An Evolutionary Model of DNA Substring Distribution

Meelis Kull[1,2], Konstantin Tretyakov[1], and Jaak Vilo[1,2]

[1] Institute of Computer Science, University of Tartu,
Liivi 2, 50409 Tartu, Estonia
[2] Quretec Ltd. Ülikooli 6a, 51003 Tartu, Estonia
{meelis.kull,konstantin.tretjakov,jaak.vilo}@ut.ee
http://biit.cs.ut.ee/

Abstract. DNA sequence analysis methods, such as motif discovery, gene detection or phylogeny reconstruction, can often provide important input for biological studies. Many of such methods require a *background model*, representing the expected distribution of short substrings in a given DNA region. Most current techniques for modeling this distribution disregard the evolutionary processes underlying DNA formation. We propose a novel approach for modeling DNA k-mer distribution that is capable of taking the notions of evolution and natural selection into account. We derive a computionally tractable approximation for estimating k-mer probabilities at genetic equilibrium, given a description of evolutionary processes in terms of fitness and mutation probabilities. We assess the goodness of this approximation via numerical experiments. Besides providing a generative model for DNA sequences, our method has further applications in motif discovery.

1 Introduction

From the very early days of bioinformatics, the computational analysis of DNA sequences has been one of its primary focuses. Genomic sequence is believed to literally define most of the key aspects of each organism's life and development. However, the sheer size of genomic data makes a manual human analysis practically impossible.

The information in the DNA can be viewed and analyzed at various levels of abstraction, from the large modules such as chromosomes and genes corresponding to perceivable phenotypic traits, down to short codes or motifs guiding the low-level chemical processes. In this work we turn our attention to the latter, and address the problem of modeling the distribution of short substrings (*i.e.* k-mers) in the genomic regulatory regions.

Understanding and modeling the distribution of short substrings is often the key element in the analysis of DNA regulatory regions, because it provides a concise description of the most relevant "regulatory codes" exploited by the organism. For example, some short substrings are generally known to directly induce the expression of the nearby genes. Others act as repressors or indirect

T. Elomaa et al. (Eds.): Ukkonen Festschrift, LNCS 6060, pp. 147–157, 2010.

switches [1]. Therefore, the problem of modeling the low-level substring distribution is an important step for further analysis, such as motif discovery and phylogenetics. This problem is most commonly referred to as *background modeling* and so far quite often overlooked in favor of rather simplistic approaches limited to 1-mer (*i.e.* single-nucleotide) frequencies only.

So far most of the probabilistic k-mer models have been based on either purely phenomenological ideas (*i.e.* HMMs), or loosely related to chemical binding energy models (*i.e.* PWMs) [2]. Despite the undoubted practical usefulness, the abovementioned models are, however, incapable of incorporating the question of *how could such a distribution arise* in the regulatory region in the first place.

In this work we present a novel modeling framework for relating the process of evolution and natural selection to the k-mer distribution expected to arise in the corresponding population as a result of this process. More precisely, we examine the situation where the *fitness*, *i.e.* the expected number of offspring of the individual with a given regulatory sequence is related to the features present in the sequence, such as the *counts of certain substrings*. We then derive a computationally tractable way of inferring the expected k-mer distribution for the given fitness function and *sequence mutation rates*.

The corresponding model is reasonably general and can be used to incorporate more complex evolutionary assumptions into various DNA analysis methods. The reasons for including these assumptions are twofold. Firstly, introducing a strong inductive bias into low-level models (background) can result in better precision of the higher-level pattern analysis and motif discovery algorithms [3]. Secondly, the background and higher-level patterns can be handled by a single model as both are products of evolutionary processes. By matching the model to data it is possible to learn something about these processes. Mustonen and Lässig provide a cross-species model with transcription factor binding sites under selection and background in neutral evolution [4]. Similar methods have been used for analysing binding site turnover [5,6,7]. Our approach uses a single species but allows for more complicated evolutionary models by defining the fitness and mutation functions.

The incorporation of evolution, even in its most simple form, can lead to computationally expensive procedures. We propose simplifications, which make the computations tractable and yet still provide a close approximation when tested numerically.

