新的基因表達規劃法架構在時間序列問題之應用

林文修
輔仁大學資訊管理學系助理教授
wslin@im.fju.edu.tw

蘇渝翔
輔仁大學資訊管理學系
include2md@gmail.com

摘要

時間序列(Time Series, TS)的研究目的，是透過歷史紀錄訓練而成的模型去預測未來的趨勢。根據人工智慧中的基因表達規劃法(Gene Expression Programming, GEP)的強健能力，求解 TS 是相當有效且快速的，不過目前，此法用於時間序列問題的架構，尤其使用多基因結構時，還未有一個完善的解決方案。據此，本研究提出一個新的 GEP 架構，結合並改良自動定義函數與隨機常數限制等演化概念，使演化過程中的一些機制或結果有再利用(reuse)的機會，亦即能重新加入新的演化過程，藉此提高演化效率與效果。本研究利用太陽黑子年平均數做為實驗對象，模擬結果發現本研究所提之方法，除了演化收斂上有不錯的表現外，也提高了模型演化的穩定度。此外，多基因架構的基因數量也會有所影響，本實驗對象，採用 3 個基因數量，其實驗效果比 5 個基因數量更好。

關鍵詞：基因表達規劃法、時間序列、自動定義函數、隨機常數。

壹、緒論

時間序列(Time Series, TS)研究目的，是對具有時間性的歷史資料進行分析，再利用這些分析結果去預測未來的趨勢。在這個變化快速的時代中，有效且快速地預測是非常重要的。但在現實情況下的時間序列問題，有著大量不規則資料，且無法得知明確的因果關係。面對這樣具有高度複雜且非線性的問題，本研究將採用強韌的基因表達規劃法(Gene Expression Programming, GEP)，作為建模工具。

基因表達規劃法屬於人工智慧技術的一種，Ferreira(2001)利用遺傳演算法(Genetic Algorithm, GA)與遺傳規劃法(Genetic Programming, GP)的基礎發展的新概念。此法不同於傳統計量方法，它除了可以高度平行的搜尋解答空間外，還可以透過樹狀結構來呈現。

目前透過 GEP，用於時間序列問題的架構，還未有一個完善的解決方案，尤其使用多基因體結構時，表現更是不理想。另外，在 GEP 演化過程中，較高適應能力的染色體，是否都具有某些特徵？能不能將其記錄下，使其對未來演化過程能帶來好處。本研究希望藉此，提出一個新的 GEP 之架構，同時結合並改良自動定義函數(Automatically Defined Functions, ADF)與隨機常數(GEP_RNC)等演化概念，使結束後的演化有機會重新加入新的演化過程(reuse 概念)。本研究期望透過這樣的重複再利用，來提升多基因架構模型的穩定度與模型的演化效率。

貳、文獻探討

一、 基因表達規劃法

基因表達規劃法(Gene Expression Programming, GEP)是由葡萄牙學者 Ferreira 於 2001 年 12 月首次提出，並發表 Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence，基因表達規劃法(GEP)、遺傳演算法(GA)與遺傳程式規劃(GP)三者在主要演化步驟上極為相似，但 GEP 具有 GA 與 GP 的特性，GEP 實現了利用簡單編碼來解決複雜問題的目的。傳統 GA 採用固定長度的編碼方式，所以面對複雜問題時表現不佳；GP 採用非線性、動態的資料結構來編碼，可以處理複雜問題，但操作卻也相對複雜。而 GEP 透過 GA 固定長度的染色體編碼與 GP 不同大小、形狀的個體結合而成，也就是基因型的部分就由 GA 來表達，表現型的部分則由 GP 樹狀結構表示。根據 Ferreira 於 2006 年之著作「Gene Expression programming」，GEP 在演化數度與效果上比 GA 和 GP 的表現更好。

GEP 的編碼包含了基因型與表現型兩種型態，基因型的部分就由 GA 來表達，表現型的部分則由 GP 樹狀結構表示。GEP 的染色體分別由頭部(head)與尾部(tail)組成，頭部部分可以包含終端節點(terminal node)與函數節點(function node)，而尾部只可以是終端節點，函數節點符號依相關問題設計，可以是(+、-、*、/、AND、OR、XOR、sin 和 cos 等等)，另外也可能是程式設計中的 function，終端節點的符號，通常為真實資料的

屬性集或長度，所以在本研究中則為其時間序列資料，頭部長度由使用者(參數)訂定，尾部長度則透過$t = h \cdot (n_{max} - 1) + 1$而得。其中 h 為頭部的長度，t 為尾部的長度，n_{max} 表示所使用的函數需要最多量參與的個數。

