細胞自動機與 RFID 之基因密碼辨識與管理資訊系統 -以 miRNA 為例

林守羿 [1]　蔡孟勳 [1,2]　吳協昌 [1]
[1] 國立中興大學資訊管理研究所
[2] 國立中興大學基因體暨生物資訊學研究所
xup6zeo@gmail.com

摘要

近年來，隨著人們對生物多樣性的重視，世界各地為保存生物的多樣性開始收集各種生物的基因序列，隨著序列的收集增加，許多科學家開始思考著如何資訊化或運用這些基因序列，2003 年，加拿大的學者 Paul Hebert 提出了利用基因序列來當作物種辨識的條碼(DNA barcode)，基因條碼(DNA barcode)提供了一種新的辨識物種方法，相較於過去傳統分類的耗時耗力，基因條碼能更快速的辨識物種，本研究使用 miRNA 序列來做處理基因條碼，經由細胞自動機方法來轉換序列，並將序列轉換成圖像，轉換出的 miRNA 圖像除了可直接藉由肉眼觀察序列間的相同相異部分外，本研究還觀察出調控相同卵巢癌基因的 miRNA 序列有其共通性，未來可將收集這些相關的序列圖形建立資料庫，結合 RFID 與後端的 miRNA 資料庫，形成一基於 RFID 的 miRNA 條碼辨識系統，用在自動的條碼辨識上。

關鍵詞：細胞自動機、基因條碼、miRNA、RFID

1. 緒論

由於人類對基因的解碼，人類開始探討基因的功能及結構，加上世界的上的物種繁多，為了保存生物的多樣性，各國皆開始收集及建立物種的基因序列資料。2003 年，加拿大學者 Paul Hebert 提出了 DNA barcode 的概念[1]，利用 DNA 序列當作物種的辨識條件，將基因序列作為物種分類，以及辨識身分不明或物種之用。然而目前基因條碼的運用範圍，都只是特定種類及固定範圍的小區域分析，而且大部分都是藉由人力來處理資料及分類，一旦需要辨識的物種變多，基因資料量變大，手工分類及檢測將會是緩慢及沒效率的。因此，我們提出一個基於無線射頻辨識(Radio Frequency Identification，RFID)的基因條碼系統，除了能系統化建置基因條碼外，還能快速辨識圖形化基因條碼，不僅能提升辨識及管理基因條碼的效率，且藉由管理化系統資料庫的建立，未來若要將資料運用在分類及比對上，將會更加便利且迅速。

2. 研究方法與架構

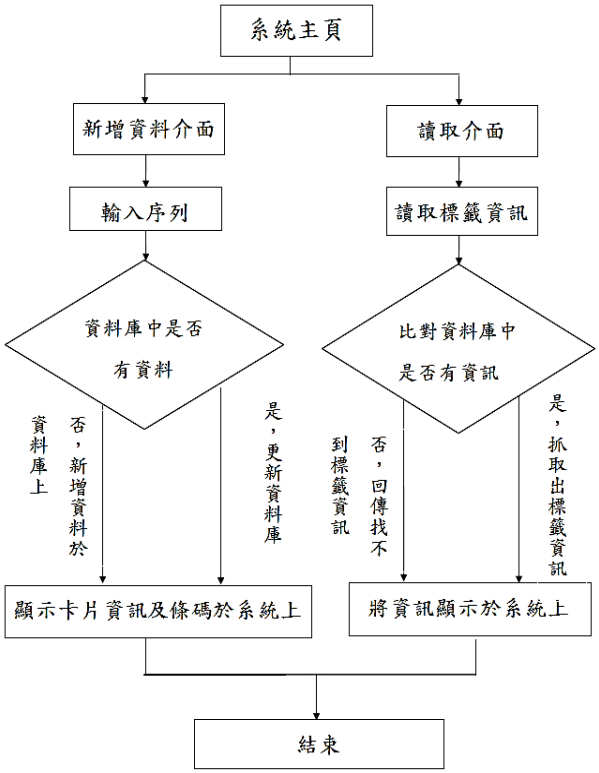

圖 1 系統流程圖

本系統主要分成兩大部分，第一部分是針對 miRNA 的圖像條碼化，將輸入的 miRNA 序列轉換成可處理的資料序列並轉換成圖形，以及把圖形存入到後端資料庫之後並管理。第二部分是系統如何與 RFID 作讀取和寫入的部分，以及和後端的資料庫作比對並將結果輸出至系統畫面上。

