Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do expectation-maximization algorithms have to make use of ‘background distribution’ for de novo motif search?

Bannikov Artyom Vladimirovich (broyler3@mail.ru)

Research Article

Keywords: expectation-maximization, motif search, regulation of transcription

Posted Date: June 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1725972/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Introduction.

The design of synthetic promoters remains a challenge for current biotechnology. By and large, the whole industry relies upon several well-characterized promoters of viral origin; for example, the CMV promoter. These promoters have high activity in a broad range of cell lines. But high expression does not invariably produce better results. Too much heterologous protein (in short time) can lead to unfolded protein response and as a consequence to diminished productivity [1, 2]. A set of eukaryotic promoters covering a wide range of activity levels would be a valuable tool for biotechnology and molecular biology.

Several recent works employed transcription-guided design of synthetic promoters [3, 4]. The approach consisted in identifying sequence motifs over-represented in promoters of highly transcribed genes. The discovered motifs were concatenated into a single sequence to obtain a highly active synthetic promoter. Obtaining a sequence with high promoter activity turned out to be a relatively simple task. Thus, taking it a step further might be feasible entertainment. It might be possible to characterize the contribution of each motif quantitatively. We would like to assign a number to each motif, so that the activity of any combination (a promoter) can be computed as the sum of its constituent motifs.

One of the most popular methods of motif search is expectation-maximization (EM) [5, 6]. The EM algorithm converges to a (minus) log-likelihood minimum, which can be any local minimum. Commonly, EM is started many times with varied (random) starting conditions in the hope of finding relevant motifs. Many EM-based algorithms score subsequences by the log-likelihood ratio. Subsequences are assumed to be produced by either the background distribution or the motif distribution. Motif sequences are required to be 'close' to each other and 'far' from the background. But what is the biological significance of this condition? Background can't change the fact that the factor will bind whenever binding is possible. The existence/importance of a 'background' is hypothetical information. The analysis is required to understand the role of a background, if any, and to meaningfully include it into the model. Scoring motifs by logL values shall be preferred because it utilizes information that is actually in our possession. In this work the scoring by logL values is shown to be practical.

Methods.

1. Expectation-maximization algorithm.

The input is a set of N nucleotide sequences with lengths $l_1, l_2, \dots, l_N$ and the length of the motif w to search for. Random subsequences of length w are picked to start expectation-

maximization (EM) iterations. The algorithm runs in ZOOPS (zero or one occurrence per sequence) mode. During the expectation step, a position weight matrix (PWM) is evaluated using defined subsequences. PWM is the $4 \times w$ matrix whose element f_{ij} is the frequency of the nucleotide j at position i. For each subsequence the value of the log-likelihood (logL) function is computed:

$$logL = \sum_{j=1}^{4} \sum_{i=1}^{w} \log (f_{ij}) \quad (1)$$

The subsequence with the maximum value of $logL$ is set as the current evaluation of the motif for a given sequence, which constitutes the maximization step. Identified motifs are cross-validated. Motifs are excluded from the motif set one at a time; PWM and then $logL$ are re-evaluated. If the motif exclusion leads to an increase in the $logL$ value, the motif is annulled; i.e., it is concluded that the sequence contains no motif occurrences. Identified motifs are masked (replaced by N's). Masked subsequences can't be included into motifs. EM iterations continue until the required number of attempts is made or until no motifs can be found. If the algorithm can't find any more motifs, but the number of attempts is below the required, the input is reset (initial unmasked sequences are set as input), and iterations continue.

Computations were done on AMD E2-9000e RADEON R2 processor (1500MHz), single core. The operating system was Debian 10 (x86_64).

2. Preparation of synthetic sequence sets.

Active promoters of the GM12878 cell line were identified by intersecting CAGE and DNAse-seq data from ENCODE portal [7]. Promoters were defined as DNAse-sensitive regions (from ENCFF748UZH.bed file) that contain transcription start sites (from ENCFF853HOH.bed file). Sequences for a motif planting experiment were prepared using following steps: (a) choose N random promoters from the set of all GM12878 promoters, (b) compute nucleotide frequencies for each promoter and create N random sequences with same nucleotide frequencies, (c) create m synthetic PWMs and for each PWM create N sequences of length w (these are motifs to be planted).

PWMs are characterized by 3 numbers: (1) motif length w; (2) the probability of insertion and (3) minimax nucleotide frequency t. For each sequence a random number from 0 to 1.0 is generated. If the random number is smaller than the probability of insertion, the motif is planted into the sequence. The minimax nucleotide frequency t is the minimum value of $\max (f_{ij})$ for all motif positions.

N synthetic sequences/promoters and $m \times N$ synthetic motifs are generated. Each motif is planted into a fraction (controlled by its probability of insertion) of sequences.

