# Information Theory—part 1: overview of key ideas

*Royal Truman*

### Appendix 1—Yockey's inappropriate use of protein probabilities

Yockey uses the Shannon–MacMillan–Breitmann Theorem to calculate the number of sequences of length N symbols, having an average entropy H:

Nr. of sequences = $2^{NH}$ (3).

This equation neglects 'low probability sequences'. For long enough sequences, there will be only a miniscule probability that some of them will be produced, meaning those using many low-probability symbols.

Yockey calculated the entropy of amino acid distribution implied by the number of their coding codons, applying this idea to variants of cytochrome c with a length N = 111. He estimated the proportion of functional cytochrome c proteins by taking the ratio of

$2^{(111 \text{ x Hcytochrome c})} / 2^{(111 \text{ x Hcodons})}$ (4),

where $H_{\text{cytochrome c}}$ refers to the average entropy at each amino acid position along cytochrome c's primary sequence and $H_{\text{codons}}$ is the entropy of all amino acids based on how many codons code for each one.

Relatively few sequences of cytochrome c were known when Yockey began publishing. Let's examine how he calculated the entropy, H, across all 111 positions.

For each position, the variety of amino acids actually observed was enhanced by postulating which other ones might also be tolerated there. To calculate the entropy –(p)log(p), a probability is needed at each of the 111 positions. Yockey simply weighted the number of potentially acceptable amino acids by the number of codons coding for them.

For example, assume only valine and leucine are acceptable at a single position. 4 codons code for valine and 6 for leucine. At that position on the protein the entropy contribution would be calculated as follows:

Number of acceptable codons: 4 + 6 = 10 (5),

$P_{\text{valine}} = 4/10$ and $P_{\text{leucine}} = 6/10$ (6).

The resulting site entropy is:

$-[(4/10) \log_2(4/10) + (6/10) \log_2(6/10)] = 0.971$ (7).

But is this notion of entropy biologically meaningful? In Shannon’s work, log(1/$p_i$) is the *surprisal,* a measure for how likely a symbol *i* will be received. The surprisals for all symbols are weighted by the probabilities, $p_i$, and summed to calculate the entropy. *Yockey’s approach assumes that if one example of a mutation is observed or presumed acceptable, then that amino acid will be present across all organisms with a probability based on the number of codons coding for it!*

That is absurd. A mutation is a rare event and whether it is tolerated will depend on what other changes have occurred. Several variants could be acceptable individually at different sites, but not mutually. The same applies if the scenario one has in mind is a naturalist origin of proteins: protein folding permits a limited combination of amino acids to be used, and this ensemble has nothing to do with how many codons code for them.

The effect of this conceptual error is to estimate far too large an entropy for the protein family, and thereby much too small a difference between $H_{source}$–$H_{protein}$.

Durston weighted the $p_i$ calculations correctly, by examining all known aligned variants. To illustrate, suppose that at a position on the protein, the dataset includes 999 cases with valine and one with leucine. In Yockey’s approach one always gets result (7) irrespective of the true distribution. Durston’s methodology produces an entirely different value:

$$–[(0.999) \log_2(0.999) + (0.001) \log_2(0.001)] = 0.008 \qquad (8).$$

Yockey calculates too much entropy and therefore much too high a proportion of acceptable/random sequences. The assumption made has no relevance to the probabilities of proteins arising abiotically nor to proteins which may have evolved later.

Truman already pointed out that essentially the same numerical values can be made using Yockey’s $p_i$ estimate, but in a much simpler manner: just multiply the probabilities of getting a suitable amino acid at each position by chance.[1] For example, if only valine and leucine have ever been found at one position, then the probability of satisfying that constraint by chance is 10/61 codons (from 64 possible codons we subtract the three ‘stop codons’). An even simpler estimate is 2/20 amino acids. Over 111 positions the two probability measures were within an order of magnitude of each other and of the result using information theory.

Yockey’s incorrect method to calculate probabilities is serious, resulting in errors of many orders of magnitude over the whole length of a protein. Therefore, performing the complex calculations serves no purpose.

## Appendix 2—Lee Spetner’s contributions

One application of Shannon’s theory is the idea of addressing specificity. A letter addressed as ‘Mannheim, Germany’ is less specific than one addressed ‘Wilhelm-Strasse 123, Mannheim, Germany’. The first location includes the latter, so the coded message to communicate the smaller region will be longer. Longer messages are able to refine the intended outcome more precisely.

