Dark Matter of the Transcription Factor Binding Site Motif Universe

Ariel A. Aptekmann
Marine and Coastal Sciences Department, Rutgers University
Denys Bulavka
Departamento de Matematica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires
Alejandro D. Nadra
Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Fisiologia, Biologia Molecular y Celular, IB3
Ignacio E. Sánchez (isanchez@qb.fcen.uba.ar)
Universidad de Buenos Aires. Consejo Nacional de Investigaciones Cientificas y Tecnicas. Instituto de Quimica Biologica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN).

Research Article

Keywords: genetic network, transcription factor binding site, transcription factor specificity, crosstalk, DNA sequence space, modified bases, sequence space usage

Posted Date: February 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-199537/v1

Dark matter of the transcription factor binding site motif universe

Ariel A. Aptekmann[1,2], Denys Bulavka[1,3], Alejandro D. Nadra[4], and Ignacio E. Sánchez[1]

[1] *Universidad de Buenos Aires. Consejo Nacional de lnvestigaciones Cientificas y Técnicas. Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN). Facultad de Ciencias Exactas y Naturales. Laboratorio de Fisiología de Proteínas. Buenos Aires, Argentina.*
[2] *Marine and Coastal Sciences Department, Rutgers University*
[3] *Departamento de Matematica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina.*
[4] *Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Fisiología, Biología Molecular y Celular, IB3. Buenos Aires, Argentina.,*

Correspondence may be addressed to Ignacio E. Sánchez (isanchez@qb.fcen.uba.ar)

1 Abstract

Background: We study the limits imposed by transcription factor specificity on the maximum size of a genetic regulatory network.

Results: Most regular expressions for natural transcription factor binding site motifs are separated in sequence space by only one to three motif-discriminating positions. This mild specificity requirement puts the number of transcription factors that can coexist with minimal crosstalk on the order of ten thousand, which would fully utilize the space of DNA subsequences. An expanded alphabet with modified bases can further raise this limit by several orders of magnitude, at the expense of sequence space usage.

Conclusions: Based on this analysis, thousands of transcription factor binding site motifs may await discovery.

Keywords: genetic network, transcription factor binding site, transcription factor specificity, crosstalk, DNA sequence space, modified bases, sequence space usage

2 Introduction

Specific interactions between proteins and nucleic acids are fundamental to the regulation of gene expression by transcription factors [1] in genetic networks [2]. Transcription factor binding sites (TFBS) are short degenerate DNA sequences of up to 30 base pairs long [3]. Characterization of TFBS usually starts by the experimental and/or computational identification of several DNA subsequences (termed TFBS instances) that perform a certain function. Once multiple instances of a TFBS are known, a TFBS motif is defined as the set of all TFBS instances that match with a given model (i.e., the set of sites to which a transcription factor binds preferentially) [4]. The computational definition of the nucleotide pattern for a TFBS motif can be a fixed consensus sequence, a regular expression, or a scoring matrix.

Our main question is how many TFBS motifs can coexist in a genome or, in other words, what is the maximum size of a genetic regulatory network. The SwissRegulon database currently contains annotations for 684 different TFBS motifs in the human genome [5], providing an empirical lower bound. From a different viewpoint, considering that each predicted human protein with DNA-binding domains recognizes a different TFBS motif suggests 2604 TFBS motifs in the human genome [2].

Theoretical estimations from first principles provide upper bounds for the number of coexisting TFBS motifs. Different transcription factors usually recognize non-overlapping sets of sequences, possibly because overlapping would lead to detrimental crosstalk between the biological signals read by the two transcription factors. When observed, the overlap of TFBS motifs is generally small [6]. We may consider as upper bound the maximum number of sequences of length n, which is $A(n) = 4^n$. A finer approach is to calculate the maximal number of TFBS motifs with a minimal Hamming distance d between sequences belonging to different motifs: $A(n, d) \leq 4^{n-d+1}$ [7]. Thus, a linear increase in transcription factor specificity leads to an exponential decrease in the maximal number of coexisting TFBS motifs. Coding theory provides a third upper bound for the number of minimally overlapping TFBS motifs: $A(n) \sim 3.5 + \sqrt{0.75 \cdot 4^n \cdot (n(4-1)-4))}$ [8]. The effects of motif length and specificity on the maximal number of TFBS motifs are thus strong. In spite of this, published work does not consider the specificity of natural TFBS motifs.