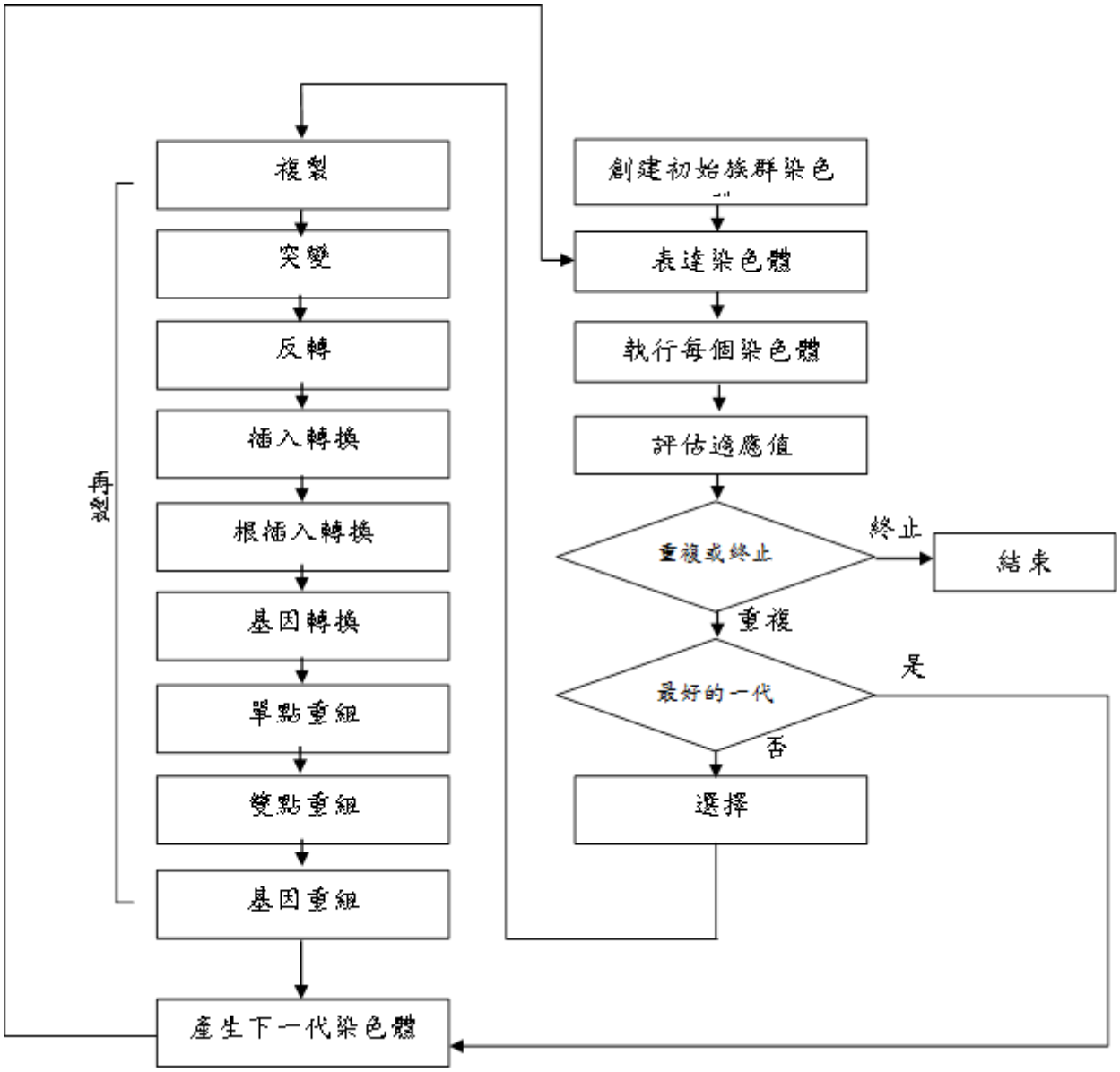

圖 1　GEP 的演化流程

資料來源：修改於 Candida Ferreira(2001)

上圖 1 是根據 Ferreira(2001)所提出的 GEP 演化流程，首先會亂數的產生初始族群，將這些初始族群的染色體轉成表示樹，透過適應函數評估每個表示樹，接著反覆地去執行演化操作(突變、反轉、插入、重組等)，產生下一代染色體，直到符合終止條件。GEP 在基因操作的方式上與 GA 類似，但另外增加了轉換及插入操作與重組操作。

二、　基因表達規劃法用於時間序列問題

由 GEP 原作者 Ferreira(2001)，利用 GEP 對太陽黑子(sunspots)時間序列資料做模型訓練，使用了多種 GEP 模型架構互相比較。由實驗結果驗證，若以 MSE 作為適應值函數，使用隨機數值限制(RNC)，會對時間序列的求解有相當幫助。而多基因架構相較於單基因架構沒有明顯優勢，不過 Ferreira 提到，會造成這樣的原因，有可能是因為其染色體結構較複雜，有更大的求解空間，在演化代數等參數必須要做調整。

不過 Barbulescu, et al.(2009)提出一個調適性基因表達規劃法(Adaptive Gene Expression, AdaGEP)，可以解決上述 Ferreira 遇到的染色體結構性問題，AdaGEP 會透過遺傳演算法(GA)來決定多基因 GEP 需要幾個基因，透過這樣的方式來自動調整基因數。原因源自，在傳統的多基因 GEP 架構上，對於基因數量的設置，是一開始在參數設定是必須先制定的，所以基因的數量設置常常是經驗取向的。

也有人將 GEP 用於財務時間序列的預測，如唐常杰等人(2004)提出 GEP-SWPM 與 GEP-DEPM。第一個方法使用移動平均時間的概念，設計適應函數來做調整，第二個方法利用差分平均的概念，設計適應函數來做調整。唐常杰等人(2005)再次利用 Fibonacci 數列的概念加入到 GEP-SWPM 與 GEP-DEPM。另外，廖勇人(2005)也有利用基因表達規劃法進行股票指數做實驗。不過可惜這些方法只採用單基因架構，並未考慮到多基因架構或是隨機常數限制等等演算機制。

