How much information does DNA instantiate?

Written by Keith Farnsworth, based loosely on the paper by Chris Adami, 2004

The meaning of Information in Molecular Biology

To answer our question, we first need to be clear about the definition of information we shall be using. This article explains how we quantify the the information-related content of a DNA or an RNA polymer of known length, solely in terms of the pattern (i.e. sequence) of nucleic acid bases (monomers), which will be considered here as indivisible wholes. More specifically, using the Floridi concept of information (based on the number of differences), we are essentially counting the total difference among nucleic acids in the polymer. We will then go on to consider the way this instantiated information interacts with other information patterns to give it 'meaning' in the sense we define on this website: meaningful information is functional. Technically, this amounts to saying it shares mutual information with at least some other pattern: the two patterns then give context to one another and some consequence may arise from the interaction. Our starting point - the total difference embodied within a system - is not what most scientists in the field of bioinformatics would call information. They may say it is *potential* information but most likely, they would agree with Chris Adami, would calls it *entropy.* To quote from Adami (2004):

> "Only in reference to another ensemble can entropy become information, i.e., be promoted from useless to useful, from potential to actual. Information therefore is clearly not stored within a sequence, but rather in the correlations between the sequence and what it describes, or what it corresponds to" ... "According to the arguments advanced here, no sequence has an intrinsic meaning, but only a relative (or conditional) one with respect to an environment".

The 'information instantiated in a piece of DNA' is Order Zero Information (I^0) and the Information to which Adami refers, that which has a context, so that it may make a difference, is what we call Order One Information (I^1). Of course, this is just a matter of labels, but the distinctions are important. We interpret information (of *any* order) as a selection from a range of possibilities, recalling that instantiating information in a physical object is equivalent to filtering, i.e. selecting from a random population of possible forms for the object.

The amount of information (in bits) is directly related to how much selection is needed to end up with this particular instantiation of the object. It is equivalent to the number of ways the object could be as *this* particular instantiation, compared with how many ways it could be as *any* instantiation. A complementary interpretation, which we may also bear in mind, helps to understand that better. In the game 20 questions a player must identify a person or object by asking a series of questions to which the answer can only be the binary yes or no. Since each answer to a binary question contributes one bit (assuming it is independent of the any previous answer), then after N questions, the player has N bits of information. So imagine a DNA sequence such as A-T-G: how many questions would you need to identify it? The answer is the same as the amount of information embodied in the sequence. Lets now work that out.

Basic calculation

Both DNA and RNA use an alphabet of 4 characters. Each character is a base, which means it is one of four possible amino acids: A, G, C, T for DNA and A, G, C, U for RNA. We can consider a molecule of either RNA or DNA (in general an information polymer) as a system comprising a string of N bases. The possibility number (P_N) of a system counts how many ways it can be arranged and it is this number that we must use to find the information content of the molecule. With just one base (not a polymer, but a monomer) obviously $P_1 = 4$. What about a dimer (two bases). For the first base, we have 4 possibilities and for each of them there are also 4 possibilities arising from the second base, so taking the whole dimer at once, there are $4 \times 4 = 16$ possibilities and $P_2 = 16$. This is the same basic maths used for working out probabilities for throwing more than one dice and it is easy to generalise. For three bases (a trimer) there are $4 \times 4 \times 4 = 4^3$ possibilities, so $P_3 = 64$ and in general $P_N = 4^N$. Now we have to turn this possibility number into a measure of the information content using the concepts suggested at the beginning. For one base, there are 4 possibilities, so 2 binary questions are needed to identify which is present (A, G, C or U). The reason it is not 4 questions is that we are specifying the minimum number, assuming the most efficient (clever) questioner. Here, you might ask 1) is it above C in the alphabet; 2) if the answer was yes, is it G?; if the answer was no, is it A? That's just 2 questions. Assuming each monomer in the sequence is independent of the previous, then we would have to repeat this operation N times for N monomers. Thus the total number of questions will be 2N, which, since these are binary questions for each of which exactly one bit of information is gained, the total information needed is 2N bits. This is how much information has been instantiated in the polymer; we will call it I_N (strictly speaking, it is I_N^0, but we are omitting the order notation for the time being).