We believe our model provides a novel view on the problem of modeling DNA substring distribution and has a potential for further development and applications.

2 Methods

2.1 Evolutionary Model of DNA Regulatory Regions

It is known that the formation of DNA sequence is mainly driven by evolution. Genomic sequence mutates from generation to generation, and the less successful variants tend to stage out in favor of the more successful ones. We consider the

influence of this process on a single regulatory region, *e.g.* a promoter of some gene. We assume that the *fitness* of a given region (*i.e.,* its expected number of offspring) is largely determined by a number of certain functional elements in the region. In this case the evolutionary process will necessarily impose some nontrivial k-mer distribution on the corresponding DNA region in the whole population. Our goal here is to compute this distribution from the information about the important features and their influence on promoter's fitness, taking into account the sequence mutation rates.

Formally, let us fix a promoter region of length n. Suppose we know that the expected average number of offspring for an individual with sequence s in this promoter region is f_s, for all $s \in A^n$, where $A = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. We shall refer to f as the *fitness function*. Suppose we also know the probability $m_{t\to s}$ of sequence t at this region mutating into a sequence s within one generation. Let the expected proportion of individuals with promoter sequence s in a population be p_s, for all $s \in A^n$. We say that this population is in genetic equilibrium, if the expected proportion of individuals with sequence s in the offspring population p'_s is equal to p_s. Note that we assume reproduction to be performed before mutation. All of the following results could also be proven for the opposite order, yet the equilibrium probabilities would be different.

2.2 The Equilibrium Distribution

We shall now prove that the equilibrium of the promoter sequence distribution exists and is uniquely determined under very general assumptions. To do that we first derive a formula to calculate p' from p. For a population of size i, the expected number of individuals with sequence t is $i \cdot p_t$. The expected number of offspring for these individuals is $i \cdot p_t \cdot f_t$. As any sequence t can mutate into sequence s with probability $m_{t\to s}$, the expected number of offspring with sequence s is $\sum\limits_{t\in A^n} ip_t f_t m_{t\to s}$. In order to get the expected proportion of sequence s in the offspring population, we have to divide by the size of the new population:

$$p'_s = \frac{\sum\limits_{t\in A^n} ip_t f_t m_{t\to s}}{\sum\limits_{u\in A^n}\sum\limits_{t\in A^n} ip_t f_t m_{t\to u}} = \frac{i\sum\limits_{t\in A^n} p_t f_t m_{t\to s}}{i\sum\limits_{t\in A^n} p_t f_t \sum\limits_{u\in A^n} m_{t\to u}} = \frac{\sum\limits_{t\in A^n} p_t f_t m_{t\to s}}{\sum\limits_{t\in A^n} p_t f_t},$$

where $\sum\limits_{u\in A^n} m_{t\to u} = 1$ because it is the probability of t mutating into any other sequence, including itself. We can now prove the following theorem.

Theorem 1. *Let $n \in \mathbb{N}$, $f_s > 0$ for all $s \in A^n$, and $m_{t\to s} > 0$ for all $s, t \in A^n$. Then there exists a unique probability distribution $p^{\mathsf{EQ}} = (p_s^{\mathsf{EQ}})_{s\in A^n}$ with $p_s^{\mathsf{EQ}} > 0$ for all $s \in A^n$, and $\sum\limits_{u\in A^n} p_u^{\mathsf{EQ}} = 1$, such that the equilibrium condition holds:*

$$p_s^{\mathsf{EQ}} = \frac{\sum\limits_{t\in A^n} p_t^{\mathsf{EQ}} f_t m_{t\to s}}{\sum\limits_{t\in A^n} p_t^{\mathsf{EQ}} f_t}, \tag{1}$$

Proof. Let us convert the formula (1) to a matrix form. For that we define an $|A^n| \times |A^n|$ matrix $E = (e_{st})_{s,t \in A^n}$ with $e_{st} = f_t \cdot m_{t \to s}$. Now (1) is equivalent to

$$\lambda p^{\mathsf{EQ}} = E p^{\mathsf{EQ}}, \tag{2}$$

where

$$\lambda = \sum_{t \in A^n} p_t^{\mathsf{EQ}} f_t. \tag{3}$$

It remains to prove that the matrix E has a unique eigenvector p^{EQ} with positive components, having the sum of components equal to 1, and the corresponding eigenvalue λ satisfies the constraint (3).

