In: Information Theory: New Research
Editors: Pierre Deloumeaux and Jose D. Gorzalka
ISBN: 978-1-62100-325-0

Chapter 1

INFORMATION THEORY AND BIOLOGICAL SEQUENCES: INSIGHTS FROM AN EVOLUTIONARY PERSPECTIVE

Ivan Erill
University of Maryland Baltimore County, Baltimore, MD, US

ABSTRACT

Information theory has been successfully applied to the analysis of biological sequences for more than thirty years. Pioneered for the study of binding sites in DNA sequences, information theoretical approaches have also become a de facto standard in protein sequence analysis and have been applied to such diverse fields as genome assembly, gene identification or the study of tRNA structure. In spite of its tremendous success at modeling the DNA sequence motifs bound by transcription factors (TF), the application of information theory to this particular field has also led to several misconceptions, such as the paradox of negative information, the formulation of simple corrections for genomic bias or the tacit assumption of a direct link with Boltzmann entropy in these molecular systems. These misinterpretations arise mainly from overlooking the evolutionary nature of the information transfer process taking place in the biological systems being studied. Adopting an evolutionary perspective on the application of information theory to transcription factor-binding motifs addresses many apparent contradictions and provides new insights into the interaction between transcription factors and their cognate binding sites.

1. LIFE AS AN INFORMATION SYSTEM

Lila Gatlin's 1972 classic manuscript on applying information theory to living systems opens with a bold statement: "Life may be defined operationally as an information processing system – a structural hierarchy of functioning units – that has acquired through evolution the ability to store and process the *information* necessary for its own accurate reproduction" [1]. Even though there is no formal proof that life is, indeed, an information process, Gatlin's statement captures the intuitive notion that life must carry out information processing at many different levels and that, ultimately, it has to maintain the physical low entropy state that

allows such information processes to operate. In fact, the concept of life as a set of self-sustaining processes lies at the core of many definitions of life. In 1974 Varela and Maturana coined the word *autopoiesis* to describe, precisely, a circular definition of life based on a network of processes that operate to generate and perpetuate themselves [2]. Such a definition captures most of the essence of Gatlin's statement and draws heavily on related concepts, such as homeostasis, advanced many years earlier to reflect the ability of living systems to self-regulate their internal states [3].

1.1. Of Life as an Information Process

The intuitive appeal of life as an information processing system suggests that parallels between life and artificial information systems should be easy to find. An obvious place to look for similarities is the nervous system. Indeed, the study of brain anatomy and function has provided invaluable insights on alternative methods of computation and organization that have later been co-opted by artificial intelligence researchers [4-6]. It is in molecular biology, however, that the parallels with concepts in computer science become most apparent. They are in fact almost impossible to miss, because computer science jargon actually riddles the pages of molecular biology journals and textbooks [1]. The most familiar example is the Central Dogma of molecular biology [7]. The Central Dogma "deals with the detailed residue-by-residue transfer of sequential information" [8] and implicitly postulates the existence of a genetic code to translate the information encoded in DNA sequences onto amino acid sequences. George Gamow's ill-fated attempt at cracking the genetic code (the "diamond code") was based primarily on information theory and became one of the first well-known instances of a formal mathematical treatment preceding and focusing biological research [9, 10]. In fact, the molecular processes of transcription and translation that animate the Central Dogma were postulated by John von Neumann as necessary for self-replicating automata several years before they were identified by molecular biologists [11, 12].

1.2. Evolution as an Information System

Autopoiesis and related concepts, like homeostasis, focus on the processes operating in living beings. That is, they deal with the set of processes that define and preserve the "alive" state in a living being. Gatlin's opening statement takes a step further by explicitly including evolution as a fundamental element in life's information processing. Even though her wording is subtle, the integration of evolution into the definition of life as an information process constitutes a decisive move that enables and empowers the application of information theory approaches to the analysis of biological sequences. When considering life as an information process, it is thus appropriate and important to distinguish between those processes taking place within the lifespan of an organism (autopoietic processes) and the long-term process of evolution by natural selection. For instance, the generation of a complex response to physicochemical stress by an *Escherichia coli* cell is obviously an information process in much the same way as the firing of a furnace by a house thermostat is an information process. Just like the thermostat will fire the furnace multiple times during a winter season, the *E. coli* stress response may be activated several times during the lifespan of

an *E. coli* cell. It is quite apparent that the continued preservation and fine tweaking of the genetic system responsible for the *E. coli* stress response is also an information process. However, this process differs substantially from the ones depicted above in that it takes place over eons, instead of minutes, and we can only observe its end result.

1.3. Information and Entropy

Gatlin's opening statement makes a further distinction that becomes essential to her main thesis; she defines life as an *information* process. This is an important point, because the literature on the application of information theory to biology is plagued by constant back-and-forth between information theory and thermodynamics or, to be more specific, between Shannon and Boltzmann entropies (see for instance Brooks and Wiley "Evolution as Entropy" [13]). In fact, Gatlin's book is among the very few that does not incur some kind of amalgamation of both theories and sticks strictly to information theory. The confusion generated by this constant back-and-forth between both fields is considerable and has led some authors, like Hubert Yockey, to state flatly that "the concept of entropy in classical thermodynamics is different from that in statistical mechanics and from that in information theory" [14]. Yockey has later moderated this claim, but remains steadfast in warning against mixing both approaches [10].

The persistent back-and-forth between both information theory and thermodynamics, and the ensuing confusion, is unfortunate and stems from two main sources. The first and most notorious is the use by Shannon of the word entropy to define his uncertainty measure, following the apocryphal advice of Von Neumann [15]. Boltzmann and Shannon definitions of entropy share both name and form and can indeed be formally linked, but the link is not as direct or evident as the similarity in name and form might lead us to believe [10, 16-18]. In the following, I will use the full terms *Shannon entropy* and *Boltzmann entropy* to make explicit distinction between both measures when necessary, but I shall revert to *entropy* or *uncertainty* for Shannon entropy whenever there is enough context to prevent misidentification. The second cause for interplay between both fields, explored here, originates in the intuitive assumption that the formalization of an autopoietic process by means of statistical thermodynamics can be mapped directly to the information theory formalization of an evolutionary process operating on it. In the example above, this would lead to equating the thermodynamic processes of the *E. coli* stress response with the information process leading to the evolution of such a response.

2. Information Theory and the Analysis of Biological Sequences

The analysis of biological sequences using information theory was pioneered by Gatlin [19, 20] and paved the way for the application of a powerful analytical framework to biological systems. They key insight in Gatlin's approach was the realization that biological sequences, such as chromosomes, are processed mostly as linear entities and in a similar fashion along their whole length. Protein coding genes, for instance, are transcribed by an RNA-polymerase holoenzyme and translated to proteins in ribosomes, regardless of their

position in the genome and of the specific protein encoded by the gene. Likewise, chromosomes are duplicated by a DNA-polymerase holoenzyme in the same manner across all their length. Even though biological sequences are embedded in tridimensional molecules that may locally modulate the processing of the sequence, the assumption of operational linearity and homogeneity is generally valid [1, 14]. This makes it possible to consider biological sequences as ergodic Markov sources and allows their analysis using the methods and concepts of information theory introduced by Claude Shannon [21].

2.1. Entropy of a Genome

For a memory-less source X that emits messages using a given alphabet $[x_1,\ldots,x_n]$ with emission probabilities $P=[p(x_1),\ldots,p(x_n)]$, Shannon [21] defined the entropy of the source as:

$$H(X) = -\sum_{i=1}^{N}\left(p(x_i)\cdot\log_2\left(p(x_i)\right)\right) \quad (1)$$

where N is the number of possible symbols for the source and $-\log_2(p(x_i))$ is the information of a given symbol x_i as defined by Shannon[1]. Hence, the entropy of a source is the mathematical expectation of the information that the source is capable of generating. Intuitively, the entropy measures the uncertainty of an observer over a symbol emitted by the source. A source emitting equally probable symbols yields, on average, maximum uncertainty because it minimizes the ability of an observer to predict the next emitted symbol.

A bacterial genome, typically in the form of a single chromosome, can be considered as a long string with an alphabet Ω of four symbols $\{A, C, G, T\}$ corresponding to the four possible bases that may occupy a given position of the DNA chain (Adenine, Cytosine, Guanine and Thymine). Treating the genome as an ergodic source, we can estimate the probability of each symbol $\{f(A), f(C), f(G), f(T)\}$ by sampling its occurrence in the linear genome sequence, as opposed to sampling the occurrence of symbols over time. Because DNA is normally a double-stranded molecule following the Watson-Crick pairing scheme (A corresponds to T in the complementary strand, C corresponds to G), it is customary to enforce $f(A)=f(T)$ and $f(C)=f(G)$. Hence, by counting the number of A, C, G and T occurrences over all the positions of a genome and normalizing them into relative frequencies we can infer the entropy of the genome as:

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

If we do this for the 4,639,675 base pairs (bp) of the genome of the bacterium *Escherichia coli*, we obtain an entropy H_G=1.999 bits that is remarkably close to the maximum possible value of 2 bits. In the *E. coli* genome the probability of observing any given base is very close to 0.25 (f(A)=f(T)=0.246, f(C)=f(G)=0.254), leading to the observed H_G=1.999 bits value, but many organisms deviate substantially from this maximum entropy

[1] Here and in the following we will be arbitrarily using base 2 for the logarithm and hence defining information in *bits*.

scenario. The heat-loving bacterium *Thermus thermophilus*, for instance, has frequencies $f(C)=f(G)=0.346$ (typically expressed as 69.2% GC), leading to an entropy $H_G=1.876$. Even though it might not seem extremely significant (6.15% difference), this is a substantial decrease in entropy with regard to *E. coli*. For a typical bacterial genome size of ~4,000,000 bp, *E. coli* could in theory encode 492,000 bits more of information than *T. thermophilus*.

