

Is the information content of transcription factors sufficient to regulate gene expression in various organisms?

M. Sheinman

Department of Physics and Astronomy, Vrije Universiteit, Amsterdam, The Netherlands

PACS numbers:

To find something one needs an information what to look for— IC —so-called *information content*. For instance to find a particular place on the DNA of length N one needs

$$IC = \log_2 N \quad (1)$$

bits of information (yes-or-no questions). The proteins—transcription factors—whose job is to locate and bind to specific sequences on the DNA do contain some information. This information is used to regulate almost all genes activity and, therefore, is extremely crucial to functioning of all organisms. It is coded in binding energy discrimination between different sequences on the DNA. For many transcription factors was measured and the data is accessible in databases.

Recently [1] it was shown that the information content in prokariotic transcription factors, indeed, is roughly equal to $\log_2 N$ prediction. However, in eukariotic cells the information content was shown to be much smaller than $\log_2 N$. Based on this findings it was suggested that the regulation strategies in prokariots and eukariots may be totally different. However, in this study the number of the transcription factor's copy number in the cell, n , and the number of its targets, m , was ignored. In fact, when n transcription factors regulate m different targets the desired amount of information is $\log_2 \frac{N}{nm}$. This easily may bring it (and brings for several transcription factors, as I found) to the measured information content of eukariotic transcription factors. This considerations also predict a correlation between the transcription factor abundance in the cell or its number of targets and its information content. This correlation was also observed by me for sevelar prokariotic transcription factors.

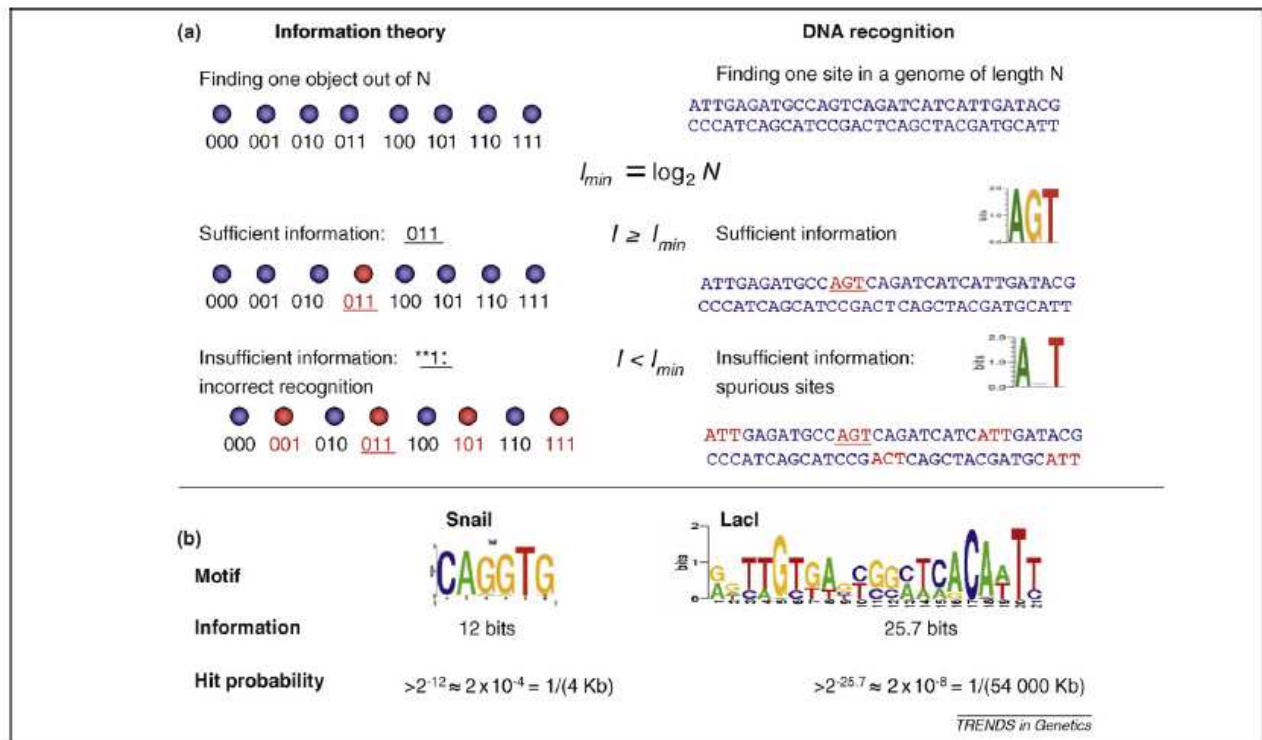


Figure 1. Information theory as applied to DNA-binding motifs. **(a)** The concepts of minimal information required in theory and in DNA recognition and the consequences of information deficiency, which results in spurious hits. **(b)** The sequence logos for low- and high-information motifs and the likelihood of a spurious hit to the motif in a 'random' genomic background.