3. Estimation of the total number of *-logL* minima.

Suppose sampling with replacement is performed from the population of animals with size N. Upon catching, an animal is marked and released. All animals have equal probability of being caught. If m unique animals are caught in M catches, the expected value of m is:

$$E(m) = N * \left(1 - \left(\frac{N-1}{N}\right)^M\right) \quad (2)$$

When the total number of $-logL$ minima are estimated, $E(m)$ is the number of unique motif sets discovered by the EM algorithm and M is the total number of discovered motif sets. The value of N was increased starting from $N_0 = m$ with step 1. When $E(m|N_{i-1}) \leq m \wedge E(m|N_i) > m$, the estimate of the population size is N_i.

4. Analysis of motif planting.

Five independent sequence sets were created for each value of N. Three motifs were planted into every sequence set. The Kullback-Leibler divergence D_{KL} between PWMs P and Q:

$$D_{KL}(P, Q) = \sum_{i=1}^{w} \sum_{j=1}^{4} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right) \quad (3)$$

A_i denotes the set of motifs planted into ith sequence set. B_i denotes the set of motifs discovered by EM in the ith sequence set. For the kth planted motif A_{ik} the closest motif $C_{ik} \in B_i$ is the one that satisfies $D_{KL}(A_{ik}, C_{ik}) = \min_p (D_{KL}(A_{ik}, B_{ip}))$. The closest motif can be found among motifs planted into other sequence sets; i.e., $D_{KL}(A_{ik}, C_{jk}) = \min_p (D_{KL}(A_{ik}, B_{jp}), i \neq j, C_{jk} \in B_j)$. The quantity $\Delta D_{ijk} = D_{KL}(A_{ik}, C_{jk}) - D_{KL}(A_{ik}, C_{ik})$ is the minimum Kullback-Leibler divergence difference. Generally, the EM would not find planted motifs exactly. The reason for this is discussed in Results. If motifs that are close to planted motifs are found by the EM, $\Delta D_{ijk} > 0$ should be expected.

5. Regression model of promoter activity.

To every promoter, two numbers corresponded, (1) expression level and (2) TSS activity. TSS activity was extracted from CAGE data (file ENFF853HOH.bed, column#7). To define the expression level, first, transcripts that originate from the promoter were identified. If the promoter end (strand-specific) was 100nt or closer to the beginning of the transcript, the

transcript was assigned to the promoter. Transcript-level expression data for GM12878 cells were downloaded from ENCODE site (see Supplementary Data for the file list). There were 32 files from 10 experiments. The expression level of the promoter was equal to the sum of abundances (in TPM) of all transcripts assigned to the promoter. The expression level was used to cluster promoters, while TSS activity was the model's predicant.

Pearson correlation coefficient r (GSL function gsl_stat_correlation) was computed between every pair of expression vectors. Promoters were clustered such that between any two promoters in the cluster the value of r was larger or equal to 0.9. Clustering was done by exhaustive search. There were 223766 clusters. Only clusters with sizes $s \geq 40$ were considered in subsequent analysis. There were 1861 clusters with $s \geq 40$ and they contained 620 promoters. From those numbers it is clear that the majority of clusters have to significantly overlap. Sets of clusters that efficiently covered all 620 promoters were obtained in the following way. A random starting cluster was chosen. A cluster was added to the group, if it shared no more than 4 promoters with any other cluster in the group. Five groups of clusters were obtained; 40 clusters in total. These clusters were analyzed by MEME (version 5.4.1). MEME options: ZOOPS mode; minimum and maximum motif widths were 6 and 12; classical scoring function; the number of motifs to find was 10. FIMO (version 5.4.1, web interface, cutoff p-value was 0.0001) was used to search 400 motifs found by MEME in all promoter sequences (N=8261). The design matrix indicating presence/absence of a motif in a promoter was obtained. The CAGE promoter activity was a predicant. The regression model was estimated using LASSO. The function glmnet() (R package glmnet version 4.1-3) performs LASSO with user-specified value of the penalty coefficient λ. The function lasso_perm() (R package adapt4pv vesion 0.2-2) performs permutation of the design matrix and chooses the best value of λ automatically.

Results.

1. The number of EM-discoverable motifs grows exponentially with the number of sequences.

Suppose we are given a collection of N nucleotide sequences. Each sequence has length l. The set of subsequences with length $w < l$, such that each subsequence belongs to different sequence, is the motif set. The number of subsequences in the motif set is the motif set size x. If one knows N and l (and the alphabet), all possible motif sets can be formulated. To every motif set a position weight matrix (PWM) and -$logL$ value can be corresponded. The knowledge of

sequences amounts to choosing a sample from this collection of points. The expectation-maximization algorithm (EM) searches for local minima in the given sample. We are interested in the number of -*logL* local minima. Specifically, how the number of local minima changes, when the number of sequences increases. Increasing the number of sequences adds points to the sample. What happens when points are added to, for example, white noise? If points are added outside of the domain, the number of local minima grows proportionally to the number of added points. When points are added inside the domain, intuition suggests that the number of minima grows slower. For the problem of motif search the domain has to be defined. If set A has a lower -*logL* value than B, and sequences of A are 'close' to sequences of B, set A is 'better' than set B. So the domain measure must be a measure of the difference between sequences. These ideas are conveyed in a more rigorous way in the following theorems.

