Anomalies Detection In DNA Sequences Using Markov Chains

Pierette M. Mastel Pamely Zantou Florent C. Bang Njenjock Cedric P. E. Manouan
Carnegie Mellon University
{pmahoro, pzantou, fbangnje, cmanouan}@andrew.cmu.edu

Abstract

Rare genetic disorders are rooted in mis-sequencing the genome in DNA [6]. Detecting anomalies in genomic sequences, finding the right genomic code, and reconstructing defective sequences represent great challenges and subjects of important and expensive research work in medical care. Many revolutionary approaches in genomic medicine, bioinformatics, and mathematics have been developed in biological sequence analysis to minimize and even completely cure genetic disorders. This work aims to model DNA Sequences using discrete-time Markov chains and apply the resulting model to detect anomalies in a given sequence.

1. Background and rationale

- **Genome**

  Genome refers to the complete set of genetic information (hereditary information) found within an individual organism [1]. It combines both the genes and the non-coding sequences of the DNA.

- **Gene**

  Gene is the basic physical and functional unit of heredity. Genes are passed on from parent to child. They are made up of DNA, represent the coded part of DNA and carry a set of specific instructions to code some molecules such as proteins. Humans for example have between 30.000 and 100000 genes [1]. There are two classes of genes:

  1. genes that are transcribed into RNAs and are translated into polypeptide chains.
  2. genes whose transcripts (tRNAs, rRNAs, snRNAs) are used directly.

- **DNA sequence analysis**

  Deoxyribonucleic acid (DNA) is the molecule that contains the genetic code needed to produce molecules of living cells such as proteins. [1]. Proteins are made of amino acids and are responsible for many vital functions such as transportation, metabolism, movement, structural support, etc. DNA is formed from 4 basic subunits or nucleotides, namely, adenine (**A**), cytosine (**C**), thymine (**T**), and guanine (**G**). [5] Although DNA is often found as a single-stranded polynucleotide, it assumes its most stable form when double-stranded. These two strands are complementary and form a double helix weakly bonded by hydrogen bonds between the nucleotides. **A** in one strand always pairs with a **T** in the other strand, and each **C** always pairs with a **G**. The synthesis of proteins is governed by certain regions in the DNA called protein-coding regions or genes. There are 64 possible nucleotide triplets ($(\text{nucleotide alphabet size})^{\text{word length}} = 4^3$) called *codons*, and mapped into 20 amino acids that bond together to form proteins. DNA sequence analysis consists in automatically interpreting DNA sequences and provide the location and function of protein-coding regions. If a given DNA sequence is not a protein-coding regions the next step will be to accurately identify the protein-coding regions and thus predict the protein that will be generated using the information in these segments. [5] Thanks to bioinformatics, many techniques have been developed to efficiently model DNA sequences.

- **DNA sequence models**

  DNA sequences analysis has been and continues to be a challenging topic in the research domain. However, to be able to do these analyzes, one needs a way to represent the DNA sequence into a suitable format that helps to extract information and do some computation.

  i. **Probabilistic approaches to DNA sequence modeling**

  a. **The Independent Identically Distributed (IID) model**: This model assigns the same chance to each nucleotide to appear at any position within the sequence (uniformly likely). In addition, each nucleotide appears independently of the others. The probability of any nucleotide X to occur in a sequence

S of length L_S is defined as following (G. Singh et al. 2003):

$$P_X = \frac{n_X(S)}{L_S} \quad (1)$$

where n_X is the number of occurrences of the base X in S.
Then the probability of a pattern p to occur can be obtained with equation 2

$$P(p|S) = \prod_{i=1}^{L_p} P(p_i) \quad (2)$$

where $P(p_i)$ is the probability of nucleotide p_i at position i along a pattern p of length L_p.