2.2. Genomic Entropy, Evolution and the Genetic Code

With a limited amount of sequence data, Gatlin analyzed for the first time the entropy of genetic sequences. The basis of her approach was the realization that the entropy of the genome as a source should reflect the coding redundancy required to compensate for the genetic noise introduced by mutations. Redundancy and uneven letter frequencies in English allow automatic detection and correction of errors, such as *thf* for *the*, leading to relatively small entropy values for the English language when analyzed as a Markov source of different orders [22]. Likewise, the genome ought to show decreased entropy due to its encoding of the genetic message through evolution. Gatlin evaluated the entropy of genetic sequences modeled as 0^{th} (memory-less) and 1^{st} order Markov sources and found that, on average, these entropies were remarkably close to maximal [20]. This surprising result is explained mostly by the large amount of redundancy present in the genetic code, but also to some extent by the extensive use of message redundancy in evolution.

The amount of redundancy embedded in the genetic translation system can be computed using information theory [14]. By introducing redundancy implicitly in the code, the genetic translation system largely detaches redundancy from the source, making it possible for genetic sequences to operate at high entropy values in spite of genetic noise. A crude analogy with English serves to illustrate the point. As noted above, the exclusion of certain letter combinations in English (e.g *thf*) decreases the entropy of an English speaker. However, the semantic redundancy of English, due mostly to its use of synonyms, allows an English speaker to increase its apparent entropy, since it allows a speaker to convey approximately the same message using a more varied set of source characters[2]. Just like synonyms are implicit to the English language, the genetic code is implicit to the translation process and is not explicitly mapped on the underlying genetic sequences, allowing them to achieve higher entropy levels[3]. In addition to the redundancy in the genetic code, the generation of multiple offspring by living systems is a plain form of message redundancy on which natural selection can operate as an error-correcting device, providing additional buffering against genetic noise.

[2] If we are allowed to use the words "beautiful", "beauteous", "lovely", "bonny", "comely", "fair", "handsome" and "pretty", and we tend to use them with equal probability, the entropy *per character* of the source will be higher than if we stick instead to using exclusively the word "lovely" whenever we are describing something as fine-looking.

[3] It can be argued that the genetic code is embedded in the genome of every single organism, in the form of ribosomal protein, rRNA, tRNA and aminoacyl tRNA synthetase genes. This corresponds, however, to a small portion of the genome and part of it (e.g. ribosomal protein genes) already uses the genetic code for expression. The entropy of tRNA genes, which cannot make use of the redundancy available in the genetic code has, indeed, been shown to be smaller than the genomic average [23].

Applications of Information Theory to Molecular Biology

The mathematical framework provided by information theory can be applied to more specific analyses of biological sequences. As described by Yockey, information theory provides a sound mathematical ground to derive similarity metrics for biological sequences [10] and the entropy of an alignment, or measures derived from it, is used today as the standard measure of conservation in biological sequence alignments [24, 25]. Information theory metrics have been applied, among other, to the analysis of protein coding genes, tRNA sequences and repeated nucleotide sequences [23, 26, 27]. To date, however, the most successful application of information theory to molecular biology has been the modeling of transcription factor-binding motifs. The remainder of this chapter describes the basic notions behind this approach and explores how frequent misconceptions can be prevented by superimposing an evolutionary perspective to the information theory approach.

3. INFORMATION THEORY MODEL OF TRANSCRIPTION FACTOR-BINDING MOTIFS

In the past two decades, the focus of genomics research has shifted steadily from prediction and analysis of coding regions to the analysis of non-coding regions, the regulatory elements therein and their interactions. This shift stems mainly from the increasing awareness that changes in regulation lie at the core of development and of most phenotypic differences within and between species [28]. In living systems, regulation takes place at almost all the steps of the information process that leads from genotype to phenotype. In fast growing bacteria, for instance, the positioning of a gene in the genome will often result in multiple copies of the gene coexisting within the cell, leading to increased expression of the gene product. Then again, in many eukaryotic cells, microRNA molecules can bind complementary messenger RNA (mRNA) transcripts and silence their expression [29]. Evolution has therefore at its disposal many mechanisms to regulate gene expression, yet transcriptional regulation is the most pervasive and well known regulatory system in both eukaryotic and prokaryotic cells [28, 30].

3.1. Transcription Factors and Binding Motifs

Regulation at the transcriptional level makes intuitive sense because transcription is the first real-time step of the information processing system for gene expression. Hence, regulation before transcription cannot be very flexible and regulation in later stages will always be more wasteful (because mRNA transcripts, at the least, will have already been produced). Transcriptional regulation is mediated mainly by a subset of proteins known as transcription factors (TF) [31]. These proteins bind DNA, typically within the promoter regions of genes targeted by the RNA-polymerase holoenzyme. By binding to these regions, transcription factors can either hinder (repressors) or promote (activators) the formation of an open complex by the RNA-polymerase holoenzyme and thus they effectively regulate gene expression [30].

Transcription factors are able to bind DNA either specifically or non-specifically. Non-specific binding is typically short-lived and associated with interactions with the DNA backbone. In contrast, specific binding involves longer association times and directed amino acid-base contacts [32, 33]. Hence, even though some transcription factors bind DNA without much specificity or rely on recognition of tridimensional properties of the DNA molecule (e.g. curvature), most transcription factors bind to DNA by recognizing specific sequence elements through definite amino acid-base contacts. These specific sequence elements are called transcription factor-binding sites. As opposed to other DNA-binding proteins, like restriction enzymes, transcription factors do not recognize a single particular sequence (e.g. `ATGGACCAT`), but target instead a number *N* of similar sequences [30]. When aligned, the collection of slightly different sequences known to be bound by a transcription factor is collectively known as the binding motif. Table 1 shows an example of a collection of 25 binding sites recognized by the hypothetical transcription factor BUH.

Table 1. Collection of binding sites (*N*=25) bound by the hypothetical transcription factor BUH

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

3.2. Transcription Factor Binding as an Information Process

Conventional communication involves a source *X* with alphabet $[x_1,\ldots,x_n]$ and a receiver *Y* with alphabet $[y_1,\ldots,y_m]$. The source and the receiver communicate through a noisy channel over which the message is transmitted. Shannon defined the equivocation of a channel as the conditional entropy[4]:

$$H(X \mid Y) = \sum_{j=1}^{M}\left[p(y_j)H(X \mid y_j)\right] = -\sum_{j=1}^{M}\left[p(y_j)\sum_{i=1}^{N}\left[p(x_i \mid y_j)\cdot\log_2\left(p(x_i \mid y_j)\right)\right]\right] \quad (3)$$

The formulation of equivocation as a conditional entropy $H(X|Y)$ can be interpreted as the uncertainty of an observer over the symbols emitted by the source given the received message. Intuitively, once a symbol $Y=y_j$ is received, the presence of noise in the channel implies a set of conditional probabilities $P(x_i|y_i)$ for every possible source symbol x_i. That is, the entropy of the source as perceived by the observer changes to $H(X|y_j)$ upon receiving a symbol $Y=y_j$, and the conditional entropy $H(X|Y)$ is the weighted average of these revised entropies for all possible values of *Y*. Equivocation is therefore a measure of the information present in the source that is lost to noise in the channel, as illustrated artfully by Shannon in his seminal paper through a worked-out example [21].

[4] The notation $P(x_i|y_i)$ stands here and in the rest of the document as shorthand for the more formal $P(X=x_i|Y=y_i)$.

Mutual Information

Having established the concept of equivocation, it is easy to derive mutual information as the difference between the source entropy and the conditional entropy (or equivocation) imposed by the channel:

$$I(X;Y) = H(X) - H(X \mid Y) \tag{4}$$

This difference is known by several names, such as mutual information, information content, information rate or mutual entropy, and it is a measure of the decrease in uncertainty brought about by the reception of messages generated by source *X* and transmitted through a channel with equivocation *H(X|Y)*. More generally, *I(X;Y)* measures the information shared by two random variables *X* and *Y*; that is, it measures how much our uncertainty on one variable (*X*) decreases upon knowing the other (*Y*). A decrease in uncertainty is typically associated with an increase in information, and here I will adhere to the term *information content* for *I(X;Y)*. Intuitively, in an error-free channel $p(x_i|y_i)=1$ (because the symbol y_i completely determines the source symbol x_i), leading to *H(X|Y)*=0 and *I(X;Y)=H(X)*. Alternatively, if a channel contains enough noise to render *X* and *Y* independent, then *H(X|Y)=H(X)* and *I(X;Y)*=0.

Transcription Factor Binding as an Information Process

The application of information theory methods to model transcription factor-binding motifs was pioneered at Larry Gold's lab in the mid 1980s [34]. The approach taken by Schneider and co-workers can be illustrated with a thought experiment regarding the uncertainty of an observer over the specific bases occupying each position of a particular DNA segment of length *L* in a given genome *G*. In this thought experiment we pick a random segment of the genome, we immerse it in a solution containing our transcription factor and we wait a reasonable amount of time to observe whether or not the segment is specifically bound by the protein[5]. Clearly, if no further information is given to us at the beginning of the experiment, our initial uncertainty is dictated by the likelihood of the occurrence of each the four possible DNA bases in the genome. Hence, if we assume independence between the positions of this segment, our *a priori* entropy for any given position of the DNA segment can be defined simply as the genomic entropy H_G (Equation 2).

$$H_{before}(l) = H_G = -\sum_{S \in \Omega} \left[f(S) \cdot (\log_2 (f(S))) \right] \tag{5}$$

Under the assumption of positional independency, an aligned collection of binding sites for a given transcription factor (Table 1) can be used to infer the base probabilities at each

[5] We are further assuming that the transcription factor is labeled somehow and that we are in possession of a reading device capable of distinguishing specific from non-specific binding. Even though this may seem farfetched, this is precisely what the experimental procedures used to determine the known collections of binding sites aim to approximate to a fair degree of accuracy [35]. Here, and in the following, we will assume that whatever methods we are using to observe binding are actually able to make this distinction. Hence, the observation of non-binding, or of an *unbound* state, at a specific position encompasses implicitly the occurrence of unspecific binding at that position.

position of the binding motif. The relative frequencies of each base at each position of the motif are usually expressed in the form of a Position-Specific Frequency Matrix (PSFM). For our hypothetical BUH transcription factor and the collection of binding sites in Table 1 we thus obtain:

Table 2. Position Specific Frequency Matrix for transcription factor 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

It is easy to see that the probability $p(S_l)$ for a given base S in each column l of the PSFM can be interpreted in our thought experiment as the conditional probability of that base occurring at that position of our query segment *given that* we have observed specific-binding of the transcription factor to the segment $p(S_l|TF_{bound})$.

