Journal of Universal Computer Science, vol. 16, no. 5 (2010), 852-861
submitted: 2/4/09, accepted: 12/2/10, appeared: 1/3/10 © *J.UCS*

Detecting Market Trends by Ignoring It, Some Days

Jessie Wenhui Zou
(Bioinformatics Solution Inc., Waterloo, Canada
wh_zou@yahoo.com)

Xiaotie Deng
(City University of Hong Kong, Hong Kong
CSDENG@cityu.edu.hk)

Ming Li[1]
(University of Waterloo, Canada
mli@uwaterloo.ca)

Abstract: The last k days of trading together tell the financial market trends. It may be inconceivable if we are told to ignore the 3rd, 6th, and 8th day, *a priori*. We introduce a novel approach to show exactly that — it pays to ignore some fixed days among the recent k days, fixed *a priori*, in order to minimize risk and maximize profit simultaneously. The theory developed here has direct implications to our common senses on how we should look at the financial market trends.

Key Words: optimal spaced seeds, market trend prediction
Category: F.0, J.0

1 Introduction

Accurate detection of market trends of stocks, interest rates or foreign exchange rates leads to profitable investments. Millions of investors, professionals and amateurs alike, rely on common sense, gut feel, or professionally advised market trend indicators. All methods indiscriminately depend on the past k days of market performance. For example, "the market has risen for the last k consecutive days". Or, "most of the last k days were rising so that the price curve has intersected with the average curve", as in the well known moving average method [Pring, 2002]. A smaller k is more sensitive but generates more false buy/sell signals; a bigger k is less risky but also less sensitive.

The problem we are interested in is at the meta level for all methods: Instead of watching all k days, are there patterns that are mathematically more likely to appear in a bull market, but less likely to appear in a bear market, than other patterns?

Such patterns do exist. We wish to show that it is better to pretend that you did not see some of the k days, fixed *a priori*. This way you minimize risk and maximize profit simultaneously. We will introduce a mathematical theory to justify this counterintuitive claim, and validate it by experiments. Our theory depends on a remarkable fact that

[1] Corresponding author.

there are provably good patterns that are inherently and disproportionally more likely to appear in the up market sequences but not in the down markets.

Market trends depend on many other factors such as trade volumes, trader psychology, political events, weather, or even rumors. These are beyond this research. We will strip off all such obstructing factors and study the essence of an observation theory of a time series.

2 Problem simplification, formalization and definition

In order to make precise and clean mathematical statements, we simplify the market movement to a 0-1 sequence, one bit per day, with 0 meaning market moving down, and 1 up. Denote $\mathcal{S}(n,p)$ to be an n day iid (independent and identically distributed) sequence where each bit has probability p being 1 and $1-p$ being 0. If $p > 0.5$, it is an up market; otherwise a down market. Now, the problem is: how do we observe a sequence $\mathcal{S}(n,p)$, not knowing p, and correctly predict $p > 0.5$ (up market) or $p \leq 0.5$ (down market)? For a fixed $\mathcal{S}(n,p)$, the *sensitivity* (*risk*) of a method is defined as the probability the method correctly (falsely) predicts $p > 0.5$ when $p > 0.5$ ($p \leq 0.5$). A true positive observation potentially leads to a profit. A false positive observation potentially leads to a loss.

Your brother-in-law Bill might tell you: "If you see eight 1's in a row, then it is an up market, buy." Your professional account manager Pam may have a different advice: "If you see eight 1's in the past 11 days (thus the price curve meets the average curve), the market is turning up. It's time to buy!" Let's use I_8^{11} to denote Pam's indicator, and I_8 for Bill's.

Bill's I_8 is too conservative for $\mathcal{S}(30,0.7)$, with only 39.7% chance to detect this trend, because 0.397 is the probability 11111111 appears in $\mathcal{S}(30,0.7)$. Pam's I_8^{11} is certainly more sensitive, but it is much too aggressive with 13.9% chance making a false positive prediction in a down market $\mathcal{S}(100,0.3)$, as compared to I_8's false positive rate 0.43% in $\mathcal{S}(100,0.3)$. Can one significantly increase I_8's sensitivity but lower its risk simultaneously? Our mathematical theory of optimized patterns will answer this question positively. We will later give one such example with a success rate of 48.5% in $\mathcal{S}(30,0.7)$ and a false positive rate of 0.32% in $\mathcal{S}(100,0.3)$ simultaneously. In order to compare methods, we must balance two factors: sensitivity and risk.