Published estimations for the maximal number of coexisting TFBS motifs assume a four letter DNA alphabet. However, many genomes harbor multiple modified bases [9] that may play a role in TFBS motifs [10]. The effective alphabet size of DNA may be over ten letters, which would significantly increase all theoretical estimates for the maximal number of coexisting TFBS motifs.

We apply regular expressions and theoretical tools developed for protein motifs (Bulavka et al., submitted) to the question of how many TFBS motifs can coexist in a genome. We consider empirical data for transcription factor sequence specificity, the effect of stable nucleotide modifications and sequence space occupancy.

3 Methods

3.1 Database of transcription factor binding site motifs

All available 684 regulatory motifs weight matrices from the SwissRegulon hg19 database were retrieved in June 2018 [5]. We converted each protein weight matrix from the original database to a regular expression. For each position of the matrix we used the observed frequencies b for A, C, G and T to calculate the effective alphabet size EAS [11]:

$$EAS = 2^{-\sum b \log_2 b} \tag{1}$$

We then assigned EAS letters to that position of the regular expression, by order of decreasing frequency. Last, we removed from the regular expression flanking positions that allow for all four bases.

3.2 Sequence specificity of transcription factor binding site motifs

We follow previous work on protein linear motifs (Bulavka et al., submitted). Briefly, we define a TFBS motif of length n as a sequence $\mathbf{A} = (A_1, \ldots, A_n)$ where each A_i is a subset of $\mathcal{A} = \{A, C, G, T\}$. A TFBS motif instance is a sequence $(a_1, \ldots, a_n)$ with $a_i \in A_i$ for all i. The structure of $\mathbf{A}$ is the sequence $(|A_1|, \ldots, |A_n|)$, i.e., the number of allowed bases at each position.

Given an alignment of two TFBS regular expressions $\mathbf{A} = (A_1, \ldots, A_n)$ and $\mathbf{B} = (B_1, \ldots, B_m)$, the number of *motif-discriminating positions* is the number of aligned positions with at most 3 allowed letters where no letter can match both regular expressions:

$$mdp\mathbf{AB} = |\{i \in \{1, \ldots, n\} : A_i \cap B_i = \emptyset \text{ with } |A_i| \leq 3 \text{ and } |B_i| \leq 3\}|. \tag{2}$$

We calculate $mdp\mathbf{AB}$ for the alignments between the two corresponding regular expressions that do not leave a hanging end for the shorter regular expression and match at least one pair of positions with less than four allowed letters. Finally, we take the minimal $mdp\mathbf{AB}$ across all relevant alignments as a lower limit for the distance in sequence space between the two TFBS motifs.

When the number of TFBS motif-discriminating positions is 0 for a given pair of motifs, we calculate an alternative measure of specificity as 1 - (number of sequences that match both regular expressions / number of sequences that match at least one of the regular expressions).

3.3 Number of potential transcription factor binding site motifs

For a given TFBS motif structure $\mathbf{e} = (e_1, \ldots, e_n)$ of length n and a number k of motif discriminating positions, $|\mathcal{M}(k)|$ denotes the maximal number of TFBS motifs in $\mathcal{M}_\mathbf{e}$ satisfying the property that every pair of motifs in it have at least k motif-discriminating positions (Bulavka et al., submitted).