2.1 新增資料介面

在新增資料介面中，主要的步驟分成三個，第一步驟是針對 MiRNA 作序列轉換將序列成可以處理的資料序列。第二步驟是對等待處理的資料序列轉換成圖形。第三步則是對於轉換成而的圖形存入 RFID 卡片，並將圖形及相關的序列資訊也存入後端資料庫中，以便之後辨識卡片之用。

2.1.1 輸入序列

目前所常用的基因條碼，主要都是由 DNA 序列或者是蛋白質序列來使用，但是 DNA 序列及蛋白質序列都相當的長(用 DNA 序列來實作的基因條碼約為 650 個字母的長度)，如此大的資料量對於存入 RFID 標籤是一個很大的問題，而且大量的資料也是不利於快速辨識，因此我們選用了長度較短，約為 20 個字母左右的 miRNA 來實作系統。

本研究使用調控影響卵巢癌基因的微核糖核酸(miRNA)來作為輸入序列，除了觀察相同的序列是否會轉換成相同圖形讓肉眼易於觀察外，另外希望能找出圖形化序列是否可以看出原始序列無法看出的共通及相似性，表 1 為影響卵巢癌基因名稱以及調控這些基因的 MiRNA 名稱。

表 1 影響卵巢癌的基因與其調控的 miRNA

影響卵巢癌的基因	miRNA
SMAD2	hsa-miR-142-5p [2] hsa-miR-105 [3]
CCND3	hsa-miR-15a [4] hsa-miR-424 [2] hsa-miR-16 [4,5] hsa-miR-195 [2]
PAPPA	hsa-miR-142-5p [2] hsa-let-7d [3,6,7] hsa-miR-15a [4] hsa-miR-200a [3,5,8,9,10]
MTMR2	hsa-miR-101 [3] hsa-miR-9 [3,11,12]
MAOA	hsa-miR-495 [7]
HELZ	hsa-miR-29a [2,5] hsa-miR-15a [4]
RARHOGAP	hsa-miR-495 [7] hsa-miR-429 [9,10] hsa-miR-15a [4] hsa-miR-424 [2]
PTPN9	hsa-miR-126 [3]
NAP1L1	hsa-miR-495 [7] hsa-let-7d [3,6,7] hsa-miR-657 [2] hsa-let-7c [3,13] hsa-let-7g [13] hsa-miR-9 [3,11,12]
ID2	hsa-miR-381 [2]

2.1.2 序列轉換

在 miRNA 序列輸入後，需要先把它轉換為二進位序列。由於 RNA 序列是由腺嘌呤(A)、鳥嘌呤(G)、胞嘧啶(C)和尿嘧啶(U)所組成[14]，每個基因也都代表著不同意義，為了讓這些基因轉換後還能保有原本意義，並且將來也能藉由系統將資料從二進位序列轉換回來，我們根據表 2，將它們各賦予一個 3 位數的二進位值，如此一來 miRNA 序列將會轉換成一組有意義的二進位序列。

表 2 微型糖核酸編碼表

miRNA	Binary Notation	miRNA	Binary Notation
A	000	U	010
G	100	C	110

2.1.3 細胞自動機生成圖像

當 miRNA 轉換成二進位序列後，之後便是將序列轉換成圖像，本研究所採用的是利用細胞自動機來對序列作轉換，將其轉換成一維的條碼圖像。細胞自動機最初是由 John von Neumann 於 1960 年代提出[15]，細胞自動機是一種離散的動態系統，每一個細胞會根據周圍細胞狀態，更改自己的狀況。細胞自動機可以說是一堆單細胞的集合，而每個細胞在每個時間點只會有一種狀況。

細胞自動機目前主要之應用主要為一維與二維，本研究中使用的是一維細胞自動機[16]來做轉換。

一維細胞自動機是由一組根據時間改變狀態的變數S_i^t組成，i 從 0 開始到 N-1，N 為變數的個數，可以將每一個變數放在一個格子中，這樣的組成就稱為一個細胞；每一個細胞有各自個狀態，在一維細胞自動機中只有將狀態區分為 0 與 1，將 0 與 1 轉換成黑(0)與白(1)。在細胞自動機中，S_0代表一開始的起始狀態，F 表示為細胞自動機之規則，每個細胞會被自己與指定範圍內的鄰居狀態影響，改變下一個時間點之自身狀態。h代表指定的鄰居個數，S_{t+1}^i 為下個時間點的狀態，最後可表示為 $S_{t+1}^i = \mathrm{F}(S_t^{i-h} \dots S_i^t \dots S_t^{i+h})$。在本研究中，鄰居個數 h 設為 1，即表示會影響自身下個時間點狀態的細胞為：自己左邊的細胞（鄰居 1）、自己、自己右邊的細胞（鄰居 2）(圖 2)。由於含自己共三個細胞，因此每個細胞會被影響的狀態組合為，每個細胞又都有兩種狀態(0 和 1)，所以總共會有種組合，而每一種組合就是一條細胞自動機之規則。而在本研究中其序列轉換圖形步驟可轉化成下列式子：