Definition 1. Let $P_{A,p}$ and $P_{B,p}$ be the probabilities of two different methods A and B having a hit in a region $\mathcal{S}(n,p)$, respectively. We say A is better than B, denoted by $A \succ B$, if for any $p > 0.5$,

$$\frac{P_{A,p}}{P_{B,p}} \times \frac{P_{B,1-p}}{P_{A,1-p}} \geq 1. \tag{1}$$

We say A is uniformly better than B with respect to $\epsilon > 0$, denoted by $B <_\epsilon A$, if

$$\min_{0.5+\epsilon \leq p \leq 1-\epsilon} \{\frac{P_{A,p}}{P_{B,p}}\} \times \max_{\epsilon \leq p \leq 0.5-\epsilon} \{\frac{P_{B,p}}{P_{A,p}}\} \geq 1. \tag{2}$$

According to this definition, any method with higher sensitivity than risk is better than predicting "buy" all the time which has 100% sensitivity and 100% risk. The definition can be equivalently expressed by the following game. A player chooses to bet a number k. He wins k dollars for a correct prediction and loses k dollars for a wrong prediction. Assume that (1) holds. If A bets after B, then for a fixed $p > 0.5$, A can always choose its bet properly to win more money than B in $\mathcal{S}(n, p)$ and lose less money than B in $\mathcal{S}(n, 1-p)$, simultaneously. This cannot be achieved by B even if he bets after A. If $B <_\epsilon A$, then A can always win more money than B in $\mathcal{S}(n, 1-\epsilon \geq p \geq 0.5+\epsilon)$ and lose less money than B in $\mathcal{S}(n, 0.5-\epsilon \geq p \geq \epsilon)$.

3 The mathematical theory of spaced patterns

Bill's I_8 is to wait for pattern 11111111 to appear in $\mathcal{S}(n, p)$. Imagine if there is a super pattern that is much more sensitive than I_8 in $\mathcal{S}(n, p)$ when $p > 0.5$ but not much more (or even less) sensitive when $p < 0.5$, our problem would have been solved. Shorten I_8 to $I_7 = 1111111$ certainly will increase sensitivity in $\mathcal{S}(n, p)$, but this is for all p, actually increasing the risk faster. Let us use the notation 11*11*1*111 to denote a pattern where a "*" represents a day we do not care if it is 0 or 1. This is like we still use I_8, but take 3 days off for vacation. Thus Pam's I_8^{11} is equivalent to looking for any of the $\binom{11}{8}$ patterns with eight 1's and three stars.

Since, $\mathcal{S}(n, p)$ is an iid sequence, all patterns with eight 1's and three stars appear in $\mathcal{S}(n, p)$ with equal frequency, which is precisely $\sum_{i=1}^{n-11+1} p^8 = (n-11+1)p^8$. Bill's 11111111 even has a bit higher frequency at $\sum_{i=1}^{n-8+1} p^8 = (n-8+1)p^8$. It seems that we are not making any progress.

Remarkably, these super patterns indeed exist. A mathematical theory for this has been introduced for the purpose of homology search programs by Ma, Tromp, and Li [Ma *et al.*, 2002], and has been recently extensively studied in the field of bioinformatics [Keich *et al.* 2004, Choi and Zhang, 2003, Buhler *et al.*, 2003, Brejova *et al.*, 2004, Li *et al.*, 2004, Li *et al.*, 2006], and implemented in PatternHunter and (mega)BLAST which are supporting thousands of DNA sequence homology search queries daily. The survey [Brown *et al.*, 2004] contains many more related references. We now extend this theory to market trend detection.

