Paper Category: genetic programming

The Impact of Training Period Size on the Evolution of Financial Trading Systems

Thomas Fernandez and Matthew Evett
Department of Computer Science and Engineering
Florida Atlantic University
Boca Raton, Florida 33431
(561)367-3459
{tfernand, matt}@cse.fau.edu

Abstract

There are many systems for optimizing and selecting between existing investment timing strategies, but there are generally few methodologies for finding new strategies that are successful. This paper investigates techniques concerning the use of GP to develop investment strategies, and in particular, the relationship between training set sizes and the profitability and risk of the strategies derived therefrom.

1 Introduction

Discovering an investment timing strategy that is successful is an important problem to the financial community. An investment timing strategy is a system that decides when to buy or sell a given item. That item is usually a bond, a stock, an option or some other financial instrument. The price of these financial instruments fluctuates over time. The goal of an investment strategy is to buy a financial instrument before it's price increases and to sell it before it's price decreases.

There are many systems for optimizing and selecting between existing investment timing strategies, but there are generally few methodologies for finding new strategies that are successful. Investment timing strategies can be precisely expressed as computer programs, and therefore genetic programming (GP) is a promising methodology for discovering new and profitable strategies. This paper investigates techniques concerning the use of GP to develop investment strategies, and in particular, the relationship between training set sizes and the profitability and risk of the strategies derived therefrom.

2 Problem Description

Chen and Yeh, in [1], use GP to study the Efficient Market Hypothesis (EMH) on the U.S. stock market, and on the S&P 500[1] in particular. Chen and Yeh use GP to see if functions can be evolved to predict future changes in the S&P 500 using the values of the S&P 500 in the recent past as input. The existence of such a function would be evidence against the EMH, and the failure of the GP to find such a function would be evidence for the EMH.

In our work, we use GP to evolve investment strategies, which are basically functions that take the values of the S&P 500 and more than 20 other financial time series as input, not just to predict the S&P 500, but actually to

[1] S&P 500 is an average of the stock price of 500 large capital companies selected by economists, and weighted by value, which is used by the financial community as a measure of the US Stock Market.

drive a trading strategy with buy and sell signals. The overall profitability of a strategy is determined by applying it over a length of time, and calculating the resulting gain or loss. (This is explained in more detail in Section 3.1.)

As Koza did in his study of financial time series in chapter 10 of [5], Chen and Yeh use a portion of an S&P 500 time series as training data for the individuals of the GP. That is, the training data is used in calculating the fitness function. After a GP run is completed, the (best-of-run) individual returned by the GP is then tested against a subsequent portion of the S&P 500 time series. The training and testing data are disjoint to ensure that the derived prediction function does not implicitly take advantage of *a priori* knowledge of how the time series will perform over the test period.

Chen and Yeh compared the use of long and short training periods in generating predictor functions. The basic result of this comparison was to demonstrate that the use of relatively small training periods tended to yield predictor functions with performance that was better *on average* than those yielded by larger training periods. On the other hand, the comparison also showed that the difference between the performance of the best and worst of a typical set of small-training set predictors was greater than in a typical set of larger-training set predictors.

In this paper we study whether training period size yields an analogous difference in the quality of investment strategies generated by GP. That is, do smaller training sets yield investment strategies that are more risky, but which could offer potentially higher returns (and potentially greater loss), while larger training periods yield investment strategies that are less risky, but offer lower average return?

3 Setting up the GP

Chen and Yeh used the S&P 500 daily time series as their only training data because of their interest in studying the Efficient Market Hypothesis. We used many additional financial time series in the hope of increasing the performance of the investment strategies resulting from the GP. With large training periods the GP uses 5 years of data from 23 financial time series to evolve a strategy for trading the S&P 500. After the Strategy is evolved, it is tested against an additional two year period to see if the strategy is robust and is successful outside of the training period. Medium training periods consist of 150 days of data (from the same financial time series) and the resulting strategies are tested against 60 days of subsequent data. Small training periods consist of 50 days of data, and the resulting strategies are tested against 20 days of data.