$$|\mathcal{M}(0)| \leq \prod_{1 \leq i \leq n} \binom{3}{e_i - 1}, \tag{3}$$

$$\prod_{1 \leq i \leq n} \lfloor 4/e_i \rfloor \leq |\mathcal{M}(1)| \leq \prod_{1 \leq i \leq n} 4/e_i, \tag{4}$$

$$|\mathcal{M}(k < n)| \leq \prod_{1 \leq i \leq n-(k-1)} 4/e_i \tag{5}$$

$$|\mathcal{M}(n)| = \min_{1 \leq i \leq n} \lfloor 4/e_i \rfloor. \tag{6}$$

3.4 Occupancy of the sequence space

In the case of zero motif discriminating positions, each motif instance may belong to multiple motifs and we were not able to find a formula for the potential occupancy of sequence space (Bulavka et al., submitted). For values of k of one or more motif-discriminating positions, motif instances belong to a single motif and the fraction of the sequence space occupied by a motif of structure $\mathbf{e} := (e_1, \ldots, e_n) \in \{1, \ldots, 4\}^n$ is:

$$PotentialOccupancy(\mathbf{e}, k) := \prod_{1 \leq i \leq n} (e_i/4) * |\mathcal{M}(k)| \quad \text{for} \quad k > 0. \tag{7}$$

4 Results

4.1 Sequence specificity of known transcription factor binding site motifs

We considered positional weight matrices for 684 TFBS motifs in SwissRegulon (section 3.1). We generated a regular expression from each matrix, using information theory to minimize the loss of information. Figure 1A shows the frequency of each motif length in the database and of the number of symbols allowed at each position. TFBS motif length ranges from 4 to 30 characters, peaking at 10 characters. We quantify the distance in sequence space between a pair of TFBS motifs as the number motif-discriminating positions (section 3.2 and Supplementary Figure 1). This number is the minimal count of positions where no symbol can match both regular expressions, for every possible alignment where the number of aligned positions is the length of the shorter regular expression (Bulavka et al., submitted). Since other positions might not fully overlap, this is a lower limit for the distance in sequence space between the two TFBS motifs. We calculated the number of motif-discriminating positions for all possible 233586 pairs of TFBS motifs in our database (Figure 1B, white bars and left Y axis). 77% of the comparisons the two regular expressions are separated in sequence space by at least one motif-discriminating position. This is in agreement with the use of regular expressions, where a mismatch at a single position is enough to rule out that a sequence belongs to a given TFBS motif. On the other hand, it is rare to find pairs of regular expressions separated by more than five motif-discriminating positions. 23% of regular expressions pairs are not separated in sequence space by a motif-discriminating position. In this case, we measure the distance in sequence space using the fraction of sequences matching any of the two regular expressions that match only one of them (section 3.2). We find that 95% of motif pairs share less than 5% of sequences (Supplementary Figure 2). We conclude that SwissRegulon motif pairs show significant separation in sequence space, in agreement with our assumption that there is little cross-talk between natural TFBS motifs.

4.2 Number of potential transcription factor binding site motifs

We used our theory based on the pigeonhole principle (section 3.3 and Bulavka et al., submitted) and the structures of TFBS motifs in SwissRegulon (Figure 1A) to estimate the number of SwissRegulon-like TFBS motifs that can potentially exist in nature. We first converted the regular expressions in our database to motif structures (section 3.1). For each structure and a number of motif-discriminating positions, we calculated the number of potential TFBS motifs. As expected from the heterogeneity in motif lengths and structures, the calculated values span several orders of magnitude (Supplementary Figure 3). We report the median of the distribution. Requiring one motif-discriminating position maximizes the number of potential TFBS motifs to over 9700 (Figure 1B, black circles and right Y axis). The lower value for two

or more motif-discriminating positions is due to higher non-overlap requirements, while the lower value for zero motif-discriminating positions arises because the overlap imposed by this condition is more restrictive than the non-overlap imposed by one or more motif-discriminating positions. It is interesting to compare bars and circles of Figure 1B. On one hand, natural TFBS motif pairs are most often separated in sequence space by a single motif-discriminating position. On the other hand, this relatively low level of sequence specificity maximizes the number of potential TFBS motifs that can coexist while fulfilling the specificity requirement.