As all the elements of matrix E are positive, we can apply the Perron-Frobenius theorem, stating that real matrices with positive entries have a unique largest real eigenvalue and that the corresponding eigenvector has strictly positive components. Furthermore, it states that this eigenvector is the only eigenvector with strictly positive components. After scaling this eigenvector so that its components would sum up to 1, we have obtained the required p^{EQ}. It remains to prove that the corresponding eigenvalue λ satisfies (3). This can be shown by summing up the components of vectors on both sides of the equation (2). On the left we get λ as the components of p^{EQ} sum up to 1. On the right we get the required expression:

$$\sum_{s \in A^n} \sum_{t \in A^n} e_{st} p_t^{\mathsf{EQ}} = \sum_{s \in A^n} \sum_{t \in A^n} p_t^{\mathsf{EQ}} f_t m_{t \to s} = \sum_{t \in A^n} p_t^{\mathsf{EQ}} f_t \sum_{s \in A^n} m_{t \to s} = \sum_{t \in A^n} p_t^{\mathsf{EQ}} f_t,$$

because $\sum_{s \in A^n} m_{t \to s} = 1$. □

The following theorem expresses p^{EQ} in terms of the fitness function and mutation probabilities.

Theorem 2. *Let $n \in \mathbb{N}$, $f_s > 0$, $m_{s \to t} > 0$, $p_s^{(0)} \geq 0$ for all $s, t \in A^n$, and $\sum_{u \in A^n} p_u^{(0)} = 1$. Further, let $p_s^{(i)}$ be defined for each $i \in \mathbb{N}$ and $s \in A^n$ as follows:*

$$p_s^{(i)} = \frac{\sum_{t \in A^n} p_t^{(i-1)} f_t m_{t \to s}}{\sum_{t \in A^n} p_t^{(i-1)} f_t}. \tag{4}$$

Then the limit $p^{\mathsf{EQ}} = \lim_{i \to \infty} p^{(i)}$ exists and satisfies the equilibrium condition (1).

Proof. As in the proof of Theorem 1 we represent the equilibrium problem as the eigenvector problem for matrix E. We have to prove that our iterative process converges to the only positive eigenvector of E, which gives us the equilibrium distribution according to Theorem 1. Since by Perron-Frobenius theorem the positive eigenvector corresponds to the largest eigenvalue, we can apply the

power iteration method to find vector p^{EQ}. At step i the original power iteration method divides the vector $Ep^{(i)}$ by its length, whereas in our iterative definition (4) we divide it by the sum of its components to obtain a probability distribution. Both methods are just different normalizations, and reach the same equilibrium eigenvector, up to a multiplicative constant. Thus, the limit distribution p^{EQ} exists and satisfies the equilibrium condition (1). □

2.3 Substring Distribution at Equilibrium

For studying the substring distribution we first introduce some notation. For two strings a and b we denote their concatenation by $a \cdot b$. For any sequence s let s_j^k denote its substring of length k starting at location j. We regard sequences as cyclic, that is, the substring that reaches the end of the sequence wraps to continue from the beginning. Further we define the shift operator "$\gg$", such that $s \gg i$ denotes the sequence obtained from s by removing the last i nucleotides and inserting these at the beginning.

In order to express the substring distribution in a usable form, we need to make assumptions about the fitness function f and mutation probabilities m. Namely, for f we assume shift invariance, that is $f_s = f_{s \gg i}$ for all $s \in A^n$ and $i \in \mathbb{N}$. This holds, for example, if fitness is measured by the number of occurences of some substring in the sequence. For m we assume that $m_{a \cdot s \to b \cdot t} = m_{a \to b} \cdot m_{s \to t}$ for all $1 \leq k \leq n$, $a, b \in A^k$ and $s, t \in A^{n-k}$. In other words, we assume that mutations at different parts of the sequence are independent. From this it follows that m is also shift invariant. As the proportions in the equilibrium population are computed directly from the fitness function and mutation probabilities, these must also be shift invariant, that is, $p_s^{\mathsf{EQ}} = p_{s \gg i}^{\mathsf{EQ}}$.