Now we should notice something mathematically important about the relation between P_N and the number of bits of information instantiated. As N increases P_N rises by the power of N, whilst I_N increases by N. The mathematical relation between these two scaling laws is the exponent / logarithm transformation, since $\log(x^y) = y \log(x)$. Thus we can say, without loss of generality, that $I_N = k \log P_N$, where k is a constant that depends on the base of the logarithm. Using a base of 2, we have the simplest possible relation:

$I_N = \log_2 P_N$ bits. Since $P_N = 4^N$, we have $I_N = N \log_2(4)$ bits, which works out as $2N$ bits. Even more generally, for any system made up of a set of N independent parts (labelled by an index $i = 1, 2, 3...N$) (e.g. a DNA molecule of N monomers 1, 2,3...N) which can each be in q_i possible states, with probability $p_i = 1/q_i$) (e.g. sequences of bases), the information instantiated in a particular realisation of that system is

$$I_N = -\sum_{i=1}^{N} p_i \log_2 p_i. \tag{1}$$

You should recognise this by now as one of the fundamental conclusions of Shannon in his theory of information and communication. Why the minus sign? Well, because probabilities are never bigger than 1, their logs are always negative, so the sum is always a negative number and that would be counter-intuitive, so we multiply it by -1. Another way of thinking of it is that this information is a (negative) change in information entropy relative to knowing nothing.

Taking account of nucleotide frequencies

So far so simple, but not quite right because the bases do not appear with equal probability, nor quite at random, in real information polymers. We have just used very simple (dice throwing) maths. It assumes equal and independent probabilities at each base in the sequence, so we now have to modify the calculation to take account of the reality that natural DNA and RNA do not have equal and independent probabilities.

To understand why, consider the more familiar case of words, made up from the English alphabet consisting of 26 letters and a space (just ignore the other symbols for the sake of this argument). In real and meaningful written English, letters E and A are far more common than letters X and Z, so, recalling the dice analogy, this time we imagine a 27 sided die, with sides loaded so that it is more likely to come up A or E than X or Z. Analysis of written English yields the following probabilities (expressed in expected frequency of use per thousand) - taken from Zernike (1972; p88):

TABLE HERE

Not only that, but you also know that Q is very nearly always followed by U, also that T is often followed by H in real written English. Thus, not only do letters appear with different probabilities, they are not independent either: sometimes knowing one predicts the next to appear. The reason for this and also the probabilities in the table above is that real English is made up of words, which are themselves a subset of all the possible sequences of letters. For example hdlc is not a word, it is a random sequence of letters that should never be found in real English. So it is with information polymers: the real ones are there for a reason and that reason is that they carry the functional information needed for constructing and maintaining an organism. They encode a specific language (yes, the genetic code), so not all sequences of bases actually make sense. Those that do, occur with different probabilities (just as words do) and to some extent it is possible to predict what base is coming next, following the rules of DNA or RNA syntax (just as U follows Q in English). In molecular biology this predictability is referred to as epistatic correlation (in analogy to the mutual dependence of genes being called epistasis).

Not surprisingly, given the importance of DNA and RNA, these probabilities have been calculated to a good degree of accuracy. Since DNA comes in a coiled pair where every monomer has its complementary partner (G with C and A with T), the monomer frequencies are usually quoted in terms of G+C pairs (note that if you know the proportion of C then you know it of G - considering both DNA strands together - and therefore you also know the proportion of T and A). These frequencies differ among organisms: in human DNA G+C is about 41% of the total, so A and T are each a little more likely than the 1/4 assumed earlier, but it is not much different. Most organisms have a G+C proportion between 30% and 50%, but it varies within the genome too and a lot of this variation is down to foreign introduced sequences (introns) and repeated sequences, the function of which remains unknown. Indeed it may really be what this sort of sequence has traditionally been called: junk.