A well-regulated gene coding for an enzyme, the activity of which occurs within a very narrow range, represents a subset of possible outcomes. That requires more information than for a poorly regulated gene leading to indiscriminate enzymatic activity (unreliable, sometimes too active, sometimes not enough).

Spetner points out that virtually all, perhaps even all, examples of adaptive mutations represent a loss in information in Shannon's sense. A mutation that sub-optimizes the functionality of a gene (such as destruction of a repressor gene) is a loss of information. And reduction of the specificity of an enzyme is also a loss of information.[2]

A system could display *i* outcomes in response to a message received, each with frequency $f_i$. Information generated is given by the difference between what potentially could have occurred and what actually did. To illustrate,[2] the entropy of an ensemble of *n* bio-chemicals with fractional concentrations $f_1,\ldots,f_n$ is given by

$$H = -\sum_{i=1}^{n} filogfi \qquad (9),$$

and for base 2 logarithms, the units of entropy are *bits.*

The input entropy, $H_I$, for a uniform distribution of n elements is, from (9), given by

$$H_I = -n[1/n \text{ x } \log(1/n] = -\log(1/n) = \log n \qquad (10),$$

since the *fi's* have a value of 1/n.

If the messages always lead to only one outcome, there is no entropy:

$$H_0 = 0, \qquad (11)$$

since -1 x log(1) = -1 x 0 = 0.

The decrease in entropy brought about by choosing one outcome is then the difference between (9) and (10), or

$$H = H_I - H_0 = log\ n \qquad (12).$$

Eqn. (12) shows that if a message is corrupted during transmission, a smaller information gain, *H*, results.

### *Application to an example of deliberate selection*[2,3]

The information content of an enzyme is the sum of many aspects, including:

- level of catalytic activity
- specificity with respect to the substrate
- strength of binding to cell structure
- specificity of binding to cell structure
- specificity of the amino-acid sequence devoted to specifying the enzyme for degradation.

The range of possible outcomes of catalytic activity, substrate specificity and so on, must be constrained through the information provided by genes. Substrate specificity can be used to illustrate.

Ribitol is a naturally occurring sugar that some soil bacteria can live on. An enzyme ribitol dehydrogenase, is the first step in its metabolism. Xylitol is a similar sugar, but does not occur in nature. Bacteria cannot normally live on xylitol, but when a large population was cultured on only xylitol, mutants appeared that were able to metabolize it. The wild-type enzyme could process xylitol, but not enough for the bacteria to live on. The mutant enzyme had an activity large enough to permit the bacterium to live on xylitol alone.

This might seem like an example of neo-Darwinian evolution, whereby an enzyme might arise and replace a former one to permit processing a new available sugar. But Spetner showed that something else was actually going on. The mutant enzyme was now processing ribitol less effectively, xylitol better, and a third sugar, L-arabitol, more. In other words, the mutant is unable to discriminate as well among substrates. Note that the enzyme can also catalyze back reactions, and continuing a trend of loss of specificity would not benefit the bacteria. Straightforward application of the equations (9) and (12) to the mutant case showed[3] a net loss of information had occurred.

Spetner knows, of course, that in principle a mutation could occur which increases information in the Shannon sense. For example, after mutations had degraded genomes for many generations, a back-mutation could occur. But a vast number of information-increasing mutations would have had to occur if neo-Darwinian evolution were true. However, not even a single example has been demonstrated to date:

> "The failure to observe even one mutation that adds information is more than just a failure to support the theory. It is evidence *against* the neo-Darwinian theory. We have here a serious challenge to neo-Darwinian theory."[4]

## Appendix 3—Thomas Schneider's work

In cells, recognizers such as proteins and other bio-chemicals attach to specific DNA-binding sites to regulate various processes. The specific ensemble of nucleobases provides a steric and chemical microenvironment which the conjugate portion of proteins recognize. Proteins which turn genes on or off must identify binding patterns which are typically 10 to 20 base pairs long.[5]

The contribution of each nucleobase to correctly define a binding site can be calculated (see below) using principles of Shannon's Information Theory, and the sum from all nucleotides defines the 'individual information', $R_i$, of individual binding sites.[6]

$R_i$ values are useful for two reasons:

1. Values above zero indicate a binding site of some kind.

2. The value, in bits, can be used to determine how many of that particular kind of binding site are present in a genome. One bit of information permits selection between two choices; two bits between four choices; and n bits between $2^n$ choices.