The GP used twenty three inputs (terminals), and ephemeral random numbers. These inputs were:

1. The relative change in the S&P 500 from one day ago.
2. The relative change in the S&P 500 from five days ago (one week).
3. The relative change in the S&P 500 from twenty one days ago (one month).
4. The relative change in the S&P 500 from one 63 days ago (one quarter).
5. The relative change in the S&P 500 from one 252 days ago (one year).
6. The relative change from today's S&P 500's open and close.
7. The relative change from today's S&P 500's low and close.
8. The difference between today's S&P 500's high and low.
9. The relative change in the American Index from one day ago.
10. Today's volume of the American Index.
11. The number of new highs on the NY Exchange.
12. The number of new lows on the NY Exchange.
13. The Ratio of Advancing Issues to Declining Issues on the NY Exchange.
14. The relative change in 30 day CD's from 21 days ago (one month).
15. The relative change in 6 month CD's from 63 days ago (one quarter).
16. The relative change in Fed. Funds rate from 63 days ago (one quarter).
17. The relative change in London Gold Price from 63 days ago (one quarter).

18. The relative change in Dow Commodity Futures Index from 21 days ago (one month).
19. The relative change in Dow Commodity Spot Index from 21 days ago (one month).
20. The relative change in Dow Industrial Average from 5 days ago (one week).
21. Today's volume of then Dow Industrial Average.
22. The relative change in Dow Transportation Average from 5 days ago (one week).
23. Today's volume of then Dow Transportation Average.

The GP function set includes the basic arithmetic functions (addition, subtraction, multiplication and protected division) plus three others:

1. $DELTA\%(a, b)$ — defined as $(b - a)/a$ if $a \neq 0$, else 100.0, representing the relative change of b to a.
2. $AVE(a, b, c)$ — defined as $(a + b + c)/3$, representing the average of a, b and c.
3. $IFGTZ(a, b, c)$ — defined as b if $a > 0.0$, else c, and representing conditional evaluation.

All terminals and the inputs and results of the functions are double precision real numbers. The final resulting real number is considered a buy signal if it is greater than zero and a sell signal otherwise.

3.1 Fitness function and termination condition

As with most GP problems, the design of the fitness function is rather complicated. The overall goal of the fitness function is to measure the relative success of the investment timing strategies. There are many factors that can be considered in determining the success of a strategy. These include the ratio of risk to return relative to other investments, as well as the system's potential for draw-downs or volatility. These factors might be built into the GP's fitness function. In this research the fitness function is simply the net result or profitability of all the buy and sell transactions.

The goal of a GP investment strategy is that it be profitable in periods of time outside the training period. We believe that lower volatility in a strategy's profitability during its training period indicates a higher probability that the strategy will be applicable outside of the training period. We intended the termination criteria to encourage the GP to evolve strategies with lower volatility. To this end, the training period is divided into sub-periods, and the termination criterion is that system be profitable (or at least have no losses) consistently over all sub-periods. To help the GP reach the termination criteria, the fitness function exaggerates the negative value of periods of loss by multiplying those losses by four.

The system uses the buy and sell signals to remain fully invested at all times in either a long or short position. Both long and short positions are allowed so that the system is completely symmetrical, meaning that it can profit equally from increases and decreases in the S&P 500. A long position simply means that all available assets are fully invested in the S&P 500. A short position is a little more difficult to explain. Theoretically it is the equivalent of paying negative money for negative shares. It can also be understood as borrowing shares of the S&P 500 and selling them immediately. Later when the short position is closed or liquidated the number of shares that were borrowed must be bought at the current market price and returned to the lender. Thus any drop in the price results in profits and any increase in the price during the time of the short position will result in losses. It is not the intention of this paper to create a practical trading system, but merely to investigate the use of GP's in this area. Mechanisms for shorting the S&P 500 may not be readily available to individual investors, and may even involve derivatives or other mechanisms for institutional investors, but are used as part of this research because we believe that a non-symmetrical system would adversely impact the study.

