Figures Supporting Information

Figure S1. Sequence walker promoter analysis

A gentle introduction to information theory. Before describing the figures, in this section we will give a brief but gentle introduction to information theory, followed by how we used it to discover the location of the LexA site at the P_{LIT} promoter. Information theory is a mathematics developed by Claude Shannon (Shannon, 1948; Shannon and Weaver, 1949). It plays an essential role in modern society since it is the basis of modern communications systems. In particular, your cell phone receives a signal from a radio tower, but the signal is not the pure clean high and low voltage pulses generated at the other person's cell phone; each pulse is contaminated with noise that sounds like a hiss. Shannon discovered that there is a way to protect the signal so that most of the noise can be removed at the receiver. That is why your cell phone communications seem to be noise free. Interested readers can learn the mathematics of information theory used here from Tom Schneider's short "Information Theory Primer" (Schneider, 2013). The best introductory book is by Pierce (Pierce, 1980). For this supplement we will just present the concepts essential for understanding the figures: **bits**, **sequence logos**, **individual information** and **sequence walkers**.

Bits in information theory are a measure of choice. A coin set on a table can have either heads or tails upwards, and so it can store 1 bit of information. A person has two possibilities to chose from when setting a coin down.[1] One could ask the person a question that is answered by either 'yes' or 'no' such as, 'Is the head pointing up?'. The answer tells us 1 bit of information.

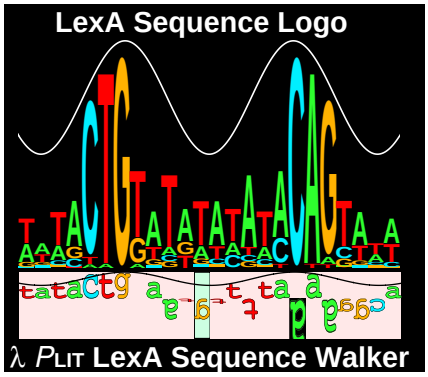
If we align a set of DNA sequences that a protein like LexA binds to, then in some positions along the site there will be more sequence conservation and in others less. Consider a position that is always a G, as is the case for position -5 of LexA (Papp et al., 1993). When the protein is searching by Brownian motion for a binding site, the amino acids that contact position -5 (Zhang et al., 2010) require that it be a G or the entire protein will not bind. That is, those amino acids chose one out of the four possible bases. If we arrange the bases in a square: $\begin{array}{c|c} \mathrm{A} & \mathrm{C} \\ \hline \mathrm{G} & \mathrm{T} \end{array}$ then we can pick one of them using two questions: "Is it on top?" and "Is it on the left?". For example, "Is it on top?" NO and "Is it on the left?" YES specifies G. So 2 yes-no questions are needed in this case which means that the information to specify one base in four is 2 bits.

Many times a protein will have flexibility in the bases it demands at a position. So for example a protein may only require A or G, and it doesn't care which base it is in contact with. So it picks 2 of the 4 bases, which is the same as a 1 in 2 selection and therefore only a 1 bit choice. Shannon figured out how to handle more complicated frequencies; you can read the Primer to see how that works. In any case, it is consistent with what we have covered so far.

Sequence logos. To create a sequence logo (see the top part of the Graphical Abstract) from a set of aligned sequences one builds stacks of letters corresponding to the positions in a binding site (Schneider and Stephens, 1990). The height of a stack is the information (number of bits) needed. So a fully conserved position is 2 bits high and one that is either A or G is only 1 bit high. If one adds up all the heights, one gets the total information in the binding site. In simple binding systems, this total evolves to equal the information needed to find the binding sites in the genome

[1] We neglect the possibility of the coin sitting on its edge.

(Schneider et al., 1986; Schneider, 2000).



Graphical Abstract for LexA sequence logo and one sequence walker.

Individual information. Mathematically a logo stack height turns out to be an average (see the Primer for details). What could this be the average of? Wouldn't it be convenient if it were the average of the individual sequences that make up the aligned sequences? Surprisingly, one can write out an equation for a weight matrix that will evaluate the aligned sequences and assign to each of them a number of bits (Schneider and Spouge, 1997). By definition, the average of all the assignments is the area under the sequence logo. John Spouge proved that there is only one way to do this, so the assignments are unique.

