In: WINGENDER, EDGAR AND HOFESTÄDT, RALF AND LIEBICH, INES: *Proceedings of the German Conference on Bioinformatics 2001*, 192-194. German Research Center for Biotechnology (GBF), Mascheroder Weg 1, 38124 Braunschweig, Germany, 2001.

On the Effects of Transcription Factor Properties on the Information Content of Binding Sites

Jan T. Kim, Thomas Martinetz, Daniel Polani

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

phone: +49 451 3909-585, fax: +49 451 3909-545
email: {kim,martinetz,polani}@informatik.mu-luebeck.de

1 Introduction

Networks of genes which encode transcription factors (regulatory networks) play a central role in the realization of phenotypic traits based on genetic information. Sequence-specific recognition of DNA subsequences by proteins is a key mechanism in constituting regulatory networks. Understanding the information theoretic principles underlying the evolution of transcription factors and their binding sites is therefore a major challenge in bioinformatics [1]. Advances in this field are expected to provide a basis for improving algorithmic binding site identification and promoter analysis [2], and for deciphering regulatory codes.

Previous studies [3] have suggested that the information content deduced from binding site sequence sets (R_{sequence}) approximately equals the information content deduced from relative binding site abundance ($R_{\text{frequency}}$). Here, we investigate the relation between these two information quantities using a maximum entropy approach.

2 Outline of the Model

We formally model genomes of length N by vectors of words $\vec{w} = (w_1, w_2, \ldots, w_N)$ where $w_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^l$. Transcription factors are represented by binary vectors $\vec{\tau} = (\tau_{w_1}, \tau_{w_2}, \ldots, \tau_{w_K}), \tau_{w_i} \in \{0, 1\}$, where $K = 4^l$ is the number of possible words and $\tau_{w_j} = 1$ if the factor binds to w_j and $\tau_{w_j} = 0$ otherwise. The number of binding sites in the genome is denoted by n and the number of words recognized by the transcription factor is denoted by k.

Within this modelling framework, R_{sequence} and $R_{\text{frequency}}$ are given by the equations

$$R_{\text{frequency}} = -\log \frac{n}{N}, \qquad R_{\text{sequence}} = -\log \frac{k}{K}. \tag{1}$$

Thus, if $R_{\text{sequence}} = R_{\text{frequency}}$ we expect

$$k = \frac{Kn}{N}. \tag{2}$$

From a probabilistic point of view, the expectation to find $R_{\text{sequence}} \approx R_{\text{frequency}}$ is to be understood to mean that $(\vec{\tau}, \vec{w})$ tuples in which $k \approx Kn/N$ are the most common type for a given value of n. We therefore derived a formula for calculating $\Omega'(n,k)$, the number of $(\vec{\tau}, \vec{w})$ tuples composed of a transcription factor $\vec{\tau}$ binding to k different words and recognizing n sites on $\vec{w}$:

$$\Omega'(n,k) = \underbrace{\binom{K}{k}}_{\text{(a)}} \cdot \underbrace{\binom{N}{n}}_{\text{(b)}} \cdot \underbrace{k^n (K-k)^{N-n}}_{\text{(c)}}$$

As a motivation of this formula, note that

- term (a) calculates the number of factors binding to k words
- term (b) calculates the number of binding site / non-site patterns with n binding sites
- term (c) calculates the number of word sequences realizing a particular binding pattern with n binding sites given particular factor binding to k words.

A more detailed derivation and discussion of this equation will appear in a forthcoming paper.

3 Results

Fig. 1 shows results of an analysis based on Ω' for genome length $N = 10^6$ and binding site word length $l = 10$ (hence, $K = 65536$). The surface plot in Fig. 1 may appear quite even-levelled, but one should notice the logarithmic scale: The Ω' values span four million decimal orders of magnitude. Thus, the probability for observing k values other than the one maximizing Ω' for a given n practically vanishes. For each value of n, the maximal Ω' value is highlighted by a diamond.

The bottom left plot displays the coordinates of these maxima on the n, k plane, showing a clear and significant deviation from the line expected if $R_{\text{sequence}} = R_{\text{frequency}}$ (equation 2).

The bottom right plot in Fig. 1 reveals the discrepancies between R_{sequence} and $R_{\text{frequency}}$ directly. Here, the n values shown in the middle plot were translated into $R_{\text{frequency}}$ values and the corresponding k values that maximize Ω' were translated into R_{sequence} values according to eq. 1. The deviation from $R_{\text{sequence}} = R_{\text{frequency}}$ is particularly prominent in the range of larger $R_{\text{frequency}}$ values. This finding is especially interesting, as binding site frequencies are usually in the order of magnitude of 10^{-3} or below, so cases of $R_{\text{frequency}} \geq 8$ are biologically most relevant.

