利用高通量定序資料預測老鼠微核醣核酸轉錄起始位置

方浩宇[1] 許郁彬[1] 朱虹瑄[1] 謝立青[1,2,*]
[1]國立中興大學基因體暨生物資訊研究所 [2]國立中興大學生物科技發展中心
[*]liching@nchu.edu.tw

摘要

微核醣核酸 (microRNA; miRNA) 是一種非編碼小片段核醣核酸 (non-coding small RNA)，它的主要功能為抑制訊息核醣核酸 (messenger RNA; mRNA) 的轉錄進而影響蛋白質表現與否。確認微核醣核酸的轉錄起始位 (transcription start sites; TSSs) 對於研究微核醣核酸上游調控網路相當重要。目前已經有少數分析人類微核醣核酸轉錄起始位的研究正在進行，在本研究我們則專注在老鼠微核醣核酸轉錄起始位位置的預測，因為預測老鼠微核醣核酸轉錄起始位對於微核醣核酸上游調控網路演化的研究將有很大的助益。我們整合了 Cap Analysis of Gene Expression (CAGE) 與 Transcription Start Sites Sequencing (TSS Seq) 這些利用高通量定序得到與基因起始位置相關的研究資料，並利用支援向量學習機 (Support Vector Machine; SVM) 方法有系統地預測老鼠微核醣核酸轉錄起始位。最後我們利用支援向量學習機找到可能含有轉錄起始位的區域，並利用表現序列標籤 (Expression Sequence Tag; ESTs) 與序列保守性 (Sequence Conservation) 作為判斷轉錄起始位的依據。

關鍵詞：微核醣核酸；轉錄起始位；老鼠；高通量定序；支援向量學習機

Abstract

MicroRNAs (miRNAs) are non-coding small RNAs that inhibit protein coding gene expression by hybridizing with messenger RNAs (mRNAs). MiRNAs are involved in a lot of diverse biological processes and various diseases. To identify miRNA transcription start sites (TSSs) is important for studying the upstream regulatory networks of miRNAs. Up to now the studies regarding miRNA TSS identification are all focus on human miRNAs. We are interested in other species and our aim in this study is to identify mouse miRNA TSSs and the result would contribute to understanding the evolution of upstream regulatory networks of miRNAs.

In this study, we integrated two types of high-throughput sequencing data, i.e. transcription start sites sequencing (TSSseq) and Cap Analysis of Gene Expression (CAGE), as the evidence of miRNA TSSs. A machine-learning-based Support Vector Machine (SVM) was developed to identify mouse miRNA TSSs. In addition, we also incorporated the ESTs (expression sequence tag) and sequence conservation information to provide evidence for mouse miRNA TSSs.

Keywords: miRNA; transcription start sites; mouse; Support Vector Machine

1. 介紹

微核醣核酸 (microRNA; miRNA) 為一段大約 22-nt (nucleotide) 單位大小的核甘酸序列，微核醣核酸會藉由結合在訊息核醣核酸 (messenger RNA; mRNA) 的 3’端未轉譯區 (3’ - untranslated regions; 3’UTR) 來抑制訊息核醣核酸的表現或使訊息核醣核酸降解，進而使得蛋白質無法被合成。而這種調控基因表現的功能，在生理上可能與細胞代謝、細胞凋亡以及其他生物功能息息相關，同時其功能也牽涉到癌症的病程發展 [1-3]，因此有關於微核醣核酸的研究非常重要。

微核醣核酸的合成是一個複雜的過程，當一段表現微核醣核酸的基因序列藉由核糖核酸聚合酶 Ⅱ開始進行轉錄，其轉錄後的核糖核酸序列上會有部分區域形成髮夾狀的結構，此時這段核糖核酸序列稱為初始微核糖核酸 (primary-miRNA; pri-miRNA)。之後這一段序列上的髮夾結構會被酵素 Drosha 切下生成前驅微核糖核酸 (precursor-miRNA; pre-miRNA)，再經由酵素 Dicer 將前導微核糖核酸上的環狀結構切掉，形成一個雙股微核醣核酸 (miRNA-miRNA* duplex)，而其中的一段便是參與基因調控的成熟微核醣核酸 (mature-miRNA)。成熟微核醣核酸會與一些蛋白質形成核糖核酸誘導沉默複合體 (RNA-induced silencing complex; RISC) 的蛋白質複合體，核糖核酸誘導沉默複合體與蛋白質編碼基因 (protein coding gene) 上的微核醣核酸標靶位置 (miRNA target) 結合後會使被結合的訊息微核醣核酸被降解或是無法進行轉譯。這種獨特的功能使微核醣核酸在基因調控網路上扮演著重要角色，進而影響到各種生物的生理與代謝功能。