If we just consider the coding regions, then we can turn to the epistatic correlation. This can be quantified for neighbouring monomers (the ith and $i+1$th); next-neighbours (ith with $i+2$nd) and so on. For nearest neighbours, we are looking at the 16 possible dimers: AA, AC, AG, ..., TG, TT. The probabilities of dimers occurring in real DNA deviates from the expectation of 1/16 each (for example AA tends to be quite common and CG rather rare), but not by a large amount (this is species dependent too). More significant are the frequencies of trimers because they code for amino acids (and have a few punctuation functions too, such as ATG which says 'start'). It turns out that when measured, the frequency of these trimers (e.g. in E coli) vary over a range of about 3 times, with the most common occurring about twice as often as expected at random and the least common about half as often as expected. The most common trimer is AAA and this goes some way to explaining why AA is a common dimer. As a result, every time an A is encountered in a DNA molecule, it is more likely than 1/4 that the next monomer will also be A, and if it is that, then also more likely that the next will be an A. In fact there is a lot more to this and it is all well studied within Bioinformatics. All we need to know at present is that these correlations among monomers in real DNA (also true for RNA) are sufficiently significant to require us to take account of them in estimating the information content.

First let us deal with the different frequencies of occurrence in single monomers (we will go on to account for the epistatic correlations). Recall the dice throwing analogy and note that deviations from equal probabilities of monomers (deviations from $p_i = 1/4$) can be represented through weighted dice. Let the weightings be p_C and $p_A = 1 - p_C$, to represent the frequencies of C and G, then A and T, respectively. If there were just one monomer being described, then you would still have 4 possibilities (A,C,G,T) and, in 20 questions you would still ask if it was above C in the alphabet. But when you asked the second question, you would know one answer is more likely than another. That is, if $p_C > 1/2$ and the answer to the first question was yes, then it is more likely that the monomer is G than T (obviously if the first answer was no, then it more likely the monomer is C than A). To put it simply, knowledge of the biased probabilities (weighted dice) gives you some information before the questions commence. In practice, this information will have been gathered by observation of previous rounds of the game, or in real DNA analysis, observations of monomer frequencies in DNA extracted from cells. The effect of this extra information is much clearer when we extend to a string of DNA: a word of length L monomers. It is easiest to understand this, using the other favourite tool of probability: the bag of coloured balls. Imagine the nucleic acids as coloured balls, e.g. red for A, blue for C, green for G and yellow for T. You have a thousand of them in a cloth bag, with the number of each colour exactly reflecting their frequency of occurrence in the real DNA of say a human (so 205 blue, 205 red, 295 green and 295 yellow). Furthermore the balls are given a serial number 1 to 1000, without reference to their colour. In front of you is a long wooden board with holes down its centre, each hole lets one of the balls nestle in it and these holes are numbered 1 to 1000 to represent the positions that nucleic acids can take along a DNA strand. Without looking, you take one ball at a time out of the bag and place it in the holes in sequence to simulate the making of a DNA strand. The first ball out and placed in position 1 on the board could be any one of ball number 1 to number 1000. With 1000 balls to choose from, there are 1000 ways you could have done this and the possibility number for the first placement is 1000. Now you place the second random ball into hole 2 and there are 999 ways you could have done that (because one has already gone). After 999 balls have been placed, only one remains in the bag, so there is only one way you could place the last ball in the last hole. For every one of the 1000 ways you could have placed the first ball, there are 999 ways you could have placed the second. Therefore there are 1000 x 999 ways you could have filled the first and second hole. For each of these 999000 combinations, there are 998 ways you could have placed the third ball. All in all, then, the number of ways you could place all 1000 balls is $1000 \times 999 \times 998 \cdots \times 3 \times 2 \times 1$, which we write 1000! (one thousand factorial). However, although the balls are given serial numbers in this thought experiment, they are otherwise identical and real nucleotides of the same kind (A or C etc..) are indistinguishable in real DNA. In other words, all the balls of a given colour are effectively equivalent if they represent nucleotides that cannot be distinguished, other than by their molecular species (A, C, G or T). There are 205 red balls, so when you laid out the first red ball (which ever in the sequence that was), there were 205 ways in which you could have done it. Having done that there were 204 ways you could have laid out the second red ball, whenever it came out of the bag, and so on. Using the same maths to combine these, we find that there were 205! ways to lay out the red balls and in nucleotides, all of these are indistinguishable. So we have effectively over-counted by 205! in relation to red balls. Of course we also over-counted by 205! in blue, 295! in green and also in yellow). To correct for each of these over-counts, we must divide, so the true possibility number for laying coloured balls sequentially onto the wooden board is