This weight matrix is general and can be used to predict binding sites. Unlike other weight matrices, this one is related to the Second Law of Thermodynamics which says that if the total information of an evaluated binding site is greater than zero, the protein should bind there (Schneider and Spouge, 1997).

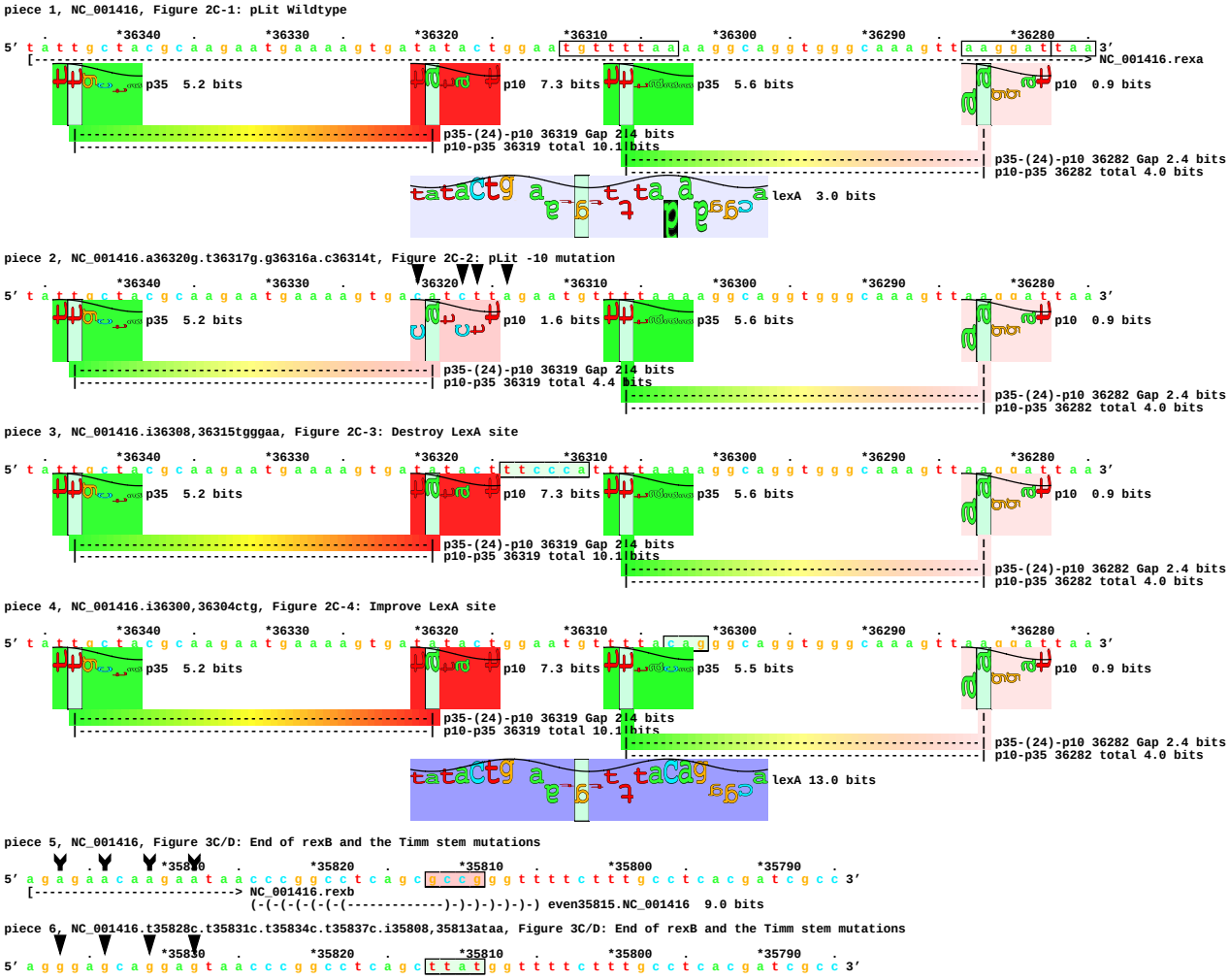
The weight matrix is a model for how a protein binds DNA. To find binding sites, the computer scans the model over the sequences and if the information reported at a position is larger than zero, that identifies a likely binding site. These sites are conveniently shown by sequence walkers.

