**DETECTION OF CANCER CAUSING DNA SEQUENCES USING SIMILARITY MEASURES**

**A PROJECT REPORT**

***Submitted by***

I.Iswarya (950815104016)

S.Nandhini (950815104037)

S.B.Yugapriya (950815104059)

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF ENGINEERING**

**IN**

### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**GOVERNMENT COLLEGE OF ENGINEERING, TIRUNELVELI**

**ANNA UNIVERSITY: CHENNAI 600 025**

**MARCH 2019**

# ANNA UNIVERSITY: CHENNAI 600 025

## BONAFIDE CERTIFICATE

Certified that this project titled **“DETECTION OF CANCER CAUSING DNA SEQUENCES USING SIMILARITY MEASURES”** is the bonafide work of **ISWARYA.I(950815104016),NANDHINI.S(950815104037),YUGAPRIYA.S.B(950815104059)**who carried out the project work under my supervision.

**SIGNATURE SIGNATURE**

**Dr. P.LATHA, M.E., Ph.D.** **Dr. G.TAMILPAVAI, M.E., Ph.D.**

**HEAD OF THE DEPARTMENT SUPERVISOR,**

Department of Computer Science and Department of Computer Science and

Engineering, Engineering,

Government College of Engineering, Government College of Engineering,

Tirunelveli-627007. Tirunelveli-627007.

###### ACKNOWLEDGEMENT

First of all, we thank the Almighty for the blessings and maneuvering to accomplish this project successfully, without whom this work would be impossible.

We render our special thanks and profound gratitude to **Dr. M.Isaac Solomon Jebamani,** M.E, Ph.D, Principal, Government College of Engineering, Tirunelveli in providing us the perfect platform for undertaking this project work.

We would like to acknowledge the constant support provided by Dr.P.Latha, M.E, Ph. D, Head, Department of Computer Science and engineering, Government College of Engineering, Tirunelveli who shared the valuable thoughts for the betterment of this work. The innovative ideas and the encouraging words provided by our guide Dr.G.Tamilpavai, M.E, Ph.D, Assistant Professor (Sr. Gr.), Department of Computer Science and Engineering, Government College of Engineering, Tirunelveli boosted us every moment throughout the course of our work. The assistance of our guide was remarkable that promoted crisp to our project.

And we extend our sincere thanks to faculty advisor**, Prof. Dr. D. Anitha** M.E, Assistant Professor, Department of Computer Science and Engineering, Government College of Engineering, Tirunelveli and all other staff members of Computer Science and Engineering department.

We also express our special thanks to our APPA and our AMMA who have sacrificed greatly in our education and welfare. We thank all those who have held directly and indirectly in doing this project and bringing out this report successfully.

**ABSTRACT**

We analyze different similarity measures of gene expression profiles and assess their usefulness and robustness in detecting biological relationships by comparing the similarity scores with results obtained from databases of interacting proteins, promoter signals and cellular pathways, as well as through sequence comparisons.

In this paper we study and compare different similarity measures between sequences based on expression data. We assess their accuracy and sensitivity in distinguishing between sequences which are functionally nearby and faraway and evaluate their effectiveness in detecting experimentally verified functional relationships extracted from pathway data, protein–protein interaction data, sequence data and promoter data. Our methodology can also applicable to new similarity measures and datasets.

Classiﬁcation of cancer tissue types and estimation of cancer stages, based on high-dimensional microarray data, in order to support clinical decisions making. Searching database with a DNA sequence rely heavily on sequence comparison techniques. Because of the importance of research into similarity measure, a number of efficient algorithms have been developed for searching genetic databases for biologically significant similarities in DNA sequences.

**CHAPTER 1**

**INTRODUCTION**

This chapter discuss about the definitions of Bioinformatics, DNA Sequences, and Similarity measures.

* 1. **Definitions**

## Definition of Bioinformatics

## Bioinformatics refers to the use of computer science, statistical modelling and algorithmic processing to understand biological data. Bioinformatics is an example of how computer science has revolutionized other fields. Bioinformatics is used heavily in fields that require processing techniques for large data sets, such as sequencing a genome. There are many natural links between biology and computer science, resulting in other exciting fields of study like artificial neural networks (ANN), DNA computing, evolutionary programming and so on.

**Definition of distance-based similarity measures**

In bioinformatics, sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution. Methodologies used include sequence alignment, searches against biological databases, and others. **Similarity:** Degree of likeness between two sequences, usually expressed as a percentage of similar (or identical) residues over a given length of the alignment, can usually be easily calculated.

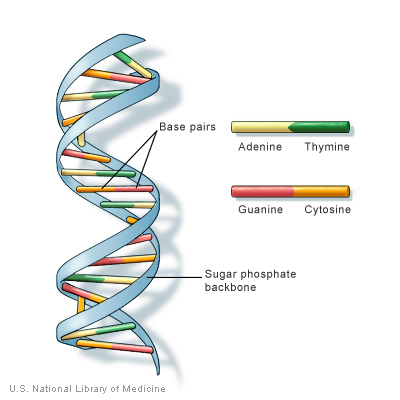
**LCS:** Longest Common SubString (LCS) algorithm considers the similarity between two strings is based on the length of contiguous chain of characters that exist in both strings.

## NW ALGORITHM: Needleman-Wunsch algorithm is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison. It performs a global alignment to find the best alignment over the entire of two sequences. It is suitable when the two sequences are of similar length, with a significant degree of similarity throughout.

## SW ALGORITHM: Smith-Waterman is another example of dynamic programming. It performs a local alignment to find the best alignment over the conserved domain of two sequences. It is useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context.

## Definition of DNA sequences

DNA sequencing is the process of determining the exact sequence of nucleotides within a DNA molecule. This means that by sequencing a stretch of DNA, it will be possible to know the order in which the four nucleotide bases – adenine, guanine, cytosine and thymine – occur within that nucleic acid molecule. The precise ordering of the bases (A, T, G, C) from which DNA is composed. Base pairs form naturally only between A and T and between G and C. DNA sequencing played a pivotal role in mapping out the human genome, completed in 2003, and is an essential tool for many basic and applied research applications today.



**CHAPTER 2**

**LITERATURE REVIEW**

## 

## 