The PSFM does not give us the reverse conditional probabilities $p(S_l|TF_{unbound})$; that is, the probability of each base at position l of the segment *given that* we have *not* observed specific binding of the transcription factor. However, it can be shown that for a genome size substantially larger that the number of sites in our collection ($G_S >> N$), these probabilities will converge to the genomic frequencies (i.e. $p(S_l|TF_{unbound}) \rightarrow f(S)$). For instance, even if a transcription factor completely specified its recognition sequence (e.g. `ATGGACCAT`), the lack of specific binding to a given genome segment would only be telling us that base `A` is marginally ($\delta \rightarrow 0$) less likely to be seen at the first position of that segment. Since $G_S >> N$, for a random genome segment we can assume that $p(TF_{bound}) \rightarrow 0$ and $p(TF_{unbound}) \rightarrow 1$. Hence, if we compute the equivocation of the channel following Equation 3, we obtain:

$$H(S_l \mid TF_{state}) \approx -\left[0 \cdot \sum_{S \in \Omega} \left[p(S_l) \cdot (\log_2(p(S_l))\right] + 1 \cdot \sum_{S \in \Omega} \left[f(S) \cdot (\log_2(f(S))\right]\right] \approx H_g \quad (6)$$

from which we obtain, by applying Equation 4, the value $I(X;Y) \approx 0$ for the mutual information.

The result in Equation 6 tells us that trying to identify the positional base frequencies of a random DNA segment by experimentally evaluating the binding of a particular transcription factor to it is an extremely inefficient way of proceeding. This is because the vast majority of the genome is not bound by the transcription factor and, therefore, on average we gain very little information about the frequencies of the bases occupying each position of a random segment[6]. Seen from a communications theory point of view, this looks rather obvious. Our source has four possible states with associated emission probabilities relatively close to equiprobability[7], yet we only have two states at the recipient and one of them (the unbound state) has probability close to one. Rather than being a futile distraction, however, this

[6] Here, and in the following unless otherwise stated, we will assume for simplicity that the transcription factor only binds specifically to the known binding sites, and that it does not bind specifically elsewhere in the genome.

thought experiment serves to illustrate the point that transcription factor-binding can be effectively thought of as the noisy channel of a communication process and that, as such, it entails an equivocation.

3.3. Information in a Transcription Factor-binding Motif

In their landmark paper, Schneider and co-workers focused on the information contained within the transcription factor-binding motif. Essentially, they evaluated the uncertainty of an observer over the base occupying each position of a DNA segment *once it is known* that the transcription factor binds that segment. As we have seen, the conditional entropy $H(S_l|TF_{bound})$ is fully specified by the conditional probabilities $p(S_l)$ available on the PSFM for a given transcription factor. Hence, it follows that for a given position of the DNA segment:

$$H_{after}(l) = H(S_l \mid TF_{bound}) = -\sum_{S_l \in \Omega} \left(p(S_l) \cdot \log_2 \left(p(S_l) \right) \right) \tag{7}$$

which leads to the expression for mutual information:

$$\begin{aligned} R_{sequence}(l) &= H_{before}(l) - H_{after}(l) \\ R_{sequence}(l) &= \left[-\sum_{S \in \Omega} f(S) \cdot \log_2 \left(f(S) \right) \right] - \left[-\sum_{S_l \in \Omega} p(S_l) \cdot \log_2 \left(p(S_l) \right) \right] \end{aligned} \tag{8}$$

The term $R_{sequence}(l)$ is the information that an observer gains over the base occupying position l in a DNA segment once binding of a particular transcription factor to the segment has been observed and is conventionally referred to as the *information content* of that position. Alternatively, one may say that it is the information content encoded within a position of the TF-binding motif, or the information gained from aligning the TF-binding motif sequences from an initial, unaligned state [34]. Since we have assumed positional independency in its derivation, the contributions of each position can be added up to yield $R_{sequence}$, which is the information content of the transcription factor-binding motif.

$$R_{sequence} = \sum_{l=1}^{L} R_{sequence}(l) \tag{9}$$

The information content of a TF-binding motif is typically represented using information logos [36, 37], which combine cleverly the frequency information of the PSFM with the conservation information derived from $R_{sequence}(l)$. For the case of our hypothetical transcription factor BUH, the information logo is shown in Figure 1.

[7] The H_G=1.876 bits result obtained for *T. thermophilus*, due to its extreme %GC content (~70%), is among the lowest genomic entropies recorded so far in a sequenced genome.

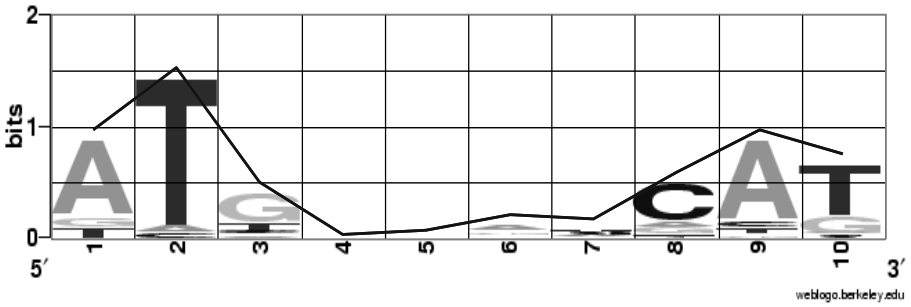

Figure 1. Sequence logo for the TF-binding motif of transcription factor BUH. The $R_{sequence}$ function is superimposed on the logo. At each position (x-axis), the height of the stack corresponds to the $R_{sequence}(l)$ value, while the height of each letter corresponds to the frequency of that particular base, normalized to the $R_{sequence}(l)$ value (i.e. $p(S_l) \cdot R_{sequence}(l)$). The logo was generated using the WebLogo server [37].

Relationship to Other Measures

A major insight of Schneider *et al*. was the realization that $R_{sequence}$ had to be related somehow with the amount of information required to find the TF-binding sites in their native genome. To pursue this hypothesis, they proposed $R_{frequency}$, a measure of the information required to find a site that is not based on the specific TF-binding site sequences, but on the size proportion between genome and TF-binding sites. The rationale behind $R_{frequency}$ is relatively easy to follow. If we are given no additional knowledge, a circular genome of size G_S must be assumed to contain G_S possible biding sites for a given transcription factor. If we assume that all these potential sites are equally likely to be bound, then we obtain the *a priori* entropy:

$$H_{Gs} = -\sum_{G_S}\left(\frac{1}{G_S}\cdot\log_2\left(\frac{1}{G_S}\right)\right) = \log_2(G_S) \qquad (10)$$

If we are told, however, that this transcription factor binds only M specific sites in the genome, and if we further assume that these M sites are equally likely to be bound and that no other genomic positions are bound, H_{Gs} simplifies[8] to the *a posteriori* entropy:

$$H_M = -\sum_{M}\left(\frac{1}{M}\cdot\log_2\left(\frac{1}{M}\right)\right) = \log_2(M) \qquad (11)$$

The *a priori* entropy H_{Gs} measures our initial uncertainty over any position of the genome being bound by a single copy of the transcription factor. Likewise, the *a posteriori* entropy H_M measures our uncertainty over any position of the genome being bound once we know that the transcription factor targets only a given number M of positions and is not binding anywhere else in the genome. It follows logically that we should experience a decrease in

[8] At the limit, $x \to 0$, we obtain $x \cdot \log(x) = 0$.

uncertainty when we move from not knowing anything to knowing the *M* sites bound by the transcription factor. This decrease in entropy is again expressed as mutual information:

$$R_{frequency} = H_{Gs} - H_M = \log_2(G_S) - \log_2(M) = \log_2\left(\frac{G_S}{M}\right) \qquad (12)$$

and provides an estimate of the information needed to locate *M* binding sites in a genome of size G_S. As expected, the larger the number of binding sites, the less information that is required to find them and vice versa. This is obvious for the extreme cases. For a transcription factor targeting a single site in the genome, the *a posteriori* uncertainty is zero and we need $\log(G_S)$ bits of information to locate the site. In contrast, if a transcription factor is completely non-specific and binds anywhere in the genome, we do not gain any information from knowing the list of sites it binds.

Schneider and co-workers reasoned that $R_{sequence}$ should be proportional to $R_{frequency}$, and they showed that both measures gave similar values for several transcription factors on genomes with equiprobable base compositions (f(A)=f(C)=f(G)=f(T), H_G=2 bits). When departing from equiprobability, however, the equality no longer holds. This is because the genome will now be a biased source and the *a priori* entropy H_{before} will always be smaller than the maximal 2 bits. Consequently, $R_{sequence}$ (which is the difference between H_{before} and H_{after}) will decrease consistently. In contrast, $R_{frequency}$ can be seen as increasing or decreasing depending on the base composition of the TF-binding motif. If, for instance, a transcription factor targets AT-rich sites in a GC-rich background, it stands to reason that the relative frequency of binding sites in the genome, as expected by chance, should be smaller (and G_S/M thus larger), leading to a larger $R_{frequency}$ value. In other words, the AT-rich sites are less frequent in a GC-rich background, leading to a larger reduction in uncertainty when identified, and thus they can be said to contain more information.