Suppose we now pick a substring of length k from a random location in the sequence of a random individual from the equilibrium population. The probability to get substring a can be calculated as follows:

$$\Pr(a) = \sum_{s \in A^n} p_s^{\mathsf{EQ}} \sum_{j=1}^{n} \frac{1}{n} \cdot [s_j^k = a] = \sum_{s \in A^n} \frac{1}{n} \sum_{j=1}^{n} p_s^{\mathsf{EQ}} \cdot [s_j^k = a],$$

where $[s_j^k = a]$ is defined as 1 if $s_j^k = a$ and 0 otherwise. Note that $s_j^k = a$ if and only if $s = (a \cdot t) \gg j$ for some $t \in A^{n-k}$. The above equality can now be rewritten:

$$\Pr(a) = \sum_{t \in A^{n-k}} \frac{1}{n} \sum_{j=1}^{n} p_{(a \cdot t) \gg j}^{\mathsf{EQ}} = \sum_{t \in A^{n-k}} \frac{1}{n} \sum_{j=1}^{n} p_{a \cdot t}^{\mathsf{EQ}} = \sum_{t \in A^{n-k}} p_{a \cdot t}^{\mathsf{EQ}}.$$

In other words, the probability of k-mer a in the equilibrium substring distribution is equal to the total proportion of all sequences starting with a. Due to this fact we introduce the following notation:

$$p_a^{\mathsf{EQ}} := \Pr(a) = \sum_{t \in A^{n-k}} p_{a \cdot t}^{\mathsf{EQ}}. \tag{5}$$

Computing the equilibrium substring distribution directly from (4) and (5) is intractable as the time complexity is exponential in n. Therefore, we propose an alternative method for estimating this distribution. For each k-mer a, we can write:

$$p_a^{\mathsf{EQ}} \overset{(5)}{=} \sum_{s\in A^{n-k}} p_{a\cdot s}^{\mathsf{EQ}} \overset{(1)}{=} \sum_{s\in A^{n-k}} \frac{\sum_{t\in A^n} p_t^{\mathsf{EQ}} f_t m_{t\to a\cdot s}}{\sum_{t\in A^n} p_t^{\mathsf{EQ}} f_t} \overset{(t=b\cdot u)}{=} \frac{\sum_{s\in A^{n-k}} \sum_{b\in A^k} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u} m_{b\cdot u\to a\cdot s}}{\sum_{b\in A^k} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u}} =$$
$$= \frac{\sum_{s\in A^{n-k}} \sum_{b\in A^k} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u} m_{b\to a} m_{u\to s}}{\sum_{b\in A^k} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u}} = \frac{\sum_{b\in A^k} m_{b\to a} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u} \sum_{s\in A^{n-k}} m_{u\to s}}{\sum_{b\in A^k} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u}} =$$
$$= \frac{\sum_{b\in A^k} m_{b\to a} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u}}{\sum_{b\in A^k} \sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u}}.$$

We now approximate

$$\sum_{u\in A^{n-k}} p_{b\cdot u}^{\mathsf{EQ}} f_{b\cdot u} \approx \frac{p_b^{\mathsf{EQ}}}{|A^{n-k}|} \sum_{u\in A^{n-k}} f_{b\cdot u}\,, \tag{6}$$

which essentially means replacing all terms $p_{b\cdot u}^{\mathsf{EQ}}$ within the sum by their average over $u \in A^{n-k}$. Although not strongly supported by theoretical considerations, the numerical experiments indicate that this approximation is quite good. We further denote $f_b := \frac{1}{|A^{n-k}|} \sum_{u\in A^{n-k}} f_{b\cdot u}$, as this is the average fitness of all sequences starting with substring b. Note that this does not conflict with the notation of the original fitness function, as for the full sequence s the sum on the right has only one element, f_s itself. The approximation (6) can now we be written down as follows:

$$p_a^{\mathsf{EQ}} \approx \frac{\sum_{b\in A^k} m_{b\to a} p_b^{\mathsf{EQ}} f_b}{\sum_{b\in A^k} p_b^{\mathsf{EQ}} f_b}.$$

which matches exactly the original equilibrium condition (1), yet now it is for substrings of length k. According to Theorem 1 there exists a unique distribution p^{EQ_k}, which for any k-mer a satisfies the following condition:

$$p_a^{\mathsf{EQ}_k} = \frac{\sum_{b\in A^k} p_b^{\mathsf{EQ}_k} \cdot f_b m_{b\to a}}{\sum_{b\in A^k} p_b^{\mathsf{EQ}_k} f_b}.$$