參、研究方法

在現實情況下時間序列預測問題，最重要的就是找出一個數學模型，且這個數學模是要根據歷史紀錄訓練而成，藉此可以推測未來某一個時段的結果。而本研究將提出一個新的 GEP 架構，使其更有效地發掘出好的數學模型。

一、 新的基因表達規劃法架構

新的演算法架構除了保有原先 GEP 的概念外，同時結合其演化機制，像是自動定義函數 (ADF, Automatically Defined Functions)概念，與隨機常數限制(GEP-RNC)演算等等。自動定義函數的概念是由 Koza(1992)提出，最早運用於 GP(Genetic Programming)，其目的在於自行搜尋出常用且重要的變數、模型或程序，使其自動演化，藉以提高基因表達規劃法的學習效率。本研究將會針對自動定義函數做更進一步的改良，改良部分為下圖 2 框起來的部分，設計出 ADF 資源庫結合 ADF 突變事件的模組。有別於傳統 GEP 中 ADF 的做法，將表現較好的染色體之 ADF 基因片段特徵保存再利用，以實現類似共同演化的概念(co-evolution)。此外，還會搭配隨機常數(GEP_RNC)之演算機制，在演化時間序列的數學模型時，藉由常數項加入來增加解答空間的多樣性，並改善數學模型微調效率。

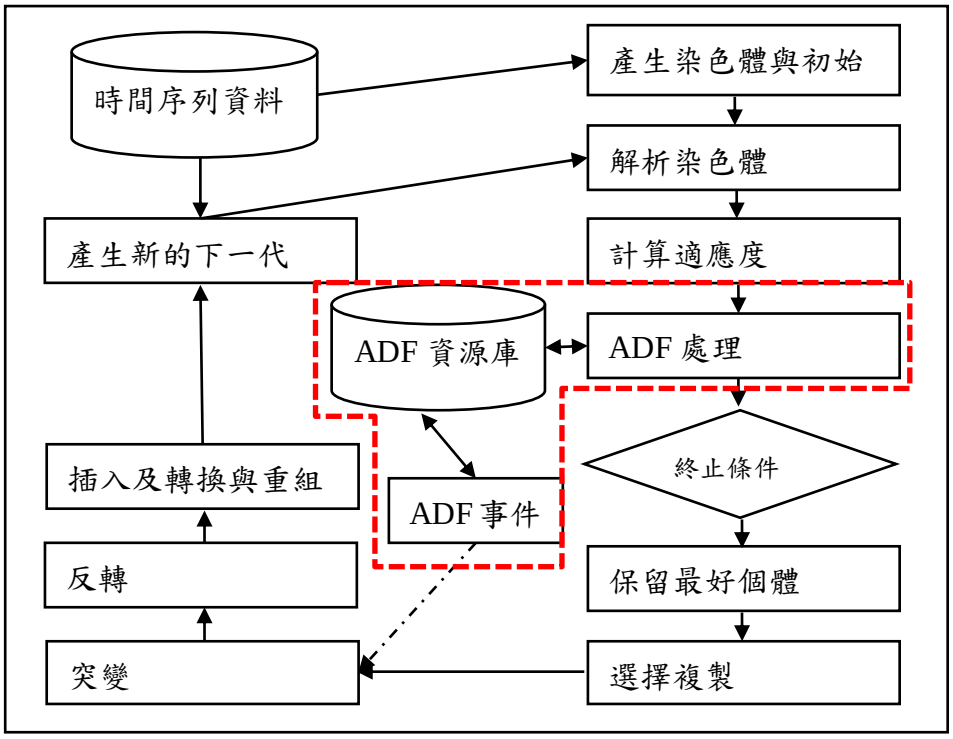

圖 2　本研究演算法流程圖

資料來源：本研究整理

（一）　ADF 處理

圖 3 為 ADF 處理流程，就是將選定的染色體其基因碼，分析成可理解的數學式，並把數學公式中函數的部分(+,-,*,/,...)保留，而非函數的部分(a,b,c,常數,...)，轉換成問號，藉此可以萃取出數學特徵式。最後，將 ADF 處理後的資訊都記錄到 ADF 資源庫中。