It must be stated that the above argument applies only when the number of functional sites is essentially the number of binding sites. This is the case of restriction enzymes, for which a functional site is actually defined as a binding site. As noted by Schneider *et al.*, for a transcription factor in a given genome the number of functional sites *M* is fixed, and later research has shown that this is true also for transcription factors across genomes, in spite of changes to the %GC composition of the genome [38]. Nonetheless, the argument still stands that a fixed number of TF-binding sites deviating from the genome bias would face reduced competition from similar, non-functional pseudo-sites in the genome, thereby facilitating their detection, and this can be interpreted as an increase in the information they encode. To maintain the equality between both information measures, Schneider and co-workers proposed the use of the Kullback–Leibler divergence or relative entropy [39] as a measure of positional information content:

$$RE(l) = R^*_{sequence}(l) = \sum_{S_l \in \Omega}\left(p(S^l)\cdot\log_2\left(\frac{p(S_l)}{f(S_l)}\right)\right) \qquad (13)$$

As in the case of $R_{sequence}$, positional relative entropy can be extended to a global TF-binding motif measure (*RE*) by assuming again positional independency and adding the contribution of each of the *RE(l)* terms. The use of relative entropy as a measure of positional information content was introduced without formal derivation, but it is easy to provide an intuitive understanding on how *RE* operates with regard to $R_{sequence}$ by making some algebraic manipulations to Equation 13:

$$RE(l) = \left[-\sum_{S_l \in \Omega} \left(\frac{p(S_l)}{f(S_l)} \cdot f(S_l) \cdot \log_2 \left(f(S_l) \right) \right) \right] - \left[-\sum_{S_l \in \Omega} \left(p(S_l) \cdot \log_2 \left(p(S_l) \right) \right) \right] \quad (14)$$

In this new formulation, *RE(l)* can be seen as a modification of $R_{sequence}(l)$ in which the background genomic entropy (H_{before}) is weighted according to the ratio between the each base in the motif $p(S_l)$ and in the background $f(S_l)$. Hence, if a base is underrepresented in the genome but used heavily in one of the positions of the TF-binding motif, the ratio $p(S_l)/f(S_l)$ increases and the *a priori* entropy H_{before} is revised upwards. This has the net effect of increasing the apparent information content of that position, in agreement with the estimates provided by $R_{frequency}$. The effect of such an up-weighting on a TF-binding motif can be observed in Figure 2.

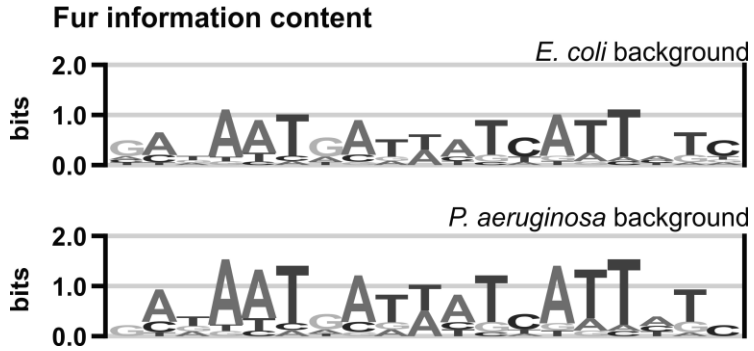

Figure 2. *RE* sequence logo for the TF-binding motif Fur of *Pseudomonas aeruginosa* (64 sites), when seen against a nearly equiprobable genome background (*E. coli*) and against the heavily biased (66% GC) *P. aeruginosa* background. The logos were generated using the enoLOGOS web server [40]. It can be directly seen the information content of A and T dominated positions increases in the *P. aeruginosa* background.

3.4. Transcription Factor Binding as an Evolutionary Process

Schneider and co-authors decided arbitrarily to use only the conditional entropy $H(S_l|TF_{bound})$ when computing $R_{sequence}$ as a measure of the information content of transcription factor-binding motifs, but they provided no formal or intuitive justification for this decision[9]. As we have seen above, this turns out to be a reasonable assumption, since computation of the full channel equivocation (Equation 6) leads to close to zero information

[9] Recall that Shannon formulation would suggest that both terms $H(S_l|TF_{bound})$ and $H(S_l|TF_{unbound})$ be used to compute the mutual information (Equation 6).

content values. Then again, it seems quite obvious that the observation of non-binding by a transcription factor must also be considered formally part of the information process. Intuitively, we can see that if we consider only a 64 bp long genome sequence, the fact that a particular segment of this sequence is not bound by a transcription factor can actually provide a substantial amount of information on the probability of a base occupying a certain position of the segment.

The main problem of dealing with transcription factor-binding as an information channel is that the recipient alphabet is extremely biased towards one of the two possible states (i.e. binding and non-binding), leading to high equivocation. In other words, the fact that the genome contains mostly non-binding sites makes them highly uninformative and dilutes the highly-specific information conveyed by the few binding sites present. The solution to this apparent conundrum, as with many other aspects of the application of information theory to biology and, in general, of bioinformatics, is to take advantage of the underlying information process operating in all living systems: evolution.

Evolution as the Informed Observer

So far we have been postulating the presence of a human observer to our information process thought experiment, but it is quite obvious that transcription factors bind to their binding sites in the absence of any human observers. We have also assumed that the information process that concerns us is the short-lived (i.e. seconds or minutes) binding or non-binding of a transcription factor to a particular genome segment. Both assumptions, even though helpful in formulating the thought experiment, are misguided. As it turns out, for all the information processes carried out by a living being there is an implicit observer in the form of evolution by natural selection. As opposed to a human observer, evolution does not focus on individual short-lived information processes, but on the average of such processes over the life-time of a population of genetically related organisms.

When Shannon defined the concept of equivocation in a channel, he used the metaphor of an informed observer to illustrate the concept. This observer is "able to see both what is sent and what is recovered (with errors due to noise). This observer notes the errors in the recovered message and transmits data to the receiving point over a 'correction channel' to enable the receiver to correct the errors" [21]. Shannon used the metaphor of the informed observer to point out the fact that the equivocation in a channel can be seen as the additional information that should be supplied to correct the received message. It seems quite obvious that most of the traits and actions attributed by Shannon to the informed observer are those implemented by evolution as an information process. Evolution is able to detect and correct "errors" by making use of extensive redundancy (i.e. populations) and eliminating poorly adapted individuals. The only crucial difference with Shannon's observer is that evolution will only correct those errors that hinder survival or reproductive fitness (i.e. the definition of error becomes context dependent), and that is limited in doing so by the size and variability of the population it operates on. By *correcting* these "errors" evolution embeds both the source *and* the recipient within the genetic sequences that encode the information processes carried out within an individual organism lifetime. It must be noted that this does not constitute a circular definition, because source and recipient are temporally and physically separated by generations within populations.

Having introduced evolution as Shannon's informed observer in disguise, it is easy to see why we should only take into account the conditional entropy $H(S_l|TF_{bound})$ when computing

$R_{sequence}$ as a measure of the information content of transcription factor-binding motifs. If we maintain the assumption that the genome is basically non-binding as a background, transcription factor-binding sites are defined *functionally* by evolution as those segments of the genome that enact some kind of selectable action upon binding of a transcription factor. Therefore, evolution cannot and will not *observe* the non-binding of the transcription factor at all the remaining positions in the genome. This is because there is no selectable component to the non-binding of the transcription factor at those positions. There is, obviously, a selectable component to the binding of the transcription factor at these non-site genome positions (if it happens in sufficiently large numbers), but we have implicitly ($R_{sequence}$) and explicitly ($R_{frequency}$) assumed that the transcription factor will bind almost exclusively to the known collection of binding sites.

In a way, we can see the "functional blindness" of evolution with regard to the non-binding of the genomic background as a kind of spatial filter. The source entropy H_G is not reduced because the genome still encodes a full variety of genetic messages[10]. With regard to the particular information process that concerns us, however, evolution tunes in exclusively to certain spatial regions, making the conditional entropy pertaining to the non-bound state $H(S_l|TF_{unbound})$ largely irrelevant. The unbound state is obviously relevant in those genome segments that are functionally targeted by evolution, but when analyzing a genome we are always seeing the end-result of the evolutionary information process. Hence, we can safely assume that all the TF-binding sites in our collection will be bound by the transcription factor (as this is exactly the reason why they have been declared members of the collection) and we are entitled to use only the bound-state conditional probability $H(S_l|TF_{bound})$ to derive $R_{sequence}(l)$ as we did in Equation 8. In doing we should properly interpret $R_{sequence}(l)$ as the minimal amount of information that must be encoded in a set of transcription factor-binding sites in order for them to perform their required function. This viewpoint has led some authors to label $R_{sequence}$ as a Redundancy Index [42], since it basically measures the amount of redundancy (in maintaining a defined state departing from the genomic background) preserved in sites by evolution.

4. RETHINKING THE INFORMATION THEORY MODEL OF TF-BINDING MOTIFS

The use of an evolutionary perspective to explain the omission of $H(S_l|TF_{unbound})$ when computing $R_{sequence}$ may seem contrived and circuitous to the casual reader, but it provides a fundamental and, so far, missing foundation for the information theory approach to transcription factor-binding motifs. By explicitly introducing evolution as a primary element of information theory as applied to living systems and their processes, this new standpoint also enables us to pose new questions and revisit the validity of certain assumptions. Furthermore, the introduction of evolution as the core information process contributes also to

[10] It is difficult here to establish a definite parallel with conventional communication systems, but Turing's attack on the Enigma cipher provides a decent analogy [41]. The Bletchley Park team would tune in at the same time every morning to listen to the German weather forecast, knowing beforehand that one the first words in the bulletin would be "wetter" (German for weather). Hence, the alphabet of the source was not constrained as it continued to emit during the day, yet the recipient's alphabet had been functionally trimmed down to the possible encodings of the word "wetter".

elucidate which information processes are being discussed and how they might relate to the underlying thermodynamic processes.