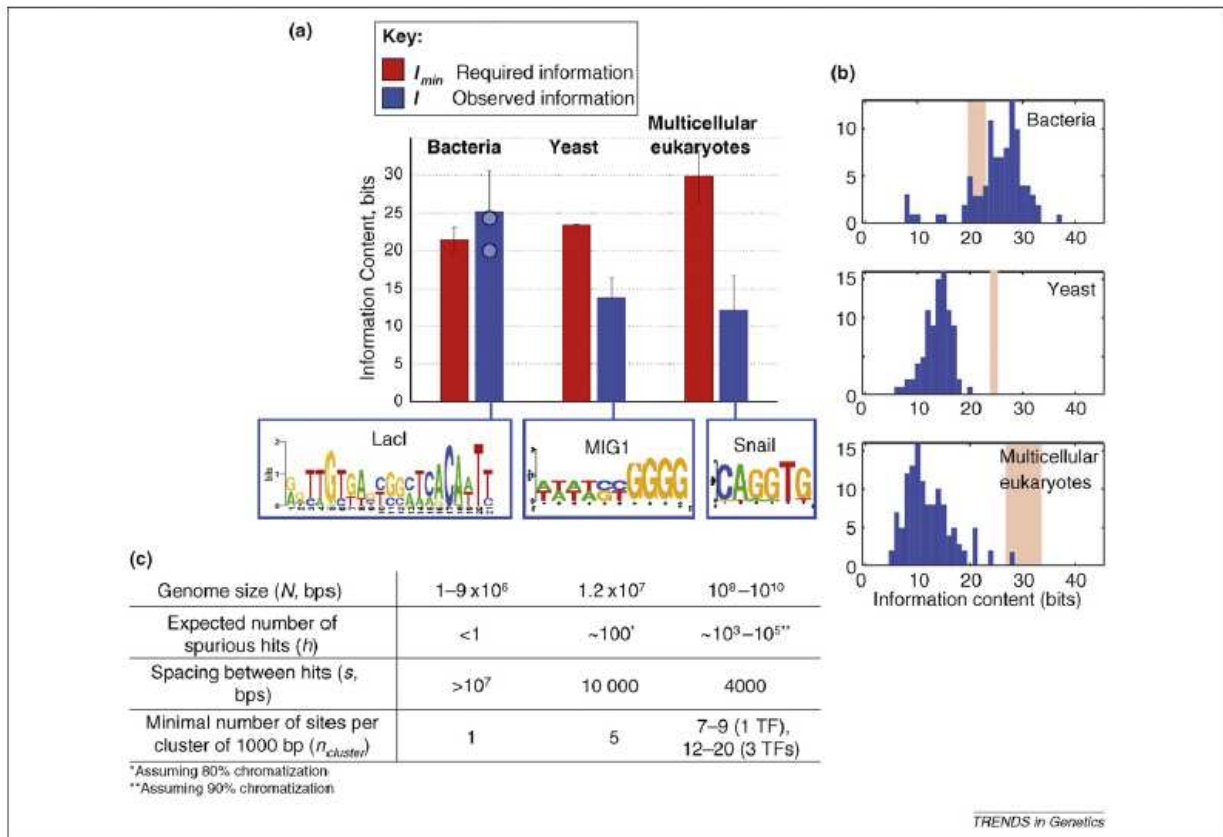


Figure 2. Properties of binding motifs for bacteria, yeast and multicellular eukaryotes. (a) The bar chart displays the minimum required information content for bacteria, yeast and multicellular eukaryotes (red), and the mean information content of TF-binding motifs (blue) for 98 bacterial [21], 124 yeast [22] and 123 multicellular [20] eukaryotic motifs. The error bars are ± 1 SD for the information content and for I_{min} , the error bars represent the variability in that quantity because of the range of genome sizes N . The blue circles on top of the bar for bacteria indicate the average information content from several other TF-binding motif databases (Table S6). Below each series in the bar chart, we display an example of the sequence logo for a binding motif with close to average information content. The chart demonstrates that bacterial TF-binding motifs are informative enough to make spurious hits to the genomic background unlikely, in contrast to yeast and multicellular eukaryotic motifs. (b) The distributions of information content of motifs from the three representative databases cited above. The ranges of required information (I_{min}) are marked in red. Most bacterial motifs have $I > I_{min}$, whereas almost all eukaryotic motifs do not. (c) The average properties of TF-binding motifs, and the expected number and spacing between the spurious sites per genome in bacteria, yeast and multicellular eukaryotes.

Based on these consideration I suggest a project in which one download from existing databases information contents of transcription factors this copy numbers in the cell and the number of their targets and compare IC to $\log_2 \frac{N}{nm}$. Good agreement would solve a very important puzzle in the field and contribute to our understanding of gene regulation both in prokariots and eukariots.

[1] Z. Wunderlich and L. A. Mirny, Trends in Genetics **25**, 434 (2009).