*Mathematical methodology*. The combination of nucleobases used by known examples of a particular binding site are aligned in a manner which provides the highest overall information content. This dataset is used to create a matrix $f(b,l)$, where the four rows $b$ represent each nucleobase A, C, G, or T. The columns represent each position of the binding site. The cells of the matrix contain the frequency in which each base is found at that position in the dataset.

A weight matrix can now be developed:

$$R_{iw}(b,l) = 2 + \log_2 f(b,l) - e(n) \qquad (13)$$

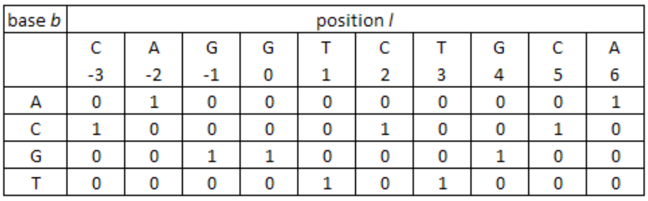
bits per base, where e(n) is a statistical correction factor due to a small dataset. $R_{iw}$ stands for 'Rate of information transmission, Individual Weight'.

> To evaluate a DNA sequence, the bases of the sequences are aligned with the matrix entries and the $R_{iw}(b,l)$ values corresponding to each base are added together to produce the total $R_i$ value.[7]

Note that $R_{iw}(b,l)$ values range between $-\infty$ and 2 bits. This is easy to see from (13). If in the dataset only one nucleobase always appears in a given position, then the frequency will be one, and $2 + \log_2(1) = 2$. If a nucleobase never shows up in the dataset, we obtain a weighting factor $2 + \log_2(0) = 2 - \infty = -\infty$. An example is given in table 1.[8] Adding the bits of information provided by each base for sequence 5' CAGGTCTGCA 3', 0.58 + 1.25 + 1.61 + 1.99 + 1.98 – 3.68 – 1.59 + 1.71 – 0.51 + 0.05 = 3.12 bits for that particular binding site. The total information gained is the total decrease in uncertainty.

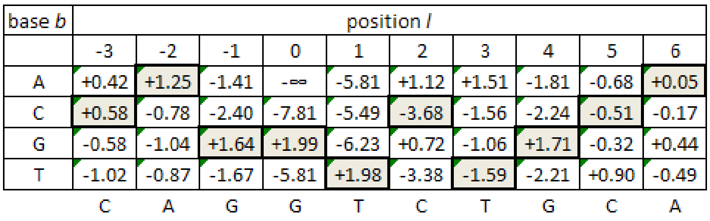
**Table 1.** Sequence (A) and weight matrix (B).

A)



| base *b* | position *l* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C<br>-3 | A<br>-2 | G<br>-1 | G<br>0 | T<br>1 | C<br>2 | T<br>3 | G<br>4 | C<br>5 | A<br>6 |
| A | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| C | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| G | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| T | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

B)



| base *b* | position *l* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| A | +0.42 | +1.25 | -1.41 | -∞ | -5.81 | +1.12 | +1.51 | -1.81 | -0.68 | +0.05 |
| C | +0.58 | -0.78 | -2.40 | -7.81 | -5.49 | -3.68 | -1.56 | -2.24 | -0.51 | -0.17 |
| G | -0.58 | -1.04 | +1.64 | +1.99 | -6.23 | +0.72 | -1.06 | +1.71 | -0.32 | +0.44 |
| T | -1.02 | -0.87 | -1.67 | -5.81 | +1.98 | -3.38 | -1.59 | -2.21 | +0.90 | -0.49 |
| | C | A | G | G | T | C | T | G | C | A |

A) The information content for sequence 5’ CAGGTCTGCA 3’ is to be determined. The position where the base is found is marked with a ‘1’.
B) The individual weight matrices are based on frequencies in which each base is found for known members of a specific binding site. The marked cells show the values assigned to each base in A).

The average $R_i$ for all the binding sites in the dataset used to create the $R_{iw}(b,l)$ matrix is the average information content, $\mathrm{R}_{\mathrm{sequence}}$ of that particular kind of binding site.

