Reveal Motif Patterns from Financial Stock Market

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

Birendra Kumar Nayak
Retd. Prof., Department of Mathematics,
Utkal University, Bhubaneswar, India.

Abstract

In this paper we propose a novel approach for identification of stock motif patterns inform of DNA patterns and study how they behave. As we know the stock market fluctuates (i.e. Day to day increasing and non-increasing of market index) is as natural as the rising and falling down. How many times in a month the stock market index rises or falls or remains steady and if there is any consistency occurs in the months the queries addressed by motif patterns in this paper. The fluctuation 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. Using such sequences, we formulate the motif finding problem as a stock motif pattern search algorithm to generate consensus string which maximizes the consensus score. Here we study the behavior of these consensus strings which can be mapped into the behavior of stock market in day to day basis. Possibility using this similarity to predict the stock market behavior is explored.

Keywords: *Alignment matrix, Consensus Score, Profile matrix, Motif finding problem*

1. Introduction

In financial market, the stock price of a company is influenced by a wealth of internal and external factors. An internal factor may be the perceived potential of the company to be successful in the future (e.g., competent management or ability to generate profit), and an external factor could be the future expectations of a market in which the company operates. There have been several studies addressing the influence of external factors such as news on stock market behavior. We do not restrict our analysis by the assumption that there is a single external factor, such as a news report, effecting stock prices. It is our objective to observe the affects of combinations of external influences that have an impact on the stock prices of two or more companies. Note that we do not attempt to identify the nature of any factors but rather observe their affects. We presume that stocks of two companies may show a similar shape when major influences or economic pressures on these companies are similar. For example, if the future expectations of a particular market (e.g., e-commerce) are very positive (or negative), then the stock time series of companies that operate in this market is likely to show a very similar shape.

We know that a group of stock series shares a pattern over a long period of time, while other stock series show a related pattern over a much shorter interval can provide valuable insights into the price developments of stocks. The relationships between patterns have important information content by themselves. It is our goal to capture the similarities between stock market time series such that their sequence–subsequence relationships are preserved. We identify patterns representing collections of contiguous subsequences that share the same shape for a particular interval [1, 4, 5].

The mining of time series data has gathered much of attention from the research community in the last 15 years. These studies have change in many fields, ranging from biology, physics, medical, financial and stock market analysis, between others. The research in mining time series has been mainly focused in four problems: indexing or query by content, clustering, classification and segmentation. Lately, the problem of mining unusual and surprising patterns has also been enthusiastically studied. Other challenging and recently proposed problem in the context of financial market is to find motif of previously unknown patterns [2, 6]. These patterns, here referred as DNA sequences, consist of subsequences that appear, in a unique and longer sequence or are subsequences that occur simultaneously in more than one sequence from a set of related sequences, here called motifs or sequence patterns. These motifs have a wide range of applications. They can be used in the clustering and classification of stock market [3]. They can also be applied in the generation of sequence rules and in the

detection of interesting behaviors, which can give the user/domain expert valuable insights about the problem that is being studied.

In this work we are interested in the extraction of stock market motifs. We derive a motif pattern by using motif finding algorithm that given as input the symbolic representation of a set of comparable DNA sequences; it finds all the patterns that occur several times in a matrix format. A sequence pattern or motif consists in a set of subsequences that share between them a similarity among the input sequences. Motif sequences we find in year by year and study their similarity among each year.

2. Related work

The problem of discovering previously unknown frequent patterns in time series, also called motifs, has been recently introduced. A motif is a subseries pattern that appears a significant number of times. Results demonstrate that motif may provide valuable insights about the data and have a wide range of applications in data mining tasks. The main motivation for this study was the need to mine time series data from protein folding/unfolding simulations. We propose an algorithm that extracts approximate motifs, i.e., motifs that capture portions of time series with a similar and eventually symmetric behavior. Preliminary results on the analysis of protein unfolding data support this proposal as a valuable tool. Additional experiments demonstrate that the application of utility of our algorithm is not limited to this particular problem. Rather it can be an interesting tool to be applied in many real worlds' problems [6].

