THE WORLD'S GONE MAD! (OR CAN INFORMATION BE NEGATIVE?)

Information theory and a priori entropy

If you are familiar with information theory as applied to transcription factor-binding site analysis (if you are not, please read this for an introduction to the field), you will probably have noticed that mutual information, as expressed in sequence logos, decreases whenever we decrease the *a priori* (or genomic) entropy H_{before}. This is a natural result of the derivation of mutual information:

$$I(l) = R_{sequence}(l) = H_{before}(l) - H_{after}(l) = \left[-\sum_{S \in \Omega} \left(f(S) \cdot (\log_2 \left(f(S) \right) \right) \right] - \left[-\sum_{S_l \in \Omega} \left(p(S_l) \cdot \log_2 \left(p(S_l) \right) \right) \right] \quad (1)$$

Where,

$$H_{before}(l) = -\sum_{S \in \Omega} \left[f(S) \cdot (\log_2 \left(f(S) \right) \right] \quad (2)$$

is the *a priori*, or genomic, entropy and

$$H_{after}(l) = -\sum_{S_l \in \Omega} \left(p(S_l) \cdot \log_2 \left(p(S_l) \right) \right) \quad (3)$$

is the *a posteriori* entropy (the entropy once we know our protein binds the site).

The intuitive interpretation of mutual information is as follows. If we are given a sequence from a genome and we don't know anything else beforehand, our best guess at what base occupies a particular in the sequence will be based on the genomic background frequencies for each of the four DNA bases. In other words, our uncertainty about the base occupying each position will be large and placated only by strong biases in the genomic composition (which will give us additional insight to make our call in the same way knowing that a die is biased will provide us with better chances at betting). This is the so called *a priori* entropy.

The picture changes if we are told that a certain protein binds that particular sequence. Since proteins can be quite picky in the sequences they chose to bind to, our guess can now be more informed. If proteins were perfect recognizers, our uncertainty would in fact decrease to zero. We would be completely sure of what base occupies each position in the sequence. But proteins are still messy biological entities and their recognition process is noisy by definition. This noise is measured by the *a posteriori* entropy, which is the residual amount of doubt we still have after knowing that the protein binds our sequence. The difference, of course, is the decrease in uncertainty that our knowledge of the protein binding the site has provided us with. Or, in other words, the information we gain by knowing that a certain protein binds that particular sequence.

Base distribution and a priori entropy

In genomes with nearly equiprobable composition, the *a priori* entropy is close to its maximum, 2. This is easy to see for the *Escherichia coli* genome. Its relative base frequencies are:

%A	%T	%C	%G
24.6	24.6	25.4	25.4

And the *a priori* entropy is:

H(Eco)= 0.246·2.023 + 0.246·2.023 + 0.254·1.977 + 0.254·1.977 = **1.999** bits

When the genomic frequencies *f(S)* become biased, either towards high or low %GC content, the *a priori* entropy H_{before} becomes smaller. This is because there is less variability in the genomic background, so less uncertainty (i.e. information). We can see this for the *Thermus aquaticus* genome, which has adapted to extremely hot environments by reaching a GC content of 69.2%. In this case, the relative genomic frequencies are:

%A	%T	%C	%G
15.3	15.3	34.6	34.6

And the *a priori* entropy becomes:

H(Taq)= 0.15·2.73 + 0.15·2.73 + 0.35·1.51 + 0.35·1.51 = **1.876** bits

So how does this affect mutual information?

Fabricating negative information

Suppose we have a protein, named BUH, which recognizes the following targets in the *E. coli* genome:

ATGACATCAT	ATTCGCTAAT	ATTGCGAGAT	GTGTGATCAT	ATGTTGCCAG
ATGCGACAAT	GCTAGCTCAG	ATGCTGATAT	GTACTGACAT	ATGAGATTAT
ATGCTGCCAA	TAGCTAGCAT	TTGTGATGAT	ATGCATTCAG	ATCAGACCAT
ATGCGATAGG	ATCGCGCCAT	TTAGCATGCC	ATGAATACTT	ATGACAGCAT
ATCGACGTAC	ATCGCTACAT	ATTGCATCAG	ATGGACCCCT	ATGATGACTT

Table 1 – List (or collection) of binding sites for the hypothetical protein BUH.

We can now derive the PSFM for BUH:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0.76	0.04	0.08	0.28	0.12	0.44	0.24	0.12	0.80	0.04
C	0.00	0.04	0.12	0.32	0.28	0.12	0.28	0.68	0.08	0.04
T	0.12	0.92	0.16	0.16	0.28	0.12	0.40	0.08	0.08	0.68
G	0.12	0.00	0.64	0.24	0.32	0.32	0.08	0.12	0.04	0.24

Table 2 – Position Specific Frequency Matrix for transcription factor BUH.