b. **Discrete-time Markov chains**: In this model, the value of the random variable X any given time step depends only on the previous value. This model has 20 parameters: the probabilities of each nucleotide to occur (P_A, P_T, P_C, P_G) and the probabilities of each *dinucleotide* to occur $(P_{AA}, P_{AC}, ...)$. The first set of parameters is computed using equation 1. And the second set of parameters (the transition probabilities) is obtained [8] using Bayes Rule (3):

$$P(\alpha|\beta) = \frac{P(\alpha\beta)}{P(\beta)} \quad (3)$$

where α and β are different nucleotides and $\alpha\beta$ is a dinucleotide (a compound with two nucleotides).
Thus the probability of occurrence of a pattern p is given by equation 4:

$$P(p|S) = P(p_1) \prod_{i=2}^{L_p} P(p_i) \quad (4)$$

2. Literature review

Mutations in gene may lead to severe diseases. In the case of Cystic Fibrosis (CF), studies were conduced and revealed that some Single Nucleotide Polymorphisms (SNPs) may be associated with presence of chronic rhinosinusitis (CRS) in certain populations, which is almost always developed by CF patients [3, Bejamen Hull et al., 2017]. Identifying those mutations quickly is crucial since they might have a high influence on disease severity in different manners [4, Franziska Gisler, 2013]. Thus, strategies for newborn screening for CF have been evaluated and as a result, it was shown that immunioreactive trypsinogen (IRT), a test that screens for a protein made by the pancreas, seems to be the most cost-efficient among others [7, Masja Schmidt et al., 2018]. In case of a positive initial IRT test, another test or a DNA is carried out. This DNA test can lead to the identification of modifier genes (genes than single mutated gene) which affect disease expression. And the most common method to study association between genetic variations in modifier genes and clinical phenotypes is SNPs identification [9, Martijn Slieker et al., 2005].

3. Research questions, aims and objectives

Our main research question is: *how can we model DNA sequences using Markov chains in order to use the resulting representation to detect disorders/anomalies in a given DNA sequence?*
Trying to answer the above question boils down to two problems that need to be addressed:

- How to model DNA sequences using Markov chains?
- How to use this model to detect anomalies in DNA sequences [2]?

4. Methodology

4.1. Modeling

The overall goal of our work is to be able to identify specific patterns within a DNA sequence in order to tell whether or not that sequence has some anomalies.
For the purpose of this research, we will be using discrete-time Markov chains as DNA sequences model [8]. Our model will be based on the four nucleotides (A, T, G, C) which will be considered as the states of the system. This approach to modeling DNA sequences is divided into two parts: on the one hand, we will work on representing nucleotides and the relationships between them in a given DNA sequence (graph of states); and on the other hand, we will compute the transition probabilities within the resulting graph to end up with the probabilistic graph representing the chain.

4.2. Anomaly detection

In this work, an *anomaly* is any nucleotides sequence that is known as the sequence of a particular disease. Since this pattern must occur at a precise location of the DNA sequence to be considered as an actual anomaly, we are considering using a threshold before returning the binary detection output.
To achieve anomalies detection given a DNA sequence we plan to use the following steps:

a. Compute the first set of parameters (individual nucleotides)

b. Compute the second set of parameters (dinucleotides)

c. Select anomaly pattern to match

d. Get the probabilities of possible matches at a given position of the sequence

e. Apply a threshold to the resulting probabilities

f. Return a binary signal (Anomaly detected or not)

References

[1] 1

[2] "what is known about the function of introns, the nonencoding sequences in genes?". *Scientific American*, 1999. 2

[3] H. BP et al. Single nucleotide polymorphisms related to cystic fibrosis in chronic rhinositus—a pilot study. *Int Forum Allergy Rhinol*, 7, 2017. 2

[4] F. Gisler et al. Identification of snps in the cystic fibrosis interactome influencingpulmonary progressionin cystic fibrosis. *Eur J. or Human Genetics*, 21, 2013. 2

[5] C. Nirnjan et al. "autoregressive modeling and feature analysis of dna sequences". *EURASIP Journal on Applied Signal Processing*, 1(13-28), 2004. 1

[6] J. E. posey. "genome sequencing and implications for rare disorders". *Orphanet J Rare Dis*, 14(153), 2019. 1

[7] M. Schmidt et al. Strategies for newborn screening for cystic fibrosis: A systematic review of health economic evaluations. *Journal of cystic fibrosis*, 17, 2018. 2

[8] G. B. Singh. *"Statistical Modeling of DNA Sequences and Patterns. Introduction to Bioinformatics"*. Humana Press, Totowa, NJ, 2003. 2

[9] M. Slieker et al. Disease modifying genes in cystic fibrosis. *Journal of cystic fibrosis*, 4, 2005. 2