Similarities between subsequences are typically regarded as categorical features of sequential data. We introduce an algorithm for capturing the relationships between similar, contiguous subsequences. Two time series are considered to be similar during a interval if every contiguous subsequence of a predefined length satisfies the given similarity criterion. Our algorithm identifies patterns based on the similarity between sequences, captures the sequence–subsequence relationships between patterns as a directed cyclic graph (DAG), and determines pattern conglomerates that allow the application of additional meta-analysis and mining algorithms. For example, our pattern conglomerates analyze time information that is lost in categorical representations. We apply our algorithm to stock market data as well as several other time series data set and show the richness of our pattern conglomerates through qualitative and quantitative evaluations. An exemplary meta-analysis determines timing patterns representing relations among time series intervals and demonstrates the merit of pattern relationships as an extension of time series pattern mining [4].

Time series motifs are pairs of individual time series, or subsequences of a longer time series, which are very similar to one another. As with their discreet analogies in computational biology, this similarity hints at structure which has been conserved for some reason and may therefore be of interest. Since the formalism of time series motifs in 2002, dozens of researchers have used them for diverse applications in many different domains. In this proposed paper, using motif finding algorithm we reveal a new idea to words stock market. Like human pattern likes genomes here we generate a stock market pattern using passed stock values which can be mapped into stock DNA patterns [5]. In Section 3, we formulated DNA sequences and find their motif patterns. In section 4, we the experimental results how stock motif patterns are behave. The ending remarks are given in Section 5 and also we propose future work for prediction of stock market.

3. Proposed Work

In this section we will formulate some DNA sequences and find their corresponding stock market motif pattern which gives a result in next session.

3.1. DNA Sequence Formulation from Stock Market

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 i.e. the non- increasing behavior and "1" means that the next day's index is higher than today's index i.e. the increasing behavior.

Suppose 'V_i' is the value of the i^{th} day and V_{i+1} is the value of the next to i^{th} day.

$C_{ji} = 0$, if $V_i \geq V_{i+1}$

1, if $V_i < V_{i+1}$

Considering average 20 trading days in a month if it is unbalanced then assume that these states are non-increasing days. Similarly for a year we have 240 trading days in place of 365 days. Generate a matrix of 12 rows and 20 columns for each year. Where first column represent first trading day of each month, second column represent second trading day of each month and so on. For each year generate a binary matrix having order 12 rows and 20 columns. Generate DNA sequences of each row using following rules:

"00" for A,

"01" for C,

"11" for T,

"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 [9].

3.2. Motifs and Profile Matrices in Bio-informatics

A motif is denoting to be a small length of code that occurs frequently in a DNA sequence, but it is not required to be an exact copy (i.e. we allow some of the bases to deference between the occurrences). The deference between copies of a regulatory motif cause deference's in regulation rates, which can lead to either beneficial or detrimental behaviors. This is in stark contrast to the restriction enzymes discussed in the digestion problem where cutting the DNA in the wrong place even on rare instances will most likely result in death [7, 8, 10, 12].

Motif Finding Premises:

• Start with a collection of upstream regions and suspect a motif is present

• Locations of the motifs are unknown

• Have an idea of the number of bases included in motif.

• Expect that the strings should look very similar

To analyze the sequences, we align the DNA sequences $\{S_1, S_2, ..., S_t\}$ along the rows of a t× n table called an Alignment Matrix. Each string is positioned so that the rest element in row j is the s^{th} element in string j. From this a 4 × n Profile Matrix can be created by counting the number of times each base {A, T, C, G} appears in each column. The Consensus Sequence is then denned by taking the base with the highest occurrence from each column with the consensus score is denned as the sum of the number of times the consensus base appears in each column. The best possible consensus score is ten while the worst score possible is tn/4. We now want to and the best parole and consensus for the set of t strings [10, 12].