Even with the same number of 1's, not all patterns are created equal. This counterintuitive observation of [Ma *et al.*, 2002] have been further studied in [Keich *et al.* 2004, Choi and Zhang, 2003, Buhler *et al.*, 2003, Li *et al.*, 2004]. Figure 1 shows the remarkable sensitivity comparison of 11111111 and the optimal 11*11*1*111 for $\mathcal{S}(30, p)$, where the x-axis is p and the y-axis is the probability the pattern appears in $\mathcal{S}(30, p)$. The probabilities are computed by a recursive formula given in [Keich *et al.* 2004] in exponential time. The positions of stars are important. A simple swapping of the second 1 with the next star drastically reduces the sensitivity from 0.5168 for 11*11*1*111 to 0.4915 1*111*1*111 by 2.5%, in $\mathcal{S}(30, 0.7)$. The problem of finding the optimal seed is NP-hard [Li *et al.*, 2004, Ma and Li, 2007, Li *et al.*, 2006]. It is easy to verify that $I_8^{11} <_{0.1}$ 11*11*1*111, as 11*11*1*111 wins the betting game against I_8^{11} by betting 4 dollars for each dollar Pam bets. Then 11*11*1*111 always wins more money than I_8^{11} in $\mathcal{S}(n, 1 \geq p \geq 0.6)$ and loses less money than I_8^{11} in $\mathcal{S}(n, 0.4 \geq p \geq 0.1)$. It also can be verified that 11*11*1*111 $\succ$ 11111111. 11*11*1*111 is the optimal pattern. All other patterns with eight 1's and no more than three stars have smaller hit probabilities, with 11111111 having roughly the smallest probability.

Now we see why both Bill and Pam's advices are bad, and why we should ignore the 3rd, 6th, and 8th day.

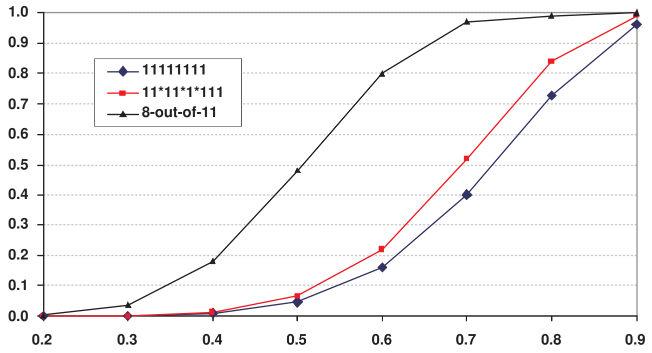

Figure 1: Hit probabilities in $\mathcal{S}(30, p)$ of I_8^{11}, 11*11*1*111, and 11111111.

Pam's strategy can now be viewed as unnecessarily using all $\binom{11}{8}$ patterns of eight 1's and three stars. If we select wisely, it is possible to design a combination of two patterns that achieve significantly higher probability when $p > 0.5$ and significantly lower probability when $p < 0.5$ than 11111111, as shown in Figure 2. 1111*1*1111 and 11*11111*11 combine to have hit probabilities 48.5% in $\mathcal{S}(30, 0.7)$ and 0.32% in $\mathcal{S}(100, 0.3)$, as promised earlier.

It is also possible to design optimal combination of more patterns to approach Pam's sensitivity while remaining at low risk. Figure 3 shows how 3 optimal patterns approach the sensitivity of I_7^9 in $\mathcal{S}(30, p)$ for $p > 0.5$ and still remain at significantly lower risk for $p < 0.5$. Searching for optimal combination of patterns is also NP-hard [Li *et al.*, 2004].

Empirical patterns of market behavior have been widely studied and implemented as computer systems [Pring, 2002, Jobman 1995]. Financial analysts dream of finding patterns that distinguish up and down markets accurately. Our theory, although for a

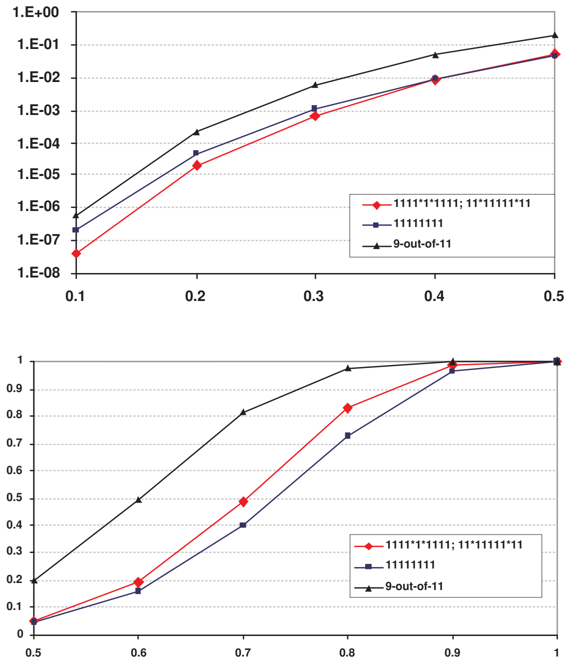

Figure 2: Hit probability of the two weight nine spaced patterns vs one weight 8 consecutive pattern in $\mathcal{S}(30, p)$. The x-axis is p. The y-axis is hit probability. The left figure is in logarithmic scale for clearer display.

simplified model, provides mathematically justified good patterns.