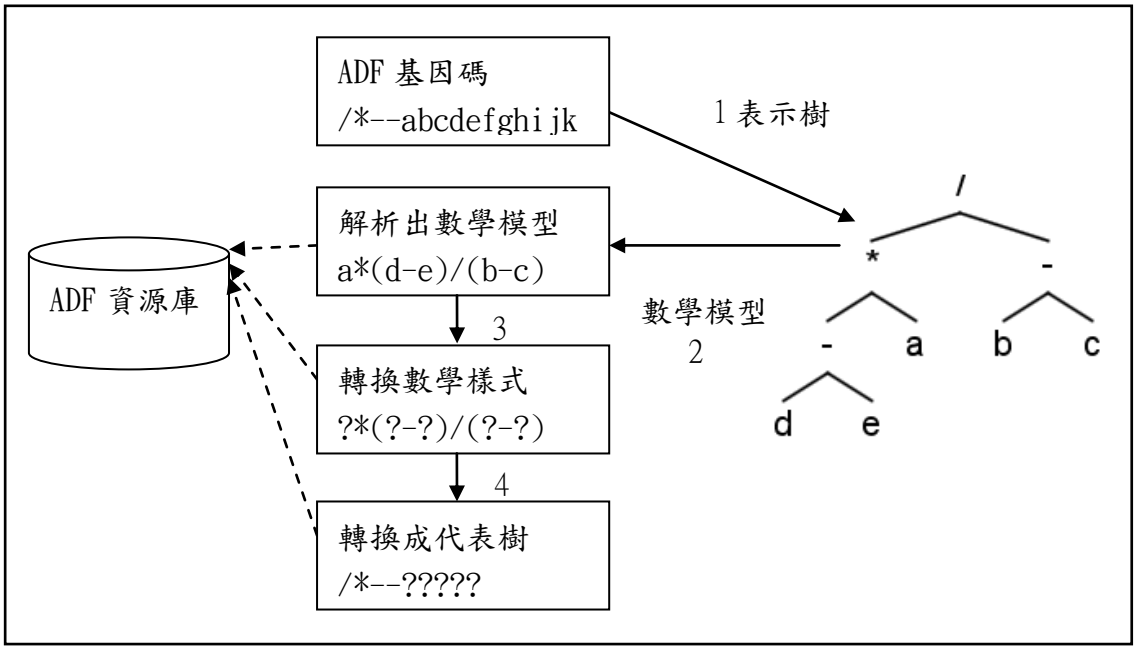

圖 3　ADF 處理機制運作流程圖

資料來源：本研究整理

1. ADF 基因碼會先轉成表示樹
2. 將表示樹解析出合法的數學模型。
3. 透過正規化保留數學運算式，其餘(常數與變數)用問號來取代。
4. 將上一步驟，再轉成基因碼格式。

（二） ADF 資源庫

ADF 資源庫的目的，是儲存每次運行或是世代中，表現較好的染色體之相關資訊，也就是記錄 ADF 處理過程中的樹狀結構、數學樣式、出現次數以及被歸納的樣式分類等等訊息，藉由這些資訊，我們可以把針對某類問題的常用數學式歸納出來，透過 ADF 事件機制以提升演化效率。

（三） ADF 事件

ADF 事件被觸發的機率，是根據參數中 ADF 突變事件率的數值而定。ADF 事件的機制，並非要完全取代 GEP 原先本來突變操作，而是提供一個新的突變的策略。ADF 事件與原先的演化操作(突變、反轉、轉換和插入等等)的性質不相同，它是一個會從過去的經驗中，也就是 ADF 資源庫所記錄的資訊，去猜測怎樣的突變更具有效率。

當 ADF 事件被觸發後，其運作流程一開始會先去 ADF 資源庫中，找出目前為止最常出現的 50 個數學特徵，透過類似輪盤法的方式，隨機選擇一個數學特徵式做為多基因架構染色體的一個基因片段。數學特徵式中的問號部分會被置換成終端節點的符號或式 GEP_RNC 架構中的常數。ADF 事件簡易處理機制示意圖如下圖 4。另外，當 ADF 事件產生一個新的基因片段，假設這個基因片段樹的長度與原先樹的長度有差異，則根據原先的 ADF 樹截長補短。如下圖 5

