Journal of Theoretical Biology **220** (2003): 529-544

Bioinformatic Principles Underlying the Information Content of Transcription Factor Binding Sites

Jan T. Kim, Thomas Martinetz and Daniel Polani

Institut für Neuro- und Bioinformatik
Seelandstraße 1a, 23569 Lübeck, Germany

phone: +49 451 3909-586, fax: +49 451 3909-545
email: `{kim,martinetz,polani}@informatik.uni-luebeck.de`

Keywords: DNA binding protein, transcription factor, binding site, coevolution, information theory, maximum entropy principle

Abstract

Empirically, it has been observed in several cases that the information content of transcription factor binding site sequences (R_{sequence}) approximately equals the information content of binding site positions ($R_{\text{frequency}}$). A general framework for formal models of transcription factors and binding sites is developed to address this issue. Measures for information content in transcription factor binding sites are revisited and theoretic analyses are compared on this basis. These analyses do not lead to consistent results. A comparative review reveals that these inconsistent approaches do not include a transcription factor state space.

Therefore, a state space for mathematically representing transcription factors with respect to their binding site recognition properties is introduced into the modelling framework. Analysis of the resulting comprehensive model shows that the structure of genome state space favours equality of R_{sequence} and $R_{\text{frequency}}$ indeed, but the relation between the two information quantities also depends on the structure of the transcription factor state space. This might lead to significant deviations between R_{sequence} and $R_{\text{frequency}}$. However, further investigation and biological arguments show that the effects of the structure of the transcription factor state space on the relation of R_{sequence} and $R_{\text{frequency}}$ are strongly limited for systems which are autonomous in the sense that all DNA binding proteins operating on the genome are encoded in the genome itself. This provides a theoretical explanation for the empirically observed equality.

1 Introduction

Biological systems store genetic information in DNA sequences, and transfer of this type of biological information into electronic media is nowadays occurring at genomic scale. Having gained access to sequence information of several complete genomes, understanding the fundamental principles by which genetic information controls and organizes complex biological processes now turns into a primary challenge for bioinformatics.

All these processes necessarily involve sequence specific contacts between DNA subsequences and molecules with a biological activity, and this function is performed by transcription factors (Pabo & Sauer, 1992). Transcription factors and regulatory gene networks are known to play key roles in fundamental biological processes, such as metabolic dynamics, development and morphogenesis (Kappen & Ruddle, 1993; Shore & Sharrocks, 1995; Theißen & Saedler, 1995).

Consequently, transcription factors and their binding sites have received much scientific interest during the last several years, as evidenced by the development of various specialized databases (Wingender *et al.*, 2001; Kolchanov *et al.*, 2000) and tools for recognizing transcription factor binding sites and other sequence motifs involved in sequence specific protein-DNA interactions (Frech *et al.*, 1997). However, a theoretical and principled basis for understanding binding site sequences and their evolution has not yet been developed.

The analysis of binding sites developed by SCHNEIDER (Schneider *et al.*, 1986; Schneider, 2000) is based on information theory, and thus it recommends itself as a point of departure for building such a basis. In this article, we revisit and extend this theoretic approach by employing maximum entropy considerations. This principle has proven useful in other bioinformatic contexts (Schmitt & Herzel, 1997).

SCHNEIDER *et al.* observed an approximate equality of the information content of binding site sequences (R_{sequence}) and the information content of binding site positions, calculated on the basis of binding site frequency within the genome ($R_{\text{frequency}}$) (Schneider *et al.*, 1986). Intuitively, this equality appears plausible: R_{sequence}, the amount of information required to identify one out of $2^{R_{\text{sequence}}}$ sequences as a binding sequence, may be expected to equal the amount of information necessary to address one site out of $2^{R_{\text{frequency}}}$ possible sites on the genome. However, this is just a vague plausibility argument, and counterexamples can easily be constructed. With this contribution, we aim to extend the theoretical basis for studying transcription factors, their binding sites, and their coevolution within the context of regulatory gene networks.

The centerpiece of our investigations is a comprehensive modelling framework which is accessible to formal analysis. A state space for the genome sequence and a state space for the transcription factor are defined. The Cartesian product of these two state spaces yields the state space of the entire system. We assume that, after coevolution, only those states will occur in which the transcription factor binds to binding sites and does not bind to non-binding sites. For mathematical analysis we consider an approximation in which we disregard the fact that binding sites may overlap because they span several nucleotides. In computer simulations we show that the mathematical description we obtain is in very good agreement with the original model system allowing overlapping binding sequences. With the mathematical model it is then possible to study systems of realistic sizes and to calculate their probability distribution over the state space of the transcription factor-genome system. For realistic sizes of this system the probability distribution of admissible states is strongly peaked. The peaks provide the states which are almost certainly observed and have to be expected as the outcome of a coevolution process. We then show that already for an unbiased (within the chosen coding scheme) a priori probability for the binding behaviour of the transcription factor the quantity R_{sequence} can significantly deviate from $R_{\text{frequency}}$. However, by using biological principles we can show that

Symbol	Meaning
$\mathcal{A}$	alphabet of letters occurring in the genome
$a = \|\mathcal{A}\|$	alphabet size, in this paper either 2 ($\mathcal{A} = \{0, 1\}$) or 4 ($\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$)
N	Genome length
$\mathcal{G} = \mathcal{A}^N$	set of all possible genome sequences
$\mathbf{d} = (d_1, d_2, \ldots, d_N)$	an individual genome, $\mathbf{d} \in \mathcal{G}$
$\mathbf{b} = (b_1, b_2, \ldots, b_N)$	pattern of binding sites ($b_i = 1$) and non-sites ($b_i = 0$) along the genome
n	Number of binding sites on the genome
L	Binding site length
$\mathcal{W} = \mathcal{A}^L$	the set of all words of binding site length, i.e. the set of words divided into binding words and non-binding words by the transcription factor
$K = \|\mathcal{W}\| = a^L$	the size of the set of all words of length L
$\mathbf{w}_i(\mathbf{d}) = (d_i, d_{i+1}, \ldots, d_{i+L-1})$	the word occurring at position i within a genome $\mathbf{d}$, $\mathbf{w}_i \in \mathcal{W}$
$(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$	a sequence of words, $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N) \in \mathcal{W}^N$
$\mathcal{T} = \{0, 1\}^K$	the set of all possible binding behaviours of transcription factors operating on words of length L
$\boldsymbol{\tau} = (\tau_{\mathbf{w}_1}, \tau_{\mathbf{w}_2}, \ldots, \tau_{\mathbf{w}_K})$	an individual transcription factor described by its binding behaviour $\boldsymbol{\tau} \in \mathcal{T}$
k	the number of words recognized by the transcription factor $\boldsymbol{\tau} \in \mathcal{T}$ as binding words

Table 1: Symbols for use in equations.

these deviations can expected to be limited.

2 General Modelling Framework

The modelling concepts which are used throughout this paper are introduced in this section, and the mathematical notation for these concepts is explained. A summary of these notational symbols, along with brief explanations, is shown in Table 1.

2.1 Genome Model

Genomes are represented by vectors of letters (strings). Let $\mathcal{A}$ be an alphabet, i.e. a finite set of symbols, and let $a = |\mathcal{A}|$ denote the size of the alphabet. We will either consider the case of $a = 2$ (binary strings, $\mathcal{A} = \{0, 1\}$) or $a = 4$ (nucleotide sequences, $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$). Let $\mathcal{G} = \mathcal{A}^N$ denote the genome space, i.e. the set of all possible genomes of fixed length $N \in \mathbb{N}$. Genomes are denoted by $\mathbf{d} = (d_1, d_2, \ldots, d_N)$, $N \in \mathbb{N}$, $d_i \in \mathcal{A}$. Obviously, the total number of genomes of length N is $|\mathcal{A}^N| = a^N$.