4 Experiments

Single indicator systems based solely on market price are inferior to those with multiple indicators using comprehensive market information including, for example, trade volumes, seasonal fluctuations, and historical patterns. Nevertheless, the intention for our two simple experiments here is to demonstrate our theory that gapped obervation is better than consecutive observation.

We first consider an artificial example. Suppose a market is modeled by a 3-state hidden markov model described in Figure 4. The UP state represents a raising market which generates a 1 with 0.7 probability and a 0 with 0.3 probability, independently. The DOWN state represents a declining market which generates a 1 with 0.3 probability and a 0 with 0.7 probability, independently. The EVEN state represents a leveled

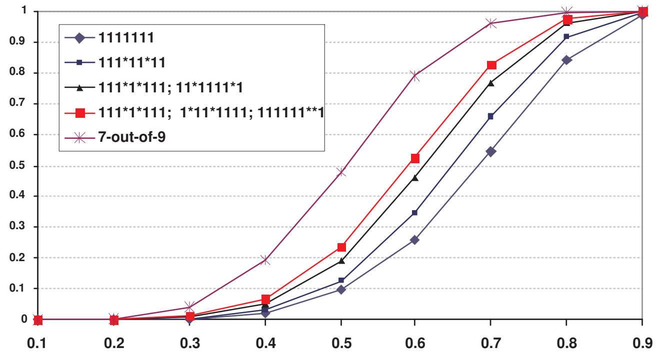

Figure 3: Hit probabilities of 1,2,3 optimal patterns vs I_7^9.

Table 1: The average of 250 simulations

	R	# Hits	Final MTM	Min MTM	Max MTM	# Bankrupts
I_7=1111111	30	12	679	84	719	16
I_7^{11}	15	47	916	76	959	14
5 Spaced Seeds	25	26	984	83	1027	13

market with 1 and 0 at 0.5 probability independently generated. With some low probability a state transfers to another, as indicated in the transition probabilities in Figure 4. Let the hidden markov model generate a sequence of 1's and 0's of 5000 days, containing short "even" regions, long "down" regions and slightly longer "up" regions, representing a generally upward market. We compare I_7^{11}, I_7, and 5 random spaced seeds (111*1*1**11, 1**111*11*1, 11***111*11, 1*11**1*111, 11*111*1**1). We did not use optimal patterns as it takes too much time to find them. Each strategy starts with \$100 and fixes a bet R. The goal is to maximize mark-to-market (MTM) and minimize number of bankruptcies simultaneously. For simplicity, every time there is a match, we buy, and then sell precisely after 5 days. Let W be the number of ones during these 5 days and $L = 5-W$ the number of zeroes. The reward will be $R \times (W-L)$. Fixing the bets, Table 1 averages 250 simulations. The not-even-optimal 5-spaced-seeds strategy is the clear winner, making more money and having fewer bankruptcies than both I_7 and I_7^{11}.

The second experiment compares the simple moving average crossover method, I_7, I_7^9, one optimal spaced seed, and two optimal spaced seeds, using the actual S&P 500 Index data. We chose I_7^9 because I_7 is the best performer among the I_k indicators. Historical data of S&P 500, from Oct 20, 1982 to Feb. 14, 2005 (which is the date when

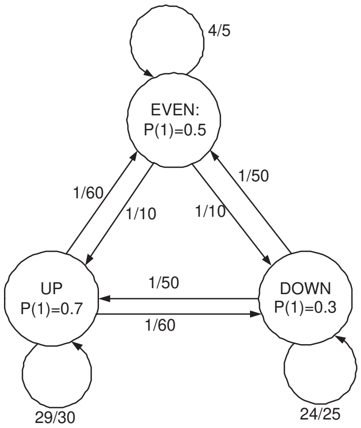

Figure 4: The HMM used to generate the market trend sequence.

this experiment was performed) were downloaded from Yahoo.com. A 0-1 sequence was generated, one bit per day, with 1 at the k-th day if the price at the k-th day is greater than at the $k-1$st day, and 0 otherwise. No other manipulation is performed on the data. The initial investment is $10,000. All profits are reinvested. Commission costs and taxes were ignored.