$$D_{(i,j)} = F\left[D_{(i-1,j-1)}, D_{(i-1,j)}, D_{(i-1,j+1)}\right]$$
$$If\ 1 \le j < S-1; 1 \le i < S-1 \quad (1)$$
$$D_{(i,0)} = F\left[D_{(i-1,s-1)}, D_{(i-1,0)}, D_{(i-1,1)}\right]$$
$$If\ j = 0; 1 \le i < n \quad (2)$$
$$D_{(i,s-1)} = F\left[D_{(i-1,s-2)}, D_{(i-1,s-1)}, D_{(i-1,0)}\right]$$
$$If\ j = S-1; 1 \le i < n \quad (3)$$

其中，$D_{(i,j)}$表示轉換後該細胞自動機序列陣列之某一(i,j)位置之 0 或 1 之狀態值;i 為選定行數；S 代表該二進位序列之長度；F 為該序列重排演算之規則函數。

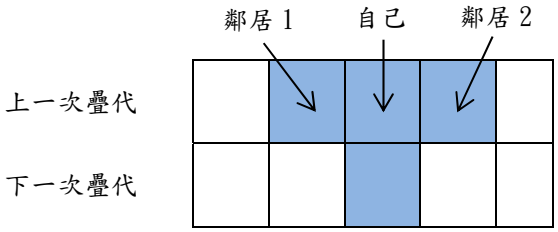

圖 2 疊代示意圖

2.1.4 存入 RFID 標籤與資料庫

將 MiRNA 轉換成圖形之後，系統將會把圖形存入 RFID 標籤及後端資料庫之中，本研究中所存入的 RFID 標籤為 MIFARE 卡，除了存入卡片之外，我們也需將資料寫入資料庫中，本研究所使用的資料庫為微軟的 SQL server 2008 ，因為 SQL server 2008 能與 Microsoft Visual Studio 的專案做快速整合與連結，而且介面淺顯易懂。在本研究實作的資料庫中主要包含幾個欄位，卡片序號(TagID)、miRNA 名稱(NAME)、原始 miRNA 條碼(barcode)、經細胞自動機轉換後的序列圖像(barcodeimage)，並將具有唯一性卡片序號設為 Primary Key。

2.2 RFID 標籤讀取與辨識界面

在處理序列的圖形轉換及 RFID 標籤和資料庫的寫入後，另一個部分便是有關 RFID 標籤讀取的部分，以及如何對後端資料資庫做比對。接下來將會介紹如何實作 RFID 標籤存取以及後端資料庫的比對部分。

在 RFID 標籤抓取資料出來之後，需要辨識 RFID 標籤的內容以及將 RFID 標籤的內容與後端資料庫做比對，來觀察 RFID 標籤資料是否與後端資料庫一致，若是抓出的 RFID 標籤辨識碼與抓出的圖像矩陣與後端資料庫 RFID 標籤辨識碼一致，則顯示出 RFID 標籤的卡號、miRNA 名稱、原始條碼和圖像條碼。整個卡片辨識流程如下：

1. 系統會抓取電腦上所有的外接裝置，找尋 RFID 裝置。

2. 搜尋到裝置後便可選擇與 RFID 裝置連線。

3. 連線後可選擇開始執行程式。

4. 系統會每隔一段就偵測讀卡機上 RFID 標籤內容。

5. 若是讀取到 RFID 標籤則將 RFID 標籤資料抓取出來並與後端資料庫做比對，若是 比對發現資料庫有此資料則將相關資訊呈現在系統介面上，反之，則回傳找不到結果。

6. 按下停止鍵結束搜尋。

3. 實驗結果與討論

3.1 開發環境

本系統使用VB語言撰寫，並使用Visual Studio 2008做為開發工具。另外，為方便處理圖形轉換，本系統後端使用Matlab負責處理條碼圖形的部分，並用SQL server 2008來建立後端資料庫。