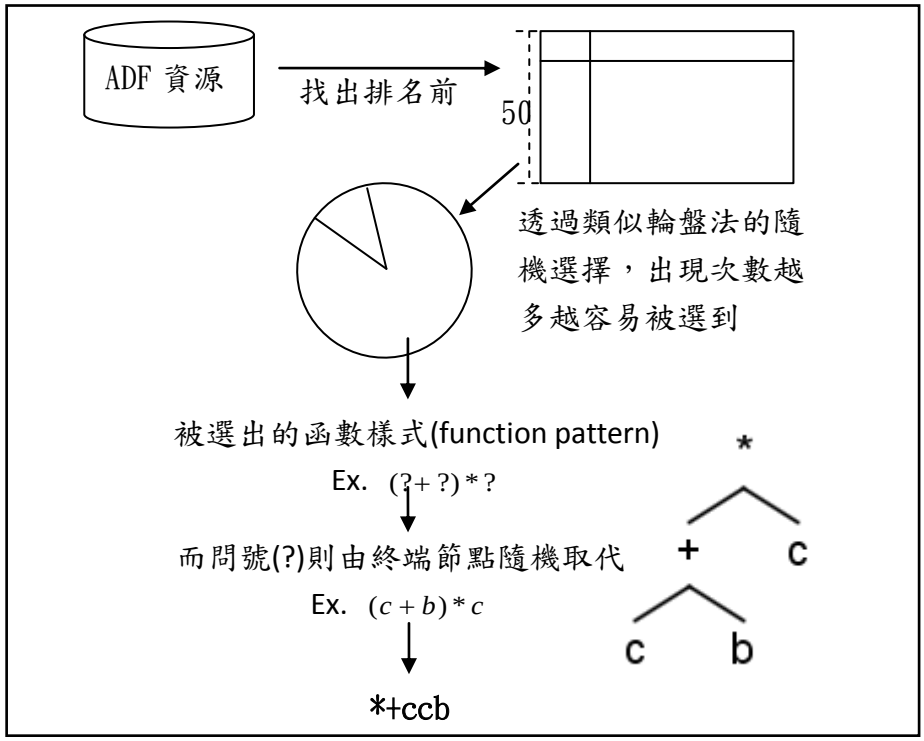

圖 4 ADF 事件處理機制

資料來源：本研究整理

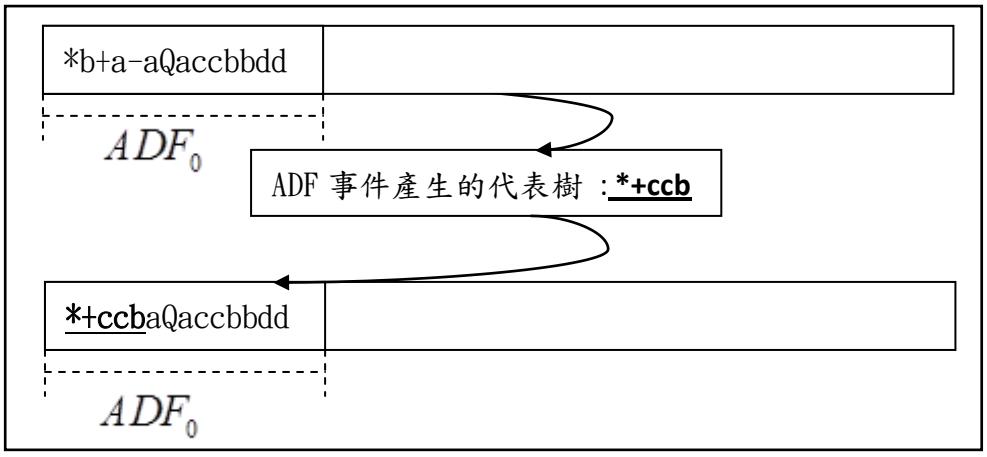

圖 5　ADF 事件截長補短

資料來源：本研究整理

二、 染色體編碼與節點設計

本研究的染色體設計，與單基因架構的染色體有很大的不同，這是因為加入了 ADF 概念與 GEP_RNC 機制造成的。首先，加入了 ADF 機制後，每個 ADF 必須遵循基礎 GEP 長度公式的規定，至於總共需有多少個 ADF，則取決於研究狀況，若有三個 ADF，依序為 ADF_0 、ADF_1 和 ADF_2 ，染色體中還必須有個基因，把 ADF 連結起來，稱為同源基因 (homeotic genes)，本研究命名為 $Cell$ 。$Cell$ 的長度規則也遵循 GEP 長度公式，在 $Cell$ 中，終端節點集合改為 ADF 之集合{ ADF_0 、ADF_1 和 ADF_2 }，函數節點之集合為{+、-、*、/}。除了 ADF 機制外，GEP_RNC 也會造成染色體設計改變。在 ADF 頭部與尾部基因，有機會出現問號符號取代原先編碼。尾部後面會加入隨機常數基因片段(RNC_0 、RNC_1 、RNC_2)，其長度與尾部長度相同，假使原先每個 ADF 基因長度為 15，必須再增加 8，改為 23。本研究中隨機常數限制基因，其值會隨機產生且為常數，將會依序對應到頭部、尾部中的問號符號。另外，每個 ADF 基因還會有一個 Dc-Domain(C_0 、C_1 、C_2)，其值根據參數設定範圍隨機產生，依序對應到隨機常數限制基因片段。表 1 為本研究染色體是意圖。

節點設計部分包含函數節點與終端節點兩部分。函數節點集合為 $F=\{+,-,*,/\}$ ，而在終端節點集合為 $T=\{a,b,c,d,e,f,g,h,i,j\}$ ，終端集合將依序對應到實驗對象之時間序列資料上，t-10,t-9,t-8,t-7,…,t-1。

三、 適應值函數設計

在適應值函數設計上，主要考慮我們透過演化的數學模型求出來的數值，與目標數值之差距。本研究採用誤差均方的指標做為演化的適應函數。其公式如下所示。

$$E_i = \frac{1}{n}\sum_{j=1}^{n}\left(P_{(ij)} - T_j\right)^2 \tag{1}$$

接著將上述公式，調整成如下公式，其範圍為 0 到 1000，1000 為理想狀態值，當我們的求出的誤差均方越大，則其適應值越小。

$$f_i = 1000 \cdot \frac{1}{1+E_i} \tag{2}$$

表 1　本研究染色體編碼示意圖

ADF-Gene0 (ADF0)	012345678901234 56789012 *a?++c?efajaacd **22154311**
	$C_0 = \{0.6314, -0.2223, 1.4432, -1.1145, 0.9992\}$
ADF-Gene1 (ADF1)	012345678901234 56789012 -**?+b?aaceadee **21153251**
	$C_1 = \{-0.5214, -0.1113, 0.3332, 1.9945, 1.9992\}$
ADF-Gene2 (ADF2)	012345678901234 56789012 -ja*+?cd?aaddce 11**225125**
	$C_2 = \{-0.9999, -0.0882, 1.8722, -0.9945, -0.4432\}$
Homeotic Gene (Cell)	012345678901234 /*+010211212000

