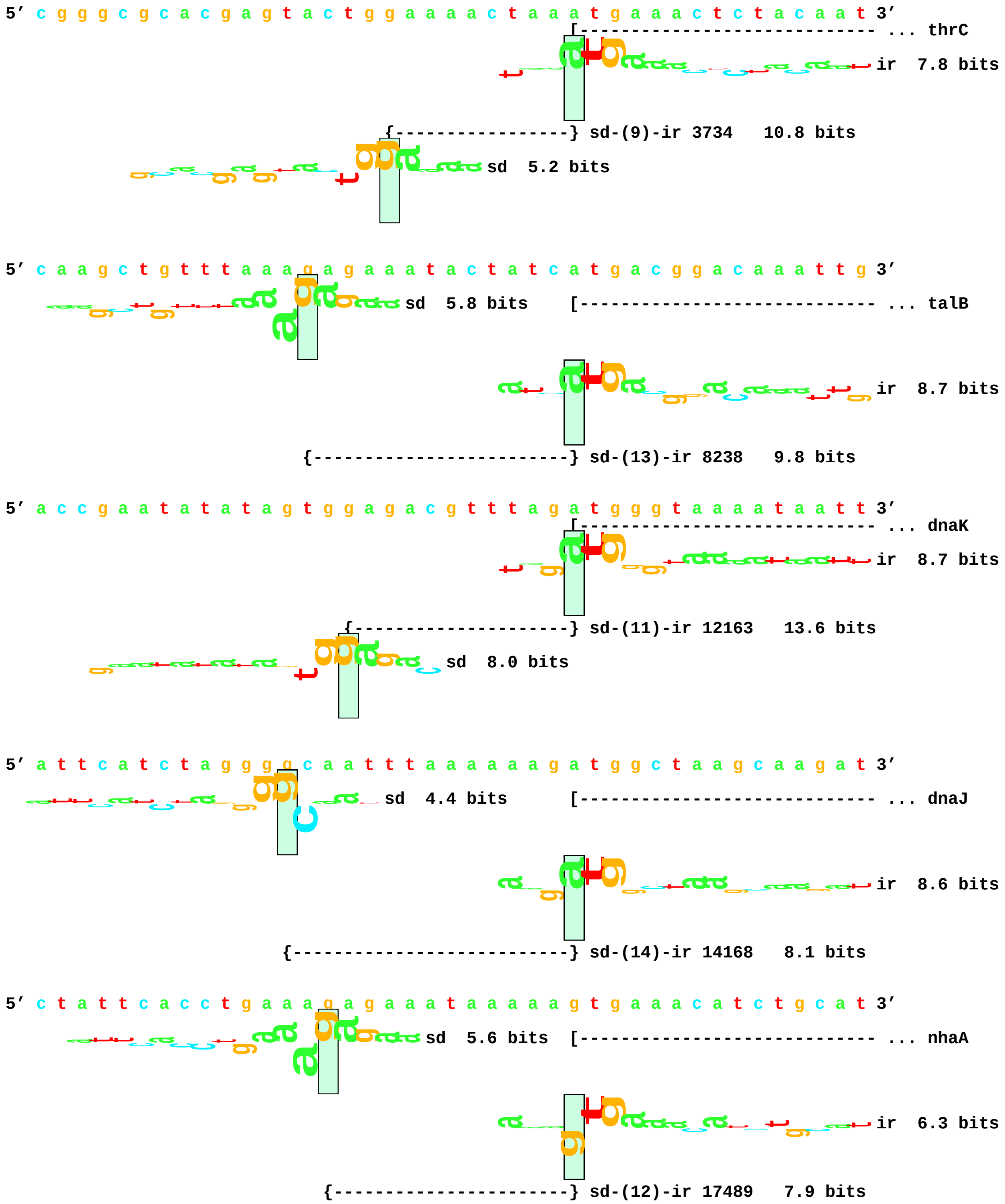


Figure 7: Flexible ribosome binding site model: sequence walkers.
The first 5 sequences in Fig. 1 were analyzed by a flexible sequence walker for ribosome binding sites. Each flexible walker consists of two sequence walkers connected by a linking bar that indicates which SD is connected to which IR. (In this figure, there is only one case per sequence.) After the bar the distance between the walkers and the coordinate of the IR walker are shown, along with the total information.

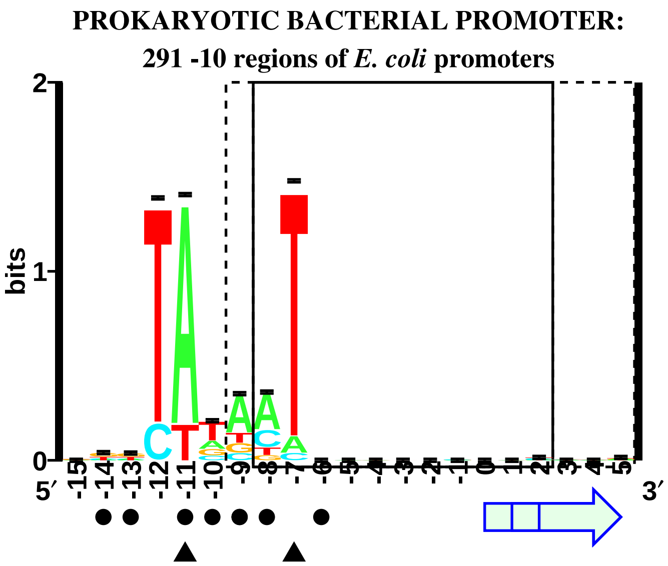


Figure 8: Sequence logo for the Pribnow 'box'.
The arrow indicates start points for transcription. The circles and triangles are data that localize the site [Schneider, 2001].