BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 13(9)

Hypothesis

Use of a neural network to predict normalized signal strengths from a DNA-sequencing microarray

Charles Chilaka[1,5], Steven Carr[2,3,*], Nabil Shalaby[3,4], Wolfgang Banzhaf[3,6]

[1]Program in Scientific Computing; [2]Department of Biology; [3]Department of Computer Science; [4]Department of Mathematics and Statistics Memorial University of Newfoundland; St. John's, Newfoundland, Canada A1C 5S7; [5]Department of Mathematics, FUT, Owerri, Nigeria [6]Present address: Department of Computer Science and Engineering, Michigan State University, East Lansing MI 48824. Steven Carr – E-mail: scarr@mun.ca; Tel: 1 (709) 764 4776; *Corresponding author

Received July 13, 2017; Accepted July 18, 2017; Published September 30, 2017

Abstract:
A microarray DNA sequencing experiment for a molecule of N bases produces a 4xN data matrix, where for each of the N positions each quartet comprises the signal strength of binding of an experimental DNA to a reference oligonucleotide affixed to the microarray, for the four possible bases (A, C, G, or T). The strongest signal in each quartet should result from a perfect complementary match between experimental and reference DNA sequence, and therefore indicate the correct base call at that position. The linear series of calls should constitute the DNA sequence. Variation in the absolute and relative signal strengths, due to variable base composition and other factors over the N quartets, can interfere with the accuracy and (or) confidence of base calls in ways that are not fully understood. We used a feed-forward back-propagation neural network model to predict normalized signal intensities of a microarray-derived DNA sequence of N = 15,453 bases. The DNA sequence was encoded as n-gram neural input vectors, where n = 1, 2, and their composite. The data were divided into training, validation, and testing sets. Regression values were >99% overall, and improved with increased number of neurons in the hidden layer, and in the composition n-grams. We also noticed a very low mean square error overall which transforms to a high performance value.

Keywords: Neural networks, n-grams, Performance, Regression values.

Background:
DNA sequences although letters contain a lot of information. They are not numeric in nature but their conversion to numerical values en-ables the application of powerful digital signal processing techniques to them. Some desirable properties of a DNA numerical representation are given in [3]. Some forms of DNA numerical representations include: Z-curves and DNA walks [4], Voss method, quaternion technique and paired nucleotide/atomic number representation [5], paired numeric representation [6], double curve and structural profile method [7] and electron-ion interaction potential [8]. N-gram method used in this paper was first introduced by C.E Shan-non in 1948 [9], and makes use of data in a sliding window fashion and neural network learning methods provide a robust approach to approximating real-valued, discrete-valued and vector-valued target functions [12] like DNA numerical. The study of artificial neural networks has been inspired in part by the observation that biological learning systems are built of very complex webs of interconnected neurons [10, 11, 12], where the neurons communicate through a large set of interconnections with variable strengths (weights) in which the learned information is stored [13]. Each neuron computes a weighted sum of its y input signals. The activation function for neurons is the sigmoid function defined [12] as

$$\delta(y) = 1 / (1+e^{-y}) - (1)$$

Where y is the weighted sum of the inputs. The output of the sigmoid function ranges from 0 to 1, increasing monotonically with its input and the weights of the interconnections between the different neurons are adjusted during the training process to achieve a desired input/output mapping. The ideas from artificial neural net-works have led to computational analysis of human DNA sequence [14], single base pair discrimination of terminal mismatches [15], biological phenomena through computational intelligence [16], human donor and acceptor sites prediction [17], coding region recognition and gene identification [18], predicting transmembrane domains of proteins [19] and the prediction of nucleotide sequences using genomic signals [20, 21].

In this paper, an Affymetrix [1] experiment output, which has numerical values, is normalized and partitioned into training; testing and validation set using a Matlab [2] neural network with 4 and 16 numbers of nodes in the input layer. The influence of the

ISSN 0973-2063 (online) 0973-8894 (print)