2.2 Binding Site Model

Along the genome, there are *binding sites* at which binding of the transcription factor is biologically required. The number of binding sites is denoted by n. In our modelling approach, we assume that at all other sites, which we call *non-binding sites* or *non-sites*, the transcription factor must not bind.

With this approach, we imply that failure of the transcription factor to bind at a binding site or binding occurring at a non-binding site uniformly result in an evolutionary disadvantage. This obviously is a gross simplification which neglects all differences in biological effects of mutations affecting the binding site or non-site status. In particular, this approach disregards the fact that there is a substantial amount of genomic sequences in which such mutations are neutral because these sequences are not accessible to transcription factors at all (e.g. heterochromatin) or because binding or non-binding of the transcription factor at these sites does not make any difference. It should be noted, however, that all these effects can conceptually be accounted for by considering the genome sequences in our modelling framework to represent only the "effective" part of the genome in which mutations which affect binding have, on average, a substantial deleterious effect. Consequently, our genome length N has to be interpreted as the effective genome length.

A pattern of n binding sites on a genome consisting of N symbols total can be represented by a bit string of length N in which n bits are set. Binding sites and non-sites induce a bit pattern along the genome, which we refer to as the *binding site pattern*. Formally, we denote such a binding site pattern by a binary vector $\mathbf{b} = (b_1, b_2, \ldots, b_N), b_i \in \{0, 1\}$ where $b_i = 1$ if the i-th site is a binding site and $b_i = 0$ otherwise. Trivially, the number of different possible binding site patterns with a given number of binding sites n amounts to $\binom{N}{n}$.

2.3 Transcription Factor Model

The transcription factor is conceptually modelled as a device for recognizing binding sites by locally inspecting words (substrings) of length L along the genome. The set of words of length L is denoted by $\mathcal{W}$. Its cardinality is $K = |\mathcal{W}| = a^L$. Out of these K words, the transcription factor accepts k words as *binding words*. All other words are called *non-binding words*. We assume that the transcription factor does not discriminate between different binding words, i.e., mutations that change one binding word into another one are neutral since they do not change the binding site pattern, and, likewise, replacements of a non-binding word with another one is a neutral mutation, too.

At this stage, the modelling concept is not yet sufficient to specify a state space for transcription factors. Such a state space is introduced in Section 5.

2.4 Approximations

2.4.1 Number of Sites in a Genome

The precise number of sites of length L in a linear genome is $N - L + 1$, and it is even smaller if the genome consists of linear segments (e.g. chomosomes). However, the relative difference between N and the exact number of sites is small if the segment lengths are large. Biological genome segments are sufficiently long to neglect "boundary effects" and to warrant approximating the number of sites with N, which we will do henceforth.

2.4.2 Genomes as Sequences of Independent Words

Within our modelling framework, it is convenient to think of a genome as a sequence of words. We define $\mathbf{w}_i(\mathbf{d}) := (d_i, d_{i+1}, \ldots, d_{i+L-1})$ as the word at site (position) i of genome $\mathbf{d}$. The sequence of words $(\mathbf{w}_1(\mathbf{d}), \mathbf{w}_2(\mathbf{d}), \ldots, \mathbf{w}_N(\mathbf{d}))$ provides a complete description of $\mathbf{d}$. We denote words up to $\mathbf{w}_N(\mathbf{d})$ following the approximation just introduced in Section 2.4.1.

Evidently, words beginning at consecutive positions on a genome are not independent as the last $L-1$ letters of $\mathbf{w}_i(\mathbf{d})$ are the first $L-1$ letters of $\mathbf{w}_{i+1}(\mathbf{d})$. We will, however, neglect this dependency in some of the following analyses, and consider sequences $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$ in which $\mathbf{w}_i$ is not related to $\mathbf{w}_{i+1}$ through an underlying genome $\mathbf{d}$. Generally, substituting genomes with sequences of independent words is a good approximation as long as $L \ll N$ because most words are independent in this case. A more specific justification based on an enumeration study is presented later in this paper.

3 Aspects of Information in the Context of Transcription Factor Binding Sites: R_{sequence} and $R_{\text{frequency}}$

Information can be measured in the context of sequence specific protein-DNA interaction in various respects which we briefly outline in this section. Details can be found in (Schneider *et al.*, 1986; Schneider, 2000; Stormo & Fields, 1998; Stormo, 1998). For principles of information theory, see (Cover & Thomas, 1991).

The binding of a transcription factor to its binding sites can be viewed as a process in which n specific positions out of all N positions in the genome are identified as binding sites. In this respect, observing a binding site in the genome is equivalent to receiving a message that identifies one out of N/n positions (on the genomic average). The information content of such a message can be quantified in bits by

$$R_{\text{frequency}} = -\log_2 \frac{n}{N}. \tag{1}$$

This quantity can also be interpreted as the amount of bits required to send the address of a binding site within N/n nucleotide positions across a communication channel (see (Schneider *et al.*, 1986)).

In another perspective, one may consider the binding behaviour of the transcription factor to be known and ask how much information about the word at some position in the genome can be gained by observing specific binding of the transcription factor at that position. If nothing is known about the sequence before observing the binding process, all of the K possible words could be present at that position. After observing binding, it is known that the word in question must be one of the k binding words recognized by the transcription factor. The amount of information gained through this observation can be calculated by

$$R_{\text{sequence}} = H_{\text{before}} - H_{\text{after}} \tag{2}$$

(see (Schneider, 2000)), where H_{before} and H_{after} denote the entropy of the probability distribution to encounter a particular word $\mathbf{w}$ before and after the observation of binding, respectively. If nothing is known about the sequence, the entropy is $H_{\text{before}} = \log_2 K = L \log_2 a$, since each word is equiprobable. After observing binding, entropy is generically given by

$$H_{\text{after}} = - \sum_{\mathbf{w} \in \mathcal{W}} P_{\mathbf{w}} \log_2 P_{\mathbf{w}}$$

where $P_{\mathbf{w}}$ denotes the probability of encountering word $\mathbf{w}$ as a binding word. Biochemically, $P_{\mathbf{w}}$ can be linked to the energy gained upon binding of the transcription factor to $\mathbf{w}$. If the nucleotides in $\mathbf{w}$ contribute independently to the total binding energy, the probability distribution over the symbols at a given position is independent of the symbols at other positions. As a consequence, H_{after} can then be calculated by adding the entropies over the individual positions. One obtains

$$R_{\text{sequence}} = L \log_2 a + \sum_{i=1}^{L} \sum_{b \in \mathcal{A}} P_{b,i} \log_2 P_{b,i} \tag{3}$$

where $P_{b,i}$ denotes the probability of encountering base b at position i. Empirically, the probability $P_{b,i}$ can be estimated by the frequency with which base b is observed at position i at a binding site (Berg & von Hippel, 1987). As direct estimations of the $P_{\mathbf{w}}$ are usually not possible since not enough data is available, Equation (3) assuming independency in the word positions is used in empirical studies.