Population size	2000
Percentage of individuals created by ramped complete growth	100
Percentage of individuals created by ramped partial growth	0
Crossover Percentage	90
Mutation Pertcentage	0
Max Number of Runs	10
Max Number of Generations	75
Max Nodes per Tree	200
Selection Strategy	Graduated Elitist
Initial Tree Minimum Depth	3
Initial Tree Maximum Depth	7
RandomSeed	0

Table 1: The Tableau used to Develop Investment Strategies.

For large training sets, the raw fitness score is calculated by totaling the profit or loss from each fiscal quarter, and then dividing by the total number of years, to produce an annualized return. Regardless of profit or loss from previous quarters, the system starts each quarter with $100,000.00. When the system's signal changes to *buy*, all previous positions are liquidated and all cash is used to buy shares of the S&P 500 Index. When the system's signal changes to *sell* all previous positions are liquidated and the number of shares of the S&P 500 Index that could be bought with the available cash are borrowed and sold. (This is referred to as *short selling.*) At the end of each quarter all positions are liquidated and the difference between the final cash position and the original $100,000.00 is returned as the score for that quarter. As previously mentioned, to encourage systems that avoid periods of loss, and so are consistently profitable, any quarterly losses are multiplied by four. For medium training sets, thirty day periods are used in place of quarters. For small training sets, ten day periods are used.

The reason for starting each subperiod with $100,000.00 is that it allows each to be evaluated independently, and avoids excessive exponential rollover of profits. No consideration is made for margins, slippage or interest charges. A commission charge of $25 (round trip) is added for each change of position, to discourage excess trading. Each subperiod that has no loss is considered a hit. A run is halted after it reaches 20 hits or 75 generations. The tableau for the experiments is shown in Table 3.1.

We used an *elitist graduated overselection strategy* to select individuals from the population for reproduction and crossover. The strategy starts by sorting the population by fitness and placing the top five individuals into a pool of candidates for the next generation. Individuals are selected from this pool of candidates with equal probability. After each candidate is selected, the next fittest individual is added to the pool of candidates. The pool of candidates continues to grow until all members of the current generation are included. Thus all elements of the population have some chance of being selected for the next generation, but the fittest members have many more chances while the least fit elements will have few chances and even those few chances consist only of an opportunity to be selected from a large pool of candidates.

This research was implemented using G-Framework[3], an object oriented genetic framerwork written in C++ for building GA and GP systems. The basic classes that make up G-Framework are the Problem class and the Population class. Spesific problems, such as the discovery of investment strategies are classes that inherit from the abstract base class Problem. The Population class contains an array of Element Objects. If the system is being used to implement a GP system, the elements include a pointer to the root of a tree of S_Expression_Node

Run#	Generation	TestScore	THits	Score	Hits	Runtime (mins.)
10	18	6574.22	4	26711.60	20	151.58
5	4	6095.32	5	18899.00	20	21.47
9	11	5571.59	5	23119.50	20	66.38
1	8	3877.09	5	23059.40	20	54.95
6	10	3385.01	4	23596.20	20	54.82
2	32	596.80	5	32741.80	20	282.77
4	12	351.04	4	16518.40	20	77.48
3	18	-32.95	4	27760.20	20	126.42
8	9	-1104.68	4	20258.20	20	80.52
7	19	-5168.67	2	31003.90	20	117.33

Table 2: Performance of the best individuals of 10 runs of the GP, using a large training period.