四、 參數設計

下表為本研究設計的參數配置。

表 2　參數設定一覽表

運行次數	20	ADF 事件突變率	0.1
演化代數	300	反轉率	0.1
族群大小	100	IS 位移率	0.1
函數節點	{+ - * /}	RIS 位移率	0.1
終端節點	{a, b, c…j}	單點重組率	0.3
隨機常數陣列長度	10	雙點重組率	0.3
隨機常數型態	有理數	基因重組率	0.3
隨機常數範圍	[-2, 2]	基因位移率	0.1
頭部長度	7	Dc 突變率	0.1
基因長度	15	Dc 反轉率	0.1
染色體基因數	3	Dc 位移率	0.1
同源基因函數節點	{+ - * /}	隨機常數突變率	0.01

同源基因終端節點	{ADF0 ADF1 ADF2}	同源基因突變率	0.1
同源基因頭部長度	7	同源基因反轉率	0.1
同源基因長度	15	同源基因 IS 位移率	0.1
染色體長度	60	同源基因 RIS 位移率	0.1
突變率	0.1	適應值函數	$f_i = 1000 \cdot \frac{1}{1+E_i}$

肆、實驗結果

一、 太陽黑子年平均數時間序列實驗

本研究採用 Wolfer Sunspots 資料集作為實驗對象，透過「移動時間視窗」概念進行模組的訓練，該資料集共有 100 筆觀測資料，將移動時間長度設為 11 個單位，將訓練期設為 10 個單位(根據 GEP 的終端節點而定)，測試期設計為 1 個單位，而每次往後滑動 1 個單位，藉此來驗證本研究所提出的架構。實驗程式由 python 實現，資料庫使用 Mysql，實驗設備為 Inter(R) Core(TM) i5-2400 CPU @ 3.10GHz，8GB 記憶體，作業系統為 Microsoft Windows 7。

本研究架構在模型穩定度與演化效率上較具有優勢。可以從下圖 6 列出從 1 到 300 個世代中，來了解最佳解的收斂情況與系統設計的穩定度。藍色的線為使用本研究所改良架構的收斂情況。反之，紅色的則是未使用的情況。兩者在相同參數下各跑 15 次，且都有使用 GEP_RNC 與 ADF 的機制，從圖中可以明顯發現，紅色部分相對於藍色部分非常地不穩定，最佳解的收斂情況時好時壞，反觀藍色部分，大約 30 代以前就差不多收斂完，且結束後的最佳適應值也都達 840 以上，而且每次執行結果都很穩定。

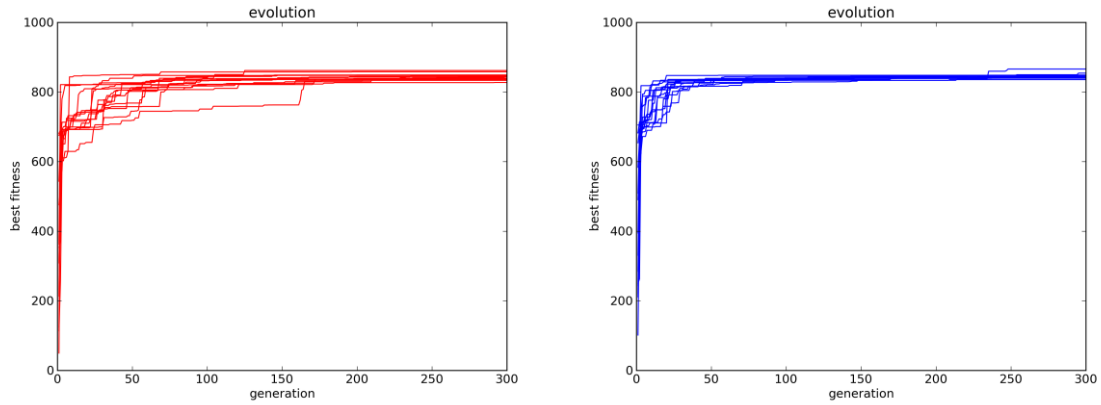

圖 6　每代最佳收斂圖(左:未使用本研究架構，右:使用本研究架構)

除了上述的兩種演化設計外，本實驗也比較了其它幾種常見的設計，包含單基因架構且不考慮 GEP_RNC、單基因架構考慮 GEP_RNC、多基因架構不考慮 GEP_RNC 與多基因架構考慮 RNC。一樣在相同參數下(不包含特殊結構差異的參數)各跑 15 次，且

每次都有 300 個世代繁衍，將其每代最佳解的適應情況呈現於下圖 7。右上圖與左下圖，每一次運行的第一代最佳適應值就到 600 多，這是因為單基因架構染色體結構比較單純，不過相對的，由於模型太過簡單，容易造成不穩定的狀況。此外，從這四張圖還可以發現，在時間序列問題，有考慮 RNC，也就是左下圖與右下圖，確實會使每次模型演化更佳趨於穩定。