In our model, however, we only distinguish between binding words and non-binding words, and we assume that all binding words are equally suitable to recruit a transcription factor molecule to a binding site. Hence, after observing binding, we know that one of the k binding words is present at the binding site. No information beyond this is gained. Without further prior knowledge, one has to assume that each of the k binding words can be present at this site with the same probability $1/k$ and obtains $H_{\text{after}} = \log_2 k$. Plugging this into Equation (2) yields

$$R_{\text{sequence}} = \log_2(K) - \log_2(k) = - \log_2 \frac{k}{K}. \tag{4}$$

This result corresponds to a maximum entropy assumption that it minimizes the information gain (2) by maximizing the entropy H_{after}. The entropy of a state probability distribution in a system state space is a measure for our uncertainty about the actual system state. In our case the state of the "system" is given by the word $\mathbf{w}$ at the site of the genom we are looking at. Without any prior knowledge, our uncertainty about which word is present at this site is maximum, i.e., one has to assume probabilities $P_{\mathbf{w}}$ for each word to be present at this site which maximize the entropy $H = - \sum_{\mathbf{w} \in \mathcal{W}} P_{\mathbf{w}} \log_2 P_{\mathbf{w}}$. In our case, without prior knowledge, this is the case for $P_{\mathbf{w}} = 1/K$, i.e. each word is equally probable to be present. After having gained the information and only the information that the site is a binding site, we again have to maximize our uncertainty, but now under the constraint that the word at the site is a binding word. This means that for each non-binding word $P_{\mathbf{w}}$ is equal to zero. The entropy as a measure for uncertainty is now maximized by $P_{\mathbf{w}} = 1/k$ for all binding words, i.e., all binding words are equally probable. Certain prior knowledge about the given sequence structure, e.g., information about varying CG-content or CpG suppression, can be incorporated by allowing only word probability distributions $P_{\mathbf{w}}$ for maximizing the entropy which comply with these constraints. This would lead to deviations from a uniform probability distribution, as one would expect. For an introduction to the maximum entropy principle as a means for deducing state probability distributions which can be used for state estimations see (Jaynes, 1952).

4 Initial Approaches to $R_{\text{sequence}} = R_{\text{frequency}}$

In this section, we review and introduce some information theoretic and probabilistic treatments of binding site recognition and evolution. This provides some background and motivation for our comprehensive model which we present in Section 5.

4.1 Random Sequence / Maximum Entropy Model

A binding site is biologically defined as a position in a genome at which a word is found which is recognized by the corresponding transcription factor. As an approach to deriving a relation between R_{sequence} and $R_{\text{frequency}}$, one can assume k, the number of binding words recognized by the transcription factor, to be given and set out to calculate n, the number of binding sites among all N positions along the genome.

Generically, there is very little constraint on n given a fixed value of k. For all values of k, genomes with an abundance of binding sites covering the full range of $0 \leq n \leq N$ can be constructed. (Only in the biologically irrelevant extreme cases of $k = 0$ and $k = K$, it trivially follows that $n = 0$ or $n = N$, respectively.)

Thus, additional assumptions must be made in order to arrive at more specific results for the relation between R_{sequence} and $R_{\text{frequency}}$. It is a standard approach to assume the abundance of all words in the genome to be (approximately) equal, i.e., by making a maximum entropy assumption about the word probability distribution. Without any prior knowledge, the maximum entropy approach suggests equal word abundance. With prior knowledge about e.g. non-uniform occurrence probabilites of bases in the sequence under inspection the maximum entropy approach leads to deviations from equal word abundance. In the following, since we are interested in the general principles of the different views, we always assume the case of no special prior knowledge.

These assumptions lead to an average word abundance of N/K. Consequently, the expected number of binding sites amounts to $n = k \cdot N/K$. By rearranging this equation, we obtain

$$\frac{n}{N} = \frac{k}{K} \quad \Leftrightarrow \quad R_{\text{frequency}} = R_{\text{sequence}}. \tag{5}$$

It is important to emphasize that this treatment involves modelling the genome as a fully random sequence. Of course, evolution of genomes may lead to non-random sequences in the sense that a uniform abundance of each word on the genome can not be expected anymore. This may lead to $n \neq k \cdot N/K$. For example, a transcription factor may accept only one binding word, but evolution may nonetheless produce genomes with high binding site densities by producing genomes in which the binding word is overrepresented.

It is important to notice that evolution of the transcription factor is neglected by this approach. It is assumed that evolution samples at random from the genome space as a state space (i.e. neutral evolution is assumed), while the transcription factor, which structures the genome state space, is assumed to be constant. This is a substantial difference from biological evolution in which transcription factors and genomes coevolve. This motivates us to introduce a more comprehensive model in Section 5.

4.2 Robustness to Mutations

A point mutation which transforms an existing binding word into another binding word does not create a new binding site on the genome. Thus, such a mutation does not change the pattern of binding sites and non-sites on the genome, and, therefore, it can be expected to be neutral, at least with respect to the pattern of binding sites of the particular transcription factor. Likewise, a mutation converting a non-binding word into another one is neutral. On the other hand, mutations transforming binding words into non-binding ones or vice versa result in changes of the binding site pattern.

According to this consideration, and since we regard only the "effective" part of the genome on which binding or non-binding of the transcription factor makes a difference for the organism (see Section 2.2), robustness against changes of the binding site pattern, i.e. a low probability for mutations which change a non-site into a binding site or vice versa, confers evolutionary stablility. A low probability of change means a high probability for an offspring to inherit all sites in their functional state. Thus, minimizing the probability of mutations which result in alterations of the binding site pattern confers an evolutionary advantage and might be prefered.

Assuming that the k binding words are randomly drawn from the set of K words, we obtain

$$\begin{aligned} P_{\text{site}\to\text{nonsite}} &= \frac{n}{N}\frac{K-k}{K} \\ P_{\text{nonsite}\to\text{site}} &= \frac{N-n}{N}\frac{k}{K}. \end{aligned}$$

As sites and nonsites are mutually exclusive sets, the probabiblity of changing the binding site pattern is the sum of these two probabilities:

$$\begin{aligned} P_{\text{change}} &= P_{\text{site}\to\text{nonsite}} + P_{\text{nonsite}\to\text{site}} \\ &= \frac{n}{N}\frac{K-k}{K} + \frac{N-n}{N}\frac{k}{K} \\ &= \frac{n}{N} + \frac{N-2n}{NK}\cdot k. \end{aligned}$$

In order to optimize evolutionary stability, k should adapt such that P_{change} is minimized. P_{change} is a linear function of k which ascends monotonically in the biologically relevant case of $n < N/2$. It follows that for minimizing P_{change}, the value of k should be minimal. This is achieved by choosing $k = 1$. Thus, as exemplified by this consideration, evolutionary forces may well induce deviations from $R_{\text{sequence}} = R_{\text{frequency}}$.

It should be noted that in addition to k, there are many parameters and properties open to adaptation in molecular evolution. In particular, a mutation of a single nucleotide position does not transform a word into an entirely unrelated word. Rather, the original and the mutant word have identical nucleotides at $L - 1$ positions. Therefore, the risk of mutating a binding word into a non-binding word or vice versa for a transcription factor recognizing k words can be substantially reduced by choosing the k words such that they are maximally sequence similar. It is interesting to note that this optimization can be achieved by transcription factor binding mechanisms in which each nucleotide in the binding site contributes to the binding energy independently of the nucleotides at the other positions. Indeed, this independence is observed experimentally (Stormo & Fields, 1998).

4.3 Computer Model

Recently, SCHNEIDER published a computer model simulating the evolution of information in transcription factor binding sites (Schneider, 2000), demonstrating a process of coevolution of a transcription factor and its set of binding sites that converges to a state in which $R_{\text{frequency}} \approx R_{\text{sequence}}$. This model is interesting because it allows for evolutionary adaptation of both the genome and the transcription factor. Nonetheless, the transcription factor model is rather specific in some respects. Firstly, the transcription factor is directly encoded as a profile matrix with a threshold value. However, not all binding behaviours which are formally possible (see Section 5.1.1) can be represented by a combination of a matrix and a threshold. Thus, not all possible binding behaviours can evolve in this model. Secondly, the particular encoding of the threshold by a six-digit number from a base four number system results in very fast adaptive changes in the number k of binding words recognized by the factor. This means that in the model, adaptation of R_{sequence} towards $R_{\text{frequency}}$ is much faster than adaptation of the sequences in the binding sites. This specific design makes it difficult to use the computer model for arriving at general conclusions. This was a major motivation for developing the model presented in the subsequent section.

5 A Comprehensive Mathematical Model

In this section, a model of a genome interacting with a transcription factor is described. As in the computer model by SCHNEIDER, both the genome and the transcription factor are open to evolutionary adaptation. However, the mathematical model we introduce is more generic and accessible to analytic treatment.