Objects. Leaf S_Expression_Nodes have pointers to Terminal Objects and non-leaf S_Expression_Nodes have pointers to Function Objects. When used to implement GP's the system is similar to Koza's Simple Lisp code [5] ADF's are not implemented in the current version of G-Framework, but they may be included in future releases. For more details on the analaysis and design of G-Framework see [3]

4 Results

To test the effect of training period size on the profitability of our GP-derived investment strategies, we ran the GP using three different training set sizes, as defined in Section 3.1 (*large* being 1265 data points, *medium* being 150 data points and *small* being 50 data points). We made ten runs using large training sets, and three sets of ten runs using both the medium and small training sets. All runs used populations of size 2000.

4.1 Using large training sets

The total run time to complete the 10 runs using large training sets was a little over 17 hours on an Intel 486 based system. The average number of generations before reaching 20 hits was 14.1. All runs reached 20 hits long before the cutoff of 75 generations. The results of the 10 runs are shown in Table 3.1, which is sorted by *TestScore*. *TestScore* is the result of running the best-of-run investment strategy over the two year test period from 1993 to 1994 after the evolution process is completed. In particular, *TestScore* is the annualized return, in dollars, on a $100,000 initial investment. Note that during the test period, quarterly losses are not multiplied by 4, as is done during the training period from 1988 through 1992. *THits* represents the number of profitable quarters out of the eight quarters in the two year testing period.

The best result during the test period was 6574.22 or an annualized return of 6.57%. This may seem unimpressive to investors who have seen the S&P 500 increase by over 20% a year in resent years, but, as shown in the Figure 4.1, during the test period of 1993 to 1994 the S&P 500 was relatively flat going from 435.38 to 459.11 for an annualized return of only 2.7%. This strategy developed by the GP actually represents almost two and a half times better return than the S&P 500.

Figure 4.1 shows the results of each generation during the evolution of this strategy. The *Best* curve represents the best individual's score in each generation. The *Median* curve represents the fitness score of the individual that