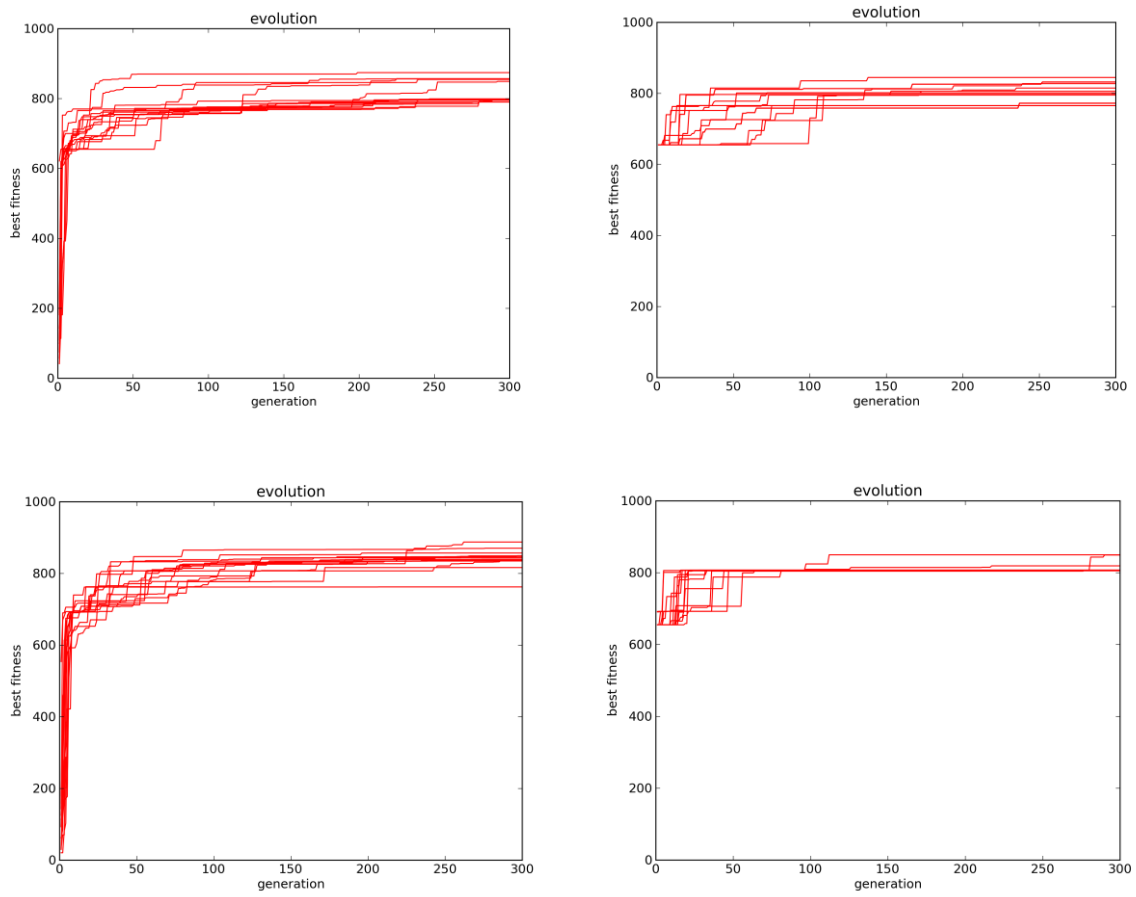

圖 7　其它演化設計每代最佳收斂圖(左上:多基因架構不考慮 RNC，右上:單基因架構不考慮 RNC，左下多基因架構考慮 RNC，右下：單基因架構考慮 RNC。)

下圖 8 顯示出 Wolfer Sunspots 的時間趨勢圖，黑色線條為真實時間序列資料，也正是本實驗之目標曲線，而藍色線條則是從圖 6 右圖中(本研究設計改良架構)，挑出演化過程中，表現最好的數學模型做預測，該數學模型之適應值達 866.24526239825，我們可以發現，所預測的曲線與真實數據的曲線非常吻合。

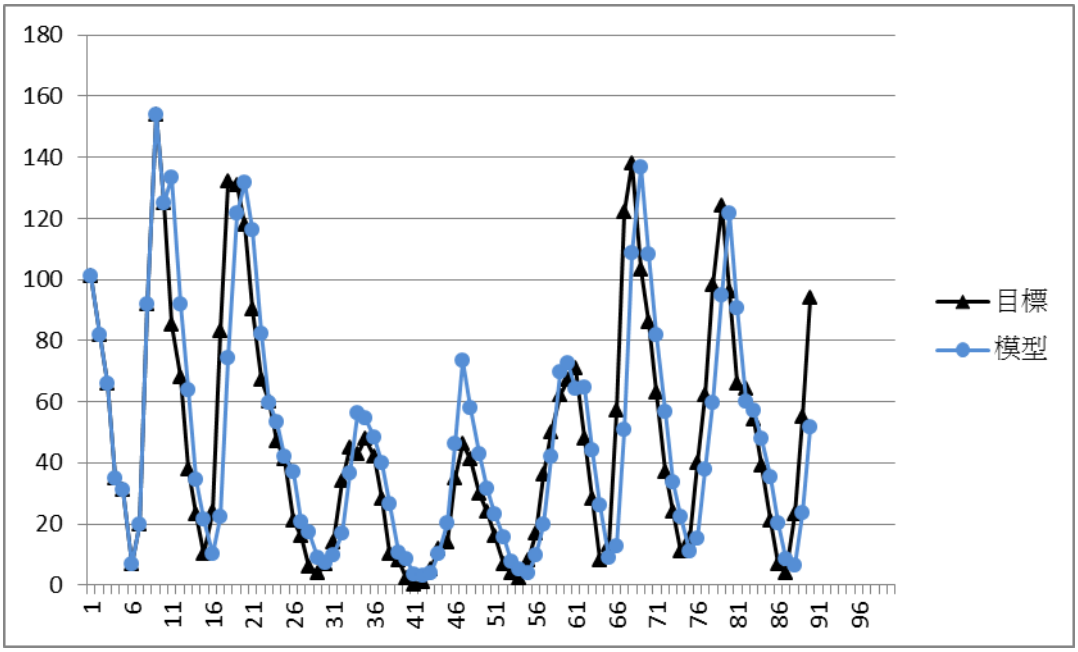

圖 8　Wolfer Sunspots 的時間趨勢圖(黑:真實資料，藍:預測資料)