5.1 Model Definition

Genome space and binding site patterns are modelled as described in Section 2. Now, we formalize the transcription factor model to arrive at a state space comprising genomes as well as transcription factors.

5.1.1 Transcription Factor state Space

The properties of a transcription factor with respect to binding site recognition are described by simply listing all words which are recognized by the factor. Unconventional as it may initially appear, this approach is the simplest representation which ensures that all possible binding behaviours can be represented. Formally, we denote a transcription factor by a vector $\boldsymbol{\tau} = (\tau_{\mathbf{w}_1}, \tau_{\mathbf{w}_2}, \ldots, \tau_{\mathbf{w}_K})$, $\mathbf{w}_i \in \mathcal{W}$, $\tau_{\mathbf{w}_i} \in \{0, 1\}$, where $\tau_{\mathbf{w}_i} = 1$ if the factor accepts $\mathbf{w}_i$ as a binding word, and $\tau_{\mathbf{w}_i} = 0$ otherwise. The set of all possible transcription factors is denoted by $\mathcal{T}$.

This representation of the binding behaviour of the transcription factor is the most general one and does not prefer any particular division of the word space into binding words and non-binding words. It is the adequate representation in case no prior knowledge about the binding behaviour is given. If we assume equal pobability for each state $\boldsymbol{\tau}$ (maximum entropy), the situation of no prior knowledge about the binding behaviour is adequately captured.

Evidently, the number of vectors $\boldsymbol{\tau}$ amounts to $|\mathcal{T}| = 2^K = 2^{(a^L)}$. The value of k, the number of different words recognized by the factor, can be computed by

$$k = ||\boldsymbol{\tau}||_1 = \sum_{\mathbf{w}\in\mathcal{W}} \tau_{\mathbf{w}} \tag{6}$$

where $||\boldsymbol{\tau}||_1$ denotes the L_1 norm (also referred to as the Manhattan norm) of $\boldsymbol{\tau}$. It is useful to divide $\mathcal{T}$ into subsets according to k. We denote these subsets by

$$\mathcal{T}_k := \{\boldsymbol{\tau} \in \mathcal{T} : \sum_{\mathbf{w}\in\mathcal{W}} \tau_{\mathbf{w}} = k\}$$

and we note here that

$$|\mathcal{T}_k| = \binom{K}{k}. \tag{7}$$

5.1.2 The Full State Space

The full state space which we now use as a basis for our analyses is the Cartesian product of the genome space, as defined in Section 2.1, and the transcription factor space

$$\mathcal{S} := \mathcal{G} \times \mathcal{T} = \{(\mathbf{d}, \boldsymbol{\tau}) : \mathbf{d} \in \mathcal{G}, \boldsymbol{\tau} \in \mathcal{T}\}.$$

5.2 Structuring the State Space According to n and k

According to Equation (4), R_{sequence} is determined by k, and $R_{\text{frequency}}$ is determined by n according to Equation (1). By structuring $\mathcal{S}$ according to n and k we will be able to relate R_{sequence} and $R_{\text{frequency}}$ via maximum likelihood considerations.

We start by observing that for a given tuple $(\mathbf{d}, \boldsymbol{\tau})$, the value of k is given by Equation (6). The value of n can be calculated by

$$n = \sum_{i=1}^{N} \tau_{\mathbf{w}_i(\mathbf{d})}.$$

Thus, unique values for both n and k can be assigned to all elements of $\mathcal{S}$, and we can now set out to determine the abundance of $(\mathbf{d}, \boldsymbol{\tau})$ tuples with fixed n and k values. Formally, we structure $\mathcal{S}$ into subsets of tuples which are labelled by the same n and k values

$$\mathcal{S}_{n,k} := \{(\boldsymbol{\tau}, \mathbf{d}) : \boldsymbol{\tau} \text{ recognizes } k \text{ words and binds to } n \text{ sites on } \mathbf{d}\}$$

and require a method for calculating the cardinality of these subsets, denoted by

$$\Omega(n, k) := |\mathcal{S}_{n,k}|.$$

This quantity can be calculated by further dividing $\mathcal{S}_{n,k}$ into subsets of individual factors. Formally, let $\mathcal{S}_{n,\boldsymbol{\tau}}$ denote the subset of $\mathcal{S}$ in which $\boldsymbol{\tau}$ recognizes n sites. Since all $\mathcal{S}_{n,\boldsymbol{\tau}}$ are mutually exclusive,

$$\Omega(n,k) = \sum_{\boldsymbol{\tau} \in \mathcal{T}_k} |\mathcal{S}_{n,\boldsymbol{\tau}}|. \tag{8}$$

We have written a computer program for calculating $\Omega(n,k)$ using this method. Unfortunately, use of this program is limited to rather small systems ($a = 2$, $L \leq 4$, $N \leq 22$), as the number of states which need to be enumerated amounts to $a^N \cdot 2^K$ (where the latter factor is most severe as $K = a^L$). The quantity $|\mathcal{S}_{n,\boldsymbol{\tau}}|$ depends on the particular structure of $\boldsymbol{\tau}$ in a complex manner. Due to this complexity, the program had to operate by exhaustive enumeration of $\mathcal{S}$.

5.2.1 Independent Word Approximation for Large Systems

For deriving a method that enables investigation of larger systems, we approximate genomes by sequences of independent words (see Section 2.4.2). Thus, we substitute $\mathcal{S}$ by $\mathcal{S}' := \mathcal{W}^N \times \mathcal{T}$. The sets $\mathcal{S}'_{n,\boldsymbol{\tau}}$ as well as $\mathcal{S}'_{n,k}$ and their cardinalities $\Omega'(n,k)$ are defined analogously to the definitions above. But differently from $\mathcal{S}_{n,\boldsymbol{\tau}}$, the independence of words allows a combinatorial construction of $\mathcal{S}'_{n,\boldsymbol{\tau}}$ which depends on k only while the particular structure of $\boldsymbol{\tau}$ does not matter. All genomes comprised within $\mathcal{S}'_{n,\boldsymbol{\tau}}$ are composed of n words drawn from the k binding words accepted by $\boldsymbol{\tau}$ and $N - n$ words out of the $K - k$ non-binding words. There are

$$B_{n,k} := k^n (K-k)^{N-n}$$

different choices of words to construct such a genome, and as each of these choices can be assembled according to $\binom{N}{n}$ different binding patterns (see Section 2.2), one obtains

$$|\mathcal{S}'_{n,\boldsymbol{\tau}}| = \binom{N}{n} B_{n,k} = \binom{N}{n} k^n (K-k)^{N-n}.$$

Now, $\Omega'(n,k)$ can be calculated analogously to Equation (8). As all values in the sum are identical we obtain

$$\Omega'(n,k) = |\mathcal{T}_k| \binom{N}{n} k^n (K-k)^{N-n}, \tag{9}$$

and by plugging in Equation (7) we arrive at the closed form equation

$$\Omega'(n,k) = \binom{K}{k} \binom{N}{n} k^n (K-k)^{N-n}. \tag{10}$$

Values of $\log \Omega'(n,k)$ can be computed for systems of realistic sizes. $\Omega'(n,k)$ is not intended to approximate $\Omega(n,k)$ in terms of absolute values. Evidently, $|\mathcal{S}'| = |\mathcal{W}|^N \cdot |\mathcal{T}| = a^{LN} \cdot |\mathcal{T}| \gg |\mathcal{S}| = a^N \cdot |\mathcal{T}|$. Considering that $\sum_{n,k} \Omega(n,k) = |\mathcal{S}|$, and analogously, $\sum_{n,k} \Omega'(n,k) = |\mathcal{S}'|$, it follows that $\Omega'(n,k) \gg \Omega(n,k)$ for biological plausible values of L. Nonetheless, we will show that the positions of the maxima in $\Omega'(n,k)$ can be used to estimate the positions of the maxima in $\Omega(n,k)$.