It is therefore natural to use $p_a^{\mathsf{EQ}_k}$ as an approximation to p_a^{EQ}.

The practical calculation of p^{EQ_k} can now be performed in two steps:

- estimating $f_b = \frac{1}{|A^{n-k}|} \sum_{u \in A^{n-k}} f_{b \cdot u}$ for all $b \in A^k$ by random sampling over u;
- finding p^{EQ_k} using Theorem 2, *i.e.* using power iteration on a $|A|^k \times |A|^k$ matrix.

For the case of DNA sequences ($|A| = 4$) this calculation is realistic up to $k = 8$ or so.

3 Experiments

In order to check the applicability of our approximations we have performed experiments to compare the distributions p^{EQ} and p^{EQ_k}. As calculating the exact distribution p^{EQ} is computationally very demanding, we restricted ourselves to the alphabet of size 2, *i.e.* $A = \{\mathsf{A}, \mathsf{C}\}$, and to the promoters of length $n = 8$. At each site independently, the probability of a mutation from A to C or from C to A was r, we tested values $r = 10^{-0.4}, 10^{-0.6}, 10^{-0.8}, \ldots 10^{-3.0}$. As explained later, smaller values of r lead to very similar results due to convergence of the distribution. Because of independent point-mutations the probability of a sequence s mutating into sequence t was $m_{s \to t} = r^{\Delta(s,t)}(1-r)^{n-\Delta(s,t)}$, where $\Delta(s,t)$ is the number of positions where s and t have a different nucleotide. The fitness of a sequence was dependent on the number of times a certain substring q occurred in the sequence, we tested $q = \mathsf{AC}, \mathsf{AAC}, \mathsf{AACA}$. Altogether we used nine different measures f as for each of the substrings q we tested the following three strategies:

(S1) sequences with i occurrences of q had fitness $i + 1$;
(S2) sequences with 0 or 1 occurrences of q had fitness 1, others had fitness 2;
(S3) sequences with 0 occurrences of q had fitness 2, others had fitness 1.

For each mutation rate r (14 values), each fitness function f (9 values), and each $k = 1, 2, \ldots, 6$ we found the k-mer distribution for the exact equilibrium p^{EQ} and the approximation p^{EQ_k}, altogether $14 \cdot 9 \cdot 6 = 756$ pairs of distributions.

To evaluate the approximated distribution we found the Pearson correlation coefficient with the exact distribution as well as the Kullback-Leibler divergence per position (KLdpp) defined as follows:

$$\mathsf{KLdpp} = \frac{D_{\mathsf{KL}}(p^{\mathsf{EQ}} || p^{\mathsf{EQ}_k})}{k} = \frac{1}{k} \sum_{a \in A^k} p_a^{\mathsf{EQ}} \log \frac{p_a^{\mathsf{EQ}}}{p_a^{\mathsf{EQ}_k}}.$$

The results did not show significant dependence of approximation error on the choice of the fitness function and value k. The precision of approximation was mainly dependent on the mutation probability r. Figure 1 plots the correlation and KLdpp for all experiments for different values of r.

