DNA sequence modeling based on context trees

DNA sequence modeling based on context trees

Lieneke Kusters Tanya Ignatenko
Eindhoven University of Technology
Dept. of Electrical Engineering, SPS group
Eindhoven, The Netherlands
c.j.kusters@tue.nl t.ignatenko@tue.nl

Abstract

Genomic sequences contain instructions for protein and cell production. Therefore understanding and identification of biologically and functionally meaningful patterns in DNA sequences is of paramount importance. Modeling of DNA sequences in its turn can help to better understand and identify such patterns and dependencies between them. It is well-known that genomic data contains various regions with distinct functionality and thus also statistical properties. In this work we focus on modeling of such individual regions of distinct functionalities. We apply the concept of context trees to model these DNA regions. Based on the Minimum Description Length principle, we use the estimated compression rate of a genomic region, given such models, as a similarity measure. We show that the constructed model can be used to distinguish specific genes within DNA sequences.

1 Introduction

The human genome contains information about human evolution and physiological properties. The genetic research community put a lot of effort in projects like the human genome project, the 1000 genomes project and the HapMap project, in order to collect, analyze and understand the human genome. These efforts resulted in the human reference genome sequence (that is a general representation of the human genome) and many new insights regarding population evolution, functional properties of the genome, as well as genetically inherited diseases and disease predispositions, and their treatment.

It is known that certain regions of the genome encode for proteins. In these regions, triplets of nucleotides (codons) encode for the amino-acids that together construct a protein of specific shape. Research on automatic detection of protein-coding regions in the genome, includes spectral analysis techniques [1],[2],[3] and Markov models [4],[5]. However, besides the protein coding regions, there are also regions in the genome with other functionalities, such as e.g. regulatory elements (control transcription of a nearby gene). To the best of our knowledge, there exists no general model that can be used to automatically identify and distinguish between various regions of different functionalities within genomic sequences.

It is our goal to construct a generic statistical model for genetic sequences. Since various regions in the genome have different functionality, their statistical properties also differ within the genome. Therefore, as a first step in constructing a generic model, we focus on determining individual models corresponding to the different functional regions in the genome. In [6] it was shown that context trees can be used to model and distinguish between the human chromosomes. We propose to use a similar approach, and construct models that can help discriminate between smaller regions of different functionality.We propose to use context trees [7] to model genetic sequences. We show that the context tree model can be used to distinguish regions of similar statistics within a sequence.

This paper is organized as follows. In the next section we first explain the proposed methods for constructing and evaluating the model. In Section 3 we present our experimental results for modeling of different types of sequences. Finally, we discuss our findings and future work in Section 4.

2 Methodology

In this work we propose to use a two-pass method, to construct the model that we can use for DNA modeling. With the two-pass method, we first construct the maximum a posteriori model corresponding to a given sequence, and then apply the constructed model to estimate the compression rate of a sequence given the model. We use the resulting compression rate as criteria to make a decision whether the sequence was generated by the given model, and thus is functionally similar to the sequence(s) used to estimate this model. In the following, we first summarize the properties of the DNA data. Next, in Sections 2.2 and 2.3 we introduce the context tree model and describe the two-pass method that we use to construct the maximum a posteriori tree model of a sequence. Finally, we describe the application of this two-pass method to DNA modeling.

2.1 DNA Sequences

Human genetic information is encoded in deoxyribonucleic acid (DNA) sequences. The DNA sequence is composed of four different symbols that correspond to the DNA building blocks, called nucleobases, i.e. Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). DNA sequences vary across populations and generations. These variants occur due to mutations and generally occur once per thousand nucleotides in the sequence. Typical genetic variations include *substitution* of one nucleotide for another and *insertion* or *deletion* of a short subsequence of nucleotides.