In summary, for genome and binding site sizes in the order of magnitude encountered in prokaryotic systems, our model predicts substantial deviations from $R_{\text{sequence}} = R_{\text{frequency}}$.

4 Discussion

Our results calls for explanations in two respects. In a theoretical respect, the question arises why previous analyses implied that $R_{\text{sequence}} \approx R_{\text{frequency}}$ was to be expected. Differently from previous models, our model explicitly comprises the space of protein binding behaviours within the state space. The deviations from $R_{\text{frequency}} = R_{\text{sequence}}$ which we have observed with our model are to be ascribed to evolutionary effects originating from the protein side. More detailed analyses of these effects are currently underway.

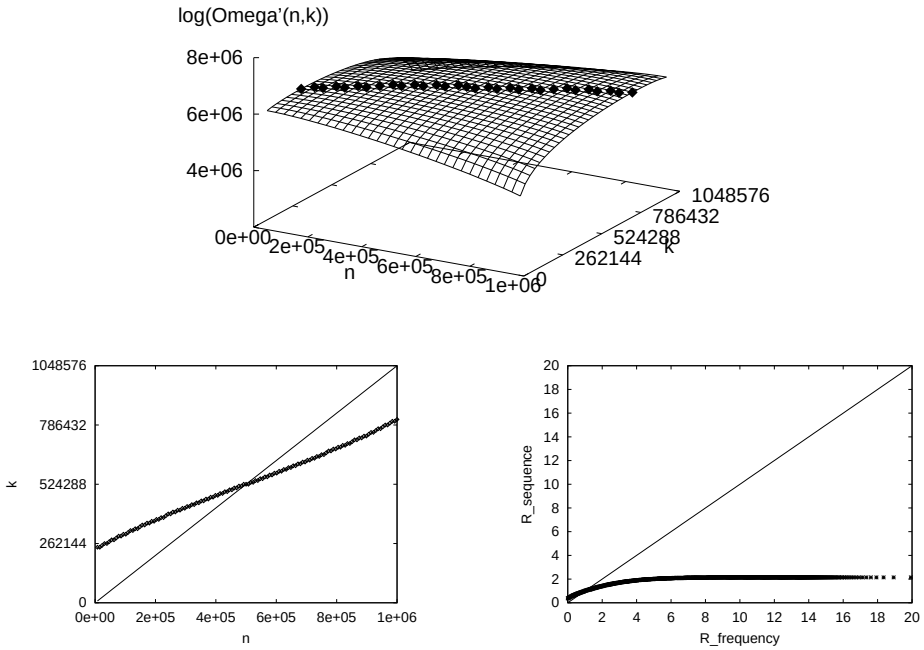

Figure 1: Top: plot of $\log_{10}(\Omega')$ for $a = 4$, $N = 10^6$ and $l = 10$, bottom left: coordinates of maximal Ω' values for each n value (diamonds) and graph of $k = \frac{Kn}{N}$ (line), bottom right: plot of the $(R_{\mathrm{frequency}}, R_{\mathrm{sequence}})$ values calculated from the coordinates plotted above according to equation 1 (asterisks) and graph of $R_{\mathrm{sequence}} = R_{\mathrm{frequency}}$ (line).

In an empirical respect, our findings call for revisiting the cases in which $R_{\mathrm{sequence}} \approx R_{\mathrm{frequency}}$ was observed, paying particular attention to deviations from equality and possible regularities detectable therein. Such analyses may provide information about the biological structure of the influence which DNA binding proteins have on the information content of their binding sites.

In a longer perspective, we expect this direction of research to lead to a deepened understanding of the evolutionary biological forces shaping protein-DNA interactions, which in turn may serve as a basis for developing tools with improved performance for the detection of biologically significant binding sites and for the analysis and characterization of regulatory mechanisms and networks.

References

[1] Gary D. Stormo and Dana S. Fields. Specificity, free energy and information content in protein-DNA-interactions. *TIBS*, 23:109–113, 1998.

[2] Kornelie Frech, Kerstin Quandt, and Thomas Werner. Software for the analysis of DNA sequence elements of transcription. *CABIOS*, 13:89–97, 1997.

[3] Thomas D. Schneider. Evolution of biological information. *Nucleic Acids Research*, 28:2794–2799, 2000.