4.2. Revisiting the Entropy Equality Assumption

A primary assumption of the information theory model for TF-binding motifs advanced by Schneider *et al*. was the intuitive notion that $R_{sequence}$ should equal $R_{frequency}$. As mentioned above, this assumption led to the introduction of an alternative measure (*RE*) to replace $R_{sequence}$ in biased genomes in order to maintain the equality between the motif-based and frequency-based measures of information (Equation 14). The use of relative entropy over $R_{sequence}$ has been advocated later by several authors [40, 43, 44], but recent work has shown that *RE* assumes by default an overly simplistic model of the genomic background and, in doing so, it can lead to substantial inaccuracies [45].

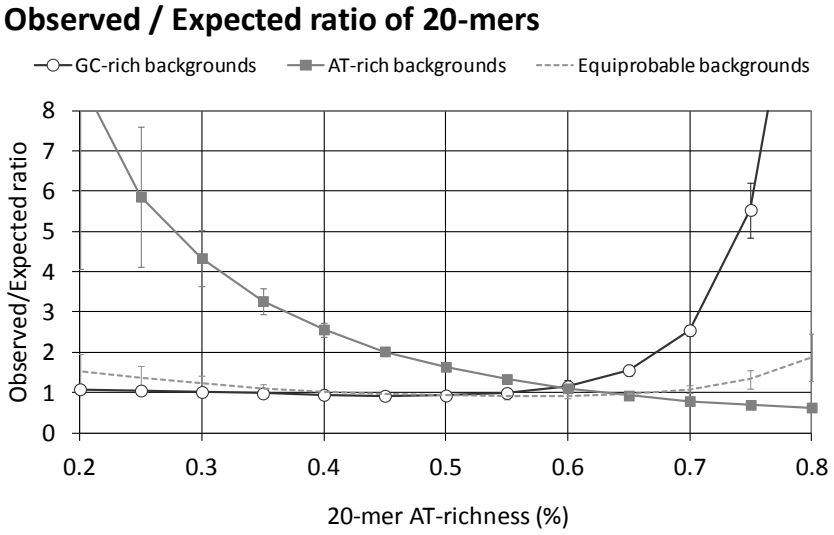

Figure 3. Average ratio of observed vs. expected 20-mers in real genomes versus randomly generated sequences. Three independent genomes were used to compute the ratios in each %GC category (GC-rich, AT-rich and equiprobable): *E. coli* str. K-12 substr. MG1655 [50.8% GC], *P. aeruginosa* PAO1 [66.6% GC], *Haemophilus influenzae* Rd KW20 [38.1% GC], *Colwellia psychrerythraea* 34H [38.0% GC], *Salinibacter ruber* DSM 13855 [66.2% GC], *Thiobacillus denitrificans* ATCC 25259 [66.1% GC], *Enterococcus faecalis* V583 [37.5% GC], *Anaplasma marginale* str. St. Maries [49.8% GC] and *Nitrosococcus oceani* ATCC 19707 [50.3% GC]. Adapted from [45].

Comparing $R_{sequence}$ and RE

Site scoring indices can be formally derived from the $R_{sequence}$ and *RE* closed forms of Equations 8 and 13, respectively [46-48]. These scoring indices can be then be used to analyze genomic sequences with a sliding window approach, searching for instances of sites that conform to the transcription factor-binding motif [45]. Site search using these indices provides the means to uncover previously unidentified sites in a genomic sequence, which can be validated afterwards by experimental means [38, 49]. As with any other classification

system, these indices can then be ranked according to their rates of false positives and false negatives when operating under different classification thresholds [45, 50]. This information is typically conveyed by means of a Receiver-Operating Characteristic (ROC) curve. In addition, and because they are strongly and formally tied to the TF-binding motif models on which they are based, these scoring indices can be used to gauge the accuracy and validity of their underlying models.

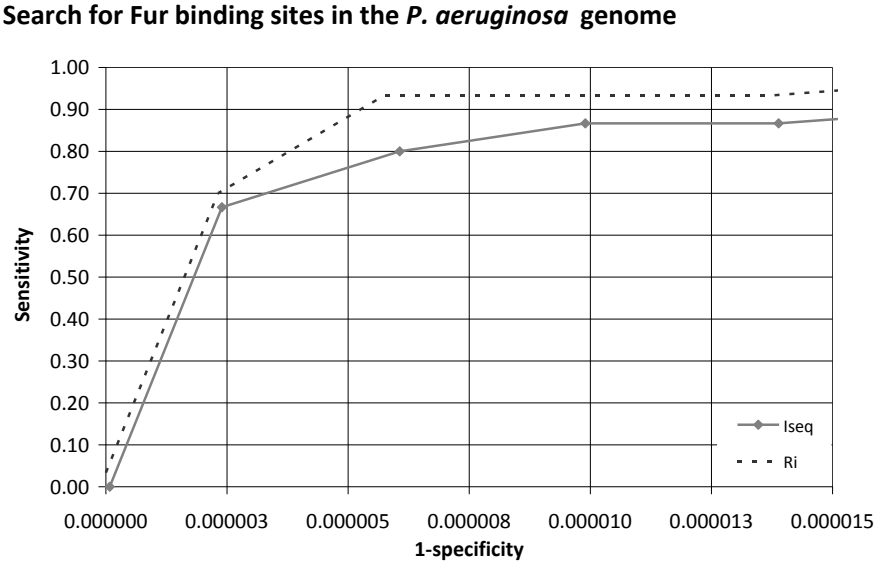

Figure 4. ROC curves for a *RE*-based search method (I_{seq}) and an $R_{sequence}$-based one (R_i) trying to locate 32 known *P. aeruginosa* Fur binding sites on the *P. aeruginosa* genome. The *RE*-based method performs worse than the $R_{sequence}$-based one despite the apparent boost in information content provided by the genomic bias ($R_{sequence}$=13.69 bits, *RE*=20.72 bits). The plot is scaled to encompass a 1/10 true to false positive ratio (320 false positives).

A main problem of *RE* and its derived indices is that the correction imposed on the background genomic entropy assumes a uniform distribution. As it can be seen in Figure 3, this turns out to be a poor estimate of the distribution of short DNA segments in %GC-biased genomes. Due to functional constraints, AT-rich 20 bp segments are notably overrepresented in GC-rich genomes, as compared to the expectation under a uniform hypothesis. The same is true for GC-rich 20-mers in AT-rich genomes and, to a lesser degree, of AT-rich segments in AT-rich backgrounds [45]. Most prokaryotic transcription factors target AT-rich binding motifs, typically in the 60-80% AT range. *RE* weights up the contribution of AT-rich positions in a GC-rich background and, as a result, *RE*-derived methods yield also higher scores for positions with conserved A and T bases. Due to the overrepresentation of AT-rich segments in GC-rich genomes, this has the effect of generating a larger amount of false positives for *RE*-based indices Figure 4, leading to poorer results than those obtained by $R_{sequence}$-derived methods [45].

The Search and Ranking Problems

The use of *RE* as a model for transcription factor-binding motifs is misguided by the assumption of background uniformity, but the notion of increased information content due to a departure from the genomic background composition is still intuitively appealing. Indeed, one can reason from a mechanistic point of view that a transcription factor targeting GC-rich binding sites in an AT-rich background should spend less time engaged in semi-specific binding at GC-rich segments (less likely to occur by chance) other than those under functional selection, thereby speeding up the detection of its binding sites. In a computational search for TF-binding sites on a sequenced genome, this translates into a lower false positive rate. This line of reasoning led Schneider *et al*. to postulate that a transcription factor sitting in a genome that is slowly evolving a %GC bias would tend to evolve a binding motif *opposing* the bias. Furthermore, they reasoned that in doing so the TF-binding motif would likely lose part of its information content, because the recognized motif would carry additional information in terms of its composition against the bias.

The available evidence indicates that TF-binding motifs do not evolve against the genomic bias [45]. A collection of 45 CRP-binding sites from *Haemophilus influenzae* (38.1% GC genome) shows that the CRP-binding motif in this organism is 69.89% AT-rich, versus 64.68% AT in *E. coli*. Likewise, the *P. aeruginosa* (66.56% GC genome) Fur collection shown in Figure 2 indicates that the Fur-binding motif in this organism displays a 70.72% AT composition, versus 74.71% AT in *E. coli*. Hence, it appears that TF-binding motifs tend to evolve, if anything, *with* the genomic bias, instead of against it. In accordance with this fact, neither of the above motifs has evolved lower information content. In fact, it seems that TF-binding motifs might evolve towards higher information values in biased genomes. The *P. aeruginosa* Fur motif maintains roughly the same information content as the *E. coli* one, while the *H. influenzae* CRP motif shows a significant increase (17.83 bits for 10.09 bits in *E. coli*) that cannot be attributed solely to undersampling in the *H. influenzae* collection [45].

The results outlined above are difficult to accommodate in a search-centered view of transcription factors. Because it yields simpler mechanistic and mathematical models, transcription factor activity has been traditionally conceptualized as an ON-OFF system. Nonetheless, many biochemical studies have established that transcription factors exhibit a wide variety of binding affinities for different binding sites [51-54]. Hence, a transcription factor must accomplish two different tasks: (1) locate its binding sites and (2) bind to them with a certain affinity. Even though both tasks are partially intertwined, it is easy to see that they are not strictly equivalent. One can, for instance, conceive of a transcription factor-binding motif with 20 positions, 10 of them fully conserved and 10 of them equiprobable, leading to $R_{sequence}$=20 bits. A transcription factor recognizing such a motif would be able, in principle, to find without much problem four binding sites in the *E. coli* genome ($R_{frequency}$=20.15). However, it seems apparent that it would not be able to distinguish among each of the four sites and, therefore, it would bind them all with the same affinity.