6 Results and Discussion

The model defined in Section 5 enables us to assess the relation of R_{sequence} and $R_{\mathrm{frequency}}$ in a framework that is more general than those underlying the treatments reviewed in Section 4. We address this issue by minimizing any prior assumptions and assuming, according to the maximum entropy principle, that all individual $(\mathbf{d}, \boldsymbol{\tau})$ tuples occur equiprobably. We will show that the maxima in Ω, or the maxima in Ω' as an approximation, are extremely peaked and that they indicate which value of k can be expected for a given value of n. This value is denoted by $k_{\max}$.

The following treatment is similar to maximum entropy analysis of physical many particle systems where so-called macroscopic states are associated to subsets of microscopic states. The microscopic states are assumed to be equally probable according to maximum entropy. The likelihood to encounter the system in a certain macroscopic state is then given by the relative number of microscopic states which realize this particular macroscopic state. It turns out that for combinatorial reasons there is always a macroscopic state which is realized by an overwhelmingly large number of microscopic states. Thus, the probability distribution in the space of macroscopic system states is extremely peaked. The probability to find the system outside of the maximum is practically zero. In the following treatment, we will subject the state space defined in Section 5.1.2 to such a maximum entropy analysis with $(\boldsymbol{\tau}, \mathbf{d})$ as the microscopic and (n, k) as the macroscopic states. With our approximation of independent words the genome can be regarded as a so-called Potts spin system which is well-known from statistical physics (Wu, 1982). Each word corresponds to a Potts spin with a^L states, and the system consists of N such spins. The number of "particles" N is not as large as in typical many particle systems in physics, but large enough to provide extremely peaked probability distributions in the macroscopic state space.

Strictly, maximum entropy analysis applies to systems in equilibrium when a uniform probability distribution in the microscopic state space has evolved. In evolution, equilibrium is attained only in the case of neutral selection, infinite population size and infinite time. Clearly, biological evolution deviates from these conditions. Nonetheless, neutral evolutionary models provide the basis for important analysis methods. For example, most models underlying phylogeny reconstruction algorithms assume neutral evolution. In many respects, biological evolution is known to exhibit relaxation towards maximum entropy when constraints are removed, e.g., third codon positions and pseudogenes tend to become saturated with mutations. Based on these considerations, maximum entropy analysis can only provide an approximate description of the statistical properties of binding site sequences. Nonetheless, maximum entropy and maximum likelihood analysis can be a valuable tool for gaining insight into biological processes, as the results we present in the following exemplify.

Our point of departure for our likelihood calculation are unbiased distributions, i.e., all genome sequences (modelled by symbol vectors $\mathbf{d}$) and all transcription factors (modelled by bit vectors $\boldsymbol{\tau}$) occur equiprobably. Formally, $P_{\mathbf{d}} = 1/|\mathcal{G}|$ and $P_{\boldsymbol{\tau}} = 1/|\mathcal{T}|$ is valid. As a consequence all $(\mathbf{d}, \boldsymbol{\tau}) \in \mathcal{S}$ also have equal probabilities. From a biological perspective, the assumption of *a priori* equiprobability of all $\mathbf{d} \in \mathcal{G}$ is well justified by the structural correspondence between symbol vectors and nucleotide sequences on DNA strands. The transcription factor space, on the other hand, does not reflect biological or biochemical structures of transcription factors. Therefore, there may be states in $\mathcal{T}$ which are impossible to be implemented by actual transcription factor molecules. Such states should have probability zero, and, thus, with todays knowledge (prior knowledge) one may rightfully demand that it should be possible to consider the case of deviations from equiprobability in $\mathcal{T}$.

With $\mathcal{S}$ being the Cartesian product $\mathcal{G} \times \mathcal{T}$, the probability of a $(\mathbf{d}, \boldsymbol{\tau})$ tuple is generally given by $P_{\mathbf{d}} \cdot P_{\boldsymbol{\tau}}$. Thus, our likelihood considerations in the (n, k)-space based on the probability distribution in the $(\mathbf{d}, \boldsymbol{\tau})$-space allows us to treat such deviations in both the genome and the transcription factor space. For example, our approach could straightforwardly be adapted to study effects of variations in GC content or specific CpG statistics. We will see in Section 6.4 that indeed strong deviations from equiprobability in $\mathcal{T}$ can be expected.

6.1 Comparison of Ω and Ω'

The computer program described in Section 5.2 was applied to compare Ω and Ω'. Fig. 1 shows plots of Ω and Ω' (see Equations (8) and (10)) for $a = 2$, $N = 22$ and $L = 4$. The order of magnitude of $\log \Omega'$ is approximately three times the order of magnitude of $\log \Omega$, as expected (see Section 5.2.1). Other than that, the surfaces have similar characteristics. In particular, the maxima of Ω' and Ω for given values of n are found at identical positions. Differences are only observed at $n = 0$ (no binding site on genome) and $n = N$ (all sites on genome are binding sites). Obviously, these extreme cases are biologically irrelevant.

As explained above, these maxima indicate the k values expected for a given n. It should be noted that the distributions are shown in logarithmic scale. The maxima in Ω and Ω' are extremely peaked. These maxima become even more extreme for increasing system size.

The bottom part of Fig. 1 shows the coordinates of these maxima on the (n, k) plane, along with a graph of $k = nK/N$. This equation follows from Equation (5), and thus the graph indicates where the maxima would be expected if R_{sequence} was equal to $R_{\text{frequency}}$. The actual locations of the maxima clearly deviate from this expectation.

Two conclusions are drawn from these results: Firstly, we have confirmed that our approach of using the positions of maxima in Ω' to estimate the locations of maxima in Ω works well and can be employed for further analysis. Secondly, the k values which are most probable to be observed for given n values are not those which would be expected if R_{sequence} was equal to $R_{\text{frequency}}$. There are significant deviations. This issue is the subject of the following analytic approach.

6.2 An Analytic Approach

By treating k in Equation (9) as a continuous variable, we can approach the determination of the maximum in Ω' for a fixed value of n analytically by virtue of the fact that $\partial\Omega'(n, k)/\partial k = 0$ at $k = k_{\text{max}}$. It is equivalent to analyze $\log \Omega'$ because this function increases strictly monotonically with Ω' and thus $\log \Omega'$ and Ω' have identical maxima. From Equation (9) we obtain

$$\log \Omega'(n, k) = \log |\mathcal{T}_k| + \log \binom{N}{n} + n \log k + (N - n) \log(K - k)$$

and the partial derivative is

$$\frac{\partial \log \Omega'(n, k)}{\partial k} = \frac{\partial \log |\mathcal{T}_k|}{\partial k} + \frac{n}{k} - \frac{N - n}{K - k}. \tag{11}$$