1000!/(205! × 205! × 295! × 295!).

Let us generalise a little now and consider a DNA strand of length L and nucleotide frequencies p_A, p_C, p_G, p_T. The possibility number for this DNA molecule is:

$$P_L = L!/((Lp_A)!(Lp_C)!(Lp_G)!(Lp_T)!), \tag{2}$$

and the associated total information content is the $\log_2$ of this which, unfortunately, is a bit of a mathematical monster. There is, however a nice way to simplify it using Stirlings approximation (named after an 18th Century Scottish mathematician) which states that:

$$\log_e N! = N \log_e N - \frac{1}{2} \log_e(2\pi N)N, \tag{3}$$

and with large N, this is very close to $N \log_e N - N$ (see http://hyperphysics.phy-astr.gsu.edu/hbase/math/stirling.html) Note also that we can easily convert between $\log_e$ and $\log_2$, by dividing by $\log_e 2$ if we want to

So, we can safely take the log of Equation 2 to find the information content of the DNA molecule, taking account of differences in the frequency of nucleic acids. Using Stirling's approximation to do that, we get:

$$\frac{1}{\ln 2} I_L = (L \ln L - L) - 2L[(p_A \ln(Lp_A) - p_A + p_C \ln(Lp_C) + p_C)], \tag{4}$$

(noting that ln is shorthand for $\log_e$).

The right hand side of this equation is multiplied by $1/\ln 2$ to change the log base to 2 for information in bits. All the terms in the square bracket had L in front of them, so we took that out as a common factor and also we noted that, due to base-pairing, $p_A \ln(Lp_A) - p_A = p_G \ln(Lp_G) - p_G$ and $p_C \ln(Lp_C) - p_C = p_T \ln(Lp_T) - p_T$, so we just count 2 lots of terms in A and C, hence the 2 outside the square bracket. Next we note that $2(p_A + p_C)$ must equal 1, since this is the total probability of *any* base appearing: $p_A + p_C + p_G + p_T = 1$.

The information content now simplifies to:

$$\frac{1}{\ln 2} I_L = L[(\ln L - 1) - 2(p_A \ln(Lp_A) + p_C \ln(Lp_C)) + 1], \tag{5}$$

The 1s cancel and we now expand what is in the logs:

$$\frac{1}{\ln 2} I_L = L[\ln L - 2(p_A \ln p_A + p_A \ln L + p_C \ln p_C + p_C \ln L)], \tag{6}$$

but $2(p_A \ln L + p_C \ln L)$ is just $\ln L$, because again, $p_A + p_C + p_G + p_T = 1$. The $\ln L$ terms cancel, so we now find that Equation 2 simplifies to:

$$\frac{1}{\ln 2} I_L = 2L(p_A \ln p_A + p_C \ln p_C). \tag{7}$$

The information per base is this divided by the number of bases L. If we convert the right hand side into $\log_2$, by dividing it by $\ln 2$, which cancels with the same term on the left, we get:

$$I_L = 2(p_A \log_2 p_A + p_C \log_2 p_C), \tag{8}$$

and as a check, if we let $p_A = p_C = 1/4$, you can see that the terms in the brackets become $1/2 \log_2(1/4)$, so the right hand side reduces to -2 bits per monomer, (noting that we interpret the negative sign as information gained: see our comments about Equation 1). If you substitute the nucleic acid frequencies found in human DNA into equation 8, you will find a little less than 2 bits per monomer (but it only falls short by less than three percent).