近年對於微核醣核酸的研究也著重於微核醣核酸標靶，例如 TarBase, miRecords, miR2Disease 和 miRTarBase 都有許多關於微核醣核酸標靶位置的預測或實驗資料 [4-7]。而研究微核醣核酸的上游例如微核醣核酸的轉錄起始位置相對來說卻很少。

微核醣核酸轉錄起始位置的預測是相當困難的，miR-34a 是一個已知的微核醣核酸，老鼠微核醣核

酸 miR-34a 髮夾結構所在與轉錄起始位置有一段距離 (~2 萬 nts)，微核醣核酸與轉錄起始位置的距離可能很遠，這就是預測微核醣核酸轉錄起始位置困難的原因[8]。

由於科技發展與計算機科學的進步，傳統的定序方法隨著機械學、生物資訊、電腦資料庫和儀器發展，起了革命性的變化。次世代定序 (next-generation sequencing) 技術，又稱為高通量定序 (high-throughput sequencing) 技術的誕生對生物基因體學有著重大的影響；高通量定序大幅度地降低定序的成本；隨著新平台的發展，目前一次可定序多達超過數十億個鹼基。高通量定序的方法對於泛基因體 (genome-wide) 核醣核酸相關的分析和非編碼核醣核酸 (noncoding RNA) 的發現等研究有很大的幫助。次世代定序可以同時定序大量的序列資料，藉由對定序後的小片段序列進行比對、組合，次世代定序對於基因體的研究有很大的幫助 [9, 10]。

高通量定序在近年的發展中，有許多衍生而出的應用，像是可以知道蛋白質與序列關係的 ChIP-seq 法，以及可以建立轉錄體資訊的 RNA-seq 法 [11]等等。由於轉錄起始位置的資訊位於訊息核糖核酸的 5 等端，所以本研究所使用的資料需要帶有訊息核醣核酸 5’端的資訊；TSSseq (transcription start sites sequencing) 和 CAGE (cap analysis of gene expression) 也是高通量定序延伸出來的定序應用，而這兩種資料皆帶有靠近基因 5 序端的資訊。

目前已經有一些對於微核醣核酸上游調控區域的相關研究；由於微核醣核酸基因起始表現與蛋白質編碼基因相同需要有核糖核酸聚合酶 II (RNA polymerase II) 的結合，所以有些要尋找微核醣核酸轉錄起始位置 (transcription start sites; TSSs) 之研究會使用核糖核酸聚合酶 II 結合 ChIP-seq (Chromatin immunoprecipitation sequencing) [12]方法的實驗資料，來為微核醣核酸轉錄體 5’端建立模型，藉此判斷微核醣核酸轉錄起始位 [13, 14]。除了核糖核酸聚合酶 II 的 ChIP-seq 資料，現在也有一些高通量定序是專注於研究基因轉錄起始位，諸如 TSSseq, CAGE [15]和組蛋白修飾作用 (histone modification) [16]等研究，這些特徵都與基因轉錄起始區域有關；曾有研究利用上述的特徵為基因的轉錄起始位置建立模型，並依此為依據來預測人類微核醣核酸的轉錄起始位得到很好的結果[17]。

本研究利用上述類似方法針對老鼠的微核醣核酸作轉錄起始位置的預測，希望藉由研究人類以外的物種來觀察微核醣核酸的上游調控網路與物種演化的關係。

2. 研究方法

2.1 資料收集

本研究從 DBTSS [18]中收集老鼠有關 TSSseq 的研究資料，從 FANTOM4 [19, 20]中收集有關老鼠 CAGE 的研究資料，這兩組資料中有多種細胞株，TSSseq 有 4 種細胞株 (3T3,10Thalf, ATDC5, embryo)；CAGE 中則有 14 種細胞株 (cerebellum, Embryo, lung, liver, macrophage, blood , heart, prostate, muscle, heart, visual cortex, somatosensory, preadipocyte, brain)，採用多種的細胞株能夠找到更多樣化的基因表現。