We employ simple and single trading strategies in order to make clean comparisons. Optimization by combining multiple indicators is not our goal here. For the simple MA crossover method, we have tested n-day moving averages, for $n = 10, 11, \ldots, 400$. When n is around 250 (12 months), as also used in [Pring, 2002], the profit is maximized. We have also tested using k-day averages, for $k = 1, 2, 3, ...22$. The profit is maximized when $k = 1$. For the I_k indicator group, when $k > 8$ there is no hit (not trade); as k goes smaller than 7, the results deteriorate; when $k = 7, 8$ the results are the best. For the I_{L-2}^{L} indicator group, L_{11}^{13} is the best policy for S&P 500 index, at $71,000, but it performed poorly for the NASDAQ index. Since L_{11}^{13} is not as good as simple optimal seed anyways, we choose to present I_7^9 in Table 2 because it is a more balanced indicator. On the other hand, our optimal seeds are unique, taken directly from Figure 3, and independent of the S&P 500 and NASDAQ market data. For pattern-based methods, a simple strategy is uniformly applied: buy when a subsequence of 1's matches the given pattern(s); and sell when a subsequence of 0's matches (the ones in) the given pattern(s). We did not do short sales which would further improve the profit. The result is given in Table 2. The buy-sell pattern for one optimal seed is in Figure 5. Its last buy was on June 5, 2003 at 990.14. At that point the S&P index had grown $990.14/139.23 = 7.1$ times. Our asset had grown $74,698/10,000 = 7.47$ times. For comparison, trading was also simulated on the NASDAQ index (downloaded from Yahoo.com), using the same

Table 2: Comparison of the simple moving average (MA) crossover method with the spaced seed methods, using the actual S&P 500 index and NASDAQ index data.

Trading Details	Trading Indicators				
	12 Month MA crossover	I_7^9	I_7= 1111111	1 Optimal Seed 111*11*11	2 Optimal Seeds 111*1*111, 11*1111*1
Initial Investment	10,000	10,000	10,000	10,000	10,000
S&P 500	20-Oct-82: 139.23 to 14-Feb-05: 1206.4				
Mark-to-Market	68,923	29,384	32,343	74,689	80,582
# Trades	43	51	3	8	10
# Trades with Profit	12	25	2	7	8
# Trades with Loss	31	26	1	1	2
Avg Gain per $1,000 per trade	29.4	18.8	309.3	210.0	178.8
NASDAQ	2-Jan-85: 353.20; 3-Jan-05: 2152.15				
Mark-to-Market	88,436	104,208	110475	111,105	144,496
#Trades	41	73	32	18	22
# Trades with Profit	15	40	19	13	17
# Trades with Loss	26	33	13	5	5
Avg Gain per $1,000 per trade	51.3	24.7	68.6	125.2	126.6

parameters. As the NASDAQ index is more chaotic, we averaged each 2 consecutive days when the 0-1 sequence is produced.

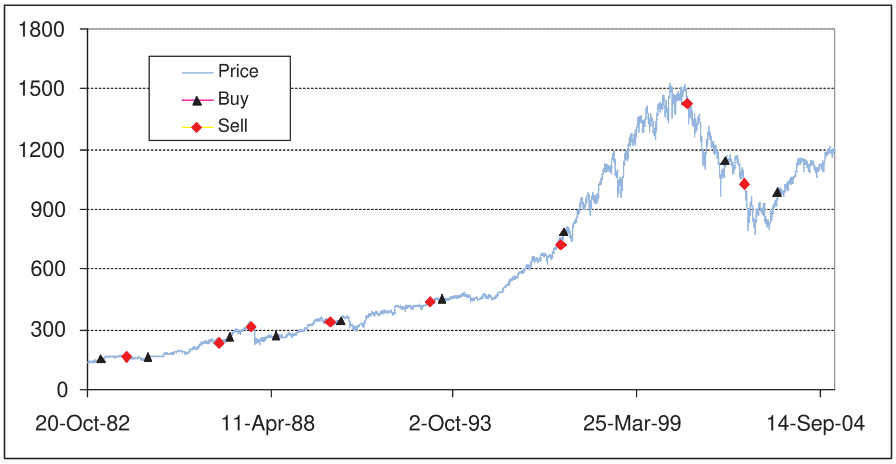

Figure 5: The buy-sell points of 111*11*11 on S&P 500 index.