The information content of several binding sites has been calculated. For protein Fis (Factor for Inversion Stimulation) it was calculated to be 7.86 bits. The 4,638,858 bp genome of *E.coli* has about 4,289 genes, or about 1,082 bp per gene. This implies that about 1,082/232 = 4.7 Fis sites are present per gene.[9]

For leucine-responsive regulatory protein (Lrp) 10.8 bits of information was reported, based on a dataset of 27 known sites.[10] This protein is known to both activate and repress transcription. Remarkably, there are no major differences in the binding site details to distinguish between activation and repression.[11]

Many other studies of a similar nature are referenced in a recent publication.[12] The number of binding sites on a genome can vary greatly.[13] On *E. coil* there are two *LacI* sites and about 2,574 ribosome binding sites.

The quantitative information extracted from a collection of known binding sites has been used to source for additional sites missed by searches based on a consensus sequence (which only uses the most frequent base found at each position). This approach was used to identify additional binding sites for OxaR transcription factor, which regulates the expression of numerous genes, especially in the presence of damaging hydrogen peroxide.[14]

### *Derivation of equation (13)*8

Using ideas from Shannon’s work, it requires no more than $\log_2 M$ binary decisions to pick one out of M symbols. If the likelihood of each symbol is the same, the probability is P = 1/M of guessing the correct one by chance. We can then write

$$\log_2 M = \log_2(1/P) = -\log_2 P \qquad (14).$$

However, the probabilities may not be equal. If for some reason a particular symbol is more likely (perhaps because it is present more frequently), then we are less surprised if we guessed correctly. The 'surprisal' that symbol *i* is the correct one, $h_i$, is expressed in information theory as

$h_i = -\log_2 p_i$.  (15).

The average surprisal over all symbols for a sequence *L* is given by

$H(L) = \sum_i p_i h_i = \sum_i p_i \log_2 p_i$  (16).

This is how Shannon described uncertainly. H is often referred to as the entropy of a list of symbols.

### *Calculating $R_{sequence}$, the average information content of a binding site*

Now, to a good approximation the sequences of the four DNA nucleotides are random. Since any nucleotides could appear at a particular position, 2 bits of information would be needed to identify each nucleobase (the before state). To illustrate, the number of nucleotides in the composition of an *E. coli* transcript library was calculated as: A = 29,526, C = 25,853, G = 27,800, T = 28,951, from which an entropy H = 1.99817 bits/base results for the genome.[15] However, the variety of bases which can appear at the location of a particular kind of binding site is far more restricted, and the surprisal upon discovering that a particular base is found at each position along the site is usually less than two bits. The decrease in surprisal is given by

$R_{iw}(b,l) = 2 - -\log_2 f(b,l)$  (17).

If a particular base at some position was not found in the data set used to create the weighting matrix, and later a candidate binding site displayed that base, then a value of $-\infty$ would be used (e.g. see table 1 part B) at position 0. In this case a weighting value of $1/(n+2)$ is recommended, where *n* is the number of examples in the data set.[16] Note that for large datasets the frequency of base occurrence approximates its probability. This is equation (13) except for the same sample correction. The total surprisal is the sum of bits using (17) across the binding site.

Another way of doing these calculations is to replace $p_i$ by $f(b,l) + e(n(l))$ in equation (16), permitting the information content of the binding site to be calculated[17] by

$R_{sequence}(l) = 2 - H(l)$  (18).

Note that the information provided by a binding site pattern, $R_{sequence}$, is the decrease in uncertainty from before to after binding[18]:

$R_{sequence}(l) = H_{before} - H_{after}$  (19).

A dataset of 20 or more examples is considered sufficient to provide a reasonable weight matrix for the binding site.[19]

## *Calculating $R_{frequency}$, the frequency of binding sites*

We can define all positions in the genome as G and the number of binding sites as γ. If any position on the genome is acceptable, the number of bits of choice is $\log_2 G$, and for the binding sites, $\log_2 \gamma$. The decrease in uncertainty is given by

$\log_2 G - \log_2 \gamma = \log_2(G/\gamma) = -\log_2(\gamma/G) = R_{\text{frequency}}$ (20).

The name $R_{\text{frequency}}$ reflects the fact that γ/G is the frequency of the sites.

