Anomalies Detection In DNA Sequences Using Markov Chains

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

Abstract

Rare genetic disorders are rooted in mis-sequencing the genome in DNA [3]. 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 Sequence using Markov Chains and apply the resulting model to detect anomalies in a given sequence.

1. Background and rationale

Over the years several models have been used for modelling DNA sequences among those include: Wright-Fisher model, Infinite alleles model, Infinite sites model, Moran Model and so on [2].

2. 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 [1]?

3. Methodology

For the purpose of this research we will be using Markov chains to model DNA sequences [4]. Our model will be based on the four building blocks of a DNA sequence which are the nucleotides: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). Our approach to modeling DNA sequences is divided into two parts: on the one hand, we will work on representing nucleotides and the relation between them in a given sequence (graph of states); and on the other hand, we will compute the transition probabilities within the resulting graph of states

References

[1] What is known about the function of introns, the nonencoding sequences in genes? *Scientific American*, 1999. 1

[2] R. Durrett. "basic models," in probability models for dna sequence evolution. pages 1–50, 2008. 1

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

[4] G. B. Singh. *Statistical Modeling of DNA Sequences and Patterns. Introduction to Bioinformatics*. Humana Press, Totowa, NJ, 2003. 1