3.2 新增修改資料

當進入新增修改資料頁面後，會進入卡片輸入介面。在卡片輸入界面中，一開始需要選擇與電腦連接的 RFID 裝置來啟動連線，當選擇完 RFID 讀取器後便需輸入卡號、miRNA 名稱以及 miRNA 的原始序列條碼，之後選取寫入便可將條碼及資訊寫入資料庫和 RFID 標籤之中。

在卡片的寫入過程中，由於本研究中是將miRNA序列的矩陣圖像存入卡片之中，程式會呼叫Matlab將所需要寫入的序列轉換成圖像，但是在序列轉換成圖像前需先進行二進位序列轉換(圖3)。在轉換成圖像矩陣後，Matlab會將其結果回傳給系統做後續的寫入卡片與資料庫部分。

```
case {'A', 'a'}
    seqBinVec(seqBinVecCount-1) = 0;
    seqBinVec(seqBinVecCount) = 0;
    seqBinVec(seqBinVecCount+1) = 0;
case {'U', 'u'}
    seqBinVec(seqBinVecCount-1) = 0;
    seqBinVec(seqBinVecCount) = 1;
    seqBinVec(seqBinVecCount+1) = 0;
case {'G', 'g'}
    seqBinVec(seqBinVecCount-1) = 1;
    seqBinVec(seqBinVecCount) = 0;
    seqBinVec(seqBinVecCount+1) = 0;
case {'C', 'c'}
    seqBinVec(seqBinVecCount-1) = 1;
    seqBinVec(seqBinVecCount) = 1;
    seqBinVec(seqBinVecCount+1) = 0;
```

圖 3 序列轉換成 2 進位條碼 程式碼片段

當系統接收細胞自動機轉換圖像後，系統會根據RFID標籤的識別碼去搜尋資料庫中是否有RFID的識別碼存在，若是RFID標籤已經存在於資料庫中，系統將會更新資料庫的資料，並更新卡片的資料。資料庫之中沒有讀取到的RFID標籤識別碼資訊，則會將此卡片資訊新增到資料庫之中。

當系統更新完資料庫之後，之後便將圖像寫入RFID標籤之中，當整個程序都完成之後系統會把轉

換出的條碼顯示在系統界面上，表示系統對資料庫及RFID卡片標籤更新完成。

3.3 RFID 標籤辨識

在RFID標籤辨識部分，程式會一直偵測RFID讀取器是否有讀取到卡片，如果有讀取到RFID標籤，會將RFID標籤資訊抓取出來與後端資料庫進行比對，如果後端資料庫有其資料，會將其相關資訊(RFID標籤、miRNA名稱、原始條碼)以及圖形條碼顯示在系統界面上(圖6)，若是後端資料庫找無RFID標籤資訊，則會在系統上顯示找無此miRNA條碼的資訊(圖7)。

3.4 細胞自動機圖形條碼之探討

當序列從字串轉換為二進位數字後，就可以開始將數字序列轉換為圖形，根據指定的細胞自動機之規則進行疊代。所謂細胞自動機的規則由於細胞本身改變後狀態會受到自身改變前狀態(細胞的狀態只有 0 或 1)及改變前左右鄰居的影響，所以三個為一組共有 000、001、010、011、100、101、110、111 這八種狀態，而每一種狀態在改變後都只有 0 或 1 兩種狀態 所以共有 28=256 種組合，意即有 256 種規則，而根據規則不同，轉換後得到的圖形也會不相同。底下的圖為根據 hsa-let-7g-3p 這組 miRNA 序列並搭配不同的規則所轉換得到之圖形。

圖 4 序列 HSA-LET-7G-3P 經規則 2 轉換之圖形

圖 5 序列 HSA-LET-7G-3P 經規則 35 轉換之圖形

圖 6 序列 HSA-LET-7G-3P 經規則 187 轉換之圖形

圖 7 序列 HSA-LET-7G-3P 經規則 230 轉換之圖形

圖 8 序列 HSA-LET-7G-3P 經規則 255 轉換之圖形

由圖 4~圖 8 可知，雖然疊代的規則有 256 種，但是並非各種疊代方式都能有效的表現出各序列的差異，舉例來說，越是後面的規則，由於規則轉換後值為 1 的部分將會越來越多，所以整體的圖像將會偏白，較無法藉由圖像看出序列的差異，反之，越前面的規則在轉換後值為 0 的部分會越來越多，整體圖像會偏黑，也較無法看出圖像間的差異，所以在規則的選擇上需要多加測試。在本研究中經由測試選出了規則 187 來當作鑑別物種的規則，接下來將會使用 187 結果出來的圖形作接續的探討。

