DEEP LEARNING FOR CANCER PREDICTION

[1]Dr P Kiran Sree, [2]N Usha Devi

[1]Department of CSE, Shri Vishnu Engineering College for Women, Bhimavaram, [2]Dept of CSE, University College of Engineering, JNTU Kakinada

DISCUSSION

Many problems in Bioinformatics can be solved by various computer techniques. Cancer prediction is a real time dynamic problem existing in the real world. The cause and reason for drastic spread of this disease is very difficult to understand. We have various forms and types of cancers, an abstract model to predict cancer is difficult. Although many papers are available to trace cancer, there is still room for evolving a new strategy for predicting cancer.

We propose a novel Deep Learning based associative memory which learns from various case studies analyzing the data and predicts the Bone Cancer. We have taken datasets from ICCR Datasets and processed them using Hybrid Unsupervised learning algorithm. Preliminary work was done and we have compared our work with some standard existing literature. The proposed classifier performance was found promising.

Human promoter sequences are taken from DBTSS database. We have extracted 30, 966 base pairs [-200, +50] bp around TSS from DBTSS database for training and testing. Non-promoter sequences exons, introns are taken from EID database . We have extracted 75,438 exons and 53,684 introns from EID database. Non promoter sequences of 3'UTR are taken from UTRdb database. We have extracted 80,538 data

components from UTRdb database. All the sequences for both promoter and non-promoter region are of length 251 base pairs.

Example Computation: Consider a human 251 bp length DNA sequence.

CGCAGCAAAATGCACGGGCTTCTGCAGCCCACATGACTTTATTCTGAAC

GGACACAAGTCTGCTCGCTGGGCCGTTCGCTTTTGGGCCAAAAACACGG

CTCCGTCGGTGACTTTTGGCCCGATATTGGCGACCAGAAAACACAAGTG

AAAGAGCATTTGGCCAGCCCGGAGAAGCCGAGCTGGGTGGCTTGAGTC

TACATGGTTCTCATGTCGCGTTTAAGGCCAGCCCCCTGCACGGTGTGGA

GCTTCAA

Number of 'C's= 69

Number of 'G's= 69

Number of 'CG's= 16

The ratio of 'C' content in a DNA sequence Cr is calculated as below.

Cr= Number of 'C's in the sequence/ Length of DNA sequence=69/251=.274

The ratio of 'G' content in a DNA sequence Gr is calculated as below.

Gr= Number of 'G's in the sequence/ Length of DNA sequence=69/251=.274

The ratio of 'GC' content in the DNA sequence GCr is calculated as below.

GCr= Number of 'GC's in the sequence/ Length of D`NA sequence

GC Percentage = GCp =Cr+Gr=0.274+0.274=0.548

The Observed/Expected CpG ratio =GCr/Cr+Gr= 16/138=0.11

GCp is more than 0.5 but Observed/Expected CpG ratio is less than 0.6, so the sequences above mentioned do not fall into CpG related. This is very helpful for cancer prediction.

After an extensive literature survey, we conclude that Deep Learning can be applied on many problems in Bioinformatics particularly to identify cancer traces patient body with greater accuracy.

REFERENCES

1. Siegel, Rebecca, Deepa Naishadham, and Ahmedin Jemal. "Cancer statistics, 2013" CA: a cancer journal for clinicians 63.1 (2013): 11-30.
2. Jemal, Ahmedin et al. "Cancer statistics, 2009", CA: a cancer journal for clinicians 59.4 (2009): 225-249.
3. Jemal, Ahmedin, et al. "Cancer statistics, 2010", CA: a cancer journal for clinicians 60.5 (2010): 277-300.
4. Siegel, Rebecca, et al. "Cancer statistics, 2014", CA: a cancer journal for clinicians 64.1 (2014): 9-29.
5. Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics, 2015", CA: a cancer journal for clinicians 65, no. 1 (2015): 5-29.
6. Dr Kiran Sree, Investigating an Artificial Immune System to Strengthen the Protein Structure Prediction and Protein Coding Region Identification using Cellular Automata Classifier. International Journal of Bioinformatics Research and Applications ,Vol 5,Number 6,pp 647-662, ISSN : 1744-5493. (2009) (Inderscience Journals , UK)Listed & Recognized in US National Library of Medicine National Institutes of Health .National Center for Biotechnology Information(Government of USA)PMID: 19887338 [PubMed - indexed for MEDLINE] H Index (Citation Index): 08 (SCImago, www.scimagojr.com)   (Nine Years Old Journal)
7. Dr Kiran Sree , Identification of Promoter Region in Genomic DNA Using Cellular Automata Based Text Clustering. The International Arab Journal of Information Technology (IAJIT),Volume 7,No 1,2010,pp 75-78.   ISSN:1683-3198H Index (Citation Index): 05 (SCImago, www.scimagojr.com)(Eleven Years Old Journal)(SCI Indexed Journal)
8. Dr Kiran Sree, A Fast Multiple Attractor Cellular Automata with Modified Clonal Classifier for Coding Region Prediction in Human Genome, Journal of Bioinformatics and Intelligent Control, Vol. 3, 2014, pp 1-6. DOI:10.1166/jbic.2014.1077 (American Scientific Publications, USA)
9. Dr Kiran Sree, A Fast Multiple Attractor Cellular Automata with Modified Clonal Classifier Promoter Region Prediction in Eukaryotes.Journal of Bioinformatics and Intelligent Control, Vol. 3, 1–6, 2014. DOI:10.1166/jbic.2014.1077 (American Scientific Publications, USA)
10. Dr Kiran Sree, MACA-MCC-DA: A Fast MACA with Modified Clonal Classifier Promoter Region Prediction in Drosophila and Arabidopsis. European Journal of Biotechnology and Bioscience, 1 (6), 2014, pp 22-26, Impact Factor: 1.74 .
11. http://www.iccr-cancer.org/datasets