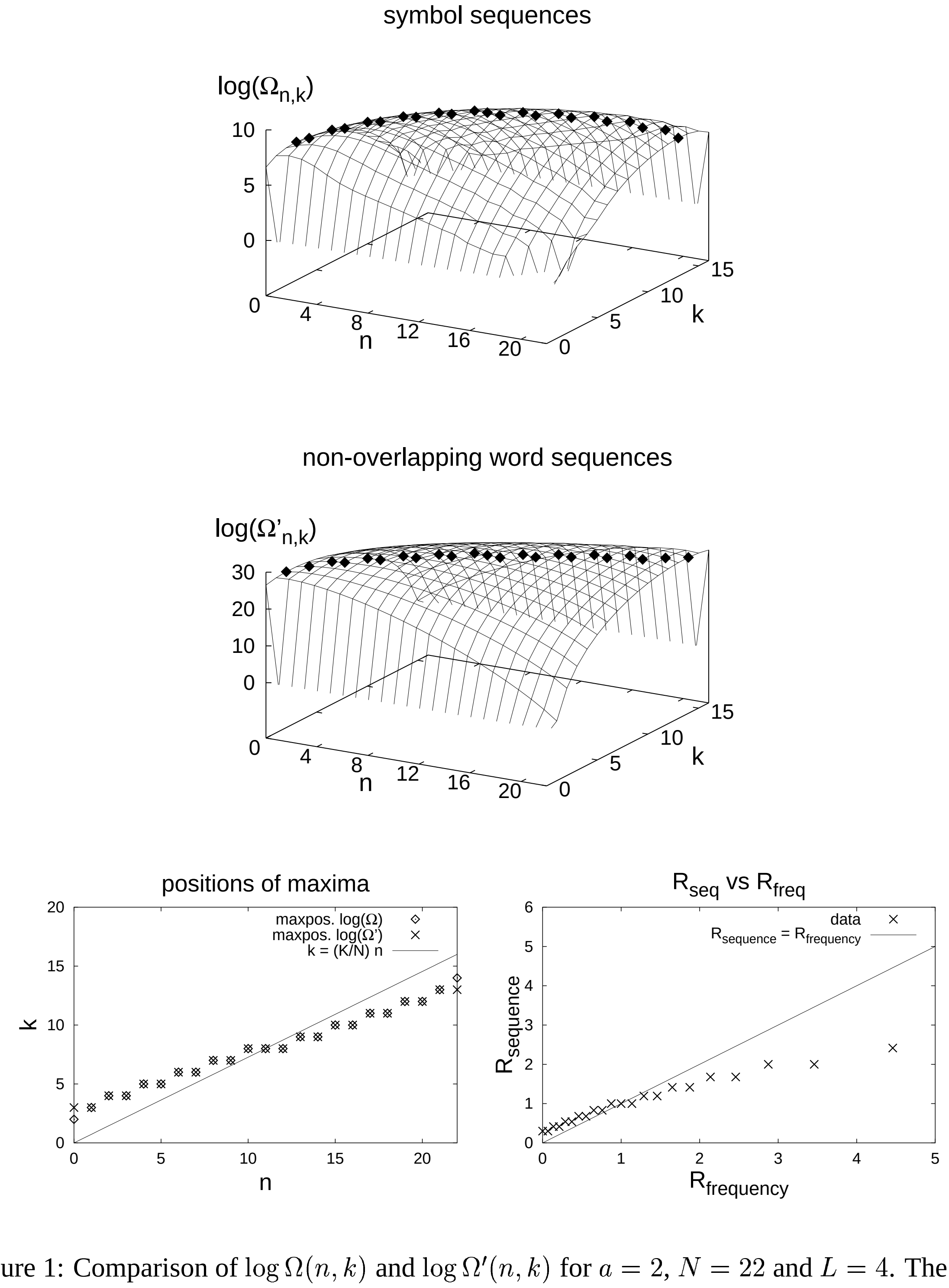

Figure 1: Comparison of $\log \Omega(n,k)$ and $\log \Omega'(n,k)$ for $a = 2$, $N = 22$ and $L = 4$. The word set size amounts to $K = 16$. Top: plot of $\log_{10} \Omega(n,k)$, middle: plot of $\log_{10} \Omega'(n,k)$. Maxima in k for fixed values of n are highlighted by filled diamonds. Except for scale, the surfaces are very similar. Bottom left: positions of the maxima of Ω and Ω' on the n, k plane, right: R_{sequence} and $R_{\text{frequency}}$ values computed from the k and n values of the data points according to Equations (4) and (1), respectively. Lines show where the data points would be expected in the case of $R_{\text{sequence}} = R_{\text{frequency}}$ (see also Equation 12).

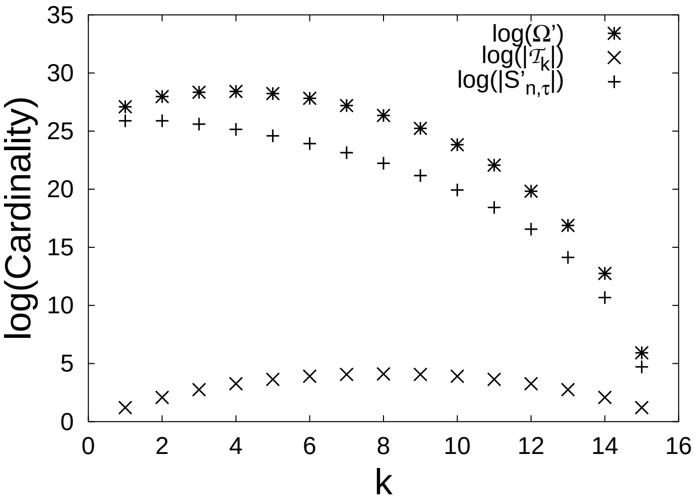

Figure 2: Plot of $\log \Omega'(n,k) = \log|\mathcal{T}_k| + \log|\mathcal{S}'_{n,\boldsymbol{\tau}}|$ (asterisks), along with the individual addends $\log|\mathcal{T}_k| = \log\binom{K}{k}$ (crosses) and $\log|\mathcal{S}'_{n,\boldsymbol{\tau}}| = \log(\binom{N}{n}k^n(K-k)^{N-n})$ (pluses), for $n = 2$. as in Fig. 1, $N = 22$ and $K = 16$. The maximum of $\log \Omega'(2,k)$ is located between the maxima of the individual addend functions.

Generally, this partial derivative is 0 at the maximum of Ω', i.e. at $k = k_{\text{max}}$. However, the term $\partial \log|\mathcal{T}_k|/\partial k$ is too generic to permit deriving an explicit expression for k_{max}. However, further analysis is possible by asking what $\partial \log|\mathcal{T}_k|/\partial k$ would have to be if $R_{\text{sequence}} = R_{\text{frequency}}$ was true. From Equation (5), we know that in this case k can be calculated as a function of n by

$$k_n := n \cdot \frac{K}{N}. \tag{12}$$

Consequently, $\partial \log \Omega'(n,k)/\partial k = 0$ has to be valid at $k = k_n$. By substituting this into Equation (11), we obtain

$$\left.\frac{\partial \log|\mathcal{T}_k|}{\partial k}\right|_{k=k_n} = \frac{N-n}{K - n\cdot K/N} - \frac{n}{n \cdot K/N} = 0.$$

This analytic result shows that R_{sequence} and $R_{\text{frequency}}$ are equal if $|\mathcal{T}_k|$ is a constant function in k or exhibits a maximum or a rather peculiar saddle point at k_n. Equation (11) also reveals that $B_{n,k}$ has a maximum at k_n. Thus, a tendendcy towards equality of R_{sequence} and $R_{\text{frequency}}$ is induced by $B_{n,k}$ while non-uniform distributions of $|\mathcal{T}_k|$ result, with exceptions of special cases, in displacements of R_{sequence} from $R_{\text{frequency}}$.

In the light of this finding, the deviations from $R_{\text{sequence}} = R_{\text{frequency}}$ seen in Section 6.1 can be explained by the fact that $|\mathcal{T}_k| = \binom{K}{k}$ (Equation (7)) has its maximum at $k = K/2$. The actual maxima seen in Fig. 1 are therefore displaced towards $K/2$. Figure 2 shows this effect for the model genome analyzed in Fig. 1 in case $n = 2$. The peak in $\log\binom{N}{n} + \log B_{n,k}$ appears around $k_n = 2 \cdot 16/22 \approx 1.45$, but as $|\mathcal{T}_k|$ becomes maximal at $k = K/2 = 8$, k_{max} is displaced from k_n towards 8.

As a summary of this formal analysis, we have seen that the combinatorial structure of the genome space, reflected by $B_{n,k}$, induces a propensity towards equality of R_{sequence} and

$R_{\text{frequency}}$, however, this equality also depends on the structure of the transcription factor space. Our approach to model the transcription factor space with $\mathcal{T}$ is too abstract in its generality to reflect any details of the biological transcription factor space. Nonetheless, a uniform distribution of $|\mathcal{T}_k|$ or a maximum at k_n would be a remarkable property and ask for explanation. For gaining further insight, we will now quantitatively analyze the relation of R_{sequence} and $R_{\text{frequency}}$ for systems with sizes in the biologically relevant range.