4.3 Role of nucleotide modifications

Current genome sequences only inform the four canonical bases, and it is often forgotten that nucleotide modifications are varied and frequent [9]. This increases the capacity of DNA to code for TFBS motifs [10]. Figure 1C shows the median number of potential TFBS motifs as a function of alphabet size for 0 to 4 motif-discriminating positions. Increasing the alphabet size from 4 to 10 increases the number of potential TFBS motifs by several orders of magnitude in all cases. When we consider an effective alphabet size of 10 letters, the increase relative to an alphabet of four letters is highest at over 9500-fold for one motif-discriminating position (Supplementary Figure 4). This effect decreases sharply with increasing motif specificity, becoming lower than ten-fold for 9 or more motif discriminating positions. This is notable since a single motif-discriminating position is the most frequent distance in sequence space between naturally occurring TFBS motifs (Figure 1B).

4.4 Sequence space occupancy

A TFBS motif of length n is a subset of the sequence space of all possible 4^n DNA subsequences. We used the size of the sequence space for each TFBS motif (Supplementary Figure 5) and the corresponding maximum number of coexisting motifs to calculate the potential occupancy of sequence space for 1 to 10 motif-discriminating positions (3.3). The calculated values span several orders of magnitude (Supplementary Figure 6). As done for the number of potential motifs, Figure 1D reports the median of the distribution. For a single motif-discriminating position, all possible DNA subsequences belong to a potential TFBS motif. The potential occupancy of sequence space drops steeply for two or more motif-discriminating positions. The commonest numbers of motif-discriminating positions (Figure 1B) maximize the potential occupancy of sequence space by the resulting TFBS motifs (Figure 1D). For a single motif-discriminating position, the potential occupancy of sequence space is 100% regardless of alphabet size (Supplementary Figure 7). For two or more motif-discriminating positions, the potential occupancy of sequence space decreases as alphabet size increases. For two or more motif-discriminating positions, increasing alphabet size leads to a trade-off between increasing the number of potential TFBS motifs (Figure 1C) and decreasing the potential occupancy of sequence space (Supplementary Figure 7).

5 Discussion

Naturally occuring TFBS motifs from SwissRegulon (Figure 1A) are commonly separated in sequence space by one to three motif-discriminating positions (Figure 1B). This level of sequence specificity not only avoids crosstalk between transcription factors but may also help coding a genetic network with several thousand TFBS motifs (Figure 1B) that maximizes sequence space usage (Figure 1D), where increasing the DNA alphabet size would allow for an even larger network (Figure 1C). This network level of TFBS motif specificity may inform

the design of new specific DNA binding proteins able to function in a cellular context, be it TALEN, Zinc-finger, CAS9 or others.

Our theory is in principle valid for any set of molecules recognizing stretches of a linear polymer, regardless of the interacting partners. The overall picture for TFBS motifs is similar to our previous results for protein-protein interaction networks mediated by linear motifs (Bulavka et al., submitted). In that case, the observed sequence specificity also maximizes the potential size of the network up to around ten thousand motifs. The main differences are that increasing the DNA alphabet size has a much larger effect than increasing the protein alphabet size and that sequence space usage is much larger for the genetic network than for the protein interaction network at the same level of specificity. These differences arise from both alphabet size and the motif regular expressions, i.e., from the physicochemical basis of protein-protein versus protein-nucleic acid complex formation [1].

TFBS motifs from SwissRegulon are commonly ten base pairs long, which corresponds to a space of $\sim 10^6$ sequences. Our theory predicts that this sequence space can be organized into a maximum of $\sim 10^4$ TFBS motifs, separated by a single motif-discriminating position. In turn, coding theory predicts a maximum of $\sim 4.5 \cdot 10^3$ minimally overlapping TFBS motifs of length 10 [8]. A similar maximum of $\sim 1.6 \cdot 10^4$ TFBS motifs can be obtained within the sphere packing approach of [7] and a minimal Hamming distance of 4 mutations between sequences belonging to different motifs. We find it reassuring that three different specificity-focused theories lead to similar estimates for the maximum size of a genetic network.