Theorem 1. The function $u = \frac{\prod_{i=1}^{w} c_i^{(1)}}{x}$, where $c_i^{(1)}$- the count for the most frequent nucleotide in the position/column i, x - size of the motif set, is the strictly increasing function of $logL$.

Nucleotide frequencies are replaced by counts because later this result is linked to the Hamming distance:

$$logL = \sum_{j=1}^{4}\sum_{i=1}^{w}\log(f_{ij}) = \sum_{j=1}^{4}\sum_{i=1}^{w}\log(f_{ij} * x) - \log(x) = \sum_{j=1}^{4}\sum_{i=1}^{w}\log(c_{ij}) - \log(x) = \sum_{j=1}^{4}\log\left(\prod_{i=1}^{w} c_{ij}\right) - \log(x) \quad (4)$$

f_{ij} - frequency of jth nucleotide at position i, c_{ij} - count of jth nucleotide at position i.

Let's write the product for the ith position as $c_i^{(1)}c_i^{(2)}c_i^{(3)}c_i^{(4)}$, where a superscript denotes the rank. When the difference between the two positive numbers increases, their product decreases. Consider two numbers a_1 and a_2, where $a_1 + a_2 = s$. Preserve the sum s and form the product:

$$(a_1 + 1)(a_2 - 1) = a_1a_2 + a_2 - a_1 - 1 \quad (5)$$

From eq.(5) $(a_1 + 1)(a_2 - 1) > a_1a_2$, iff $a_2 > a_1$. If we write $a_1 = c^{(1)}$ and $a_2 = \sum_{k=2}^{4} c^{(k)}$, it follows that $\prod_{i=1}^{w} c_i < \prod_{j=1}^{w} c_k$, when $c_i^{(1)} > c_k^{(1)}$. Since $\frac{\prod_{j=1}^{4}\prod_{i=1}^{w} c_{ij}}{x} = \prod f_{ij} \leq 1$, it follows that $-\log\left(\frac{\prod_{j=1}^{4}\prod_{i=1}^{w} c_{ij}}{x}\right) \geq -\log\left(\frac{\prod_{j=1}^{4}\prod_{k=1}^{w} c_{kj}}{x}\right)$, when $\prod_{i=1}^{w} c_i < \prod_{k=1}^{w} c_k$.

Thus, u and $-logL$ can be used to rank motif sets interchangeably.

Theorem 2. When the number of sequences N changes, the number of motif sets with given Hamming distance H changes by a factor whose value does not depend upon the Hamming distance.

When two motifs in a motif set have same nucleotide at the same position, call this a match. Let's denote p_{match} the probability of a match in the motif set. The probability of observing m matches $q(m)$ is binomial:

$$q(m) = \binom{w}{m} p_{match}^{m}(1 - p_{match})^{w-m} \quad (6)$$

In the motif set of size x the number of comparisons is $(x^2 - x)$. Assume that sequences in the motif set are independent. The probability $p(m)$ of observing a motif set with size x and the Hamming distance between all possible pairs of motifs equal to or less than m can be written:

$$p(m|N, x, w) = \frac{\sum_{k=1}^{m}[\binom{w}{k} p_{match}^{k}(1-p_{match})^{w-k}]^{x^2-x}}{\sum_{x=1}^{N}\sum_{m=1}^{w}[\binom{w}{m} p_{match}^{m}(1-p_{match})^{w-m}]^{x^2-x}} = \frac{c}{b} \quad (7)$$

Only the value of the normalization factor b depends upon the number of sequences N. After adding one sequence $p(m|N+1, x, w)$ becomes:

$$p(m|N+1, x, w) = \frac{\sum_{k=1}^{m}[\binom{w}{k} p_{match}^{k}(1-p_{match})^{w-k}]^{x^2-x}}{\sum_{x=1}^{N}\sum_{m=1}^{w}\left[\binom{w}{m} p_{match}^{m}(1-p_{match})^{w-m}\right]^{x^2-x} + \sum_{m=1}^{w}\left[\binom{w}{m} p_{match}^{m}(1-p_{match})^{w-m}\right]^{(N+1)^2-(N+1)}} = \frac{c}{b+a} \quad (8)$$

The probability $p(m)$ is proportional to the number of motif sets:

$$\frac{p(m|N+1, x, w)}{p(m|N, x, w)} = 1 + \frac{a}{b} = 1 + \frac{\sum_{m=1}^{w}\left[\binom{w}{m} p_{match}^{m}(1-p_{match})^{w-m}\right]^{(N+1)^2-(N+1)}}{\sum_{x=1}^{N}\sum_{m=1}^{w}\left[\binom{w}{m} p_{match}^{m}(1-p_{match})^{w-m}\right]^{x^2-x}} \quad (9)$$

When N increases, the increase in the number of motif sets with $H = m$ is independent of m.