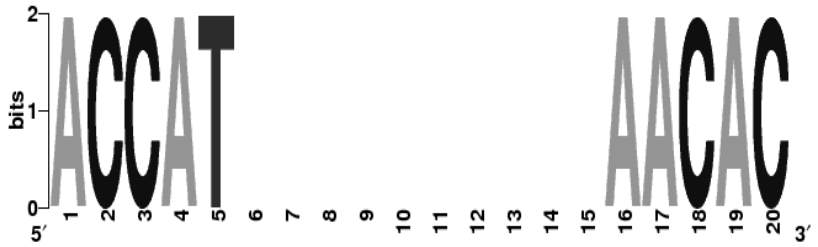

Figure 5. Sequence logo for a searchable, yet non-rankable, TF-binding motif. The logo was generated using the WebLogo server [37].

In deriving the formula for $R_{frequency}$ (Equation 12) we made the explicit assumption that all binding sites were bound with equal probability and that the transcription factor did not bind elsewhere. Hence, $R_{frequency}$ is explicitly based on an ON-OFF model of transcription factor and, as such, it deals exclusively with the search problem. As we have seen, however, the evolutionary interpretation of $R_{sequence}$ frees it from such constraints. Even though $R_{sequence}$ is based on the observed frequencies for the bound state $H(S_l|TF_{bound})$, the affinity with which the transcription factor binds each site is only bounded by the functional requirement of each particular site, as assessed by the evolutionary process over time. Hence, $R_{sequence}$ captures implicitly, in the $H(S_l|TF_{bound})$ term, the different binding affinities of the transcription factor for each of the known binding sites. Still, sites need to be located before they are bound with a certain affinity, and it follows that search requirements will also be assessed functionally by evolution and that they will also be encoded within $R_{sequence}$. As a consequence, $R_{sequence}$ must be interpreted as a *compound* measure of the search *and* ranking requirements of the binding sites targeted by a particular transcription factor.

The search and ranking problems are intertwined because, for a given concentration of the transcription factor, the amount of time spent by the transcription factor at a particular site will be impacted directly by the time spent by the transcription factor searching for the site. Based on $R_{sequence}$, one can derive sound ranking functions for transcription factor-binding sites [55] and, by plotting the rank scores of each site, it is possible to analyze the theoretical affinity range targeted by a particular transcription factor (Figure 6). The *effective* affinity range for the transcription factor, however, must incorporate the indirect effects of the search process on the perceived affinity. By normalizing the computed affinity of a binding site by the number of pseudo-sites of equal or higher affinity found in the genome (i.e. the non-functional sites that are likely to sequester the transcription factor), one can provide a first-order approximation of the effective affinity range (Figure 7).

As it can be seen when comparing Figure 6 and Figure 7, factoring in the search component can have a very substantial impact on the effective affinity range of transcription factors. The effective affinity range of TF-binding motifs with low information content, like Fis ($R_{sequence}$=5.16 bits), is strongly affected by the search process because all but the best sites become rapidly indistinguishable from the background. On the opposite end, TF-binding motifs with very high information content, like LexA ($R_{sequence}$=21.33 bits), are able to maintain their original linear range, since the search process has little impact on their effective affinity. In between both extremes, transcription factors may endorse a number of strategies to reach a compromise between site conservation ($R_{sequence}$), the desired effective affinity range and a viable concentration of transcription factor.

In equiprobable genomic backgrounds ($H_G \approx 2$ bits), search and ranking are fundamentally linked because the search and ranking processes operate on an even ground. Still, one can envision situations in which the need for a specific regulatory range imposes additional restrictions on $R_{sequence}$ (e.g. the aforementioned LexA repressor). In biased genomes, however, the interplay between both processes is unbalanced by the relative abundance of pseudo-sites in the genomic background. Based on the equivalence between $R_{sequence}$ and $R_{frequency}$, and on the replacement of $R_{sequence}$ by *RE* in biased genomes, Schneider and co-workers predicted that a transcription factor harbored by a genome evolving a %GC bias would tend to evolve against the bias and, in the process, shed positional information content ($R_{sequence}$). By explicitly including both ranking and search as evolutionarily actionable terms in $R_{sequence}$, we can now recast the evolutionary scenarios faced by a transcription factor on a genome evolving towards %GC bias.

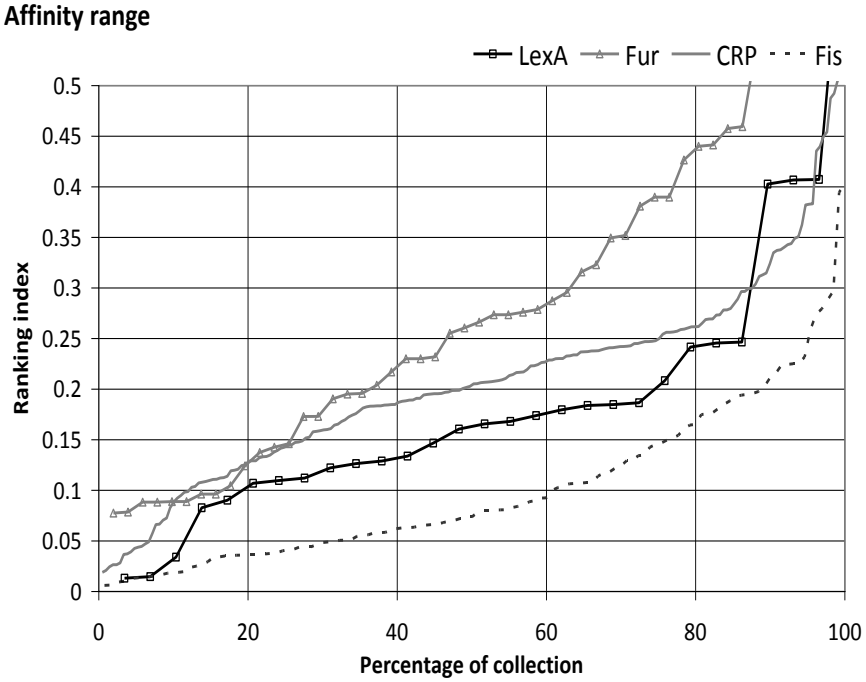

Figure 6. Computed affinity range for different transcription factors of *E. coli*. The affinity range is represented as the distribution of computationally inferred affinities for all its experimentally determined binding sites. Computed affinity values (Y-axis) are normalized to the length of the binding motif for each transcription factor and ranges (X-axis) are shown as the percentage of experimentally determined sites (collection). Adapted from [45].

When sitting in a genome evolving towards high %GC content, a transcription factor targeting an AT-rich motif faces a decreased impact of the search component on its effective affinity range. In order to maintain a similar function (i.e. a similar effective affinity range), the transcription factor may lower its cell concentration or bias its binding motif towards the genomic background, since both strategies will restore partly the impact of the search process on the effective regulatory range. It is unlikely, however, that the transcription factor would lower the positional information content ($R_{sequence}$) of its motif, since this would only further

reduce its ranking range. A transcription factor targeting an AT-rich motif in a genome drifting towards high %AT content faces the opposite scenario: its effective affinity range is heavily impacted by the search process in the new background. In such a case, the transcription factor may opt to increase its concentration, but such a strategy has a limited scope. Increasing the positional information content provides instead a gradual mechanism to match the increasing impact of the search process while maintaining an accurate ranking function.

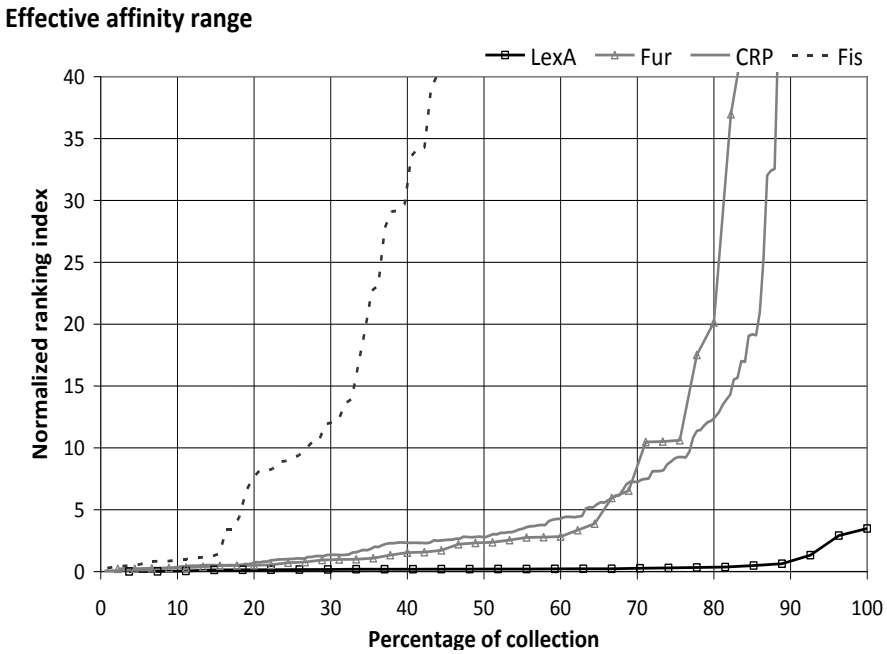

Figure 7. Estimation of the effective affinity range, represented as the distribution of normalized affinities for all its experimentally determined binding sites. Normalized affinities are estimated by normalizing the computed ranking index for each site with the number of false positives incurred locating the site. Computed affinity values (Y-axis) are normalized to the length of the binding motif for each transcription factor and ranges (X-axis) are shown as the percentage of experimentally determined sites (collection). Adapted from [45].