Since enough bits of information are needed to identify the location of binding sites, $R_{\text{frequency}}$ should be similar in value to $R_{sequence}$. An excess of information, such that $R_{sequence} > R_{\text{frequency}}$, is interpreted to mean all, or portions of, the binding site are shared by different recognizers.[20]

The information theory tools developed can be used for additional studies, such as identifying restriction recognition sites. Special enzyme can identify these kinds of locations on viral genomes, cut one or both DNA chains and thus destroy the virus.

Schneider claims concerning binding sites developed by evolutionary processes:

> "The significance of $R_{\text{sequence}}$ is that it reflects how the genetic control system evolves to meet the demands of the environment as represented by $R_{\text{frequency}}$."[21]

In other words, the genes worked fine without the promoter, and random mutations managed to introduce additional binding sites at just the right locations later. He claims using a computer program and paper calculations that this is possible, but this had been thoroughly refuted.[22, 23]

## *Implications*

The notion that a collection of nucleotides could together identify a location address for binding sites is very attractive and congenial to a design perspective. But Schneider's quantitative work does pose some questions. For example, what significance should be attached to the fact that the *average* $R_i$ for all binding sites is supposed to represent the information content of that kind of binding site? For example, in one reference[24] the bits of information in the sixty example datasets ranged between 2.5–15.7 bits, but the information content of the binding site, the overall average, was 7.86 bits. It would be easier to understand if all the examples had very similar bits of information.

In Shannon's work average values are indeed important. Entropies are calculated, but he makes clear that his results are mathematically correct for infinitely long messages. And messages transmitted across communication lines are very long, so the error has no practical significance. But a binding site is very small.

Schneider's work and results are much easier to explain using a designed genome than trial-and-error (evolution).

1) *Initial starting point*. For the *designed scenario*, let us presuppose binding sites have been conceived to be robust against future mutations. To illustrate, consider a binding site like *ACGTACGT*, representing 16 bits of information, or one site per 65,536 nucleotides. In a perfectly random genome of *g* base pairs, *g*/65,536 such locations would be present somewhere. But the nucleotides in a designed genome need not be random, and more sites than predicted by the large number of bits of information can be present. *Mutations could subsequently be tolerated,* lowering the bits of information, as long as the necessary chemical interactions between protein and binding site details suffice to identify the original site. Consistent with this view, up to 15.7 bits of information were reported for binding sites in one small sixty-member dataset, although 9.7 bits were deemed sufficient to unambiguously locate such sites.[24] In another twenty-seven-member dataset, 10.8 bits defined the binding site, but value for the individual sequences ranged between 4.1–17.3 bits.

   Mutations which destroy indispensable binding sites would be eliminated by natural selection, and over time the binding site sequences would randomize as much as permitted to still identify the binding sites.

   It is unlikely this state has been reached yet, which implies that additional sites not discovered yet would also work, including some incorporating mutations. To illustrate, in the first position perhaps *(A|G)CGTACGT* has been found, with some frequency distribution between A and G. Greater entropy at this position could still occur, also through the use of another nucleobase at that position.

   a. This designed model predicts that the binding portion on the protein was tailored to permit participation by many nucleobases and a variety of them, to offer robustness to future mutations. An alternative design, using smaller binding sites, could have been created by permitting only one nucleobase to be acceptable at each position, but these could easily be destroyed later by a single mutation. The more compact strategy would be less fault-tolerant to future mutations.
   b. The designed viewpoint predicts that mutations are increasing the randomness of binding sites and that binding sites may be disappearing over time: the genomes are degenerating. The evolutionary view is the opposite: natural selection has been continuously creating new binding sites at useful locations.
   c. This designed model predicts excess bits of information on average for the purpose of locating the site. It also makes research-orienting predictions. The original binding site designs probably took into account how long the activator should ideally remain attached to control the amount of protein to be produced, and the probabilities of subsequent point mutations.[25,26]

   For the *naturalist scenario*, an organism did fine originally, without the regulatory details provided via a binding site. At some point, a portion of a protein happened to transiently interact with a section of DNA on some organism and this interaction was beneficial. Subsequently, mutations at other nearby locations on DNA would enhance this interaction.

Initially this new binding interaction would have minimal biological functionally, not being part of a highly optimized regulatory process. The resulting benefits need to be greater than the deleterious effects of that protein also binding to a multitude of undesirable sequentially similar sites. A valuable protein should not needlessly be tied up at false binding sites, since the energy and materials costs would not be compensated for by the still suboptimal function which is to evolve.