Sequence walkers. A sequence walker is a graphic like a sequence logo but instead of simultaneously showing a set of sequences, the walker shows just one sequence at a time (Schneider, 1997). So in a walker there is only one letter per position. (See the bottom part of the Graphical Abstract.) The letter height is the information in bits for that base at that position in the binding site, and the sum of the heights is the total information of that particular binding site.

Sequence walkers pack a lot of data into a small picture. When a binding site is symmetrical, like LexA sites that are bound by the dimeric protein, the letters of the walker are oriented up, to indicate binding, and up-side down and down to indicate bad 'repulsive' binding. When a binding site is asymmetric, as in the σ^{70} promoters shown later in this supplement, the letters are rotated 90 degrees so that the direction one would read them downward is the direction the binding site is pointing.

Behind a walker is a colored rectangle called a 'petal'. The rectangle color has three parts: hue, saturation and brightness. The brightness is set fully on. The hue (*e.g.* red or green) specifies the kind of binding site. The saturation (*e.g.* white, pink or red) shows how strong the site is relative to the strongest possible site that the matrix can evaluate (the consensus sequence).

P_{LIT} promoter mutations



Sequence walker promoter analysis. The first 4 sequences above (labeled 'piece's) are from λ (NC_001416). They correspond to Figure 2C. The fifth and sixth sequences correspond to Figure 3C/D. Binding sites are displayed as sequence walkers (Schneider, 1997) using individual information theory (Schneider and Spouge, 1997). The software is freely available from `https://alum.mit.edu/www/toms`. For piece 1, the P_{LIT} promoter was located using a flexible two-part σ^{70} model (Shultzaberger et al., 2007) in which the 5.2 bit -35 walker ('p35') has a green rectangle behind it ('petal') and the 7.3 bit -10 ('p10') has a red petal. A horizontal colored bar shows that they are connected. The variable distance 'Gap' between these two components costs 2.4 bits so the total information of the site is $5.2 + 7.3 - 2.4 = 10.1$ bits. Each letter in a sequence walker is aligned with the corresponding sequence above and has a height in bits, as in a sequence logo (Schneider and Stephens, 1990), showing in detail how the model reacts to that particular sequence (sequence walkers are a logical extension of sequence logos). The top of the rectangle is at 2 bits and the bottom at -3 bits, with the letters resting on a line (not shown) at 0 bits. Positive heights contribute to binding while negative heights detract from binding. The zero base of the coordinate system of a binding site is indicated by a light green vertical bar.

LexA sites (purple petals) were identified using a LexA model built from 23 experimentally proven natural sequences and their complements (Papp et al., 1993). Notably the up-side-down A at base +6 (relative to the zero coordinate indicated by the light green bar on the walker) of the wild-type LexA site has a black background, indicating that an A at +6 was not found in the genomic *E. coli* LexA sites. Given only 23 sequences, the estimated weight for the A is −3.6 bits but since we have shown that this is a functional site, the value should be an unknown amount higher. So the total site is likely to be at least $3.0+3.6=6.6$ bits. This is lower than the average *E. coli* LexA site $R_{\text{sequence}} = 21.0 \pm 4.5$ bits (the area under a sequence logo) but positive and so should be functional according to the second law of thermodynamics (Schneider and Spouge, 1997).

Piece 1, Figure 2C-1: Wild-type P_{LIT} sequence showing the 10.1 bit P_{LIT} promoter and a 3.0 bit LexA site. Since the LexA site is functional, it is likely to be stronger than 3.0 bits as an A position +7 is apparently acceptable to LexA even though that variation was not seen in proven *E. coli* LexA sites. The end of the sequence has the *rexA* stop codon TAA indicated by a box. The other two boxes on the sequence indicate the promoter proposed by others (Landsmann et al., 1982; Pirrotta et al., 1980). That promoter has only 4 bits of information so since σ^{70} promoters are 69% efficient (data not shown) the alternative proposed promoter should be $2^{(10.1-4.0)/0.69} = 458$ fold weaker, essentially undetectable.