Theorem 3. The number of *-logL* minima is proportional to exp $(\frac{1}{2e})$.

First, let's clarify the meaning of the maximum Hamming distance H_{max}. If $H_{max} = a$, only a columns/positions can vary; other columns must have all identities. For example, consider the following motif set (variable positions are in bold):

ATG**G**CTC

ATG**C**CTC

ATG**T**CTC

A**A**G**T**CTC

For first 3 sequences $H_{max} = 1$. The 4th sequence differs in one more position, which makes $H_{max} = 2$. If a motif set has size x and $H_{max} = a$, $(w - a)$ columns have $c_i^{(1)} = x$. The remaining a variable positions have $c_i^{(1)} = k_i$:

$$\frac{\prod_{i=1}^{w} c_i^{(1)}}{x} = \frac{x(w-a)\prod_{i=1}^{a} k_i}{x} \quad (10)$$

Accordingly, for a set with size $(x + 1)$ and $H_{max} = b$:

$$\frac{\prod_{j=1}^{w} c_j^{(1)}}{x+1} = \frac{(x+1)(w-b)\prod_{j=1}^{b} k_j}{x+1} \quad (11)$$

The motif set with size $(x + 1)$ is 'better' than the motif set with size x, when $\frac{\prod_{j=1}^{w} c_j^{(1)}}{x+1} > \frac{\prod_{i=1}^{w} c_i^{(1)}}{x}$ according to the Theorem 1. Using eq.(10) and eq.(11) we arrive at the following inequality:

$$\frac{\prod_{i=1}^{a} k_i}{\prod_{j=1}^{b} k_j} < \frac{w-b}{w-a} \quad (12)$$

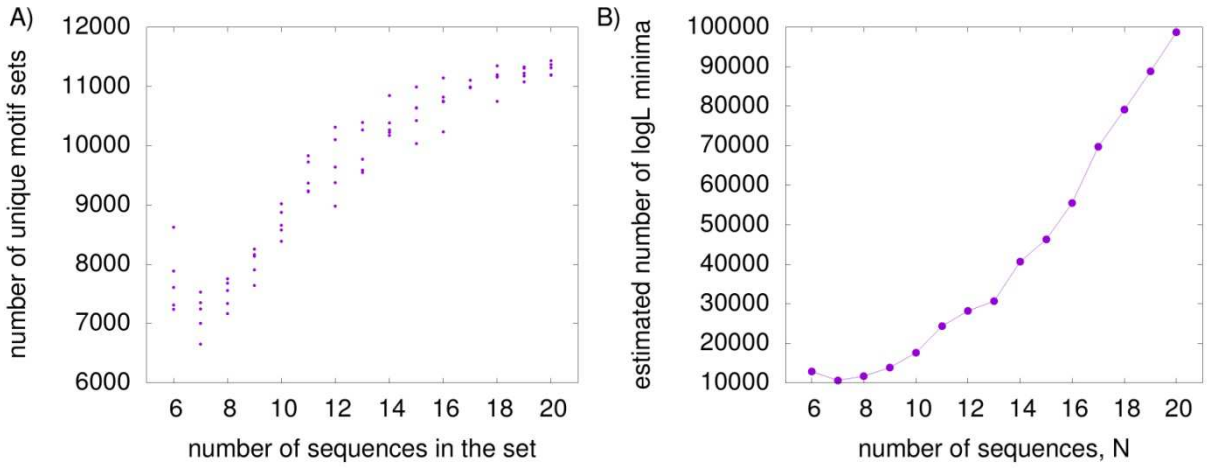
Motif sets are grouped according to the value H_{max}. A motif set is a *-logL* minimum, only if there is no 'better' motif set. Any motif set that satisfies the condition eq.(12) is a 'better' motif set. Thus, the total number of *-logL* minima is the number of motif sets for which no 'better' sets exist. Theorem 2 tells what happens, when sequences are added. The number of motif sets in each group increases ω –fold. A motif set is simultaneously a potential *-logL* minima and a potential 'better' motif set. Consequently, the number of minima does not change. Adding one sequence also spawns the group, where $H_{max} = N + 1$. Only motif sets of the maximum size can have $H_{max} = N + 1$. The fraction of motif sets of the maximum size asymptotically goes to $\frac{1}{e}$ [8]. If motif sets are independent and uniformly distributed, half of them are expected to be (local) minima. It follows that the number of *-logL* minima grows exponentially with the rate $\frac{1}{2e}$.