6.3 Results for systems of realistic size

Figure 3 shows results of analyses based on Ω' for alphabet size $a = 4$, i.e. for the biological nucleotide alphabet, and binding word length $L = 10$, which is a reasonable biological ballpark figure. Genome lengths range from $N = 10^6$ to 10^9, i.e. from the size of prokaryotic to smaller eukaryotic genomes.

The top plots, in which $N = 10^6$, are qualitatively similar to the results obtained by enumeration of $\mathcal{S}$ for small systems (Fig. 1). The $k_{\max}$ values are closer to $K/2$ than expected for $R_{\text{sequence}} = R_{\text{frequency}}$. However, as genome length increases, this displacement effect becomes less and less pronounced. Based on the observations described in Section 6.2, this can be explained by the relation of the terms in Equation (11): While $\partial \log |\mathcal{T}_k| / \partial k$ does not depend on n and N, the term $n/k - (N - n)/(K - k)$ grows linearly with N if $R_{\text{frequency}}$ is kept fixed (i.e. if the ratio of n/N is held constant). Hence, relative to $n/k - (N - n)/(K - k)$ the term $\partial \log |\mathcal{T}_k| / \partial k$ shrinks to zero with increasing N. As a result, the displacement of the maximum in $\log \Omega'$ induced by adding $\log |\mathcal{T}_k|$ decreases with N.

Interestingly, while the plots of $k_{\max}$ show almost no deviation from k_n for $N \geq 10^8$, the plots of R_{sequence} as a function of $R_{\text{frequency}}$ reveal a ceiling at $R_{\text{sequence}} = 20$. This effect is not explained by an influence of $|\mathcal{T}_k|$ on the location of the maximum of Ω'. Rather, this phenomenon is solely due to the fact that k is integer valued and cannot fall below 1 for $n > 0$ because the transcription factor has to recognize at least one binding word in order be capable to induce any binding pattern on the genome other than the empty one. Thus, R_{sequence} is strictly limited by $-\log_2(1/K) = 20$. This might become a problem for complex living systems with large genomes. It seems plausible to think that multiprotein systems appeared in evolution because they allow an effective increase in L, and hence in K. The well known fact that these systems are particularly complex in eukaryotes is well in line with this assumption.

The main finding from this analysis is that the deviation between R_{sequence} and $R_{\text{frequency}}$ is determined by the relation between $\partial \log B_{n,k} / \partial k$ and $\partial \log |\mathcal{T}_k| / \partial k$. Therefore, we perform a quantitative estimation of this relation as the final part of this investigation.

6.4 The Cardinality of Transcription Factor Space in Biological Systems

The modelling framework introduced in this article employs $\mathcal{G}$ as a model of genome space and $\mathcal{T}$ as a model of transcription factor space. Symbol strings $\mathbf{d} \in \mathcal{G}$ structurally correspond to DNA strands. Thus, $\mathcal{G}$ is a biologically adequate model, in both the qualitative and the quantitative respect. However, $\mathcal{T}$ is too generic to reflect specific quantitative properties of transcription factor space. Therefore, additional biological principles have to be considered in order to derive further quantitative estimations.

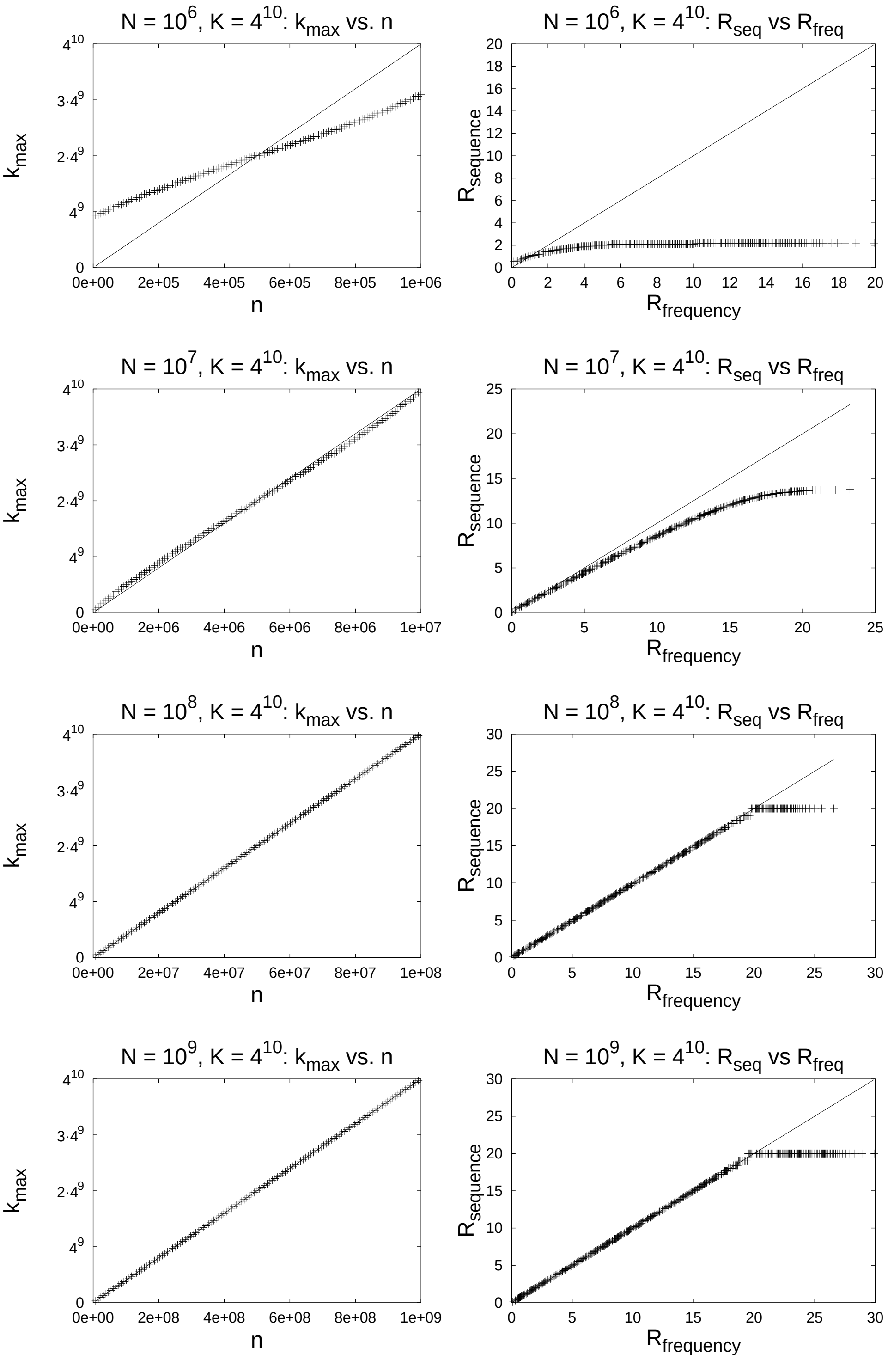

Figure 3: Plots of $k_{\max}$ vs. n (left column) and $R_{\text{frequency}}$ vs. R_{sequence} (right column). Parameter settings are $a = 4$ and $L = 10$, the value of N is 10^6 (top row), 10^7 (second row), 10^8 (third row) and 10^9 (bottom row). The straight line in each plot indicates values for $R_{\text{sequence}} = R_{\text{frequency}}$.

Following the approach used for the genome model, it is an obvious idea to represent transcription factors by their amino acid sequences. However, this would require a method for deducing the binding behaviour of a transcription factor from its amino acid sequence. Modelling this in a biochemically adequate way would demand computations of three-dimensional protein structures from amino acid sequences as well as computational predictions of protein-DNA interactions. Evidently, this demand cannot be fulfilled and, consequently, structurally adequate models of transcription factor space are not available.