Example: Considering t=12 and n=10

Sl. No.	1	2	3	4	5	6	7	8	9	10
Alignment Matrix	A	A	C	A	A	C	C	A	G	C
	C	A	A	C	G	C	A	C	T	A
	A	C	G	C	T	A	G	C	C	A
	G	C	T	A	A	C	T	A	A	C
	T	G	A	C	A	A	T	C	G	G
	A	C	A	A	C	A	T	T	T	A
	A	A	G	C	A	G	C	C	G	C
	C	G	T	A	G	C	C	G	G	A
	C	C	A	C	T	A	A	C	C	A
	T	T	G	C	C	C	G	C	A	T
Profile Matrix	A	A	T	A	T	A	T	A	G	T
	C	C	A	C	G	C	A	C	T	A
T	5	4	5	5	4	5	3	3	2	6
C	2	1	3	0	3	0	4	1	3	2
A	4	5	1	7	2	6	3	7	2	3
G	1	2	3	0	3	1	2	1	5	1
Consensus Score	A	C	A	A	A	C	T	C	G	A

3.3. Algorithm to reveal Stock Motif Patterns

Generally defined, a motif is a recurring pattern in the sequence of nucleotides or amino acids. In the DNA sequence, it is usually a short segment that occurs frequently, but not required to be an exact copy for each occurrence.

Given a list of t sequences each of length n, find the "best" pattern of length l that appears in each of the t sequences. we solve the Motif Finding Problem using Greedy technique. Here we have chosen starting position from first position of DNA Sequences.

Let s=(s_1,...,s_t) be the set of starting positions for l-mars in our t sequences. Here all si = 1, for all I=1, 2, …t and l=n. That means t number of sequences having fixed length n. The substrings corresponding to these starting positions will form: t x l alignment matrix and 4 x l profile matrix P. Given starting position s, the consensus score is define to be $ConensusScore(s, DNA) = \sum_{j=1}^{n} Mp(j)$

In above example Score(s, DNA) = 5+5+5+5+4+6+4+7+5+6 = 52

Use P-Most probable l-mars to adjust start positions until we reach a "best" profile; this is the motif [12].

3.3.1. Procedure to find Motif

- Select starting positions of each sequence.
- Create a profile P from the substrings at these starting positions.
- Find the P-most probable l-mar a in each sequence.
- Compute a new profile based on the starting position and proceed until we cannot increase the score anymore.

3.3.2. Algorithm

STOCK_MARKET_MOTIF (DNA, t, n)

1. bestScore := 0
2. For i = 1 to t
3. si= 1
4. if (Score(s,DNA) > bestScore)
5. bestScore= Score(s,DNA)
6. bestMotif =(s1,s2,…,st)
7. Return bestMotif

For s, the algorithm calculates Score (s, DNA), which requires O(n) operations since l=n. But in line-2 for loop takes O(t) operations. Thus the overall complexity of the algorithm is evaluated as O(n+t), which is linear time.

4. Experimental Work

We use daily historic data for stocks of the S & P 500 index from http://kumo.swcp.com/stocks/. The obtained data set includes stock prices for an entire year by year from 01/01/1980 to 12/31/2010(thirty years) [11]. The analysis of the stocks are done using the closing values of each day, and excludes all those stocks that have not been a member of the S & P 500 index for the entire year. Each year are having 240 trading days and each month average 20 trading days. Show that each year consists of 12 rows and 20 column matrix. Comparing their stock index (Closing price) one with next day we obtained a binary matrix having order 12X20. Generate DNA sequences for each year using above mapping. These matrices of order 12X10 called alignment matrices. Derive a motif patterns using above algorithm in each year. Similarly collect all consensus string which maximizes the consensus scores are simulated in a MATLAB and study their behavior in each pair of trading days.

4.1 Results

To show the experimental result we have select some years from daily historical data for stocks of the S & P 500 index. Suppose data set includes stock prices for an entire years start from January, 2003 to December, 2008. Using above procedure find their consensus sequences which are defined in below table-1. Next fig-1 represents the behavior of motif patterns by considering Nucleotides A, T, C, G into decimal number (refer table-2) and implemented in MATLAB simulation fig-1 which describes increasing and non-increasing behavior of each trading pair days. In fig-1 X-axis represent pair of trading days and Y-axis represent consensus motif patterns.