The actual upper bound for the number of TFBS motifs may be lower than 9700 due to phenomena not included in the theory. For example, the molecular interactions mediating protein-DNA interactions [1] may prevent some sets of DNA subsequences from becoming actual TFBS motifs. A need for mutational robustness [12] may further constrain maximal genetic network size. These factors could be accounted for in future models. Effectively reaching the upper limit may not be a requirement for the regulation of current genomes [2]. On the other hand, the gap between the 684 TFBS motifs in SwissRegulon [5] and the 2604 predicted DNA-binding human proteins [2], together with the observation of conserved DNA sequences of unknown function [13], directly point at significant amounts of dark matter of the transcription factor binding site motif universe awaiting discovery.

6 Figures and Tables

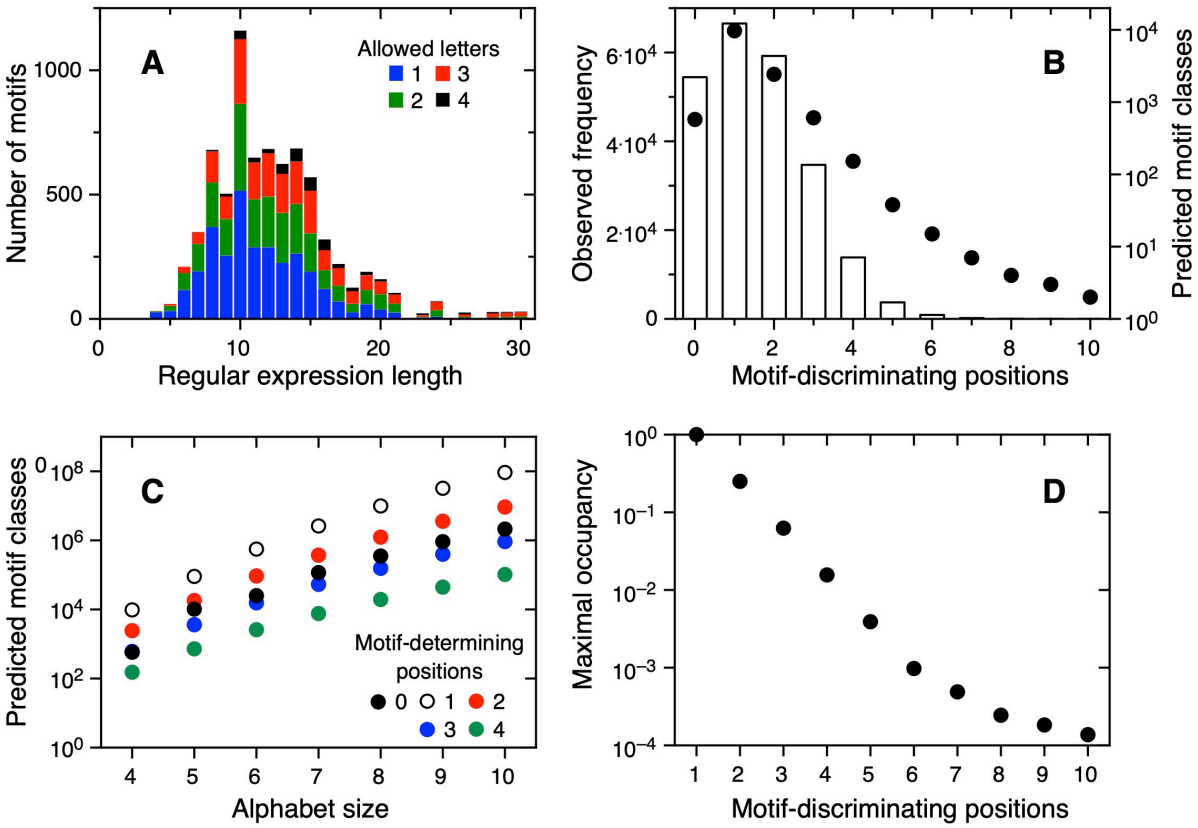

Figure 1: Known and predicted transcription factor binding site motifs. (A) Regular expression length and number of letters allowed for TFBS motifs in SwissRegulon. (B) Bars (left Y axis): Motif discriminating positions for every pair of TFBS motifs in SwissRegulon. Black circles (right Y axis): Theoretical estimation of the maximal number of coexisting TFBS motifs, as a function of the minimal requirement of motif-discriminating positions. (C) Theoretical estimation of the maximal number of coexisting TFBS motifs, as a function of alphabet size. (D) Maximal occupancy of the protein sequence space by TFBS motifs for an alphabet size of 4 as a function of the number of motif-discriminating positions.

7 Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Competing interests

There is no conflict of interest.

Funding

We acknowledge funding from Agencia Nacional de Promoción Científica y Tecnológica (PICT 2015-1213 to I.E.S.) and Consejo Nacional de Investigaciones Científicas y Técnicas (T.K., A.D.N. and I.E.S. are CONICET career investigators). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