In general, the evolutionary perspective taken here on $R_{sequence}$ provides an intuitive interpretation for the fact that $R_{sequence}$ will decrease systematically whenever the genomic background deviates significantly from equiprobability. The mathematical reason for this effect is a net decrease in the background genomic entropy (H_{before}; Equation 8), but this has long defied interpretation, because it is easy to see that a bias in the background can be used to boost recognition of certain elements that deviate strongly from the bias. In terms of information theory, however, this is akin to a free lunch proposition, because it implies that a source with reduced entropy is capable of transmitting more information and we know this to be false. We have seen above that evolution is in theory capable of partitioning the source message. Such a strategy might be used to exploit a bias effect, but the fact remains that less information is available at the source. The interaction between the search and ranking

processes for transcription factors provides an illuminating glimpse into the complex evolutionary tradeoff generated by a reduction in source entropy. Even though source bias can theoretically be used to improve search efficiency, the reduction in source entropy has a negative effect on the regulatory range available to the transcription factor and thus limits the extent to which such a strategy can be exploited.

4.2. On Negative Information

The decrease of $R_{sequence}$ with decreasing background entropy (H_{before}) has puzzled researchers for years due to the lack of a proper evolutionary perspective on which to frame this effect. The intuitive feeling that genome bias could be exploited to improve search efficiency was one of the main arguments to put forward relative entropy (*RE*) as a corrected measure of positional information content ($R_{sequence}$) in biased genomes [34]. A less voiced, but strong argument to advocate the use of *RE* was the advent of negative information in transcription factor-binding motifs when using the $R_{sequence}$ measure. This perplexing result illustrates perfectly the need for an evolutionary perspective on the application of information theory to biology and, in particular, to transcription factor binding as an information process.

Analyzing Negative Information

In Figure 1 we analyzed the information content of the hypothetical transcription factor BUH, derived from the list of sites shown in Table 1. In doing so, we implicitly assumed that the BUH transcription factor belonged to *E. coli* or another species with a quasi-equiprobable background. With hindsight, we might conclude that such an assumption was unfounded and that, for all we know, the BUH transcription factor could belong to *T. thermophilus*. As seen above, the genomic entropy (H_{before}) for *T. thermophilus* is just 1.876 bits and we should therefore expect that when we compute $R_{sequence}$ for BUH in this new background we shall obtain a lower information content value (4.36 bits, instead of the original 5.45 bits in *E. coli*). What is not so expected is the fact that several positions of the BUH-binding motif will yield *negative* $R_{sequence}$ values (Figure 8). This result is counterintuitive because $R_{sequence}$ was devised to measure the amount of information that we gain upon observing binding of the transcription factor to a given DNA segment, and it is hard to imagine how we could have a negative gain of information as a result of such a process.

The fact that $R_{sequence}$ can generate negative values for certain genomic backgrounds is seen by many as a flaw in the derivation of $R_{sequence}$ and, implicitly, as an argument for the use of *RE* (which will never generate negative values) as a superior index of information content in TF-binding motifs. As a matter of fact, however, the ability of $R_{sequence}$ to generate negative values should be taken as an argument *supporting* the use of $R_{sequence}$ as the proper measure of information content in transcription factor-binding motifs. As in the previous cases, this requires that we take an evolutionary stance in order to analyze what is wrong with the computation of the BUH $R_{sequence}$ value in the *T. thermophilus* background.

As seen in Table 2, some of the central positions in BUH are close to equiprobability. This is a relatively common phenomenon among *E. coli* transcription factors because many of them operate as dimers and perform specific recognition only at the ends of the motif. The central part of the motif is hence only loosely contacted and typically shows equiprobable or

AT-rich composition[11]. Most importantly, one should realize that if a position of a TF-binding motif is not involved significantly in binding it will not be acted upon by natural selection. In *E. coli*, this implies that such a position of the TF-binding motif will be approximately equiprobable, because the genomic entropy is close to 2 bits. This will, in turn, lead to an $R_{sequence}(l)$ value close or equal to zero. In *T. thermophilus*, however, there is an active push towards a lower entropy state all across the genome[12]. Hence, if a position is not important for binding in *T. thermophilus*, it will tend to drift towards the genomic bias (69.2 %GC), but it should *not* show an equiprobable base composition, since this would involve active selection towards such a state and we just established that this particular motif position was under no selection for binding. By generating negative values, $R_{sequence}$ is thus letting us know that our artificial transplantation of BUH into *T. thermophilus* violates the primary assumption on which $R_{sequence}$, and biology in general, rests: evolution as the core information process.

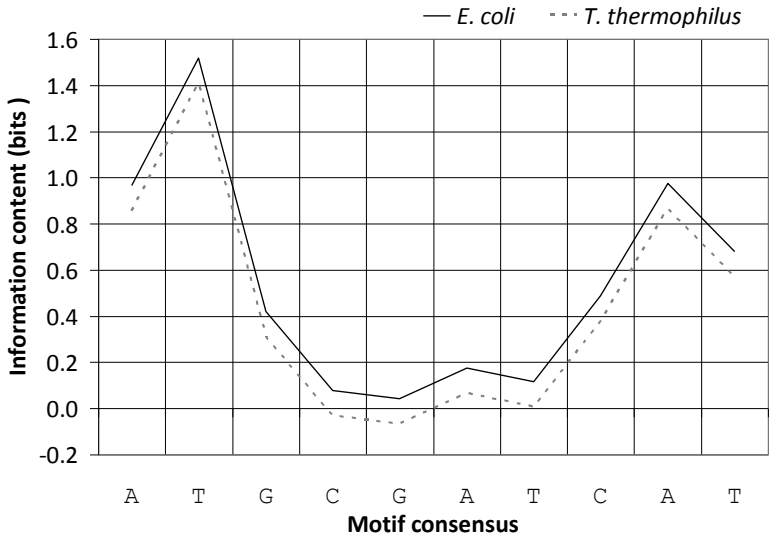

Figure 8. Negative information in the binding motif of the hypothetical transcription factor BUH. The plot shows the $R_{sequence}(l)$ profile BUH on *E. coli* and *T. thermophilus*. On *T. thermophilus*, positions 4 and 5 have, respectively, $R_{sequence}$ values of -0.03 and -0.07 bits.

4.3. Shannon Entropy and Thermodynamics

As mentioned in the introduction, the formal and nomial similarities between the Boltzmann entropy of thermodynamics and the Shannon entropy of information theory have led to much confusion when applying information theory to living systems. In particular, the assumption of a straight connection between both entropies has had a lasting impact in the

[11] The AT-rich composition is related to structural requirements on the DNA that favor bending and facilitate recognition. Information theory-based search methods are able to pick up partly such loosely conserved signals [45].

[12] The push towards a lower entropy state in the genome may be accidental or the result of active selection. In extremophiles, like *T. thermophilus*, high %GC content is often selected for because it enhances DNA stability at high temperature or salt concentrations [56]. In other species, the pull towards high or low %GC backgrounds appears mainly to be driven by a bias in mutation rate patterns [57].

field and is relevant to the application of information theory to transcription factor-binding motifs developed here.

The Link between Boltzmann and Shannon Entropies

For a physical system, the Boltzmann entropy is defined as:

$$S = -K_B \log(W) \tag{15}$$

where K_B corresponds to the Boltzmann constant, expressed in units of energy divided by temperature and W is the number of possible microstates of a physical system in thermodynamic equilibrium. Boltzmann entropy can be generalized to take into account uneven probabilities for the microstates, leading to the Boltzmann-Gibbs entropy formulation:

$$S = -K_B \sum_{i=1}^{W} p_i \cdot \log(p_i) \tag{16}$$

in which p_i is the probability that a given microstate i occurs during the fluctuations of the system [58].

Needles to say, Equations 15 and 16 present a strong resemblance to the formula for Shannon entropy derived in Equation 1. In fact, the Boltzmann-Gibbs entropy can be recast as a measurement of an observer's average uncertainty over the specific microstate description of a given physical system [58]. Jaynes noted, specifically, that "the thermodynamic entropy is identical with the information-theory entropy of the probability distribution except for the presence of the Boltzmann's constant". Indeed, the dimensionality that the Boltzmann constant introduces to thermodynamic entropy is one of the key elements that have prevented the acceptance of its interpretation by means of information theory. As Ben-Naim deftly points out, however, the involvement of energy and temperature in the definition of entropy are basically a historical accident due to the definition of temperature and its units before the acceptance of the atomistic theory. Once temperature is re-defined as the mean kinetic energy of atoms, it is easy to see that Boltzmann constant is, in essence, dimensionless [16, 59].

In spite of the obvious connection between both entropies, it is incorrect to assume that they are equivalent beyond the purely theoretical setting. Jaynes warns that "the major occupational disease of this field is a persistent failure to distinguish between the information entropy, which is a property of any probability distribution, and the experimental entropy of thermodynamics, which is instead a property of a thermodynamic state as defined, for example by such observed quantities as pressure, volume, temperature, magnetization, of some physical system" [17]. He also points out that the formal equivalence between both equations does "not in itself establish any connection between these fields" [17]. These statements are not contradictory with Jayne's earlier claim of identity between both entropies. Jaynes is simply stating that the information theory interpretation of thermodynamic entropy is namely that, an interpretation. Both entropies are measures of information (or uncertainty), but they are usually applied to measure different things. It is only when the more general definition of Shannon entropy is specifically adapted to the microstate description of a physical system that both entropies become equivalent. This may seem to be a subtle distinction, but it is extremely significant, because the equivalency is accomplished by

introducing the Boltzmann constant (~1.38^{-23} JK^{-1}). Hence, when not explicitly dealing with atomic microstates, Shannon entropy is separated from Boltzmann entropy by many orders of magnitude.