2.2 Context Tree Model for DNA sequences

A DNA sequence is a string of concatenated quaternary symbols, where each symbol can take on a value from a quaternary alphabet $(A, C, G, T) \in \mathcal{A}$, corresponding to the four different nucleobases. Let a DNA sequence $x_0 x_1 x_2 \dots x_{N-1}$ of length N be denoted by x_1^N. We assume that the DNA sequence is generated by a tree source. For a tree source the probability $\Pr\{X_t = a\}$ of a symbol X_t in the sequence to take on a value $a \in \mathcal{A}$, is determined by its context, where the context is defined by at most D preceding symbols in the sequence. Such a tree source can be described by a context tree. A context tree is a set of nodes labeled with contexts s with $0 \leq len(s) < D$, and a set of leafs that correspond to the contexts of maximum depth, $len(s) = D$, with $len(s)$ the length of the context s. An example context tree is shown in Figure 1. Given such a context tree, we can determine the probability $\Pr\{X_t = a | x_{t-D}^{t-1}\}$, by starting at the root λ of the tree and moving along the nodes $x_{t-1}, x_{t-2}, \dots$ until a leaf of the tree is reached. In this leaf, s, we find the corresponding parameter $\Theta_s = \left\{\theta_s^A, \theta_s^C, \theta_s^G, \theta_s^T\right\}$, that are the conditional probabilities of a symbol to take a certain value from the alphabet $\mathcal{A}$, given its context s. Therefore, using the context tree with parameters, we can find the conditional probability of our symbol as $\Pr\{X_t = a | x_{t-D}^{t-1}\} = \theta_s^a$. The suffix set, that represents the leafs of the tree, $\mathbf{S}$ is called the model of the source and the corresponding parameters are stored in its leafs and denoted by $\mathbf{\Theta}$. Furthermore, we define the mapping from the context of depth D, to a suffix s in the model $\mathbf{S}$ as $\omega^{\mathbf{S}}(x_{t-D}^{t-1}) = s \in \mathbf{S}$.

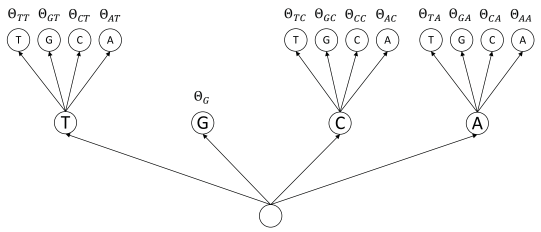


Figure 1: Context Tree $\mathbf{S}$ and model parameters $\mathbf{\Theta}$.

Now, given the example tree model with parameters, in Figure 1, the probability of a subsequence $ACGTC$, in $x_1^N = (\dots CAA\mathbf{ACGTC}GG\dots)$, can be estimated as follows:

$$\Pr\{ACGTC|\mathbf{S}, \mathbf{\Theta}\} = \theta_{AA}^{A}\theta_{AA}^{C}\theta_{AC}^{G}\theta_{G}^{T}\theta_{GT}^{C}. \tag{1}$$

In general we do not know the actual source model that corresponds to the DNA sequence. In [7] the context tree weighting algorithm (CTW) is proposed to estimate the unknown sequence distribution. In CTW they estimate a good coding distribution that can be used to compress data in the sequential way. Instead, we want to find the model that best describes the sequence and evaluate its performance. This can be achieved by the CTW two-pass method [8], which uses the techniques to determine the maximum a posteriori (MAP) model after observing the complete sequence. In [8] this MAP model is first estimated and then used for compression of the sequence.
Here we propose to use this two-pass method to first estimate an optimized statistical model corresponding to a given training sequence. Then we evaluate the model performance, based on the compression rate of a sequence, given this model.

2.3 Maximum a posteriori (MAP) model selection

In this section we summarize the algorithm that is used for the MAP model selection. Our implementation is based on the two-pass method as proposed by Willems *et al.* in [8]. First, we construct a context tree where we assume a maximum depth D. Then, we process the training sequence sequentially, and estimate the probabilities of the subsequences that correspond to each of the contexts s in the tree. Finally, we evaluate the estimated sequence probabilities at different nodes in the tree to find the MAP model, selecting the nodes as either leafs or nodes based on the estimated probabilities.
First of all, for each context that corresponds to a node in the tree, we count the symbols that occur with this context in the training sequence. We store the counts c_s^a that correspond to symbol $a \in \mathcal{A}$, occurring with context $s \in \mathbf{S}$ in the corresponding node. We use the KT-estimator from [9] to estimate the probability of the subsequence P_e^s with context s, given the counts in the corresponding node, as follows:

$$P_e^s\left(c_s^A, c_s^C, c_s^G, c_s^T\right) = \frac{\prod_{a\in\mathcal{A}}\prod_{t=1}^{c_s^a}(t-1/2)}{\prod_{k=0}^{T_s-1}(k+|\mathcal{A}|/2)}, \tag{2}$$

with $T_s = \sum_{a\in\mathcal{A}} c_s^a$, the length of the subsequence with context s. Similarly, the probability of a single symbol X_t to have value $a \in \mathcal{A}$ can be estimated, given its context s and the counts in the corresponding node of the tree, as follows:

$$\Pr\{X_t = a|c_s^A, c_s^C, c_s^G, c_s^T\} = \frac{c_s^a + 1/2}{T_s + |\mathcal{A}|/2}. \tag{3}$$