Since as a rule a large number of the same kind of binding site is found on genomes, the original protein must now mutate in a manner to recognize additional binding locations while not destroying what had already been working, and at the same time avoiding false sites.

This process appears to be absurdly improbable, relying on random mutations as feedstock to natural selection. Several things must occur simultaneously:

- A correct binding location must be recognized by a mutated protein.
- This must result in a new biological function which was not needed before and often requires additional proteins.
- Existing binding sites must remain compatible with the mutated protein.
- Deleterious, false binding sites must be avoided.
- The process must occur for the hundreds or thousands of new binding locations for that single class of binding protein.

a. This model predicts no anticipation for future eventualities. It is very unlikely the same class of binding sites could be used many times and for many purposes.
b. To avoid false binding, the initial binding sites should involve very few nucleobases.
c. There would be no justification for extra robustness in binding sites against future destructive mutations. Natural selection cannot plan ahead.

## References

[1] Truman, R. and Heisig, M., Protein families: chance or design? *J. Creation* **15(3)**:115–127, 2001.

[2] Spetner, L., A Scientific Critique of Evolution, www.trueorigin.org/spetner1.asp.

[3] Spetner, L., *Not by Chance! Shattering the Modern Theory of Evolution*, The Judaica Press, Brooklyn, NewYork, 1998; chapter 5.

[4] Spetner, ref. 3, p. 160.

[5] Schneider, T.D., Some lessons for molecular biology from Information Theory; in: *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, Springer-Verlag, New York, 229–237, 2003; www.ccrnp.ncifcrf.gov/~toms/paper/lessons2003/latex/paper.pdf.

[6] Hengen, P.N., Bartram, S.L., Stewart, L.E. and Schneider, T.D., Information analysis of Fis binding sites, *Nucleic Acids Research* **25**(24):4994–5002, 1997.

[7] Hengen, ref. 6, p. 4997.

[8] Schneider, T.D., Information Content of Individual Genetic Sequences, *J. Theor. Biol.* **189**(4):427–441, 1997; www.ccrnp.ncifcrf.gov/~toms/paper/ri/latex/paper.pdf.

[9] Hengen, ref. 6, p. 4999.

[10] Shultzaberger, R.K. and Schneider, T. D., Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX, *Nucleic Acids Research* **27**(3):882–887, 1999.

[11] Shultzaberger, ref. 10, p. 883.

[12] Schneider, T.D., A Brief Review of Molecular Information Theory, *Nano Commun. Netw.* **1**(3):173–180, 2010; www.ccrnp.ncifcrf.gov/~toms/paper/brmit/brmit.pdf.

[13] Schneider, T.D., Storno, G.D., Gold, L. and Ehrenfeuch, A., The Information Content of Binding Sites on Nucleotide Sequences, *J. Mol. Biol.* **188**:415–431, 1986; www.ccrnp.ncifcrf.gov/~toms/paper/schneider1986/latex/paper.pdf.

[14] Zheng, M., Wang, X., Doan, B., Lewis, K.A., Schneider, T.D. and Storz, G., Computation-Directed Identification of OxyR DNA Binding Sites in *Escherichia coli*, *J. Bacteriology* **183**(15):4571–4579, 2001.

[15] Schneider, ref. 13, p. 9 (on the Internet pdf version).

[16] Schneider, ref. 8, p. 9 (on the Internet pdf version).

[17] Schneider, T.D., New Approaches in Mathematical Biology: Information Theory and Molecular Machines, www.ccrnp.ncifcrf.gov/~toms/paper/trieste1996/trieste1996.pdf.

[18] Schneider, ref. 5, p. 4 (on the Internet pdf version).

[19] Schneider, T.D., Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences, *Nucl. Acids Res.***25**(21):4408–4415, 1997.

[20] Schneider, ref. 17, p. 7 (on the internet pdf version).

[21] Schneider, ref. 12, p. 4 (on the internet pdf version).

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

[23] Truman, R., The Problem of Information for the Theory of Evolution. Has Tom Schneider Really Solved It?; www.trueorigin.org/schneider.asp.

[24] Hengen, ref. 6, p. 4996.

[25] Bio-molecules attach to binding sites using a variety of chemical/electronic principles. The information theory analysis only requires that unique patterns are used, but there must also be a suitable physical basis for two members to interact.

[26] There are different probabilities of mutations between the four nucleobases A, C, T, and G.