Information Theory and the Free Energy of Binding

When dealing with transcription factors and their molecular interactions with DNA many authors have assumed that a tacit link can be established between the information content of a transcription factor-binding motif and the free energy of binding. In fact, because we are dealing with molecular ensembles, it is tempting to associate the conformational states of transcription factor binding to Boltzmann microstates and, through them, to Shannon entropy [60]. In fact, almost all information theory-based scoring indices proposed to date have been identified explicitly with binding free energy [32, 48, 60]. The fundamental problem with such an approach is not so much the lack of a formal derivation for the proposed equality, but the implicit assumption that the information process from which the respective indices are derived corresponds to the physical process of binding of the transcription factor to its binding site.

As we have seen above, when we analyze a genome sequence or a set of binding site sequences for a given transcription factor, we are witnessing the end-result of the evolutionary information process. Hence, the positional base probabilities inferred from the analysis of a collection of transcription factor-binding sites do not have a direct correspondence with the information process that takes place when a transcription factor binds one of its target sites (and which could be theoretically mapped to the thermodynamic processes involved). Instead, the positional base probabilities inferred from a binding motif correspond to the amount of information fixated by evolution in order for the transcription factor to perform its required function. As such, they correspond to the selective pressure exerted over evolutionary time on a multitude of thermodynamic binding events operating under varying environmental constraints. Thus they not only convey information on the relative free binding energies of all the binding processes involved, but also on the specific functional requirements of each binding process over a population of genetically related organisms. Undoubtedly, the evolutionary fixation of base states in the DNA sequence must ultimately bear some relationship with the thermodynamic binding processes on which selection for such fixation is exerted, but without explicit knowledge of the selective constraints involved the use of information theory estimates to predict binding free energies must be taken as an approximation of uncertain degree.

ACKNOWLEDGMENTS

I am indebted to Michael C. O'Neill and Thomas D. Schneider for extensive discussions about information theory, thermodynamics and their application to transcription factor binding sites, without which this chapter and many of the ideas therein would have never been conceived.

REFERENCES

[1] Gatlin LL. Information *Theory and the Living System*. New York: Columbia University Press 1972.

[2] Varela FG, Maturana HR, Uribe R. Autopoiesis: the organization of living systems, its characterization and a model. *Currents in modern biology*. 1974 May;5(4):187-96.

[3] Cannon WB. The wisdom of the body. New York: W W Norton & Co. 1932.

[4] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biophysics*. 1943;7:115-33.

[5] Werbos PJ. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting: Wiley-Interscience* 1994.

[6] Kohonen T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*. 1982;43(1):59-69.

[7] Crick FH. On protein synthesis. *Symposia of the Society for Experimental Biology*. 1958;12:138-63.

[8] Crick F. Central dogma of molecular biology. *Nature*. 1970 Aug 8; 227(5258):561-3.

[9] Gamow G, Rich A, Ycas M. The problem of information transfer from the nucleic acids to proteins. *Advances in biological and medical physics*. 1956;4:23-68.

[10] Yockey HP. *Information theory, evolution, and the origin of life*. New York: Cambridge University Press 2005.

[11] Von Neumann J, Burks AW. *Theory of self-reproducing automata*. Urbana: University of Illinois Press. 1966.

[12] Sipper M. Fifty years of research on self-replication: an overview. *Artif Life.* 1998 Summer; 4(3):237-57.

[13] Brooks DR, Wiley EO. *Evolution as entropy: toward a unified theory of biology*. 2nd ed. Chicago: University of Chicago Press 1988.

[14] Yockey HP. Information theory and molecular biology. Cambridge: Cambridge University Press 1992.

[15] Tribus M, McIrvine EC. Energy and information. *Scientific American*. 1971;224(September):178–84.

[16] Ben-Naim A. *A farewell to entrop : statistical thermodynamics based on information : S=logW*. Hackensack, N.J.; London.: World Scientific 2008.

[17] Jaynes ET. Papers On Probability, *Statistics and Statistical Physics*. Dordrecth: Reidel publishing Co. 1983.

[18] Tribus M, Shannon PT, Evans RB. Why thermodynamics is a logical consequence of information theory. *Journal of the American Institute of Chemical Engineering*. 1966;12:244-8.

[19] Gatlin LL. The information content of DNA. *J Theor Biol.* 1966 Feb; 10(2): 281-300.

[20] Gatlin LL. The information content of DNA. II. *J Theor Biol.* 1968 Feb; 18(2): 181-94.

[21] Shannon CE. A mathematical theory of communication. *Bell System Technical Journal.* 1948 July 1948;27:379-423 623-56.

[22] Shannon CE. Prediction and Entropy of Printed English. *Bell System Technical Journal.* 1950;3:50-64.

[23] Adami C, Cerf NJ. Physical complexity of symbolic sequence. *Physica D.* 2000;137:62–9.

[24] Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge Univ Press 1998.

[25] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics* (Oxford, England). 2007 Aug 1; 23(15): 1875-82.

[26] Fickett JW, Tung CS. Assessment of protein coding measures. *Nucleic acids research*. 1992 Dec 25;20(24):6441-50.

[27] Herzel H, Ebeling W, Schmitt AO. Entropies of biosequences: The role of repeats. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*. 1994 Dec;50(6):5061-71.

[28] Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*. 2007 Feb;8(2):93-103.

[29] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009 Jan 23;136(2):215-33.

[30] Ptashne M. Regulation of transcription: from lambda to eukaryotes. *Trends in biochemical sciences*. 2005 Jun;30(6):275-9.

[31] Orphanides G, Reinberg D. A unified theory of gene expression. *Cell*. 2002 Feb 22;108(4):439-51.

[32] Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*. 1987 Feb 20;193(4):723-50.

[33] Halford SE, Marko JF. How do site-specific DNA-binding proteins find their targets? *Nucleic acids research*. 2004;32(10):3040-52.

[34] Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *Journal of molecular biology*. 1986 Apr 5;188(3):415-31.

[35] Garner MM, Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic acids research.* 1981 Jul 10;9(13):3047-60.

[36] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*. 1990 Oct 25;18(20):6097-100.

[37] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome research*. 2004 Jun;14(6):1188-90.

[38] Erill I, Jara M, Salvador N, Escribano M, Campoy S, Barbe J. Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. *Nucleic acids research*. 2004;32(22):6617-26.

[39] Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics*. 1951;22:79-86.

[40] Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic acids research*. 2005 Jul 1;33(Web Server issue):W389-92.

[41] Singh S. The Code Book: *The Science of Secrecy from Ancient Egypt to Quantum Cryptography: Anchor Books* 2000.

[42] O'Neill MC. Consensus methods for finding and ranking DNA binding sites. Application to *Escherichia coli* promoters. *Journal of molecular biology*. 1989 May 20;207(2):301-10.

[43] Stormo GD. Information content and free energy in DNA--protein interactions. *J Theor Biol*. 1998 Nov 7;195(1):135-7.

[44] Bailey TL, Boden M, Whitington T, Machanick P. The value of position-specific priors in motif discovery using MEME. *BMC bioinformatics*. 2010; 11:179.

[45] Erill I, O'Neill MC. A reexamination of information theory-based methods for DNA-binding site identification. *BMC bioinformatics*. 2009;10(1):57.

[46] Schneider TD. Information Content of Individual Genetic Sequences. *Journal of Theoretical Biology*. 1997;189(4):427-41.

[47] Hertz GZ, Hartzell GW, 3rd, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci*. 1990 Apr;6(2):81-92.

[48] tormo GD, Fields DS. Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences.* 1998 Mar; 23(3): 109-13.

[49] Guerin E, Cambray G, Sanchez-Alberola N, Campoy S, Erill I, Da Re S, et al. The SOS Response Controls Integron Recombination. *Science.* 2009 May 22, 2009;324(5930):1034.

[50] Babu MM. Computational approaches to study transcriptional regulation. *Biochemical Society transactions*. 2008 Aug;36(Pt 4):758-65.

[51] Schnarr M, Oertel-Buchheit P, Kazmaier M, Granger-Schnarr M. DNA binding properties of the LexA repressor. *Biochimie*. 1991 Apr;73(4):423-31.

[52] Kolb A, Spassky A, Chapon C, Blazy B, Buc H. On the different binding affinities of CRP at the *lac*, *gal* and *malT* promoter regions. *Nucleic acids research.* 1983 November 25, 1983;11(22):7833-52.

[53] Gaston K, Kolb A, Busby S. Binding of the *Escherichia coli* cyclic AMP receptor protein to DNA fragments containing consensus nucleotide sequences. *The Biochemical journal*. 1989 Jul 15;261(2):649-53.

[54] Baichoo N, Helmann JD. Recognition of DNA by Fur: a Reinterpretation of the Fur Box Consensus Sequence. *Journal of bacteriology*. 2002 November 1, 2002;184(21):5826-32.

[55] O'Neill MC. A general procedure for locating and analyzing protein-binding sequence motifs in nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America*. 1998 Sep 1; 95(18): 10710-5.

[56] Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S. Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome research*. 2001 Oct;11(10):1641-50.

[57] Freese E. On the evolution of the base composition of DNA. *Journal of Theoretical Biology*. 1962;3(1):82-101.

[58] Jaynes ET. Information Theory and Statistical Mechanics. *Physical Review*. 1957;106(4):620.

[59] Kalinin M, Kononogov S. Boltzmann's Constant, the Energy Meaning of Temperature, and Thermodynamic Irreversibility. *Measurement Techniques*. 2005;48(7):632-6.

[60] Schneider TD. Theory of molecular machines. I. Channel capacity of molecular machines. *J Theor Biol*. 1991 Jan 7;148(1):83-123.

Reviewed by Michael C. O'Neill