2. The EM algorithm search time grows linearly with the number of sequences.

Synthetic sequence sets were created as described in Methods. All sequences had length $l = 120$ nucleotides. The number of sequences N in the set varied from 6 to 20. For each value of N five independent sets were created. Three motifs with length 10nt were planted into sequences. Minimax nucleotide frequencies were set to 0.9, 0.8 and 0.7; planting probability was equal to 0.6. Using the number of unique motif sets identified by the EM, the total number of

$-logL$ minima was estimated. Estimating the total number of $-logL$ minima is a problem in the domain of capture-recapture analysis (CRA) [9]. The CRA was developed to estimate the sizes of animal populations. A minimal capture-recapture experiment includes two time points. First, animals are trapped, marked, and freed. Second, trapping is repeated to find out the proportion of marked animals. In the problem under consideration 'animals' are $-logL$ minima. During EM iterations 'animals' are trapped (the algorithm converges), marked (the motif set is identified). Repeated trapping is possible (motif sets can be re-discovered). The EM algorithm is sampling from the population of $-logL$ minima with replacement. Under the assumption of equal 'catchability' the simplest estimator of the size of the population showed good performance (Fig. S1). The total number of $-logL$ minima grew exponentially (Fig.1). The rate of growth 0.17 was close to the theoretically predicted value $\frac{1}{2e}$.

Fig.1. A) The dependence between the number of unique motif sets discovered by EM and the number of sequences, *N*. For each value of *N* five sequence sets were analyzed. B) Estimation (mean, n=5) of the total number of unique motif sets discoverable by the EM in sequence sets of different sizes.



Every sequence set contained 3 planted motifs. The discovery of precisely these motifs shall not be expected for at least two reasons. First, only a fraction of the $-logL$ minima were analyzed. Second, planted motif sets might not be a $-logL$ minimum. A sequence set might contain a motif set that is 'close' to the planted motif set and is 'better'. For example, suppose the following motif was planted:

ATG

CTG

ACG

A sequence from the set, where no motif was planted, may happen to contain the substring ATG:

ATG

CTG

ACG

ATG

Or the sequence with planted CTG may contain an ATG substring:

ATG

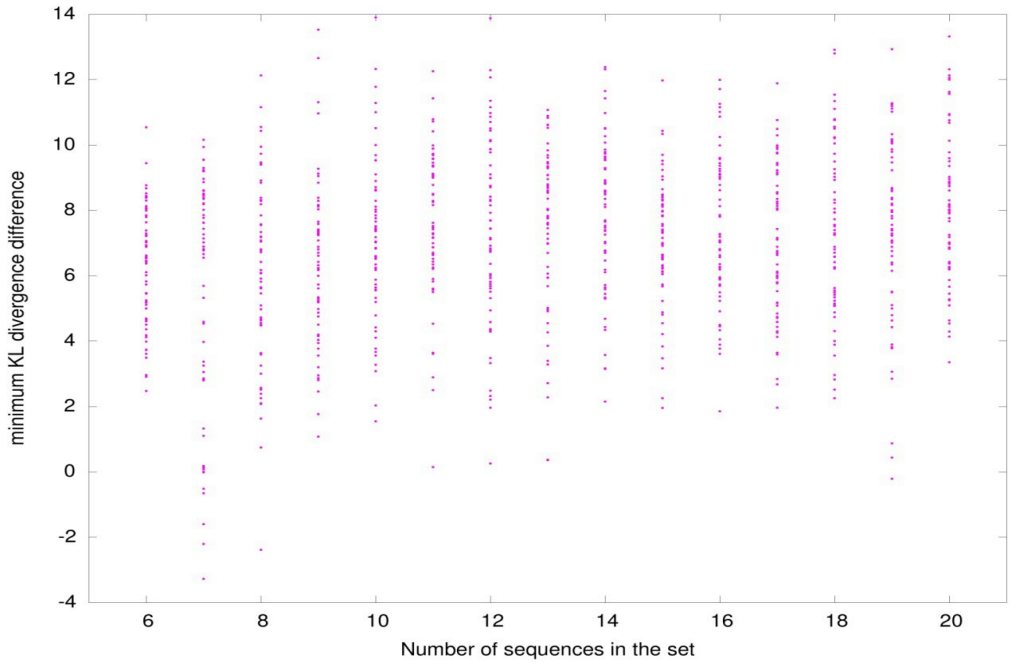
ATG

ACG

The example indicates that the EM shall be able to discover PWMs that are close to PWMs of planted motifs. Indeed, EM-discovered motifs were closer to planted vs non-planted motifs (Fig.2).

As the number of sequences increases the number of $-logL$ minima grows exponentially. The time needed to discover all minima (or a constant fraction of all minima) would also grow exponentially. In practice, we are interested only in 'statistically significant' motif sets. Motif sets have to have a certain minimum size to be 'significant'. For example, the minimum significant motif set size (MSMSS) can be defined as size x_0 such that the combined fraction of motif sets with size $x_0 \leq x \leq N$ is less than 5%. Considering only 'significant' motifs saves computation time. The cross-validation step subtracts one sequence at a time and checks how the value of *logL* changes. The execution time can be reduced by immediately stopping iterations once the number of sequences in the motif set becomes lower than the MSMSS.

