A Novel Approach to Represent Detected Point Mutation

Dhanya Sudarsan, P.R. Mahalingam, and G. Jisha
Rajagiri School of Engineering and Technology, Cochin, India
{dhanyasudarsan127, prmahalingam}@gmail.com, jishag@rajagiritech.ac.in

Abstract. Research in point mutation is ubiquitous in the field of bioinformatics since it is critical for evolutionary studies and disease identification. With the exponential growth of gene bank size, the need to intelligibly capture, manage and analyse the ever-increasing amount of publicly available genomic data became one of the major challenges faced by bioinformaticians today. The paper proposes a new method to represent point mutation by effectively reclassifying the DNA sequences on the basis of occurrence of point mutation to form a mutation hierarchy which considerably reduces the memory space requirement for storage and heavily reduces the complexity in data mining.

Keywords: Point mutation, Data warehousing, Data mining.

1 Introduction

A gene mutation is a permanent change in the DNA sequence that makes up a gene. Mutations range in size from a single DNA building block (DNA base) to a large segment of a chromosome. Gene mutation can be either small scale or large scale. Point mutation is a small scale mutation resulting in a single nucleotide base change in DNA. A point mutation may be due to the loss of a nucleotide resulting in a shorter sequence, the insertion of an additional nucleotide increasing the sequence, or the substitution of one nucleotide for another. There are many methods to detect point mutations such as:

1. denaturing gradient gel electrophoresis (dgge)
2. temperature gradient gel electrophoresis (tgge)
3. heteroduplex analysis (het)

The new approach to represent point mutation presented in the paper tries to re-classify the DNA sequences based on the obtained point mutation data, and create a hierarchy for the mutations. This can aid in effective representation of the sequences, track mutation chains and even in DNA regeneration.

2 Motivation

The Genbank size is increasing exponentially day by day. Accumulation of information into Genbank was heavily boosted by the introduction of shotgun
technology for sequencing DNA[2]. Thus the memory requirement for storage has increased drastically and the retrieval of needed information from this huge volume became a complicated task.

![Growth of GenBank](image)

**Fig. 1.** Graphical representation of growth of Genbank (Source:Wikipedia)

All these situations motivated to think of a new method to represent mutation which can save memory as well as reduce the complexity of data mining. Initially point mutation in DNA sequence is taken into consideration.

### 3 Existing Methods

Point mutations are represented in a variety of ways. Most frequent representation consists of three distinct parts: a nucleotide, a sequence position, and a mutant. A typical representation of a point mutation is $A_{113}T$, denoting a change from adenine to thymine at position 113 of a DNA sequence. Variations on this shorthand form include $A_{123}\rightarrow T$, $A(123)T$, and $A-123-T$. Three letter abbreviations were also used. For example: $\text{Ala}_{123}\text{Thr}$, $\text{Ala}_{118}\rightarrow \text{Thr}$, $\text{Ala}(118)\text{Thr}$, and $\text{Ala}-118-\text{Thr}$. Aside from this, point mutations are also represented in a sentence