The random expectation of information instantiated per base is 2 bits, but non-uniform monomer frequencies reduce that. The randomly expected information content provides us with a yard-stick (a metre rule) to compare with and it is called a '*mer*'. So you expect no more than 1 mer of information per base, but might well get less. The idea of a mer generalises, for example with a protein, which is a tangled string of amino acid molecules (forming a peptide polymer), there are 20 possible amino acids, so one mer is $\log_2(20) = 4.32$ bits.

Epistatic correlations

Now let us consider the issue of epistatic correlation: knowing a base in one position along the molecule may change the probabilities of bases in other positions. First, note that if the nucleotide code was perfectly efficient (as compressed as possible, to record the maximum information in a given length), then there would be zero epistatic correlation. Neither DNA nor RNA are quite that efficient, but they are not far off and certainly a lot better than written English. Of course the most obvious form of correlation arrises simply from the Watson-Crick base-pairing in DNA: we have not really mentioned it, but in a double helix, each nucleotide is entirely correlated with its partner on the other side of the helix, so taken as a whole, there is half the information in a whole double helix strand as there are nucleotides. Single strands tend to wrap into knots as nucleotides are attracted to one another, forming hydrogen bonds that are quite predictable. For example, the transfer RNA (tRNA) molecule wraps itself into a shape which forms some tracks, with nucleotide bases running parallel to one another so that adjacent bases are bonded in Watson-Crick pairs.

In Adami (2004), this is taken as an example to illustrate the resulting epistatic correlation. If you separate the two strands of a DNA double helix, you have all the information in each of them separately. One of the strands can easily make a perfect copy of the other and this is of course the foundation of biological reproduction. The strands therefore each instantiate all the information needed to reproduce their partner and we can say that they bind perfectly because they are perfectly correlated. If you took two random DNA strands of length L and tried to get them to form a double helix, it would not work at all, because the strands, being random, would be uncorrelated. Never the less, if you laid them side by side, bits of the length would by chance be in Watson-Crick parings and the molecules would bind over these regions and we could say that here, the molecules were correlated. Let us call one of the molecules X and the other Y. The Information (I^0) embodied in each we can call $H(X)$ and $H(Y)$ (where I am using the symbol H in deference to Chris Adami's interpretation of I^0 as information entropy: *intropy* as some would say). The information (I^1) that one strand has *about* the other is:

$$I(X : Y) = H(X) + H(Y) - H(XY), \tag{9}$$

where $H(XY)$ is the information (I^0) instantiated in the whole system of two molecules, X and Y, which you could think of as the I^0 of a molecule formed by joining Y to the end of X. Thus equation 9 reads: the *mutual information* of X and Y is the intropy in X, plus the intropy in Y, minus the intropy in both together. If we now just take a piece of X and Y over which Watson-Crick pairing occurs when they are laid side by side (let's say it is of length w), we know that the mutual information over that piece is $w + w - w = w$ mers. That is because each individual strand of length w instantiates w mers of intropy and the combined intropy of both is also w mers because they are perfectly correlated in Watson-Crick pairing (instantiated information (I^0) is half the total number of nucleotides). So the mutual information is exactly what we thought about the partnered strands in a DNA double helix of length L: each has exactly the information (L mers) needed to make the other.

Now think of those two strands X and Y again. Placed side by side, the parts that can bind do so and the molecules join together there. If the total length (in bases) over which they join is w, then the combined molecule instantiates $2(L - w)$ over the part where they don't join and w where they do. Thus their mutual information will be $2L-(2(L-w)+w) = w$ mers and the intropy of the whole molecule will be $2L - w$ mers, so the intropy (I^0) per base in the whole molecule will be $2 - w/L$ mers. This reduction by w/L is the loss of I^0 caused by epistatic correlation.

a temporary note from the author

That's it for now - I will update and this document and add the next (which deals with binding sites and translation) in about 2 months. I intend this document to become a web-page in the style of our website, but have to get the mathematical notation working in HTML first. I also need to solve a couple of small problems with this document: e.g. the table and the web-links.