Using the *E. coli* background frequencies and the BUH frequency matrix, we can now compute both H_{before} and H_{after} to obtain the mutual information ($R_{sequence}$) at each position of the BUH motif:

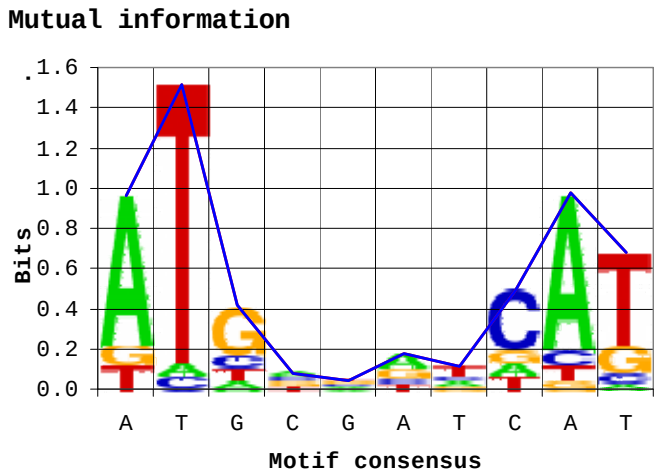

Figure 1 – Mutual information plot for the BUH binding motif using the collection in **Table 1** as input and the *E. coli* genome to compute background frequencies.

The global information content of the BUH binding motif is `0.96 + 1.52 + 0.42 + 0.08 + 0.04 + 0.17 + 0.12 + 0.49 + 0.97 + 0.68 =` **`5.45`** bits. But what happens if we now switch to *T. aquaticus*?

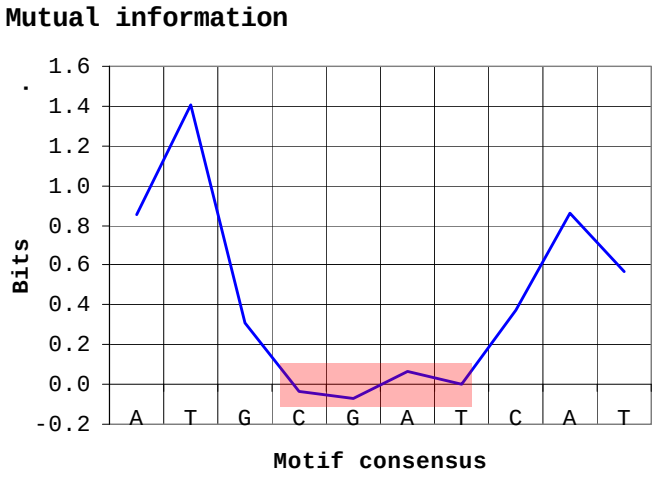

Figure 2 – Mutual information plot for the BUH binding motif using the collection in **Table 1** as input and the *T. aquaticus* genome to compute background frequencies.

Whops! Now some positions (4 and 5) are negative! Indeed, the whole information content of BUH has gone down to `0.85+1.41+0.31-0.03-0.07+0.06+0.00+0.38+0.86+0.57=`**`4.34`** bits. But if mutual information represents an information gain (or a decrease in uncertainty), what does this result mean? Has our uncertainty over certain positions *increased* by observing binding of BUH to the site? Have we *lost* information on those positions by observing that BUH binds there?

This, clearly, does not make a lot of sense. We cannot be *more* uncertain of what base occupies a position in a given site *after* knowing that a protein binds there (this is what negative information content actually means). It is thus surprising to find that many scientists apparently believe this is indeed possible, based probably on the single fact that it is the result of a well-defined mathematical expression (see, for instance, some lecture notes at the Broad Institute). In fact, negative information has been a long-standing, albeit informal, argument for proposing other informational quantities, such

as the Kullback–Leibler divergence (Stormo 1998), to measure information in binding sites (Erill and O'Neill 2009)[1].

Notice that in the above example there are two different and distinguishable effects. On the one hand, the global information content for the BUH motif goes down. This is right, and to be expected, because the *a priori* uncertainty for any position in *T. aquaticus* is lower, and thereby our information gain must also be lower (i.e. if a die is biased, we can't be as surprised by the outcome of a roll as we would be by the outcome of a roll from an unbiased die)[2]. On the other hand, the information content of certain positions has become negative. This is of course completely bananas (mutual information cannot be negative in a realistic setting) and is the subject of this treatise.

Evolution and negative information

The key to the fabrication of negative information in binding sites, as with many other mishaps in bioinformatics, is to loose sight of the strongest single unifying principle in biology: evolution. What we did above with the BUH sites is akin to crossing a zebra with a zebra fish and to then expect the offspring to swim nicely. We *transplanted* the BUH motif, defined in *E. coli*, into *T. aquaticus* and, alas, we got negative information! What is wrong with this line of reasoning, which dates back to the early days of information theory and binding sites (Schneider, Stormo et al. 1986), is to assume that motifs can be transplanted. Motifs, as everything else in biology, evolve, and all sort of quirky things can happen if we ignore evolution.

