Gene 583 (2016) 112–120

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Application of Euclidean distance measurement and principal component analysis for gene identification

Antara Ghosh, Soma Barman *

Institute of Radio Physics and Electronics, University of Calcutta, 92, APC Road, Kolkata 700009, India

ARTICLE INFO

Article history:
Received 17 August 2015
Received in revised form 27 November 2015
Accepted 7 February 2016
Available online 11 February 2016

Keywords:
Amino acid
Cancer
Discrete Fourier transform
Euclidean distance
Principal Component Analysis
Digital signal processing

ABSTRACT

Gene systems are extremely complex, heterogeneous, and noisy in nature. Many statistical tools which are used to extract relevant feature from genes provide fuzzy and ambiguous information. High-dimensional gene expression database available in public domain usually contains thousands of genes. Efficient prediction method is demanding nowadays for accurate identification of such database. Euclidean distance measurement and principal component analysis methods are applied on such databases to identify the genes. In both methods, prediction algorithm is based on homology search approach. Digital Signal Processing technique along with statistical method is used for analysis of genes in both cases. A two-level decision logic is used for gene classification as healthy or cancerous. This binary logic minimizes the prediction error and improves prediction accuracy. Superiority of the method is judged by receiver operating characteristic curve.

1. Introduction

A revolution in life science and medical science research appeared in the year of 1953, when Watson and Crick discovered the double helices structure of DNA or deoxyribonucleic acid (Watson and Crick, 1953; Brenner, 2012). This discovery leads to convergence of engineering and life sciences. DNA is the hereditary material in almost all organisms, encoded as a sequence of nucleotide bases: guanine (G), adenine (A), thymine (T), and cytosine (C). The sequence of nucleotide bases A, T, G, and C provides genetic information needed to carry out cell's activities. By analyzing the nucleotide or DNA sequence, one can find out not only the variations in the DNA sequence among species but also the variation between healthy and diseased cell. This feature of DNA sequence is successfully able to pull the attention of researchers from various fields and introduces a new interdisciplinary field of research named Genomic Signal Processing, which blends biosciences, medicine, and engineering. A DNA sequence is divided into two regions: genes and intergenic spaces. Prediction and processing of gene are important because they are responsible for the production of different proteins. Basically, a gene has two types of sub-regions called exons and introns. The gene is first copied into a single-stranded chain called the messenger RNA or mRNA molecule. The introns are then removed from the mRNA by a process called splicing. The spliced mRNA is divided into groups of three adjacent bases. Each triplet is called a codon. There are 64 possible codons, which are responsible to generate twenty amino acids as shown in Fig. 1 (Anastassiou, 2001; Vaidyanathan, 2004). The amino acid composition is important to determine protein folding type (Chou and Zhang, 1993). The proteins drive all the biological process of living organisms and deficiency of proteins may cause different types of diseases, such as Alzheimer, Parkinson, cancer disease, as well as other neurodegenerative disorders. For prediction of human genetic diseases like cancer, tumor, etc., researchers of life science and medical science investigate the role of codons and amino acids in gene (Chou and Zhang, 1992; Zhang and Chou, 1993; Zhang and Chou, 1994; Chou et al., 1996; Qiu et al., 2007). Out of all the genetic diseases, cancer is responsible for one in eight deaths worldwide (Stratton et al., 2009; Barman et al., 2011a; Chin et al., 2011; Ghosh and Barman, 2014), and breast cancer is a highly diverse disease which is very common among women. A successful treatment is very important for prediction of cancer as early as possible. Therefore, understanding the mechanism of breast cancer development, accurate detection and early prediction of breast cancer is a significant recent day's research topic (Tadayyon et al., 2014; Sfakianakis et al., 2014). Protein acts as enzymes, hormones, and antibodies catalase and regulates the chemical reactions in the body. Without amino acids, our body fails to function because protein

Abbreviation: CAAT, Controlled amino acid therapy; DFT, Discrete Fourier transform; DNA, Deoxyribonucleic acid; DSP, Digital signal processing; ED, Euclidean distance; mRNA, Messenger ribonucleic acid; NCBI, National Center for Biotechnology Information; NIH, National Institutes of Health; PCA, Principal component analysis; PseAAC, Pseudo amino acid composition; PseKNC, Pseudo K-tuple nucleotide composition; RNA, Ribonucleic acid; ROC, Receiver operating characteristic; TPR, True-positive rate; TNR, True-negative rate.

* Corresponding author.
E-mail address: barmanmandal@gmail.com (S. Barman).

http://dx.doi.org/10.1016/j.gene.2016.02.015
0378-1119/

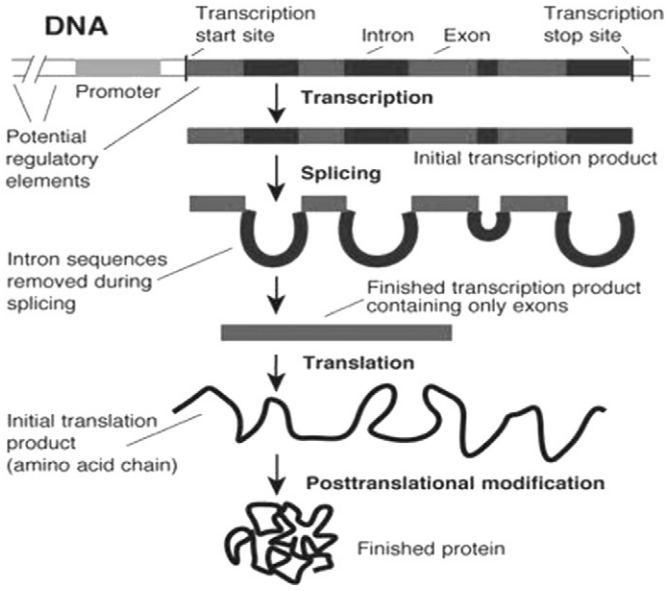

Fig. 1. From DNA to protein synthesis.

cannot be synthesized. Mutated cancer cells are also influenced by certain proteins and their associated amino acids.