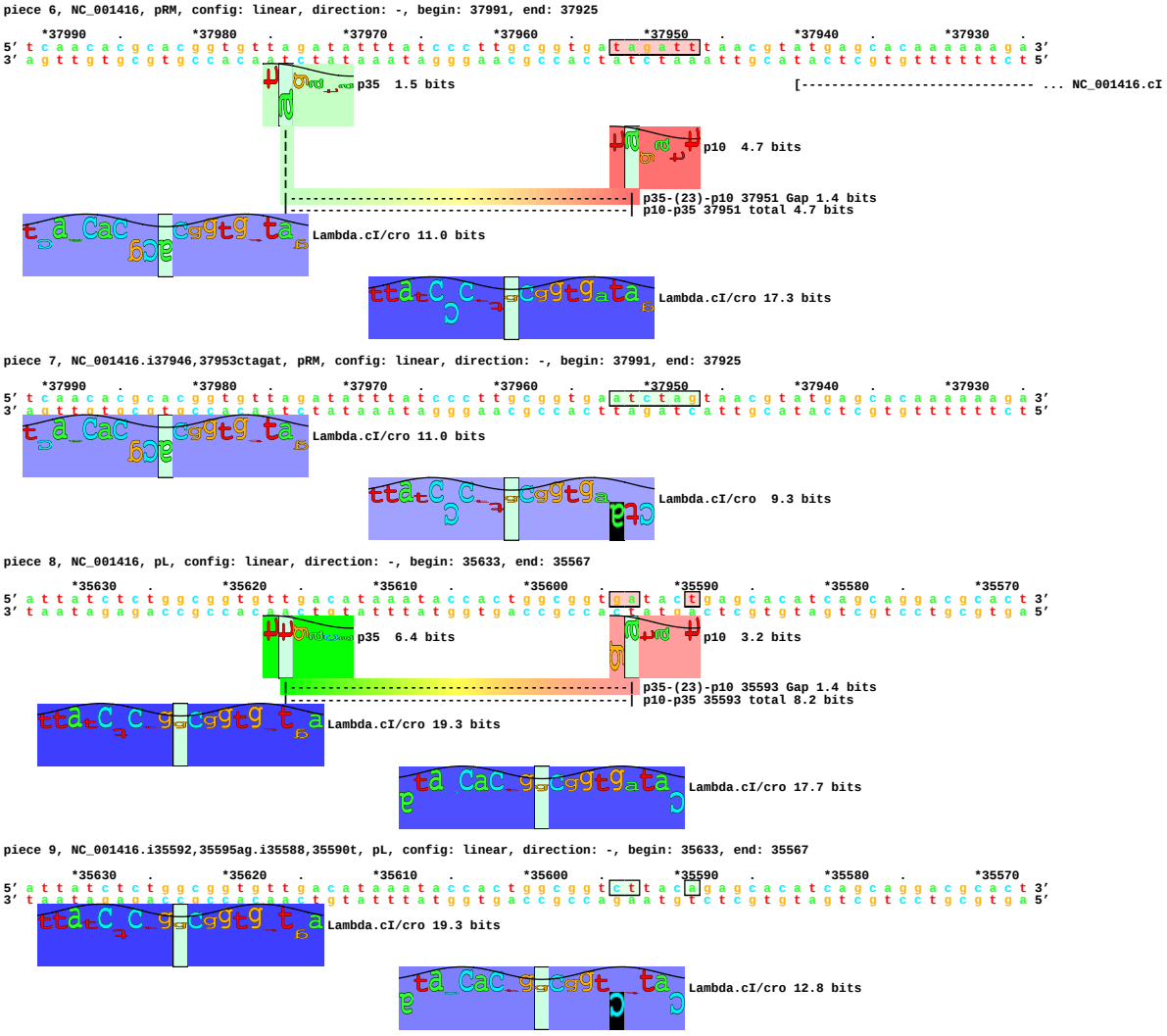
Piece 2, Figure 2C-2: Mutation in the −10 region, as indicated by the black triangles, reduces the P_{LIT} promoter to 4.4 bits which should be $2^{(10.1-4.4)/0.69} = 307$ fold weaker than wild-type. This also destroyed the LexA site.