圖 9 序列 hsa-miR-142-5p 經規則 187 轉換之圖形

圖 10 序列 hsa-miR-105-3P 經規則 187 轉換之圖形

圖 11 序列 hsa-miR-105-5p 經規則 187 轉換之圖形

hsa-miR-142-5p	cauaaaguagaaagcacuacu
hsa-miR-105-3P	acggauguuugagcaugugcua
hsa-miR-105-5P	ucaaaugcucagacuccuguggu

圖 12 hsa-miR-142-5p、hsa-miR-105-3P、hsa-miR-105-5P 之原始序列

如圖 9、10、11 所示、hsa-miR-142-5p 和 hsa-miR-105是藉由miR2Disease資料庫查詢調控有關卵巢癌基因SMAD2的miRNA基因，由圖可知，這三個調控同樣SMAD2的基因在圖像上在圖像最右下角都有一個點，藉由圖像上可發現到這些圖像的共同點，但是若是由圖12中卻無法看出這三條序列的相同或相異之處。

圖13 序列hsa-miR-15a-3p經規則187轉換之圖形

圖14 序列hsa-miR-15a-5p經規則187轉換之圖形

圖15 序列hsa-miR-429經規則187轉換之圖形

hsa-miR-15a-3p	caggccauauugugcugccuca
hsa-miR-15a-5p	uagcagcacauaaugguuugug
hsa-miR-429	uaauacugucugguaaaaccgu

圖16　hsa-miR-15a-3p、hsa-miR-15a-5p、has-miR-429原始序列

圖13~15是有關調控基因RARHOGAP的相關miRNA序列圖形，而RARHOGAP是被發現影響卵巢癌的基因之一，由圖像中可知，所有的圖像右下角都有點的存在，而從圖16的所有原始序列中，只能觀察各自序列部分的異同之處，但對於全部序列找出異同點則沒有辦法，相較於用序列判斷，用圖像觀察反而能觀察出相似之處。

圖17 序列hsa-miR-15a-3p經規則187轉換之圖形

圖18 序列hsa-miR-15a-5p經規則187轉換之圖形

圖19 序列hsa-miR-142-5p經規則187轉換之圖形

hsa-miR-15a-3p	caggccauauugugcugccuca
hsa-miR-15a-5p	uagcagcacauaaugguuugug
hsa-miR-142-5p	cauaaaguagaaagcacuacu

圖20 hsa-miR-15a-3p、 hsa-miR-15a-5p、hsa-miR-142-5p之原始序列

圖17~19是有關調控基因PAPPA的部分相關miRNA序列圖形，而PAPPA也是有關影響卵巢癌發生率的基因之一，由圖像中可知，所有的圖像右下角如同之前調控卵巢癌相關基因的miRNA一樣，右下角的都有一個點的存在。從圖20之原始序列中，並無法輕易的用肉眼觀察出序列的差異，但是藉由圖像的轉換後，反而能從圖像看出其相同之處。

綜合上述三組有關調控影響卵巢癌基因的圖像可知，原始相同的序列在轉換後會成為相同的圖形，而從結果上來看，我們可以直觀的從肉眼看出序列的相同部分，除此之外，在本研究中所使用的例子，包含影響SMAD2基因的miRNA：hsa-miR-142-5p、hsa-miR-105-3P、hsa-miR-105-5P；影響RARHOGAP基因的miRNA：hsa-miR-495、hsa-miR-429、hsa-miR-15a-5p、hsa-miR-15a-3p、hsa-miR-424-5p、hsa-miR-424-3p：影響PAPPA基因的miRNA：hsa-miR-142-5p、hsa-let-7d-5p、hsa-let-7d-3p、hsa-miR-15a-5p、hsa-miR-15a-3p、hsa-miR-200a-5p、hsa-miR-200a-3p。上述這些調控影響卵巢癌基因的miRNA序列在原始序列中無法看出的共通性，但是藉由細胞自動機轉換成圖像之後，其圖像發現了相同之處，而且不只是調控相同基因的圖像間有相同之處，這些相關於調控卵巢癌基因的miRNA都出現了相同之處，或許這些相同的序列片段對於調控卵巢癌基因可能有相同的影響，將來可藉由生物實驗去測試這些出現共同圖像的片段是否功能上也有其關聯之處，將可做為miRNA在調控基因上功能之參考。