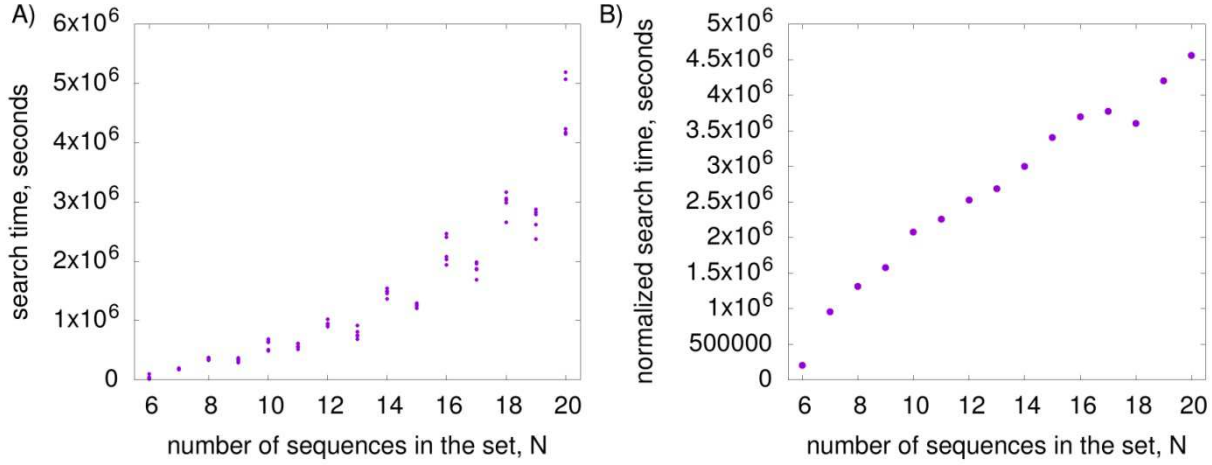
Fig.2. Minimum KL divergence differences. For definitions see "Methods".



The MSMSS value depends upon the number of sequences and their lengths. MSMSS can be estimated empirically. If the number of sequences and their lengths are decreased by the same factor, ratios between numbers of motif sets of different sizes are approximately preserved (see Supplementary Results). For example, EM is to be performed on a sequence set with $= 200, l = 1200$. We are interested in estimating the fraction of motif sets with size $x \geq 100$. To answer this question, we perform EM on the sequence set with $N = 20, l = 120$ and estimate the fraction of motif sets with size $x \geq 10$. The actual value is ~0.15.

The time to find 1000 motifs $t_{1000}(x > \frac{N}{2})$ with size $x > \frac{N}{2}$ was measured for same sequence sets, which were used in planting experiments ($6 \leq N \leq 20$). The dependence $t_{1000}(N)$~exp (N) was observed (Fig.3A). In planting experiments, 12000 motifs were identified for each sequence set. The fraction of motif sets with size $x > \frac{N}{2}$ was computed (Fig.S2). The normalized time $t^*_{1000} = t_{1000}(N) * \frac{f(N)}{f(N=20)}$ is the time needed to discover the equal fraction of motif sets with length $x > \frac{N}{2}$. The linear dependence $t^*_{1000}(N)$~N was observed (Fig.3B). Thus, we conclude that the time needed to find the constant fraction of motif sets satisfying the condition of type $x > a$, where $2 \leq a \leq (N - 1)$, increases linearly with the number of sequences in the set.

Fig.3. The EM algorithm search times. A) Time to find 1000 motif sets larger than N/2; B) Normalized time to find 1000 motif sets larger than N/2.



A motif search experiment has the following layout. Suppose we have to search for motifs in the A motif search experiment has the following layout. Suppose we have to search for motifs in the $MSMSS = \frac{N}{2}$ was accepted. For the second EM run motifs are planted into the sequence set. A fixed number of motif sets with size that is equal to or larger than MSMSS are searched for. If planted motifs are successfully identified, it is concluded that the search time is

sufficient. The linear model is then used to compute how much time would be required to obtain the same results with larger sequence set.

Finding 1000 motif sets with size $x > \frac{N}{2}$ in the sequence set $N = 20, l = 120$ allowed detecting planted motifs. The theory suggests that the sequence set $N = 200, l = 1200$ requires ~150 hours to produce similar results (Fig.4). It is worth noting that motifs with lower information content (lower values of minimax fraction t) seem to be more discoverable. Motif 'catchability' may be more important than the information content. The motif 'catchability' is proportional to the number of starting conditions that converge to the motif. Motifs with high information content might require very specific starting conditions and thus be hard to catch.