In the next step, we use the method proposed in [8] to find the maximum a posteriori tree model. That is, we estimate for each node the maximum a posteriori probability of the corresponding subsequence,

$$P_m^s = \begin{cases} \max\left(\alpha \cdot P_e^s, (1-\alpha) \cdot \prod_{a \in \mathcal{A}} P_m^{as}\right) & \text{if depth}(s) < D, \\ P_e^s & \text{otherwise.} \end{cases} \tag{4}$$

where $\alpha = \frac{|\mathcal{A}|-1}{|\mathcal{A}|}$, is a penalty for the model complexity. Note that the penalty is increasing with the depth of the tree, see also [9]. We find the nodes corresponding to the MAP model, by tracking the above maximization procedure, starting from the root. If in a node s in the context tree $\alpha \cdot P_e^s \geq (1-\alpha) \cdot \prod_{a \in \mathcal{A}} P_m^{as}$, this node is a leaf in the MAP model and the corresponding context s is added to the model $\mathbf{S}$. Otherwise this node is an internal node in the MAP model and we continue to evaluate the children that are deeper in the tree: $\{As, Cs, Gs, Ts\}$. In this way we find all the leafs corresponding to the MAP model $\mathbf{S}$. Finally, we can compute the parameters $\mathbf{\Theta}$ of our model using equation 3.

2.4 DNA sequence model evaluation

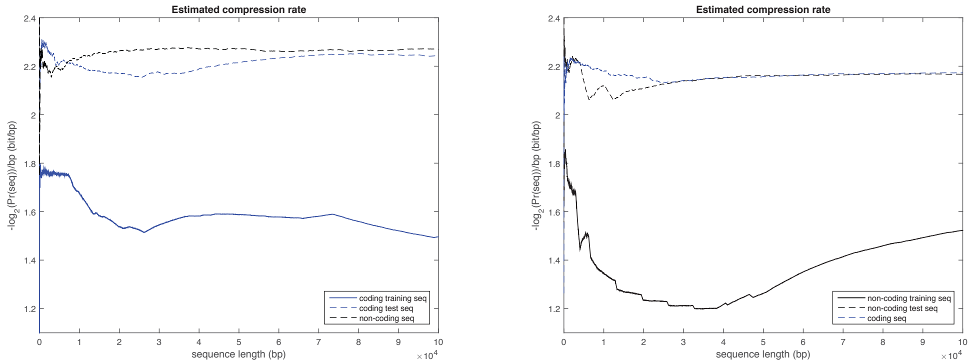
As explained in Section 2.2, the CTW two-pass algorithm for MAP model approximation, was originally developed for compression of the corresponding sequence. However, we would like to apply this model to evaluate or to detect sequences of similar functionality. The Minimal Description Length principle [10], states that the model that describes the data in the shortest possible way is the model that produced the data. Therefore, we use the estimated compression rate of a sequence, given the model, as a measure of the correctness of the model. We can estimate the compression rate, by using the constructed model $\mathbf{S}$ and corresponding probabilities $\mathbf{\Theta}$, to estimate the probability of the sequence given the model. We have shown in Section 2.2 how to estimate the probability of a sequence (or a single symbol) given the model.
We estimate the compression rate of the sequence x_1^t as follows:

$$R(x_1^t) = -\sum_{j=1}^{t} \log_2(\theta^{a_j}_{\omega^{\mathbf{S}}(x_{j-D}^{j-1})})/t, \tag{5}$$

with $\omega^{\mathbf{S}}(x_{j-D}^{j-1}) \in \mathbf{S}$ the context of X_j (a symbol in the sequence) that corresponds with a leaf in the model $\mathbf{S}$, and $\theta^{a_j}_{\omega^{\mathbf{S}}(x_{j-D}^{j-1})}$ the corresponding probability of the symbol X_j having its corresponding value a_j. Furthermore, we can also estimate the contribution of a symbol X_t to the compression rate, as

$$R(X_t) = -\log_2(\Pr\{X_t = a | x_{t-1} \ldots x_{t-D}\}) = -\log_2(\theta^{a}_{\omega^{\mathbf{S}}(x_{t-D}^{t-1})}), \tag{6}$$

with $\omega^{\mathbf{S}}(x_{t-D}^{t-1}) \in \mathbf{S}$ is the mapping of the context of X_t to a leaf in the model $\mathbf{S}$, and $\theta^{a}_{\omega^{\mathbf{S}}(x_{t-D}^{t-1})}$ the corresponding probability of the symbol $X_t = a$. Finally, we note that the compression rate is measured in bits per base-pair, which means that for our data, a compression rate smaller than 2, i.e. $\log_2(4)$, corresponds to actual compression of the sequence.