To get some idea about how the exact and approximate distributions change with r we plotted the distributions for $q = \mathsf{AAC}$, fitness strategy (S2), $n = 8$,

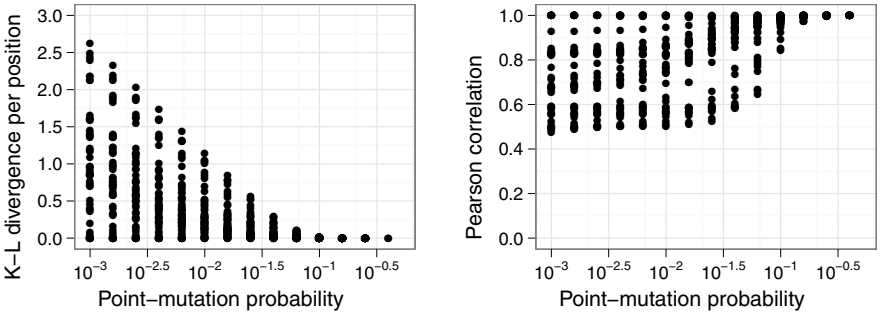

Fig. 1. The effect of point-mutation rate r on the approximation quality measured as Kullback-Leibler divergence per position and correlation between the exact and approximated distributions. Each circle denotes an experiment with a different set of parameters.

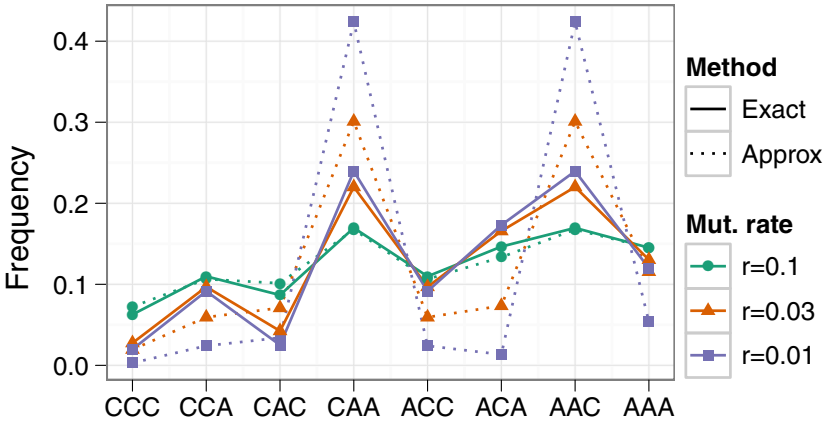

Fig. 2. The exact and approximate 3-mer distributions for point-mutation rates $r = 0.1, 0.03, 0.01$ where fitness is defined with strategy (S2) for substring $q = \mathsf{AAC}$

$k = 3$, and $r = 0.1, 0.03, 0.01$ (see Figure 2). For mutation rate $r = 0.1$ the approximation is almost perfect and is gradually becoming worse with decreasing r. Still, the correlation between the exact and approximate distribution remains quite high, which is confirmed in Figure 1 where all correlations are above 0.4. High Kullback-Leibler divergence for small values of r is apparently caused by the substrings with moderate true frequency but very low approximated frequency, such as ACA in Figure 2.

Figure 2 also illustrates the convergence of the exact distributon with decreasing r, as the distributions of $r = 0.03$ and $r = 0.01$ are highly similar. The same holds for the approximated distributions, explaining why approximation errors for $r = 10^{-2.8}$ and $r = 10^{-3}$ have extremely similar patterns in Figure 1.

4 Discussion

In this work we presented a novel approach to modeling DNA substring distribution that is based on an evolutionary model. We have derived a computationally tractable approximation for estimating the k-mer distribution from the description of the process, and verified that the approximation is fairly precise. This allows to use the model in generative settings as well as for significance computations in motif discovery procedures.

The major merit of the approach lies in the fact that it provides a sound way of incorporating evolutionary assumptions and prior knowledge into further analyses. Introduction of such *inductive bias* opens up novel possibilities of application for sequence analysis algorithms. To be more precise, consider the case of motif discovery from promoter sequences. This task is often solved by searching given DNA sequence data for short *significantly overrepresented* substrings [8]. Significance here denotes a measure of deviation of observed substring frequencies from a certain *null-model* – a presumed distribution of substrings, which could be explained using prior knowledge only. For example, if we presume that a given DNA region is inherently rich in `CG`-pairs, we shall not be surprised to find that a substring `CGCGCG` is frequent. On the other hand, detecting a similarly frequent substring `ATATAT` might be interpreted as a presence of something, which cannot be explained using previous knowledge only, *i.e.* an overrepresented motif.