4. 未來展望

本研究未來希望能夠改進一些目前系統的缺陷、及增加實用性及範圍性。第一、希望能更加改進讀寫的速率，由於目前系統在生成圖像方面是藉由 MATLAB 來執行、而且在資料庫的寫入上耗費的時間也比想像的長，希望將來能改進寫入資料的時間以及改進圖像生成的時間，將 Matlab 的方法轉成 package 給系統直接執行，不需呼叫 Matlab 來執行，相信能更減少執行的時間。第二，本文所實作的 RFID 系統所使用讀取器讀取範圍並不大(<10CM)，若未來希望增加系統的泛用與實用性，則需要將讀取器改成遠距離的讀取器，藉此提升系統的實用性。第三、本文研究所轉換的圖像主要是希望能用肉眼觀察其結果，一旦資料量越來越多，肉眼辨識也是需要耗費相當大的時間，因此希望未來希望能再新增比對系統部分，因為目前序列比對是生物序列主要的運用之一，探究物種間不同的結構、功能，找出重要的序列結構，所以希望未來能讓使用者除了對序列做辨識之外，也能藉由系統直接對不同序列直接做比對。

致謝

本人誠摯感謝審稿者提出的寶貴建議與意見，改善了本文的內容與品質。本文與台中榮民總醫院與國立中興大學合作研究計畫(榮興計畫)，計畫編號 TCVGH-NCHU 1027619，以及行政院國家科學委員會，計畫編號 NSC 102-2622-E-005-011-CC3 共同合作。

參考文獻

[1] Hebert, P. D. N., A. Cywinska, et al. (2003). "Biological identifications through DNA barcodes." Proceedings of the Royal Society of London. Series B: Biological Sciences 270(1512): 313-321.

[2] Dahiya, N., C. A. Sherman-Baust, et al. (2008). "MiRNA expression and identification of putative miRNA targets in ovarian cancer." PLoS One **3**(6): e2436.

[3] Iorio, M. V., R. Visone, et al. (2007). "MiRNA signatures in human ovarian cancer." Cancer Res **67**(18): 8699-8707.

[4] Bhattacharya, R., M. Nicoloso, et al. (2009). "MiR-15a and MiR-16 control Bmi-1 expression in ovarian cancer." Cancer Res **69**(23): 9090-9095.

[5] Nam, E. J., H. Yoon, et al. (2008). "MiRNA Expression Profiles in Serous Ovarian Carcinoma." Clinical Cancer Research **14**(9): 2690-2695.

[6] Park, S. M., S. Shell, et al. (2007). "Let-7 prevents early cancer progression by suppressing expression of the embryonic gene HMGA2." Cell Cycle **6**(21): 2585-2590.

[7] Zhang, L., S. Volinia, et al. (2008). "Genomic and epigenetic alterations deregulate miRNA expression in human epithelial ovarian cancer." Proc Natl Acad Sci U S A **105**(19): 7004-7009.

[8] Bendoraite, A., E. C. Knouf, et al. (2010). "Regulation of miR-200 family miRNAs and ZEB transcription factors in ovarian cancer: evidence supporting a mesothelial-to-epithelial transition." Gynecol Oncol **116**(1): 117-125.

[9] Yang, H., W. Kong, et al. (2008). "MiRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN." Cancer Res **68**(2): 425-433.

[10] Hu, X., D. M. Macdonald, et al. (2009). "A miR-200 miRNA cluster as prognostic marker in advanced ovarian cancer." Gynecol Oncol **114**(3): 457-464.

[11] Guo, L. M., Y. Pu, et al. (2009). "MiRNA-9 inhibits ovarian cancer cell growth through regulation of NF-kappaB1." FEBS J **276**(19): 5537-5546.

[12] Laios, A., S. O'Toole, et al. (2008). "Potential role of miR-9 and miR-223 in recurrent ovarian cancer." Mol Cancer **7**: 35.

[13] Shell, S., S. M. Park, et al. (2007). "Let-7 expression defines two differentiation stages of cancer." Proc Natl Acad Sci U S A **104**(27): 11400-11405.

[14] Quaresma, Alexandre J.; Nickerson, Jeffrey A. (2013). "Regulation of mRNA export by the PI3 kinase/AKT signal transduction pathway", *Mol Biol Cell* **8** (8): 1208–21.

[15] Von Neumann, J. and A. W. Burks (1966). "Theory of self-reproducing automata." Urbana, University of Illinois Press.

[16] S Wolfram (1983)."Statistical mechanics of cellular automata" Reviews of modern physics(55): 601-644.