Piece 3, Figure 2C-3: Intentional destruction of the LexA site (to below 0 bits) without affecting P_{LIT} , as indicated by the green box, generated a second −10 2 bp downstream that uses the same −35 . This site should be $2^{(10.1-5.3)/0.69} = 124$ fold weaker than wild-type, so for clarity it is not shown.

Piece 4, Figure 2C-4: Improving the LexA site (boxed sequence) to 13 bits. The saturation of the purple petal is darker than that in piece 1 to indicate the increased strength.

Pieces 5 and 6, Figure 3C/D: The end of the *rexB* gene (arrow) is followed by the T_{IMM} stem palindrome shown using nested parenthesis. Black arrow tails and heads indicate the mutations made in the end of *rexB*. The mutation that destroyed the palindrome is indicated by pink and green boxes.

P_{RM} and P_L promoter mutations



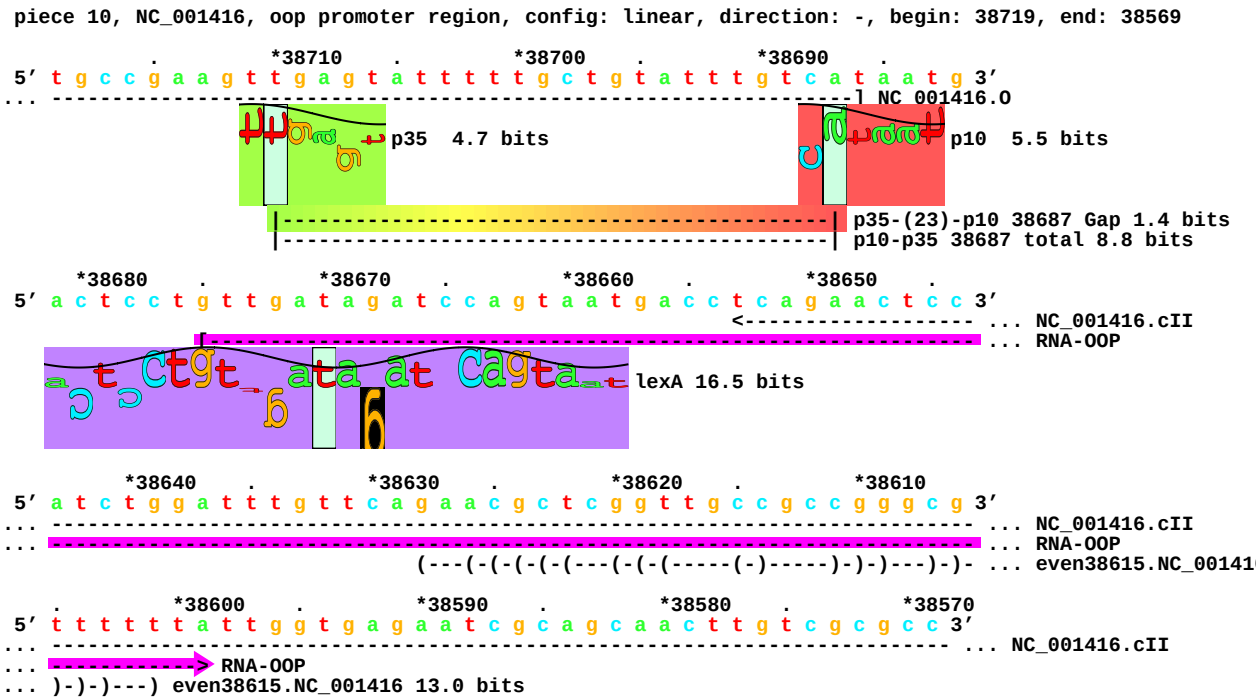
Piece 6: The 4.7 bit P_{RM} σ^{70} promoter walker is shown with connected -35 (green petals, p35) and -10 (red petals, p10). A second previously unrecognized 4.9 bit σ^{70} promoter is predicted 6 bases downstream of both the -35 and -10 (for clarity, it is not shown). The ATG start and coding region of the *c*I gene is indicated by a bracket and dashes. The -10 region sequence about to be mutated is boxed in pink. The λ CI and cro site walkers (Lambda.cI/cro) (Papp et al., 1993) are shown with purple petals.

Piece 7: Mutation in the -10 region removed the 4.7 bit P_{RM} promoter but left the 4.9 bit predicted promoter (not shown), the mutated bases are indicated with a green background box on the sequence.