AAA, DB, ADN and IES designed and performed research and wrote the manuscript.

Acknowledgments

Not applicable

References

[1] Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, and Richard S Mann. Origins of specificity in protein-dna recognition. *Annual review of biochemistry*, 79:233–269, 2010.

[2] M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14(3):283–291, 2004.

[3] Mikhail Pachkov, Ionas Erb, Nacho Molina, and Erik Van Nimwegen. Swissregulon: a database of genome-wide annotations of regulatory sites. *Nucleic acids research*, 35(suppl_1):D127–D131, 2006.

[4] Thomas D Schneider and R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.

[5] Mikhail Pachkov, Piotr J Balwierz, Phil Arnold, Evgeniy Ozonov, and Erik Van Nimwegen. Swissregulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic acids research*, 41(D1):D214–D220, 2012.

[6] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[7] Amit Marathe, Anne E Condon, and Robert M Corn. On combinatorial dna word design. *Journal of Computational Biology*, 8(3):201–219, 2001.

[8] Shalev Itzkovitz, Tsvi Tlusty, and Uri Alon. Coding limits on the number of transcription factors. *BMC genomics*, 7(1):239, 2006.

[9] Ankur Jai Sood, Coby Viner, and Michael M Hoffman. Dnamod: the dna modification database. *J Cheminform*, 11(1):30, 2019.

[10] Aaron M Fleming, Yun Ding, and Cynthia J Burrows. Oxidative dna damage is epigenetic by regulating gene transcription via base excision repair. *Proc Natl Acad Sci U S A*, 114(10):2604–2609, 2017.

[11] Claude E Shannon. A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, 27:623–656, 1948.

[12] Anirvan M Sengupta, Marko Djordjevic, and Boris I Shraiman. Specificity and robustness in transcription control networks. *Proceedings of the National Academy of Sciences*, 99(4):2072–2077, 2002.

[13] Gill Bejerano, Michael Pheasant, Igor Makunin, Stuart Stephen, W James Kent, John S Mattick, and David Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, 2004.