(a) MAP model of coding sequences.

(b) MAP model of non-coding sequences.

Figure 2: Achievable compression rates for coding and non-coding sequences, using MAP context tree model. Two models were trained, one on coding (2a) and one on non-coding (2a) DNA compound sequences. For both models the performance is shown, when applied to the sequence used for training, when applied to a sequence of similar functionality, and when applied to a sequence of the opposite functionality

3 Experimental results

We evaluate the performance of the maximum a posteriori tree model in two experiments. In each experiment we first use the techniques explained in Section 2.3 to construct the MAP model corresponding to the training sequence. Then we test the performance of the model, by estimating the resulting compression rate for various sequences.

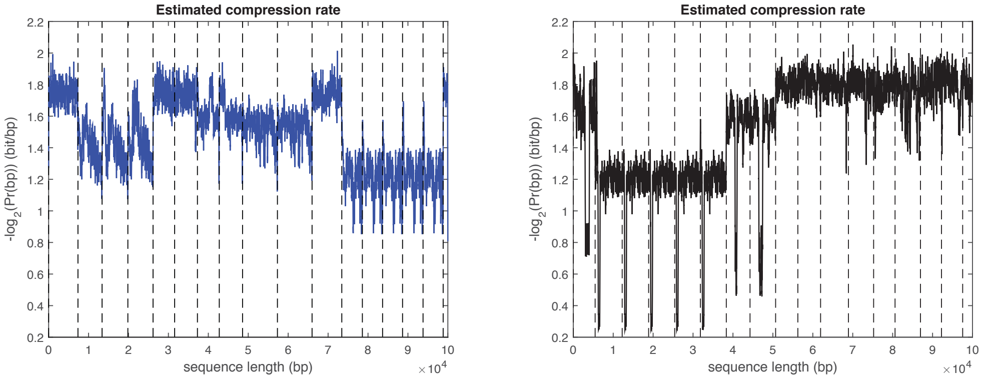
3.1 Coding and non-coding sequences

In the first experiment we construct two MAP tree models for (protein) coding and non-coding sequences respectively. We use a set of sequences from the Human Reference Genome (build: GRCh38.p2) annotated as mRNA (coding) and ncRNA (non-coding) in the NCBI Homo sapiens Annotation Release 107 [11].

First, we construct one coding and one non-coding sequence of 10^5 base-pairs long, by compounding the subsequences of corresponding functionality from the annotated set. We estimate the MAP model for each of the constructed sequences, assuming maximum tree depth 7. We estimate the compression rate of each model on both sequences, the resulting compression rates are shown in Figure 2. We can see that both models have a good compression rate on the sequence that was used for the model training, and they can be used to distinguish between the coding and non-coding training sequences. Therefore, we may conclude that the resulting model provides a good estimate of the source model of the sequences.

Now, we construct two more ('test') sequences in a similar way as before, but from other sets of the annotated ncRNA and mRNA sequences. When we apply the previously constructed coding and non-coding models to the new sequence of corresponding functionality (Figure 2), the compression rate is above 2, which means that the sequences are not compressible at all (see Section 2.4).

In Figure 3, we estimate the compression rate per symbol for the models on their corresponding training sequence. Now we can see, that the rate varies for different regions in the sequence. We conclude that, though the constructed models do have a



(a) MAP model of coding sequences.

(b) MAP model of non-coding sequences.

Figure 3: Achievable compression rates per symbol, for coding (3a) and non-coding (2a) sequences using MAP context tree model. The dotted lines mark the regions corresponding to different subsequences that were used to construct the total sequence.

sufficient overall performance to represent the entire sequence (Figure 2), the model varies for different regions in the sequences and the overall model is actually a mixture of models. Furthermore, we find that the variations in the compression rate are related to the transitions between the subsequences (marked by the dotted lines) that jointly form the total sequence.

3.2 MAP model for gene in mitochondrion

Now we concentrate on construction of the model for a sequence with a more specific functionality than just coding or non-coding functionality. In this experiment we would like to detect COX1 gene in the mitochondrial DNA and use our model to detect the gene in a set of mitochondrial DNA variant sequences.