Many contemporary motif discovery methods are rather unsophisticated in the way of modeling prior knowledge, using just the single- or di-nucleotide distribution for their *background model* [9]. This may result in spurious discoveries, such as detecting multiple versions of a single motif or just some generic sequence features. Other methods use the set of *background sequences* [10,11], or a higher-order HMM [12,3] to model prior knowledge. The drawback of this approach is that it requires many sequences "of the same kind" to estimate the model. Yet it is often not clear which regulatory sequences may be modeled as being from the same kind. Therefore further assumptions, such as co-regulation or co-expression must be made. In our method we essentially provide a set of *evolutionary* assumptions, which may be used instead. Given these assumptions only, a background model of k-mer distribution can be computed, incorporating the information about which sequence features are already known to be significant.

Another natural way of regarding our method is just as a purely generative model for DNA sequences. Indeed, the marginal k-mer distribution can be straightforwardly extended to a generator of arbitrary-length substrings satisfying this distribution [3]. The need for such generators of "random" DNA sequences arises often in connection with testing and analysis of various algorithms, and a number of tools have already been developed for this purpose. Some of these proceed by simulating evolution [13], some focus on simulating alignments [14], and yet others propose ways of planting randomized motifs [15]. Our model can account for motifs and evolutionary aspects simultanously through the fitness and mutation functions. For modelling gene promoter regions our model can in principle aggregate such information as the transcription factor binding motifs and their combinations [1], nucleosome positioning code [16] and CpG mutation

rates [17]. While defining the fitness function is mostly a biological endeavor, the mutation function can require more mathematical effort, as exemplified in the Experiments section.

The open problem which yet remains to be solved is the question of efficient estimation of model parameters from data, as this could open new possibilities and application areas both in sequence analysis as well as the study of DNA evolution. Or in other words, what information can we extract from the substring distributions, assuming genetic equilibrium? Another interesting question is related to the possibility of improving the precision of the approximation (6), especially for smaller mutation rates, perhaps by incorporating higher-order terms, and yet still keeping the computations tractable.

References

1. Davidson, E.H.: The regulatory genome: gene regulatory networks in development and evolution. Academic Press, San Diego (2006)
2. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics 16(1), 16–23 (2000)
3. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., Moor, B.D., Rouzé, P., Moreau, Y.: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17(12), 1113–1122 (2001)
4. Mustonen, V., Lässig, M.: Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc. Natl. Acad. Sci. USA 102(44), 15936–15941 (2005)
5. Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.Y., Biggin, M.D., Eisen, M.B.: Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput. Biol. 2(10), e130 (2006)
6. Doniger, S.W., Fay, J.C.: Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol. 3(5), e99 (2007)
7. Huang, W., Nevins, J.R., Ohler, U.: Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. Genome. Biol. 8(10), R225 (2007)
8. Brazma, A., Jonassen, I., Vilo, J., Ukkonen, E.: Predicting gene regulatory elements in silico on a genomic scale. Genome. Res. 8(11), 1202–1215 (1998)
9. Das, M.K., Dai, H.K.: A survey of DNA motif finding algorithms. BMC Bioinformatics 8(Suppl. 7), S21 (2007)
10. Redhead, E., Bailey, T.: Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics 8(1), 385 (2007)
11. Vilo, J.: Pattern discovery from biosequences. Thesis PhD (2002)
12. Wang, G., Yu, T., Zhang, W.: WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. Nucleic Acids Res. 33(Web Server issue), W412–W416 (2005)
13. Cartwright, R.A.: DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 21(Suppl. 3), iii31–iii38 (2005)
14. Varadarajan, A., Bradley, R., Holmes, I.: Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. Genome. Biol. 9(10), R147 (2008)

15. Rouchka, E.C., Hardin, C.T.: rMotifGen: random motif generator for DNA and protein sequences. BMC Bioinformatics 8, 292 (2007)
16. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.P.Z., Widom, J.: A genomic code for nucleosome positioning. Nature 442(7104), 772–778 (2006)
17. Saxonov, S., Berg, P., Brutlag, D.L.: A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. USA 103(5), 1412–1417 (2006)