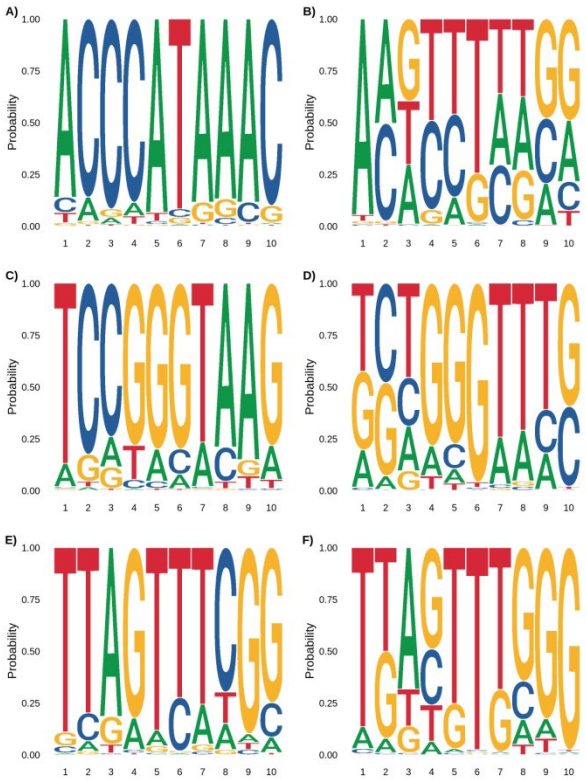


Fig.4. The result of EM search in the sequence set $N = 200, l = 1200$ for 150 hours. Logos of planted motifs and motifs discovered by the EM, which were closest to planted motifs. A) Planted motif#1 (t=0.9); B) Discovered motif, closest to the planted motif#1; C) Planted motif#2 (t=0.8); D) Discovered motif, closest to the planted motif#2; E) Planted motif#3 (t=0.7); F) Discovered motif, closest to the planted motif#3.

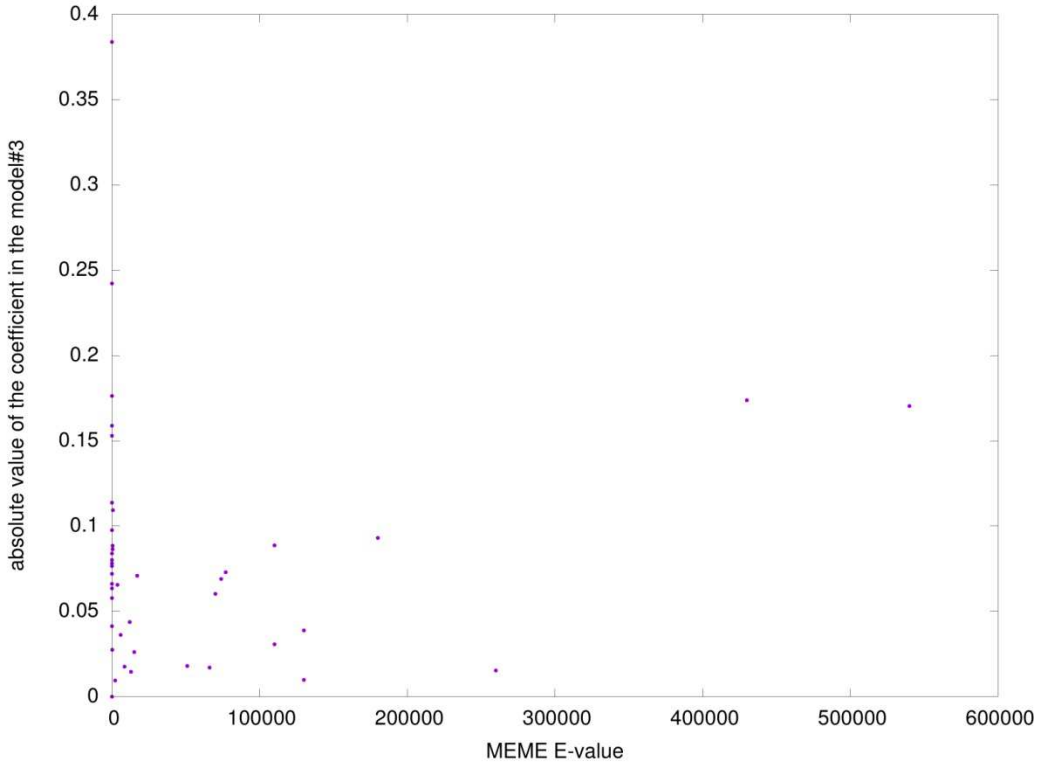
The next experiment tested how well the output of MEME [10], which ranks motifs by E-value, suits the purpose of expression modeling. Promoters were clustered by co-expression. Motifs were searched only in clusters with a size larger than or equal to 40 (see Methods for details). Analyzed clusters contained 620 promoters out of a total of 8960. Four models were tested (Table 1). For each model, a (lasso) regression was evaluated by the glmnet() function (Fig.S3). Model#3 was analyzed in more detail. With $\lambda = 0.05$ the model#3 had $R^2 = 0.207$. The Kendal correlation coefficient between model prediction and data was 0.295 (Pearson and Spearman correlation coefficients were 0.499 and 0.417, respectively). The number of significant

Table 1. The four regression models (N – the number of predictors in the model).