伍、結論

本研究所提之架構，有助於多基因架構(包含 ADF 架構)提升演化效率，以及每次模型演化的穩定度，但在模型預測的準確度，還未達到本研究預期之效果，可能在終端節點、函數節點與適應值的設計上必須在調整。此外，多基因架構的基因數量也會有所影響，本實驗對象，採用 3 個基因數量，其實驗效果比 5 個基因數量更好，但不同預測問題，可能就必須再做調整。未來本研究會嘗試使用不同類型的時間序列資料，像是股價預測、大盤指數預測與市場銷售數據預測等等。隨著不同類型問題與參數的使用，這將使得本研究設計的 ADF 資源庫更加豐富，我們將可以利用 ADF 資源庫來提升某類問題的演化效率，且因為相關數學特徵式都儲存於 ADF 資源庫中，有利我們分析甚麼問題適合哪類的數學公式。若能善加利用，即可有效的減少多基因架構，求解不穩定的現象。

陸、參考文獻

1. 王衍智，2004，台灣股票選擇權日內價格定價模型研究–比較時間序列方法與遺傳規劃方法，朝陽科技大學財務金融系碩士論文。
2. 林宜芬，2006，遺傳程式規劃為基的時間序列模型在金融市場之應用，輔仁大學資訊管理學系碩士班。
3. 陳寬裕，2006，結合遺傳演算法與支援向量迴歸於台灣股票加權指數之預測，長榮大學經營管理研究所。
4. 楊奕農，2005，時間序列分析：經濟與財務上之應用，臺北：雙葉書廊。

5. 廖勇，唐常杰，元昌安，陳安龍，段磊，2005，基於基因表達式編程的股票指數時間序列分析，Journal of Sichuan University(Natural Science Edition)第四十二卷第五期。
6. 劉瑞鑫，2003，時間序列與人工智慧方法在台股指數報酬率預測之績效比較，朝陽科技大學財務金融系碩士論文。
7. Barbulescu, A., Bautu, E., Meteorological Time Series Modeling Based on Gene Expression Programming, Recent Advances in Evolutionary Computing, WSEAS Press, 2009, pp. 17–23.
8. Bautu, E., Bautu, A., Luchian, H., AdaGEP - An. Adaptive Gene Expression Programming, Proceedings of the Ninth international Symposium on Symbolic and Numeric Algorithms For Scientific Computing (September26–29, 2007), SYNASC, IEEE Computer Society, pp. 403–406.
9. Brockwell, P., Davies, R., Introduction to time series, Springer, New York, 2002.
10. Ferreira, C., Gene Expression Programming: A New Adaptive Algorithm for Solving Problem, Complex System, Vol.13, 2001, pp.87-129.
11. Ferreira, C., Genetic Representation and Genetic Neutrality in Gene Expression Programming. Advances in Complex Systems 5(4):389-408.
12. Ferreira, C., Function Finding and the Creation of Numerical Constants in GeneExpression Programming. In J. M. Benitez, O.Cordoon, F.Hoffmann, and R.Roy, eds., Advances in Sort Computing: Engineering Design and Manufacturing, pages 257-266, Springer-Verlag.
13. Holland, J. H., Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.
14. Koza, J.R., Genetic Programming: On the Programming of Computers by Means of Natural Selection, Cambridge, MA:MIT Press,1992.
15. Koza, J.R., Genetic Programming Ⅱ: Automatic Discovery of Reusable Programs, Cambridge, MA:MIT Press,1994.
16. ZUO Jie, TANG Changjie, LI Chuan, et al. Time series prediction based on gene expression programming, International Conference for Web Information. Berling: Springer Verlag, 2004: 55-64.

An improvement Gene Expression Programming to solve the time series problem

Wen-Shiu Lin
Fu-Jen Catholic University Institute of Information Management
wslin@im.fju.edu.tw

Yu-Hsiang Su
Fu-Jen Catholic University Institute of Information Management
Include2md@gmail.com

Abstract

The time series studies is through the training historical data to build model, and then use it to predict future trends. Artificial Intelligence in gene expression programming is a very effective and fast algorithm. But for the prediction problems, especially in multicellular structure, none of method can better solve. For this reason, we propose an improvement Gene Expression Programming method, combining the concept of ADF (Automatically Defined Functions) and RNC(Random Numerical Constants). The idea is like co-evolution, but in different way, it save some information end of the evolution, and reuse it in another evolution. In this study, we are going to use the Wolfer sunspots series data as the experimental task. The simulation results showed that our proposed method, in addition to good performance on the convergence evolution, but also improve the stability of the model.

Keywords: Gene Expression Programming, Time Series, Automatically Defined Functions, Random Numerical Constants