However, while the complexities of structural biology preclude a quantitatively adequate model, the biological fact that transcription factors reside within the space of amino acid sequences is useful to derive a quantitative estimation for the upper bound of this displacement. Let Θ denote the set of amino acid sequences, and Θ_k denote the set of transcription factor proteins accepting k words as binding words. Transcription factors are globular proteins that consist of a few hundred amino acids. Typically, a domain of less than 100 amino acid residues is sufficient for binding site recognition. Therefore, it is reasonable to estimate $|\Theta| \approx 20^{100} \approx 10^{130}$.

In Section 6.3 (Fig. 3), however, we analyzed systems with $K = 4^{10} = 1048576$ and $|\mathcal{T}| = 2^{1048576} \approx 10^{315653}$. Thus, $|\mathcal{T}|$ overestimates the biological transcription factor space cardinality by many orders of magnitude. From this perspective it becomes clear that the set of binding behaviours that can actually be implemented by sequence specific DNA binding proteins is bound to be much smaller than $\mathcal{T}$, the set of all distinguishable binding behaviours.

In Section 6.2 we have seen that displacements of R_{sequence} from $R_{\text{frequency}}$ can be attributed to the distribution $|\mathcal{T}_k|$. Despite the impossibility to quantitatively specify $|\mathcal{T}_k|$, it can be stated that the amount by which R_{sequence} is displaced is determined by the variance of $|\mathcal{T}_k|$ in relation to the variance to $B_{n,k}$ at k_n. Considering that the cardinalities $|\mathcal{T}|$ and $|\Theta|$ also provide upper limits for the variances in $|\mathcal{T}_k|$ and $|\Theta_k|$, respectively, it is reasonable to expect the true displacement to be smaller than it is depicted in Fig. 3. From the view developed in Section 6.2, it may even be that the variance in $|\Theta_k|$ is so small in relation to the variance in $B_{n,k}$ that assuming $|\Theta_k|$ to be constant becomes a reasonable approximation. Thus, the approximate equality of R_{sequence} and $R_{\text{frequency}}$ in biological systems can be explained to result from the limited size of the transcription factor state space.

Having linked this equality to the quantity $|\Theta|$ invites the question whether limitations of the size of transcription factor space can be derived from biological principles, or whether this size is due to some evolutionary contingency. From structural biology, it may be that the size of globular proteins is limited to something less than 1000 amino acid residues. This observation implies that $|\Theta|$ is limited by constraints deriving from protein biochemistry. Furthermore, even if much larger globular proteins were possible, it would be very hard to conceive of ways how domains which are remote from the binding center should effect the factor's recognition capabilities. From this perspective, it appears that the physical size of the device for storing genetic information (i.e. DNA), relative to the size of the components of the machinery interpreting genetic information (i.e. transcription factors etc.), imposes a limit on $|\Theta|$.

The strongest limitation of $|\Theta|$ in relation to $|\mathcal{G}|$ comes from the principle of genetic autonomy. Genetically autonomous systems are viable without any external genetic information. Thus, genetically autonomous systems must encode their gene expression machinery, including all transcription factors, within their genome. The average cardinality of the coding space of a single gene is limited by $a^{N/G}$, where G denotes the number of genes. Since transcription factors are encoded by single genes, $|\Theta| \leq a^{N/G}$ is valid and consequently, $\log|\Theta|/\log|\mathcal{G}| < 1/G$ holds. As a minimum of around 1000 genes is generally considered necessary for an autonomous living system, $\log|\Theta|/\log|\mathcal{G}| \leq 10^{-3}$ is a biologically plausible relation. The approximate equality of R_{sequence} and $R_{\text{frequency}}$ can therefore be characterized as a bioinformatic property which invariantly applies to any genetically autonomous system even if the chemistry or the information storage and interpretation devices underlying biological systems were fundamentally different from those known today.

7 Conclusion

In this contribution, we have proposed a model for genetic systems that consist of a genome and a transcription factor (which, in fact, can more broadly be construed as any factor capable of sequence specific DNA binding). Within the framework of this model, we determined which R_{sequence} value is most abundantly represented within a set of states with a fixed value of $R_{\text{frequency}}$. According to the maximum entropy principle, this defines the relation of R_{sequence} and $R_{\text{frequency}}$.

Even initial studies showed that scenarios in which R_{sequence} is considerably different from $R_{\text{frequency}}$ do exist. If existing, this deviation is particularly prominent for small abundances of binding sites on the genome, i.e. for the biologically relevant range of larger $R_{\text{frequency}}$ values. Thus, an equality of $R_{\text{sequence}} = R_{\text{frequency}}$ cannot be deduced from information theoretic principles alone.

Further analysis revealed a link between deviations from this equality and the distribution of k, the numbers of words recognized as binding words, within the space of transcription factors. A strong variance among the $|\mathcal{T}_k|$ values may lead to a correspondingly strong deviation of R_{sequence} from $R_{\text{frequency}}$.

While the distribution of $|\mathcal{T}_k|$ cannot be calculated in detail, an upper limit of its magnitude derives from the biological principle that the genome has to encode all components of a living system, including all transcription factors, in a genetically autonomous system. Therefore, the size of the transcription factor space, and also the variance of the distribution of k values within this space, are limited to be small in relation to the size of the genome space. This constraint, which applies independently of the physical and chemical constitution of a living sytem, is the key principle which confines deviations from $R_{\text{sequence}} = R_{\text{frequency}}$ to a very small range.

From this perspective, $R_{\text{sequence}} = R_{\text{frequency}}$ is not an information theoretic, but a bioinformatic principle which is valid for all biological systems that are autonomous in the sense that they encode their gene regulatory logic within their own genomes. As a test of this hypothesis, it would be interesting to determine R_{sequence} and $R_{\text{frequency}}$ values for systems which are not genetically autonomous, such as viruses (which depend on the transcription machinery of the host) or plastids and mitochondria (which import a substantial part of their transcription machinery components from the nucleus).

References

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, New York, NY, USA.

Frech, K., Quandt, K. & Werner, T. (1997). Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.* **13**, 89–97.

Jaynes, E. T. (1952). Information theory and statistical mechanics. *Physical Review,* **106**, 620–630.

Kappen, C. & Ruddle, F. H. (1993). Evolution of a regulatory gene family: HOM/HOX genes. *Current Opinion in Genetics and Development,* **3**, 931–938.

Kolchanov, N., Podkolodnaya, O., Ananko, E., Ignatieva, E., Stepanenko, I., Kel-Margoulis, O., Kel, A., Merkulova, T., Goryachkovskaya, T., Busygina, T., Kolpakov, F., Podkolodny, N., Naumochkin, A., Korostishevskaya, I., Romashchenko, A. & Overton, G. (2000). Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Research,* **28**, 298–301.

Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053–1095.

Schmitt, A. O. & Herzel, H. (1997). Estimating the entropy of DNA sequences. *J. theor. Biol,* **1888**, 369–377.

Schneider, T. D. (2000). Evolution of biological information. *Nucleic Acids Research,* **28**, 2794–2799.

Schneider, T. D., Stormo, G. D. & Gold, L. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.

Shore, P. & Sharrocks, A. D. (1995). The MADS-box family of transcription factors. *Eur. J. Biochem.* **229**, 1–13.

Stormo, G. D. (1998). Information content and free energy in DNA-protein interactions. *J. theor. Biol.* **195**, 135–137.

Stormo, G. D. & Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA-interactions. *Trends in Biochemical Sciences,* **23**, 109–113.

Theißen, G. & Saedler, H. (1995). MADS-box genes in plant ontogeny and phylogeny: haeckel's 'biogenetic law' revisited. *Current Opinion in Genetics and Development,* **5**, 628–639.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S. & Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Research,* **29**, 281–283.

Wu, F. Y. (1982). The Potts model. *Rev. Mod. Phys.* **54**, 235–268.