Journal of
Information Technology & Software Engineering

Gupta, J Inform Tech Soft Engg 2011, 1:1
DOI: 10.4172/2165-7866.1000e101

Editorial Open Access

Prediction of Disease in Compressed DNA Sequences: An Open Problem

Ashutosh Gupta

Department of Computer Science & Information Technology, M. J. P. Rohilkhand University, Bareilly, UP, India

The deoxyribonucleic acid (DNA) constitutes the physical medium in which all properties of living organisms are encoded. The understanding of its sequence is primary concern in molecular biology. Some important molecular biology databases (ERIBL, GenBank, DDJB) are developed around the world to accumulate nucleotide sequences (DNA, RNA) and amino-acid sequences of proteins. It is well acknowledged that their size increases nowadays exponentially fast. Not as big yet as some other scientific databases, their size is in hundreds of GB [1]. For complete genomes, these texts can be very elongated. The human genome for example contains three billions characters over twenty-three pairs of chromosomes. It contains all the genetic substance of the human beings. With escalating number of genome sequences being made available, the difficulty of storing and using databases has to be addressed. The compression of genetic information as a result constitutes a very important job. Another factor which is also to be considered is the prediction of certain kind of disease by applying the searching a pattern in the compressed domain.

The importance of common compressibility for identifying patterns of interest from genomes is recognized [2] and it is also established that compressibility is a well dimension of relatedness among sequences [3]. Many compression algorithms use the characteristics of DNA like point mutation [4] or reverse complement to achieve a god compression rate. Many compression methods [5-7] have been exposed to compress DNA sequences. Invariably, all the compression methods instigate so far take benefit of the fact that DNA sequences are made of only four alphabets, together with schemes to exploit the repetitive nature of DNA. Some methods have the capability to search the disease but none of the methods investigate the prediction of diseases by searching a pattern in compressed DNA sequences.

Searching a pattern may be the repetitive structures that have been implicated in various diseases and genetic disorders. While the exact function of each of the different genes so far identified is not completely known, even less is known about the repeats, and this represents a major problem and should be investigated by sharing the ideas and views through some open access medium. Anyone can share their ideas and opinions through the OMICS group who organizes international conferences throughout the year in different aspect.

References

1. Kamel N (1991) Panel: Data and knowledge bases for genome mapping: What lies ahead? In Proc. Intl. Very Large Databases.
2. Stern L, Allison L, Coppel RL, Dix TI (2001) Discovering patterns in plasmodium falciparum genomic DNA. Molecular & Biochemical Parasitology 118:175–186.
3. Powell DR, Allison L, Dix TI (2004) Modelling-alignment for non-random sequences. Advances in Artificial Intelligence 203–214.
4. Grumbach S, Tahi F (1994) A new challenge for compression algorithms: Genetic sequences. Inf. Process. Manage 30: 875–866.
5. Grumbach S, Tahi F (1993) Compression of DNA sequences DCC 340–350.
6. Adjeroh D, Nan F (1998) On compressibility of protein sequences DCC 422–434.
7. Boulton DM, Wallace CS (1969) The information content of a multistate distribution. Theoretical Biology 23: 269–278.

Corresponding author: Ashutosh Gupta, Department of Computer Science & Information Technology, M. J. P. Rohilkhand University. Bareilly, UP, India, Tel: +91-9415351823; E-mail: ashutosh333@rediffmail.com

Received November 17, 2011; **Accepted** November 19, 2011; **Published** November 21, 2011

Citation: Gupta A (2011) Prediction of Disease in Compressed DNA Sequences: An Open Problem. J Inform Tech Soft Engg 1:e101. doi:10.4172/2165-7866.1000e101