www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 9 September 2014 Page No. 8026-8029

Stock Market Behavior Prediction Using Pattern Matching Approach

Prakash Kumar Sarangi, Birendra Kumar Nayak

Department of Information Technology
NM Institute of Engineering and Technology, Bhubaneswar, India.
Prakashsarangi89@gmail.com.

Retd. Prof., Department of Mathematics,
Utkal University, Bhubaneswar, India.

Abstract

In this paper we propose a new model for prediction of stock market behavior using pattern matching approach. The fluctuation of stock market is characterized by a number 0 and 1, '0' denoting non-increasing state and '1' denoting an increasing state. The behavior of the stock market is put into sequence of 0's and 1's which was converted to the sequence of nucleotides A, T, C, G. This sequence so obtained is matched to the text DNA sequence by using BLAST. Comparing their results using hamming distance and predict the future increasing or non-increasing behavior of stock market. Possibility using this approach to predict the stock market behavior is explored.

Keywords: Nucleotides, Multiple Sequence Alignment, Hamming distance

1. Introduction

For many years the following question has been a source of continuing controversy in both academic and business circles: To what extent can the past history of a common stock's price be used to make meaningful predictions concerning the future price of the stock? Answers to this question have been provided on the one hand by the various theories and on the other hand by the theory of random walks. Although there are many different theories, they all make the same basic assumption. That is, they all assume that the past behavior of a security's price is rich in information concerning its future behavior. History repeats itself in that "patterns" of past price behavior will tend to recur in the future. Thus, if through careful analysis of price charts one develops an understanding of these "patterns," this can be used to predict the future behavior of prices and in this way increase expected gains [1].

Due to its randomness, lots of research has going on for studying behavior of stock market. In 1959, Roberts wrote" If the stock market behaved like a mechanically imperfect roulette wheel, people would notice the imperfections and, by acting on them, remove them. This rationale is appealing, if for no its value as counterweight to the popular view of stock market "irrationality," but it is obviously incomplete. Roberts generated a series of random numbers and plot result to see whether any patterns that were known to technical analysts would be visible [2].

In another way many work has been done using intelligent computational technique like Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Genetic Algorithms (GA) forecast financial market behavior [3-5]. But in this paper we are putting a relationship between behaviors of stock market with DNA sequences. For above analysis we need support of Bio-informatics tools like BLAST.

The primary goal of bio-informatics is to increase the understanding of biological processes. Bio-informatics [6], the application of computational techniques to analyze the information associated with bi-molecules on a large scale, has now firmly established itself as a discipline in molecular biology. Bio-informatics is a management information system for molecular biology [8]. Bio-informatics encompasses everything from data storage and retrieval to the identification and presentation of features within data, such as finding genes within DNA sequence, finding similarities between sequences, structural predictions [7]. Using such BLAST software we consider very large DNA database as a text pattern and some part of stock behavior as a pattern [6]. We can also predict future aspects of stock market.

The main motivation of this paper is to propose a model to find stock market tendency and to test the predictability of the proposed BLAST software[9-10]. The rest of the study is organized as follows. The next section will describe the methodology which processing in detail like mapping, encoding, partitioning, and proposed pattern matching algorithm. In Section 3, we give an experiment scheme and Empirical results and analysis are reported in this section. The concluding remarks are given in Section 4. In last Section we conclude the order of accuracy to predict the stock market behavior.

2. METHODOLOGY

In this section, the closing price of day to day trading builds a process is presented in detail mapping with human genome. First a mapping of stock price to binary is described. Encode these binary values to nucleotides using Huffman tree to compress total data to half. Nucleotides are divided DNA sequences having a continuous distribution. Each pattern of DNA sequences matched with very large DNA database using BLAST, which finally predict the day to day behavior of stock market.

2.1. Representation of Stock behavior to Binary

This study is to map and explore the tendency of stock price index. The research data used in this study is technical indicators and the direction of change in the daily S&P500 stock price index thirty years. Considering their closing price of each day, they are categorized as “0” and “1” in the research data. “0” means that the next day’s index is lower or same to today’s index, and “1” means that the next day’s index is higher than today’s index [8].