DBTSS 中的資料庫也含有老鼠的蛋白質編碼基因起始位置，由於蛋白質編碼基因與擁有自己的轉錄起始位的微核醣核酸有著類似的轉錄起始位特性，故將這些蛋白質編碼基因起始位置用以建構 TSSseq 和 CAGE 資料在機械語言學習法之中的模型，該模型將作為預測基因起始位置的判定標準。為了確認蛋白質編碼基因起始位置是否會與基因序列重疊，我們從 Ensembl Genome Browser [21]收集老鼠的蛋白質編碼基因序列位置，若預測的位置與蛋白質編碼基因重疊則不予以採用。另外收集來自於 UCSC [22]之序列保守性 [23]和表現序列標籤 [24]的相關資料，這些資料將被用於微核醣核酸轉錄起始位置預測的可信度判斷。

miRBase 是一個收錄許多物種前導微核醣核酸或成熟微核醣核酸資訊的資料庫網站 [25]，本研究從 miRBase 中收集關於老鼠的微核醣核酸資料，在之後的研究中將預測這些微核醣核酸的轉錄起始位置。微核醣核酸依其在序列的位置有兩種分類，若微核醣核酸的位置在一個有蛋白質編碼基因序列內，稱為基因內 (intragenic) 微核醣核酸，通常這類微核醣核酸會與其宿主基因 (host gene)擁有共同的轉錄起始位置；反之微核醣核酸若位於沒有表現蛋白質的基因序列上則稱為基因間 (intergenic) 微核醣核酸，此類微核醣核酸擁有自己的轉錄起始位，而我們的主要預測目標為基因間微核醣核酸。從 mirBase 的資料庫下載的有關老鼠的微核醣核酸，經由篩選將基因內微核醣核酸的資料去除，剩下來的微核醣核酸為沒有與基因重疊的基因間微核醣核酸，總共有 1047 個微核糖核酸資料，從中篩選出有 178 個基因間微核醣核酸。

本研究採用老鼠 GRCm38 (mm10) 的基因圖譜版本，從 DBTSS 和 CAGE 收集的關於 TSSseq 與 CAGE 相關資料，其基因圖譜版本皆為 GRCm37。為了使基因圖譜版本一致，我們利用來自於 UCSC 的 liftover 軟體，將 GRCm37 轉換為 GRCm38。本實驗的 TSSseq 和 CAGE 的研究資料皆經由 liftover 轉換為 GRCm38 基因圖譜版本。

2.2 模型建立

從 DBTSS 的資料庫中下載獨立蛋白質編碼基因的轉錄起始位置，將每個轉錄起始位置取前後 1100 bp，共有 2200 的範圍內以分別 200 bp 為單位分隔出 11 格的框架，分別計算每個區塊中 TSSseq 標籤和 CAGE 標籤的密度。目前有 11415 個蛋白

質編碼基因轉錄起始位置的資料將其作為正向資訊，如圖 1 所示；另外再以每個蛋白質編碼基因轉錄起始位置的前後 10000 bp 的位置，分別計算其前後 1100 bp 位置中 2200 bp 的範圍內每隔 200 bp 區塊中 TSSseq 標籤和 CAGE 標籤的密度，這部分的資料作為負向資訊。將正向資訊和負向資訊作為支援向量學習機的訓練資料，去產生一個支援向量學習機模型，此支援向量學習機模型將用於尋找轉錄起始位置。本研究利用一套共享的向量學習機程式 LibSVM [26]來訓練與建立模型。

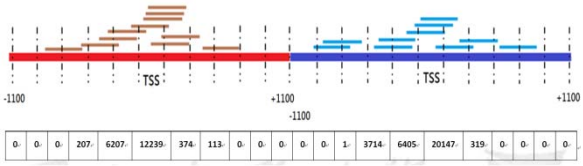

圖 3. 正向支援向量學習機模型

2.3 預測轉錄起始位置

利用從 miRBase 中得到的 178 個老鼠前驅微核醣核酸位置資訊，有研究指出 2 個微核醣核酸若距離在 50000 bp 範圍內則可能共用一個轉錄起始位 [27]，所以我們取微核醣核酸上游 50000 bp 的範圍內，從微核醣核酸的位置，每隔 100 bp 的位置，取前後各 1100 bp 的範圍內以分別 200bp 為單位分隔出 11 格的區塊，分別計算每個區塊中 TSSseq 標籤和 CAGE 標籤的密度。利用建立的支援向量學習機模型去掃描上述資料尋找符合正向資訊的位置，藉此幫助我們找出可能含有轉錄起始位置的區域。整體資料分析流程請見圖 2。