5 Discussion

We have presented and validated a mathematical theory of gapped observation of market movements. We have shown that some patterns are disproportionally more likely to appear in a bull market and less likely in a bear market, ideal for trade indicators. Our patterns were optimized against the iid sequences and worked reasonably well on the real data.

One may wonder if one could achieve similar effects by optimizing a function or a support vector machine. This can be done, based on the same principle of our theory, as now we know what to optimize. But that's missing the point. The point here is to initiate an observation theory study on such real-time-market-decision time series. We have provided a simple intuitive method for this problem, with a few patterns that can be used by commonplace traders.

While we have demonstrated that even a simple minded approach works well, the 0-1 market sequence can certainly be generalized. The levels of ups and downs and can be more accurately modeled. When there is an underlying statistical model of data, it is possible to optimize patterns relative to the data model [Brejova *et al.*, 2004, Buhler *et al.*, 2003]. The independence assumption of day-to-day market can be generalized by a hidden markov model [Brejova *et al.*, 2004] to model for example the Friday effect. Many market trend indicators potentially can be adapted using the gapped observations. For example, our method tells the weighted moving average which day should be weighted more. While no method will definitely make money, we have shown that our gapped observation method, although implemented in a most rudimentary way, does help to make more money and lose less money. In principle, the theory is adaptable to other observable time series such as the up and down of the trade volumes. The theory also extends beyond financial market analysis to an observation theory of any time series.

Acknowledgement

We would like to thank Dan Brown, Bin Ma, John Tromp, and Louxin Zhang for the enlightening discussions and the three referees who have made very useful comments. This work has been performed during 2004-2005 when the first author and the third author were visiting the City University of Hong Kong and we thank Frances Yao for creating the research environment for us; during this period, we have enjoyed visiting Derick Wood at HKUST. This work has been partially supported by the City University of Hong Kong, NSERC Grant OGP0046506, 863 Grant 2008AA02Z313 from China's Ministry of Science and Technology, Canada Research Chair program, MITACS, an NSERC Collaborative Grant, and Premier's Discovery Award, as well as a CityU SRG grant (Project No:7002308).

References

[Brejova *et al.*, 2004] Brejová, B., Brown, D.G. and Vinař, T. "Optimal spaced seeds for homologous coding regions". J. Bioinf. Comput. Biol. 1:4(2004), 595–610.

[Brown *et al.*, 2004] Brown, D.G., Li, M., and Ma, B., A tutorial of recent developments in the seeding of local alignment", J. Bioinf. Comput. Biol. 2:4(2004).

[Buhler *et al.*, 2003] Buhler, J., Keich, U. and Sun, Y., "Designing seeds for similarity search in genomic DNA", Proceedings of the 7th Annual International Conference on Computational Biology (RECOMB), pp. 67-75.

[Choi and Zhang, 2003] Choi, K.P. and Zhang, L. "Sensitive analysis and efficient method for identifying optimal spaced seeds", J. Comput. Sys. Sci., 68(2004), 22–40.

[Jobman 1995] Jobman D.R. (ed): "The handbook of technical analysis – a comprehensive guide to analytical methods, trading systems and technical indicators", Irwin Professional Publishing.

[Keich *et al.* 2004] Keich, U., Li, M., Ma, B. and Tromp, J. "On spaced seeds of similarity search", Discrete Appl. Math. 138(2004), 253–263.

[Li *et al.*, 2004] Li, M., Ma, B., Kisman, D. and Tromp, J., "PatternHunter II: highly sensitive and fast homology search". J. Bioinf. Comput. Biol. 2:3(2004), 417-440.

[Li *et al.*, 2006] Li, M., Ma, B. and Zhang, L.X. "Superiority and complexity of the spaced seeds." Proc. 17th ACM-SIAM Symp. on Disc. Alg. pp. 444-453.

[Ma and Li, 2007] Ma, B. and Li, M.: "On the complexity of spaced seeds", J. Comput. Syst. Sci. 73(2007), 1024-1034.

[Ma *et al.*, 2002] Ma, B., Tromp, J. and Li, M. "PatternHunter: faster and more sensitive homology search", Bioinformatics, 18:3(2002), 440–445.

[Pring, 2002] Pring, M.J. "Technical analysis explained. The successful investor's guide to spotting investment trends and turning points" 4th Edition, McGraw Hill Inc. 2002.