2.2. Converting Binary sequence to DNA sequence

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell [11]. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA bases pair up with each other, A with T and C with G, to form units called base pairs.

Given a DNA sequence consisting of A, C, T, G characters, we use two bits to encode each characters as follows:

“00” for A,
“01” for C,
“11” for T, and
“10” for G.

As a result, each cell of 2 bits representing one DNA character. As we know in DNA chaining “A” bonding with “T” and “ C” bonding with “G” , we generate a rule which pairs are 1’s complement with each other [10].

For an example, DNA sequence TACCTGCGCTA is encoded by binary sequence 11 00 01 01 11 10 01 10 01 11 00.

2.3. Building of DNA Patterns

Entire compressed DNA pattern of thirty years is divided into several parts; each part consider as a pattern for some month trading or years trading of stock market behavior. Since each part is a form of DNA sequence, we can say each part has a life. Now we matched with a very large DNA database and predict the future behavior of rest occurring patterns. There are many patterns matching algorithms to match pattern of any data, but in bioinformatics we have a very user friendly tool BLAST, which have the potential to find different patterns.

2.4. Multiple Sequence Alignment using BLAST

BLAST (Basic Local Alignment Search Tool): family of sequence alignment algorithms developed by Altschul et. Al. 1997. These programs are used for sequence similarity identification [7]. They identify regions of local alignment to assist in detecting relationships among sequences, which allows the user to identify similarities between the query nucleotide or protein sequence with sequences in public databases, Identifies clusters of nearby or locally dense "similar" k-tuples (number of string of letters), Used to identify whether a given sequence is novel, homologous to a known sequence, or if the sequence contains motifs which may provide clues to a possible roles of the sequence being queried [7].

The preferred query sequence format of the BLAST program is the FASTA format which takes input as DNA sequences. Now each part of patterns is aligned with the largest DNA database of BLAST. It is found that thirty years stock pattern behavior performs a match with DNA Data base of BLAST.

2.5. Hamming Distance Measure

The Hamming distance between two strings of equal length is the number of positions for which the corresponding symbols are different. Put another way, it measures the minimum number of substitutions required to change one into the other, or the number of errors that transformed one string into the other.

We define the Hamming distance between strings x and y, denoted dH(x, y), to be the number of places where x and y are different. Using the concept of Hamming distance, we can mathematically describe between two DNA sequences having equal in length how many places error mismatch occurs. For example

Let X= ATCGTCGTATAGCTAG and
Y = ATGTTCGATTGACTAG then
dH(X,Y)= 6 (i.e. X differ from Y exactly six positions)

If X is the acual DNA sequence and Y will be the predicted DNA sequence after alignment using BLAST then error occurs during alignment is total six number of positions. This can be soon by using MATLAB representation in form of graph between actual DNA vs predicted DNA.

3. EXPERIMENTAL RESULTS

The entire data set contains close price of S&P500 of thirty years which, covers the period from the first trading day of January, 1980 to last trading day of December, 2010. The data sets are divided into several patterns. Using BLAST we found that some part of stock market behavior DNA is a cent percent match with the very large DNA database. Using the concept of Hamming distance, we can evaluate error gap between DNA patterns of stock market behaviors with predicted DNA patterns. From MATLAB implementation experimental results are collecting from different stock DNA pattern vs predicted DNA sequences as follows:

Human DNA sequence from clone RP11-234K24 on
chromosome 20, complete sequence
Sequence ID: emb|AL121895.26|Length: 153192
Range 1: 104016 to 104033GenBankGraphics
Next Match Previous Match First Match