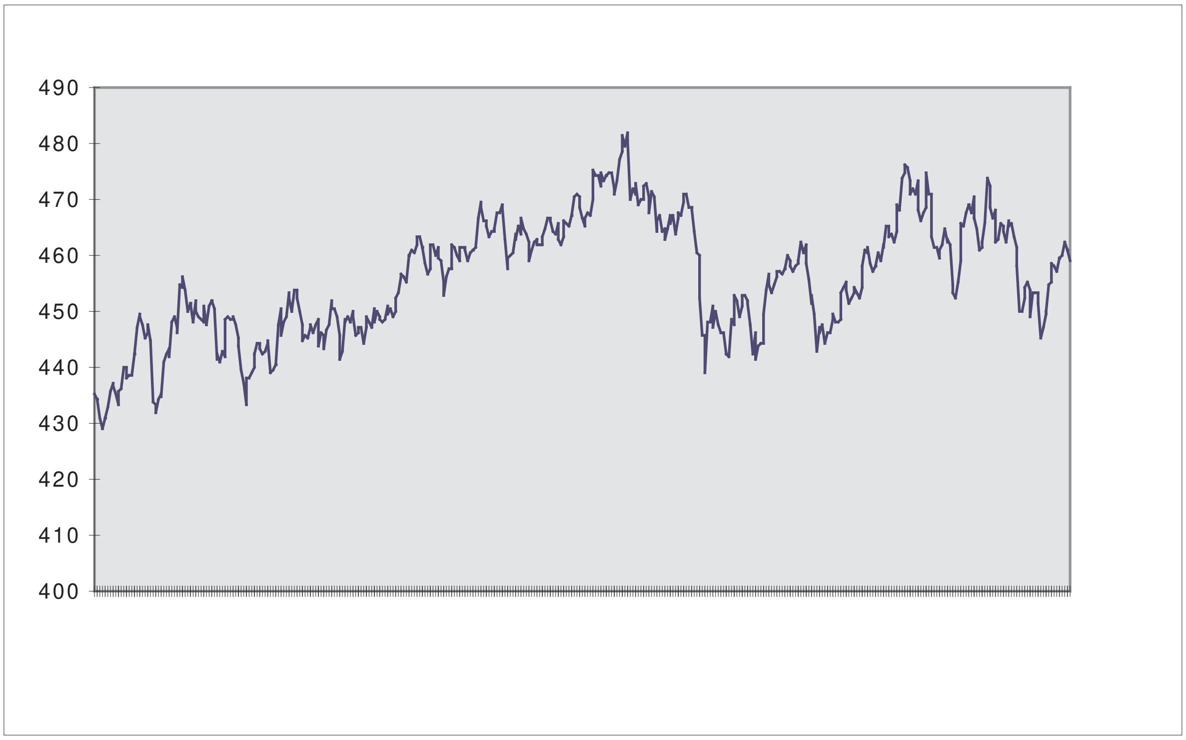

Figure 1: The S&P 500 over the range of the large training period (1993-94).

ranked 1000 out of the population of 2000. After each generation, we ran the best-of-generation individual on the two year test period. The *Test* curve is a plot of the results.

The similarity of the *Test* and *Best* curves indicate that an increase in the performance of the best individuals on the training data generally lead to a corresponding increase in the performance of those individuals on the test data. This is an encouraging sign that the GP is evolving strategies that are robust and can be extended to periods of time outside the training period.

The actual representation of the best strategy during the testing period is below. Note that this is represented with infix notation and is not a LISP expression.

```
AVE(AVE(DELTA%(DELTA%(DELTA%(AmerVol, DQ%CD6), (33.252 / DQ%FedFunds)), (AVE(D5%SPClose,
D5%DowT, -90.8142) + (NYNuHi + NYNuHi))), (DELTA%((-72.0398 + DM%CD1),
(58.606-4.11987))-((DY%SPClose * 25.9583) * (57.8735 / DowIVol))), (AVE(AVE((((DQ%LGold /
47.113)-AVE(DM%CIF, -25.9888, D5%DowI)) * (DELTA%(75.4822, DM%CIF) * AVE(DM%CD1, D5%DowI,
-2.63672))), IFGTZ(((DowTVol / DowTVol) + AVE(DowIVol, DY%SPClose, -33.2275)),
((47.6135-D1%Amer)-DELTA%(SPHi-SPLo, DowIVol)), (IFGTZ(24.2981, -19.7327,
DM%CIS)-DELTA%(-56.7261, DQ%FedFunds))), (((DowTVol-NYNuLo)-DM%SPClose) /
IFGTZ(DELTA%((DELTA%((92.5476-DM%SPClose), (74.5056 * DM%CD1)) + DELTA%(IFGTZ(0.65918,
D1%SPClose, DQ%SPClose), DELTA%(79.4373, NYADRat))), -73.4009), (DQ%LGold / -8.47778),
AVE(-43.6707, NYADRat, 53.9063)))), AVE(-60.7056, 84.906, 36.2366), (-40.9729 + 56.7322)) +
AVE(DELTA%(-67.3706, 86.8408), IFGTZ(DQ%CD6, -19.5862, -58.5815), AVE(D5%DowT, 15.8875,
59.9915)))), IFGTZ((DELTA%((92.5476-DM%SPClose), (74.5056 * DM%CD1)) + DELTA%(IFGTZ(0.65918,
D1%SPClose, DQ%SPClose), DELTA%(79.4373, NYADRat))), AVE(DQ%CD6, (AVE(DM%CIS , NYADRat,
73.2849) / (D1%SPClose * -55.072)), ((DM%SPClose + 42.3401) -DELTA%(D%SPLo:SPClose,
-74.5789))), DELTA%(IFGTZ((-47.2473 / DQ%LGold), AVE((DowTVol-NYNuLo), -14.8926, DM%SPClose),
```