Exclude promoters with zero expression \ Exclude non-clustered promoters	no	yes
no	Model#0 (N=8960)	Model#2 (N=620)
yes	Model#1 (N=5761)	Model#3 (N=547)

predictors/motifs was 43 (Fig.S4). The function lasso_perm() found 7 significant motifs. They were among the significant motifs identified by the glmnet() function. When the design matrix *m* was permuted by the command apply(*m*,2,sample), the lasso_perm() did not find significant motifs. MEME E-values were not good predictors of whether lasso identified a motif as significant (Fig.5).

Fig.5. The relation between MEME E-values and absolute values of regression coefficients for significant motifs of the model#3 (lasso penalty term 0.05).



Discussion.

It was shown that EM algorithm without background distribution is suitable for finding motifs in relatively large sequence sets (hundreds of sequences with lengths of thousands of nucleotides) in realistic time (hundreds of hours on a single core).

Several shortcomings of this work are to be mentioned. First, the estimator of population size implies sampling with replacement. The EM algorithm does not fully correspond to such a model. After a motif is found, it is masked. The motif can't be re-discovered until the algorithm resets. In the capture-recapture experiment, every day animals are trapped. In the EM algorithm, one day is the time from one reset to the other. Animals that are trapped during a single day are 'sampled without replacement', i.e., they can be re-sampled at later days, but not the same day. If the number of animals captured at any given day is low compared to the total population size, sampling can be well approximated as sampling with replacement. Nevertheless, some conditions have to be satisfied for the 'sampling with replacement' approximation to work.

Second, the argument is made that the relation between the time to find a fixed fraction of motif sets with size $x \geq a$ and the number of sequences is linear. Planted (or close to them) motifs could be discovered in experiments because (a) the fraction of discovered motifs is high or (b) motifs have unequal discoverability/catchability. If (b) is true, some questions are to be asked. How unequal discoverabilities are? How this affects the quality of the population size estimator?

Data availability.

All research data pertinent to this work including programs, source code, raw and processed analytical data were uploaded to Zenodo, DOI: 10.5281/zenodo.6678391.

Conflict of interest.

I have no conflict of interest to declare.

References.

1. Prashad K, Mehra S. Dynamics of unfolded protein response in recombinant CHO cells. Cytotechnology. 2015 Mar;67(2):237-54. doi: 10.1007/s10616-013-9678-8. Epub 2014 Feb 7. PMID: 24504562; PMCID: PMC4329310.

2. Nishimiya D. Proteins improving recombinant antibody production in mammalian cells. Appl Microbiol Biotechnol. 2014 Feb;98(3):1031-42. doi: 10.1007/s00253-013-5427-3. Epub 2013 Dec 11. PMID: 24327213.

3. Cheng JK, Alper HS. Transcriptomics-Guided Design of Synthetic Promoters for a Mammalian System. ACS Synth Biol. 2016 Dec 16;5(12):1455-1465. doi: 10.1021/acssynbio.6b00075. Epub 2016 Jun 16. PMID: 27268512.

4. Roberts ML, Katsoupi P, Tseveleki V, Taoufik E. Bioinformatically Informed Design of Synthetic Mammalian Promoters. Methods Mol Biol. 2017;1651:93-112. doi: 10.1007/978-1-4939-7223-4_8. PMID: 28801902.

5. Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984 Jun;6(6):721-41. doi: 10.1109/tpami.1984.4767596. PMID: 22499653.

6. Stormo GD. Motif discovery using expectation maximization and Gibbs' sampling. Methods Mol Biol. 2010;674:85-95. doi: 10.1007/978-1-60761-854-6_6. PMID: 20827587.

7. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, Baymuradov UK, Graham K, Litton C, Miyasato SR, Strattan JS, Jolanki O, Lee JW, Tanaka FY, Adenekan P, O'Neill E, Cherry JM. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020 Jan 8;48(D1):D882-D889. doi: 10.1093/nar/gkz1062. PMID: 31713622; PMCID: PMC7061942.

8. Garrido M, Lezaud P. Extreme Value Analysis : an Introduction. Journal de la Societe Française de Statistique, Societe Française de Statistique et Societe Mathematique de France. 2013; 154(2):66-97.

9. Otis DL, Burnham KP, White GC, Anderson DR. Statistical Inference from Capture Data on Closed Animal Populations. Wildlife Monographs. 1978; 62:3–135. http://www.jstor.org/stable/3830650.

10. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36. PMID: 7584402.

Supplementary Files

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

- supplementaryV2.docx