Therefore, cancer genomics researchers are investigating the critical role of amino acids cancer in genes. Controlled amino acid therapy (CAAT) is an efficient medical treatment used to impair the development of cancer cells by controlling the amino acids (https://www.apjohncancerinstitute.org/caat-protocol). Dr. Otto Warburg discovered that all cancer cells produce inordinate amount of lactic acid (Warburg, 1956). Lee et al. (2000) showed that the deficiency of glucose can commit suicide of cancer cells. Arginine can modulate the growth of breast cancer (Singh et al., 2000). Rock stated that aspartic acid (D), glutamic acid (E), glycine (G), serine (S), alanine (A), and cysteine (C) generated through the synthesis of glucose (Rock and King, 1968). Cysteine (C) is one particular powerful amino acid recognized to fight cancer (https://www.apjohncancerinstitute.org/caat-protocol). Realizing the significance of amino acid in genetic disease, the objective of the present paper is to identify the healthy or cancerous genes based on amino acids composition.

The prediction of genetic disease based on amino acids sequence by the statistical method has become quite popular during the last two decades. Because of DNA microarrays, a large number of high-dimensional gene expression databases are available in public domain, which helps to grow the interest among the researchers from various fields to develop fast and accurate methods for DNA/protein sequence analysis and disease diagnosis. Among these methods, amino acid sequence–based statistical method became very popular in recent years. Nakashima et al. analyzed the protein sequences in terms of 20 amino acid compositions and used this 20-dimensional composition for protein structure classification in the year 1986 (Nakashima et al., 1986). The "distance" in the space is a simple and convenient method used for classification of protein folding types. Further K. C. Chou et al. used various statistical methods like least Hamming distance, least Euclidean distance, discriminant analysis, amino acid principal component analysis, etc., to predict protein structural classes in 20-D amino acid composition space (Nakashima et al., 1986; Chou, 1989; Zhang and Chou, 1992; Chou and Zhang, 1995; Chou, 1995; Chou and Maggiora, 1998; Chou et al., 1998; Du et al., 2012). K. C. Chou, in 2001, proposed pseudo amino acid composition or PseAAC to improve prediction quality by considering sequence-order effect in protein sequence (Chou, 2001). This proposed concept is not only used for protein structure prediction but also used in disease diagnosis and drug development areas successfully (Chou, 2005; Zhanga et al., 2008; Esmaeili et al., 2010; Zhong and Zhou, 2014; Hajisharifi et al., 2014; Du et al., 2014). The PseAAC was first developed only to deal with proteins/peptides, but this method is also used for DNA/RNA sequence analysis which is named as pseudo K-tuple nucleotide composition (or PseKNC) (Chen et al., 2014; Liu et al., 2015a, b; Chen et al., 2015a,b).

Since amino acid composition method is significant in genomic research, the authors used digital signal processing along with statistical methods (Euclidean distance measurement and principal component analysis method) for discriminate analysis of healthy and cancer breast genes. Euclidean distance measurement method is applied on 20-D space of genes and measures the distance between the genes to differentiate the healthy and cancerous breast gene, and a PCA model is established based on 20-D amino acid composition for prediction of genes. The PCA approach reduces the 20-dimensional amino acid spaces to fewer dimension orthogonal space and also minimizes the random errors and redundant information in gene database. The performances of these methods are compared by receiver operating characteristic (ROC) analysis.

Euclidean distance measurement and principal component analysis are mostly used for classification of the protein structure class, but here, the authors used the techniques for predicting cancerous/healthy genes. The methods are tested on various *Homo sapiens* databases downloaded from NCBI homepage (http://www.ncbi.nlm.nih.gov), a branch of the National Institutes of Health (NIH) and one of the most important public resources for DNA and protein sequence database. The paper is structured as follows: Introduction, Methods, Results, and Conclusion.

2. Methods

Genomic and proteomic information is digital in nature. Such information either in the form of DNA or proteins is mathematically represented by character strings, where each character is represented by an alphabet. For DNA, the alphabet is size 4 and consists of letters A, T, C, and G, but in the case of protein, the size of the corresponding alphabet is 20 such as A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. The digital nature of genomic information makes it suitable for application of signal processing methods and tools for better analyzing and understanding the function of DNA, protein, and genes. A proper mapping of a character string into numerical sequence is needed prior to digital signal processing (DSP) application. Here, binary mapping techniques based on presence or absence of amino acids in a gene are attempted prior to DSP application (Barman et al., 2011b).

For example, an amino acid sequence of length n:

$$x[n] = [M\ P\ A\ G\ S\ K\ E\ R\ P\ T\ F\ F\ E\ I\ F\ K\ T\ R\ C\ N\ K\ A]$$

After mapping for amino acid A:

$$x[A] = [0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1]$$

Similarly for the rest 19 amino acids, binary mapped sequence is obtained.

For better representation of the signal, discrete Fourier transform technique is applied on spatial domain sequence $x_s[n]$.

The spectral estimation of mapped sequence ($X_s[k]$) is obtained by discrete Fourier transform (DFT) technique

$$X_s[k] = \sum x_s[n]\, e^{-j2\pi k/n} \tag{1}$$

where $k = 0, 1, 2, \ldots, N-1$, $n = 0, 1, 2 \ldots, N-1$,

$$s = A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y,$$