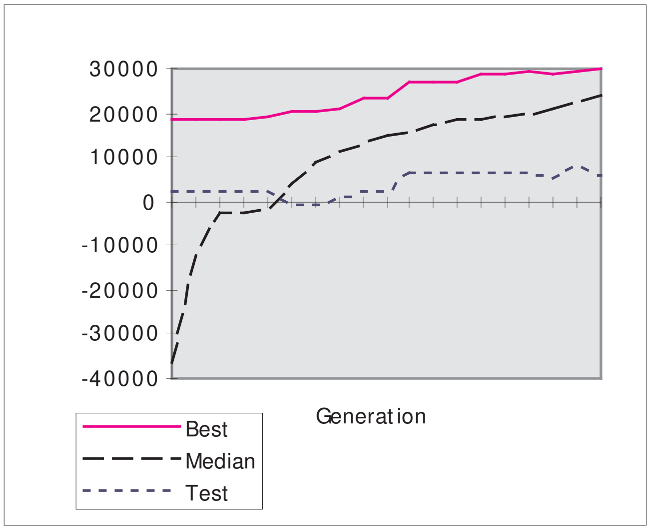

Figure 2: Performance of the GP by generation, using large training sets.

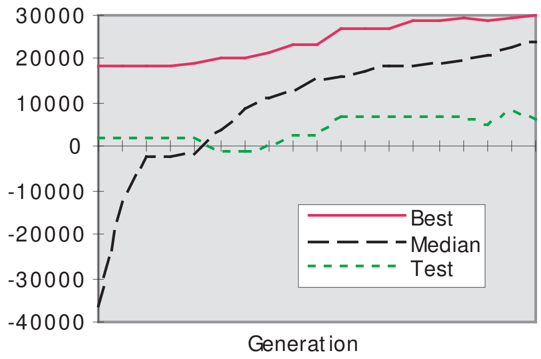

Figure 3: Performance of the GP by generation, using small training sets.

```
D5%SPClose), DELTA%(DowTVol, -3.38745))), ((27.3743-DQ%SPClose)-AVE(-87.1155, 16.1194,
7.44019)))
```

4.2 Using small training sets

We ran 3 sets of 10 runs using both medium and small training sets. The results of one of sets of 10 runs using small training sets is shown in Figure 4.2. The performance of the GP process by generation is shown in Figure 4.2.

The average profitability during the test period for short, medium and long training periods is shown in Figure 4. The results indicate that the use of a medium sized training period is more profitable on average than larger ones. It also indicates that short training periods do not provide the GP with the information needed to produce stratagies that are, on average profitable. An examination of the Standard Deviation of the test period results across these different training period lengths, shown in Figure 5, indicates that the longer training period, while less profitable on average, also entails less risk. This is also confirmed by examining the worst cases for the three training period lenghths. The worst case for the large training period was only an annualized loss of 5.2% but the worst case losses for short and medium training periods was 58.8% and 51.5% respectively.

Run#	Generation	TestScore	THits	Score	Hits	Runtime (mins.)
3	19	21242.8	1	76114.7	5	5.05
10	11	-5877.04	1	64288.9	5	2.68333
1	18	-15692.3	0	71440.1	5	4.7
9	15	-15692.3	0	68848.9	5	3.76667
5	21	-20963.6	0	72289.1	5	6.05
7	30	-30737.2	0	88412.5	5	11.9667
6	20	-35953	0	78050.9	5	6.55
8	18	-38104.5	0	68219.6	5	6.2
4	27	-39337.7	0	64478.4	5	11.25
2	22	-58849.4	0	85560.9	5	6.58333
Average:		-23996.4				
Range:		80092.2				
Std. Dev.		22008.92				

Table 3: Performance of the best individuals of 10 runs of the GP, using a small training period.