For this experiment we have used a set of mitochondrial DNA sequences from 20 individuals of various ethnicity (America, Africa, Europe, Asia)*. Between those sequences small variations occur in the form of substitutions, insertions and deletions of nucleotides (see also Section 2.1). We use the sequence from two persons to construct the MAP model, with initial context tree depth 5, for the COX1 gene (approx. 1500 bps long). Then we evaluate the performance of the model on the sequences that correspond to the other 18 individuals. The estimated compression rate per symbol is shown in Figure 4. We observe a very good compression rate of the subsequence corresponding to the COX1 gene, when the learned COX1 model is used for mitochondrion compression. On the other hand, in the other regions no compression is achieved. Therefore, we can clearly distinguish the COX1 gene in the sequences, using this model. Furthermore, the model is generic in the sense that its performance is similar for all sequences, despite the small variations that occur.

4 Discussion and Future Work

In this study we have shown that the context tree can be used to model the statistics of DNA sequences. Though a model can be constructed to represent sequences of variable length and functionality, it is not clear whether the model also implies information

*Sequences were downloaded from the mitochondrion database [12]

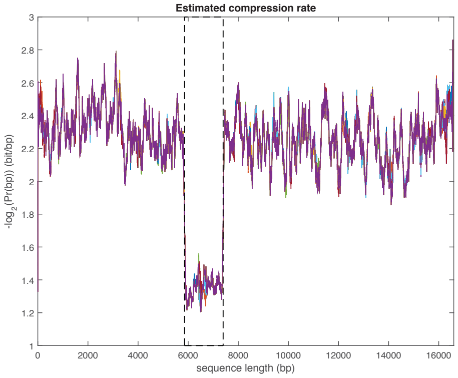


Figure 4: MAP model applied to detect COX1 gene in mitochondrion. The region corresponding to the COX1 gene is marked between the dotted lines.

about the functionality of the modeled sequence. The model can be used to recognize sequences that have similar statistics to the original sequence. Besides functionality analysis, other applications of such a model include read mapping and genome compression.
In this work we assumed that non-coding and coding regions in DNA sequences are stationary. However, our experiments imply that these regions have non-stationary statistics. As a future work we plan to develop an algorithm that automatically recognizes a change in the model and constructs multiple models to accurately represent the different regions in the source-sequence. These models can give more insight in the statistics of the different regions and can be related to the functionality of the region.
As a final remark we state that the strength of the context tree model for DNA sequences, is that it has low sensitivity to variations in the sequence. We plan to further explore this property for application in privacy-sensitive modeling of DNA sequences, since variations are hidden in the model.

5 Acknowledgment

This work has been funded by the EC via grant agreement no. 611659 for the AU2EU project.

References

[1] M. Roy and S. Barman, "Effective gene prediction by high resolution frequency estimator based on least-norm solution technique," *EURASIP journal on bioinformatics & systems biology*, vol. 2014, no. 2, 2014.

[2] S. A. Marhon and S. C. Kremer, "Gene prediction based on DNA spectral analysis: A literature review," *Journal of Computational Biology*, vol. 18, no. 4, pp. 639–676, 2011.

[3] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions," *Genome Research*, vol. 13, no. 8, pp. 1930–1937, 2003.

[4] A. M. Gupal and A. V. Ostrovsky, “Using compositions of Markov models to determine functional gene fragments,” *Cybernetics and Systems Analysis*, vol. 49, no. 5, pp. 692–698, 2013.

[5] M. Stanke and S. Waack, “Gene prediction with a hidden Markov model and a new intron submodel,” in *Bioinformatics*, vol. 19, no. Suppl. 2, 2003, pp. 215–225.

[6] T. Ignatenko and M. Petković, “AU2EU: Privacy-Preserving Matching of DNA Sequences,” in *Information Security Theory and Practice. Securing the Internet of Things (WISTP 2014 Proceedings).* Springer Berlin Heidelberg, 2014, pp. 180–189.

[7] F. Willems, Y. Shtarkov, and T. Tjalkens, “The Context-Tree Weighting Method: Basic Properties,” *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, May 1995.

[8] F. M. J. Willems, A. Nowbakht, and P. A. J. Volf, “Maximum a posteriori probability tree models,” in *Proceedings of the 4th International ITG Conference on Source and Channel Coding*, Berlin, Germany, 2002, pp. 335–340.

[9] T. J. Tjalkens, Y. M. Shtarkov, and F. M. Willems, “Sequential weighting algorithms for multi-alphabet sources,” 1993, pp. 22–27.

[10] J. Rissanen, “Modeling by shortest data description,” pp. 465–471, 1978.

[11] *The NCBI handbook [Internet].* Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 2002.

[12] M. Ingman and U. Gyllensten, “mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences,” *Nucleic Acids Res*, vol. 34, pp. D749–D751, 2006.