Table:1 Consensus motif patterns

Year	Consensus String	Consensus Score
2003	TTTTCAAAAG TTTTCAAATG	50
2004	TCTCCTGTAA TCTCCTGTAT TCTCGTGTAT TCTCGTGTAA	52
2005	TTTTATACTA TTTTCTACTA TTTTATACTG TTTTCTACTG	47
2006	TTCATTATTT TTCATTATTG	46
2007	TAAATCTTTA GAAATCTTTA TACATCTTTA GACATCTTTA	45
2008	TCAAAAGAGC TCTAAAGAGC	48

Table:2 Decimal mapping of Nucleotides

Nucleotides	Binary Mapping	Corresponding Decimal value
A	00	0
C	01	1
G	10	2
T	11	3

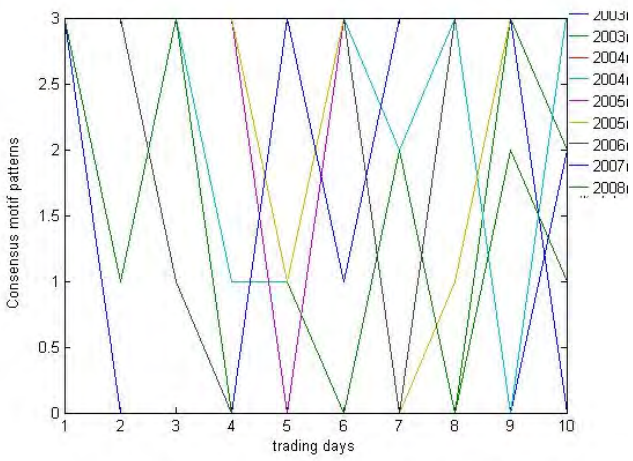

Fig:1 Trading Days vs Consensus String

5. Conclusion and future work

In this work we have formulated the financial stock market into motif finding problem considering day to day closing price. We also a simple stock motif pattern finding algorithm which takes linear time i.e., O(t n) due to fixed starting position, all sequences are start from first position. From above fig-1 we found that in some days some motif pattern are same. We also found that more than one consensus string appears in some years. Show that the consensus score of these strings are same.

To overcome such problems we have to formulate same problem into median string problem, by comparing each of DNA sequences with consensus string and find their hamming distance which is the future work. To minimize the hamming distance and find exact motif pattern who can predict the future aspects of financial stock market.

References

[1] Leung, H., Chin., F., Algorithms for Challenging motif problems, JBCB, 43—58, 2005.
[2] Stormo, G., DNA binding sites: representation and discovery, Bioinformatics, 2000.
[3] Agarwal1, P., Rizvi, S. A. M., Pattern Matching Based Technique to Solve Motif-Finding Problem, BVICAM'S International Journal of Information Technology, Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi, 2008.
[4] Dorr, D. H., Denton, A. M., Establishing relationships among patterns in stock market data, Data and Knowledge Engineering, 318–337, 2009.
[5] Xiaoxi, D., Ruoming, J., Liang, D., Victor, E. L., John, H. T., Migration Motif: A Spatial-Temporal Pattern Mining Approach for Financial Markets, Bibliometrics Data Bibliometrics, ACM, 2009.
[6] Ferreira,P. J., Azevedo, P. J., Silva, C. J., Brito, R. M. M., Mining Approximate Motifs in Time Series, 9th International Conference on Discovery Science, LNCS, 4265, springer-2006.
[7] Sinha, S., A greedy approach to the motif finding problem, Chapter 5.5, CS446.
[8] Sung, W. K., Motif Finding, Lecture 10, combinatorial methods in bioinformatics, CS5238, 2005.
[9] 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 Vol. 4 No.01, 144-147, January 2012.
[10] Jones, N. C., Pevzner, P. A., An Introduction to Bio-informatics Algorithm, The MIT press, Cambridge, London, England, 2004.
[11] http://kumo.swcp.com/stocks.
[12] Lippert, R., Brute Force Algorithms: Motif Finding, Introduction to Computational Molecular Biology, Lecture 2, September 16, 2004