Chapter 10
Use of Fuzzy Set Theory in DNA Sequence Comparison and Amino Acid Classification

Subhram Das
Narula Institute of Technology, India

Jayanta Pal
Narula Institute of Technology, India

Soumen Ghosh
Narula Institute of Technology, India

Dilip K. Bhattacharya
University of Calcutta, India

ABSTRACT

This chapter describes the use of fuzzy set theory and intuitionistic fuzzy set theory in DNA sequence comparison. It also shows an indirect application of fuzzy set theory in comparing protein sequences. In fact, protein sequences consist of 20 amino acids. The chapter shows how such amino acids can be classified in six different groups. These groups are obtained purely from theoretical considerations. These are entirely different from the known groups of amino acids based on biological considerations. Also it is known how these classified groups of amino acids help in protein sequence comparison. The results of comparison differ as the groups differ in number and their compositions. Naturally it is expected that newer results of comparison will come out from such newer classified groups of amino acids obtained theoretically. Thus fuzzy set theory is also useful in protein sequence comparison.

INTRODUCTION

This chapter highlights the importance of fuzzy set theory and intuitionistic fuzzy set theory in problems of Bioinformatics. Initially the standard Voss numerical representation of Nucleotides is interpreted by a two valued logic. While generalizing it to a polynucleotide or a whole genome, it is shown how the notion of fuzzy logic comes into play and helps in obtaining their representations finally on a 12 dimensional unit hypercube. Naturally the set of poly-nucleotides or of whole genomes may be thought of as forming a metric space under suitable metric. The metric defined this way helps in comparison of DNA sequences. Sometimes the comparison is found to be unsatisfactory. This is compensated by using intuitionistic fuzzy set theory in place of fuzzy set theory and adopting same procedure as in fuzzy

DOI: 10.4018/978-1-5225-0914-1.ch010

set theory. Next two valued logic of Voss representation is extended from Nucleotides to amino acids. Using Fuzzy set theory each amino acid is ultimately represented on a 240 dimensional hypercube. Using Euclidean measure on the set of amino acids, each having such 240 components, the amino acids are classified in different groups based on similarity measures. Such groups are finally used in protein sequence comparison. This is the general perspective behind the introduction of the chapter. However the main objective of the chapter is to show application of fuzzy set theory and intuitionistic fuzzy set theory in DNA and Protein sequence comparison.

BACKGROUND

DNA is usually presumed to be the critical macromolecular target for carcinogenesis and mutagenesis. To predict sequence changes induced by different agents, it is imperative to have quantitative measures to compare and contrast the different DNA sequences. In addition, the very rapid rise in available DNA sequence data has also made the problem more emerging and interesting too. Again the character of a whole genome is not reflected from a particular type of its gene. So for the purpose of comparison whole genomes are to be considered. But the main problem in genome sequence comparison lies in the fact that the lengths of the corresponding sequences may be too large and at the same time lengths may differ from sequence to sequence. Obviously the main target is to convert whole genome sequence of any length to a desired sequence of a manageable size. This will definitely make the process of comparison of sequences much simpler and manageable too. Let us describe how this is achieved.

Two Valued Logic in Voss Representation

It is known that DNA and RNA are made of codons, each of which is a triplet of nucleotides, having the possibility to be one of four nucleotides {T, C, A, G} in the case of DNA and {U, C, A, G} in the case of RNA (A: adenine; C: cytosine; G: guanine; T: thymine; U: uracil). In Voss representation (Voss, 1992) nucleotides T/U, C, A, G are represented as (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) respectively. It may be argued that when T/U is written as (1, 0, 0, 0), it is meant that T/U is understood fully but C, A, G are not understandable at all. Thus for T/U, C, A, G taken in this order T/U is given the value 1 and others are given value 0. The same argument may be given to C, A, G also. Thus a two valued logic using binary 1, 0 works well and a single codon (a combination of three nucleotides) is represented on a 12 dimensional unit hypercube and is expressed by crisp values 1 and 0. Naturally if it is polynucleotide or a whole genome consisting of n codons, it is represented on a 12n dimensional hypercube and the process becomes unmanageable if n is large. This is definitely a drawback in the representation procedure. The second and most important difficulty arises when one tries to compare two polynucleotides of different lengths. In fact, in this case, they are represented on spaces of different dimensions. So the process of comparison is no longer applicable. Obviously both types of difficulties could be avoided, had the representation been made on a single 12 dimensional hypercube. This is the reason why, for representation of a polynucleotide or a whole genome, always a 12 dimensional hypercube is chosen. As a matter of fact, necessity of introducing fuzzy set theory is realized in the process of representing a polynucleotide consisting of finite number of codons, n say, on a single 12 dimensional hypercube. This is the background of fuzzy polynucleotide space as introduced by Torres and Nieto (2003).