Demystifying negative information

In *E. coli*, the genomic frequency of all four bases is approximately 0.25. If we presume that there is something akin to "unimportant" positions in a site (i.e. those without positional conservation), we must assume that these positions are under no positive selection to maintain any kind of offset from the genomic background. In other words, we expect them to have, on average, background genomic

[1] If anything, the advent of negative information should be taken as an argument supporting $R_{sequence}$ as the *correct* measure for information content in a binding site. The fact that, when the most fundamental assumption in biology (evolution) is violated by transplanting binding motifs from one species to another, $R_{sequence}$ yields apparently incongruous results should not be interpreted as an erroneous result that needs to be addressed, but rather as a red flag signaling that we are really messing things up.

[2] It has been argued that a protein has no knowledge of the genomic background frequencies and that, therefore, a value of 2 could be used by default in lieu of the formal H_{before}. This is, however, difficult to justify in the light of evolution. We must not forget that a protein acting on any genome will have evolved with and within that genome. Therefore, we can assume that the genomic background entropy if needs to face has been somehow hardwired by evolution into a sort of protein instinctive knowledge of the milieu in which it must operate.

frequencies (0.25). In *T. aquaticus*, however, the *a priori* frequencies are biased towards high %GC. If, again, we presume that some motif positions are not under positive selection to be conserved, then we would expect them to follow the genomic background towards high %GC. We might argue that binding site regions require *breathing* regions and that we should expect positive selection towards *low* %GC (i.e. against the genomic bias). That is well and fine and might indeed be true for many motifs (Erill and O'Neill 2009). But it will not yield negative information. AT-rich positions, if positively selected for, will have lower entropy than the %GC-rich background[3]. Thus, what we should *never* expect to see is a site being *positively* selected to move away from the genomic background towards a higher entropy state (less conservation; equiprobable bases). This is because high entropy is precisely the result of *relaxed* (or absent) selection. There is no conceivable mechanism by which positions in a binding motif might deviate from the background (a fact that implies *de facto* selection) towards a higher entropy state.

And that is all, no more and certainly no less

To summarize, and to put it in very layman terms, if a site position is truly unimportant, it will not be under positive selection and it will remain at the background genomic frequencies (i.e. *zero* mutual information). Otherwise, it will be selected for a purpose and it will deviate from genomic frequencies to a more specific state (i.e. *positive* mutual information). *Negative* mutual information in binding site positions is simply the result of scientists forgetting about the single unifying principle of biology: evolution by natural selection.

[3] Some readers might argue that, in an 80% GC-rich genome, a 50% GC, equiprobable stretch will be distinct from its background and, therefore, that such stretches might be positively selected for in order to make part of (or even constitute the whole of) a binding site. This might sound appealing and appear correct in theory, but it is both a fallacy and impossible in practice. Indeed, one can easily create a program that locates equiprobable stretches in an 80% GC-rich genome and it will work like charm. So, there is information in the high-entropy states! Yes, there is, if they are put in the proper context (one might also argue, even more convincingly, that the information resides instead in the low-entropy states that make up the rest of the genome). In any case, the only caveat of our otherwise very nice program is that it needs to carry out a type of computation that cannot be easily performed by a protein. The only way to recognize an equiprobable stretch is to make a local computation of its base composition and then comparing it to the background composition. The last part (comparing a local composition to the background one) is relatively easy to achieve by nature, as evolution might shape a molecule to have an intrinsic "knowledge" of the genomic background. The first part, however, is an altogether different matter. Computing the local base composition of a stretch implies independently recognizing each base at each position and computing the average frequency of each base over the whole stretch. The problem is that the only way in which a protein can recognize a position regardless of what base that occupies it is by binding non-specifically, which will prevent it from recognizing the base, and thus from computing the average base composition. Even if a reader came up with a feasible way in which a protein might recognize such an equiprobable stretch, it still would not matter. The averaged nature of the measurement involves invariably forgoing the positional independence assumption in which we based our measure of information content, so the whole concept of information content would have to be reformulated before embarking anew in a discussion of negative information. My bet? Negative information would not be possible in the new formulation either.

References

Erill, I. and M. C. O'Neill (2009). "A reexamination of information theory-based methods for DNA-binding site identification." BMC Bioinformatics **10**(1): 57.

Schneider, T. D., G. D. Stormo, et al. (1986). "Information content of binding sites on nucleotide sequences." J Mol Biol **188**(3): 415-31.

Stormo, G. D. (1998). "Information content and free energy in DNA--protein interactions." J Theor Biol **195**(1): 135-7.