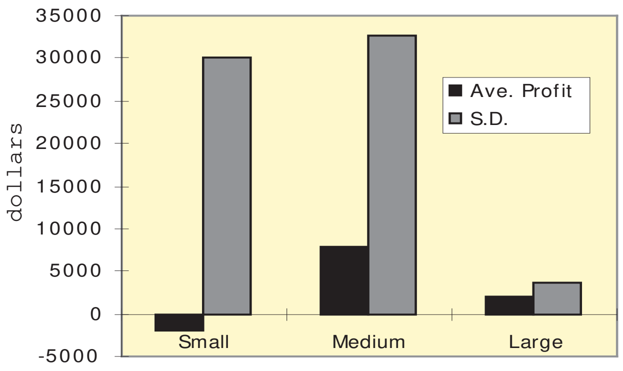

Figure 4: Comparison of the profatability versus training period size.

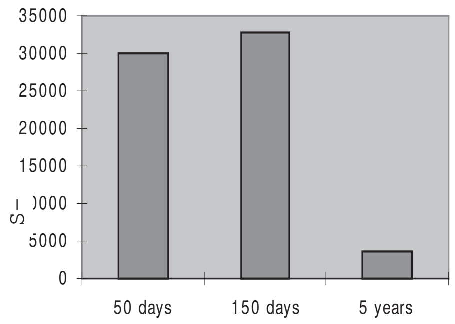

Figure 5: Comparison of the volatility versus training period size.

5 Conclusion

It is clear that the problem of discovering investment timing strategies for the S&P 500 is still an open problem. Large and multiple noise components and the seemingly chaotic nature of financial markets make the problem of market timing very difficult, but it's enormous impact on human activity and it's potetial rewards make it a problem that will not be ignored. Undoubtably it will continue to be worked on using a wide variety of techniques. We believe that it is the countinously changing nature of this problem, combined with a very large search, that makes GP an esential tool for working on Investment Timing Strategies in all Financial Markets. Darwin compares man made objects with objects created by the evolutionary process in *Origin of Species* and points out that evolutionary objects "should be infinitely better adapted to the most complex conditions of life"[2].

While this research has not closed any problems, it is our hope that we have opened some new ones and can suggest some directions for further research. In this work we preselected and preproccessed the financial time series that were used as inputs in our terminal set. Given suffiecent computational resourses perhaps the GP could be designed to select financial time series from very large sets that were not preproccessed and thus discover new econometric relationships as opposed to explointing known ones. Further along this line it would be interesting to see if the GP system could select, not just the financial time series used for inputs, but also the financial time series that it can profitably predict.

An additional area of research suggested by this work and by the work of Chen et al [1] is the determination of the optimal length of training period or combination of training periods. Perhaps the GP could determine for itself what is the optimal length of the training period. This introduces the curious question of how large a training period should be used to determine the size of the training period.

One area of research that we believe shows promise is the use of GP to develop not one investment strategy but many strategies that would combine and contine to evolve as an emergent system. Kevin Kelly states in his book *Out of Control*, "constructing a system that can adjust to new stimuli, or to change beyond a narrow range, requires a swarm — a hive mind."[4]

It would also be interesting to see GP's used to generate functions for preproccessing input data for financial neural networks.

6 Note to Reviewers

The final version of this paper will include results of additional experiments involving larger training periods. This will strengthen the Results section of the paper. (See, especially, Figures 5 and 4.)

References

[1] S. Chen and C. Yeh. Genetic programming and the efficient market hypothesis. *Genetic Programming 1996: Proceedings of the First Annual Conference*, 1996.

[2] C. Darwin. *On the origin of species by means of natural selection*. 1859.

[3] T. Fernandez. Analysis and design of the evolution of strategy objects. Technical report, Florida Atlantic University, 1997.

[4] K. Kelly. *Out of Control: the new biology of machines, social systems, and the economic world.* Addison-Wesley, 1994.

[5] J. Koza. *Genetic programming: on the programming of computers by means of natural selection.* MIT Press, 1992.