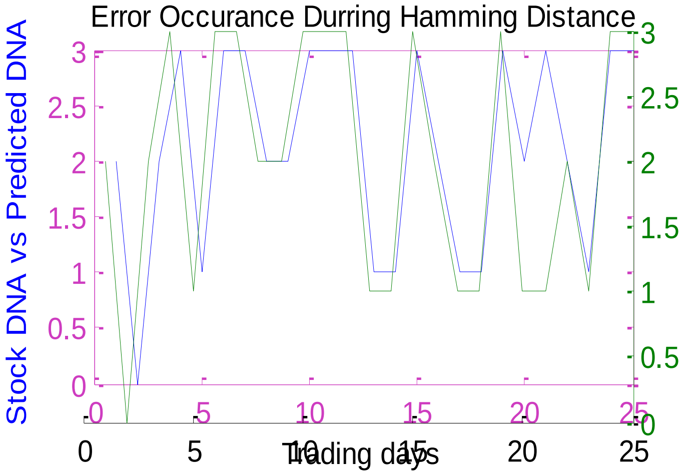
Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
36.2 bits(18)	4.6()	18/18(100%)	0/18(0%)	Plus/Plus	

Features:

```
Query  9      GAGTCTTGGTTTCCTGCC   26
              ||||||||||||||||||
Sbjct  104016 GAGTCTTGGTTTCCTGCC  104033
```

104016 GAGTCTTGGT TTCCTGCCTG TGCTT 104040
GAGTCTTGGT TTCCTGCCTC CGCTT
dH(GAGTCTTGGT TTCCTGCCTG TGCTT,
GAGTCTTGGT TTCCTGCCTC CGCTT) = 2



4. CONCLUSION & FUTURE WORK

This study proposes using bioinformatics tool BLAST that performs a relationship between stock market behaves with DNA sequences. In terms of the empirical results, we find that each five years or each decade the stock market tendency behaves like human DNA sequence. The alignment score says that between them maximum identification is 100% in each part. But for prediction purpose it is possible when continuous matching occurs. From experimental work we found that maximum identification not exactly equal to 100% in each 5 years.

As we found from the above experiment each five years stock data are pattern matched with the human genome. Also stock market is fully random, for the prediction of stock price possible for next five years. This is the future work to predict the behavior of the stock market for coming trading days.

5. REFERENCES

[1] Hasan, A., Saleem, HM., Abdullah, S., "Long- Run Relationships between an Emerging Equity Market And Equity Markets of the Developed World an Empirical Analysis of Karachi Stock Exchange," International Research Journal of Finance and Economics, 2008, 16, pp. 52-62.

[2] Robert, H., "Stock Market 'Patterns' and Financial Analysis, Journal of Finance , 1959

[3] Kamijo, K., Tanigawa, T.: Stock Price Pattern Recognition: A Recurrent Neural Network Approach. In: Proceedings of the International Joint Conference on Neural Networks, San Diego, CA (1990) 215-221.

[4] Tsaih, R., Hsu, Y., Lai, C.C.: Forecasting S&P 500 Index Futures with a Hybrid AI system. Decision Support Systems 23 (1998) 161-174.

[5] Hassan, M.R., Nath, B., Kirley, M., A fashion model of HMM, ANN and GA for stock market Forecasting, Expert Systems with Applications 33 (2007) 171–180.

[6] Li, F., Stormo, D. G., Selection of optimal DNA Oligos for gene expression arrays. Bioinformatics, 17: 1067-1076. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/17/11/1067.

[7] http://www.ncbi.nlm.nih.gov/BLAST

[8] Yu, L., Wang, S., Lai, K. K., Mining Stock Market Tendency Using GA-Based Support Vector Machines, WINE 2005, LNCS 3828, pp. 336 – 345, 2005.

[9] Karp, R., Rabin, M., An efficient randomized Pattern-matching algorithms, IBM Journal of Research and Development, 31 (2): 249–260, 1987.

[10] Rajarajeswari, p., Appear, A., Kiran, K. R., Huffbit Compress – Algorithm for Compress DNA Sequences Using Extended Binary Trees, Journal of Theoretical and Applied Information Technology, 2005.

[11] Sarangi, P. K., Nayak, B. K., Dehuri, S., A Compression- Based Technique for Comparing Stock Market Patterns Behavior with Human Genome, International Journal of Engineering Science and Technology (IJEST), ISSN: 0975-5462