Piece 8: The P_L promoter before mutation.

Piece 9: The P_L promoter after mutation.

P_O promoter



Piece 10: The P_O promoter region of λ (NC_001416). The σ^{70} P_O promoter −35 (p35, green walker) and −10 (p10, red walker) is shown overlapping the beginning of the λ *O* gene which is translated in the opposite orientation. The P_O mRNA is marked with a dashed arrow in magenta. The P_O mRNA terminates with a hairpin loop followed by 6 T residues and an A. The end of the *c*II gene terminates in the middle of P_O in the anti-sense direction. A 16.5 bit LexA binding site overlaps the 5′ end of P_O mRNA. Position +2 of the LexA site (relative to the zero coordinate indicated by the light green bar on the walker) has an upside down 'g' with a black background to indicate that there are no G residues at that position in the original sequences from which the LexA model was derived. The estimated weight for the G is −3.6 bits but since this is a functional site the value should be an unknown amount higher. So the total site is likely to be at least $16.5 + 3.6 = 20.1$ bits. This is comparable to the average *E. coli* LexA site $R_{\text{sequence}} = 21.0 \pm 4.5$ bits (the area under a sequence logo).

References

Landsmann, J., Kroger, M., Hobom, G., 1982. The *rex* region of bacteriophage lambda: two genes under three-way control. Gene 20, 11–24, https://doi.org/10.1016/0378-1119(82)90083-X.

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, https://doi.org/10.1006/jmbi.1993.1501 https://alum.mit.edu/www/toms/papers/helixrepa/.

Pierce, J. R., 1980. An Introduction to Information Theory: Symbols, Signals and Noise, 2nd Edition. Dover Publications, Inc., NY, http://store.doverpublications.com/0486240614.html, https://www.amazon.com/gp/product/0486240614/, https://archive.org/details/symbolssignalsan002575mbp.

Pirrotta, V., Ineichen, K., Walz, A., 1980. An unusual RNA polymerase binding site in the immunity region of phage lambda. Mol Gen Genet 180, 369–376, https://doi.org/10.1007/BF00425850.

Schneider, T. D., 1997. Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. Nucleic Acids Res. 25, 4408–4415, https://doi.org/10.1093/nar/25.21.4408, https://alum.mit.edu/www/toms/papers/walker/, erratum: NAR 26(4): 1135, 1998.

Schneider, T. D., 2000. Evolution of biological information. Nucleic Acids Res. 28, 2794–2799, https://doi.org/10.1093/nar/28.14.2794, https://alum.mit.edu/www/toms/papers/ev/.

Schneider, T. D., 2013. Information theory primer, with an appendix on logarithms. Published on the web 2013, https://doi.org/10.13140/2.1.2607.2000, https://alum.mit.edu/www/toms/papers/primer/.

Schneider, T. D., Spouge, J., 1997. Information content of individual genetic sequences. J. Theor. Biol. 189, 427–441, https://doi.org/10.1006/jtbi.1997.0540, https://alum.mit.edu/www/toms/papers/ri/.

Schneider, T. D., Stephens, R. M., 1990. Sequence logos: A new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100, https://doi.org/10.1093/nar/18.20.6097, https://alum.mit.edu/www/toms/papers/logopaper/.

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, https://doi.org/10.1016/0022-2836(86)90165-8, https://alum.mit.edu/www/toms/papers/schneider1986/.

Shannon, C. E., 1948. A Mathematical Theory of Communication. Bell System Tech. J. 27, 379–423, 623–656, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x https://doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Shannon, C. E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, https://books.google.com/books?id=qQHFoQEACAAJ.

Shultzaberger, R. K., Chen, Z., Lewis, K. A., Schneider, T. D., 2007. Anatomy of *Escherichia coli* σ^{70} promoters. Nucleic Acids Res. 35, 771–788, https://doi.org/10.1093/nar/gkl956, https://alum.mit.edu/www/toms/papers/flexprom/.

Zhang, A. P., Pigli, Y. Z., Rice, P. A., 2010. Structure of the LexA-DNA complex and implications for SOS box measurement. Nature 466, 883–886, https://doi.org/10.1038/nature09200.

version = 1.20 of plitsup.tex 2019 Aug 28