Figures

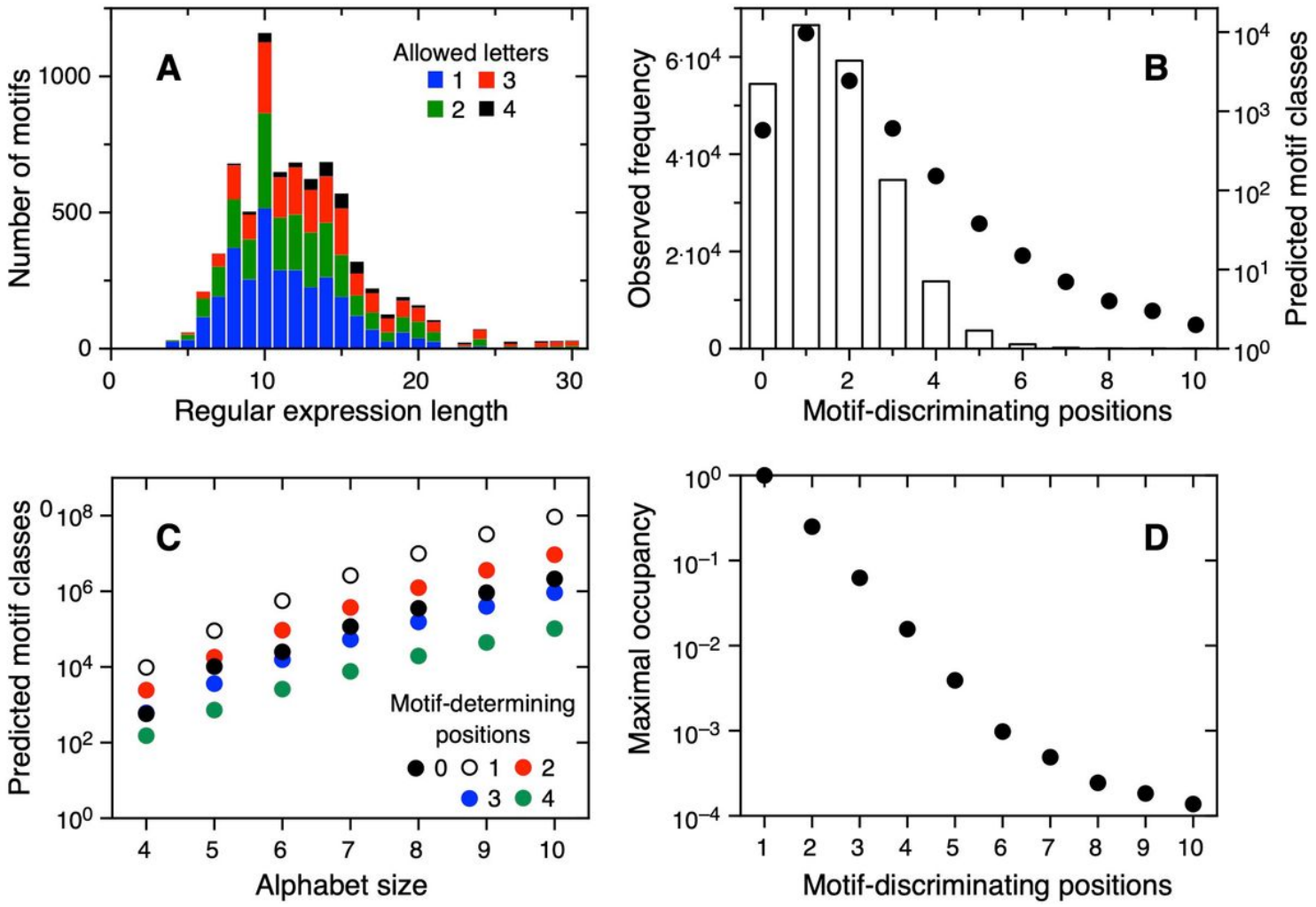

Figure 1

Known and predicted transcription factor binding site motifs. (A) Regular expression length and number of letters allowed for TFBS motifs in SwissRegulon. (B) Bars (left Y axis): Motif discriminating positions for every pair of TFBS motifs in SwissRegulon. Black circles (right Y axis): Theoretical estimation of the maximal number of coexisting TFBS motifs, as a function of the minimal requirement of motif-discriminating positions. (C) Theoretical estimation of the maximal number of coexisting TFBS motifs, as a function of alphabet size. (D) Maximal occupancy of the protein sequence space by TFBS motifs for an alphabet size of 4 as a function of the number of motif-discriminating positions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DNAmotifsBMCBioinfo20210211SupplementaryFigures.pdf