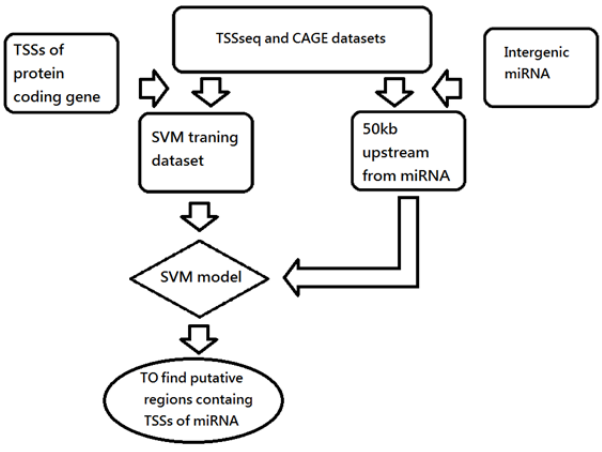

圖 2. 資料分析流程圖

同時為了提高準確性，參考相關的研究論文 [17]，還需要符合以下條件: (1) 被支援向量學習機判斷為正向。(2) 被判斷為正向區域且不能與蛋白質編碼基因序列重疊。(3) 表現序列標籤和基因保留性的特徵。(4) 距離微核醣核酸的上游位置愈近愈好。

3. 結果

3.1 TSSseq 和 CAGE 與轉錄起始位置

一般我們知道核糖核酸聚合酶 II 可以用來判斷 mRNA 轉錄的起始位置，因為核糖核酸聚合酶 II 與基因序列結合的位置與轉錄起始位置極為接近。所以同理也可以運用於微核醣核酸 轉錄起始位的預測，而除此之外，TSSseq 和 CAGE 的提供的實驗數據也可以為我們用來研究微核醣核酸的轉錄起始位，TSSseq 和 CAGE 的實驗包含有一個轉錄體的 5'端資訊，這些 5'端的資訊也是最接近轉錄起始位的所在，為了驗證此法的可行性，可以從一般的蛋白質編碼基因驗證。將總共 11415 個從 DBTSS 中收集到的蛋白質編碼基因的轉錄起始位置，分別計算 TSSseq 標籤和 CAGE 標籤在蛋白質編碼基因轉錄起始位置附近的平均表現量，從轉錄起始位置開始到其前後 1100 位置，以 200 單位為框架，如圖 3 所示，從已知蛋白質編碼基因的轉錄起始位區域都可以觀察到 TSSseq 標籤和 CAGE 標籤的表現量集中的訊號，故利用 TSSseq 標籤和 CAGE 標籤的實驗資料也可以應用於微核醣核酸 轉錄起始位置預測，這些研究資料也將用於支援向量學習機中訓練模型的資料。

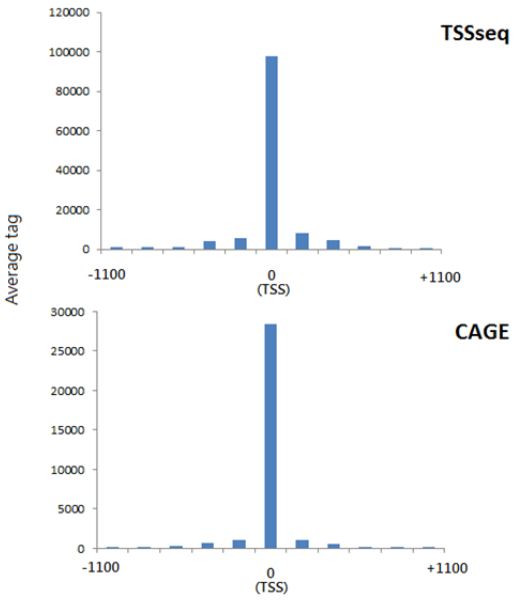

圖 1. TSSseq 標籤和 CAGE 標籤在蛋白質編碼基因轉錄起始位附近的平均的表現量

3.2 利用支援向量學習機有系統地預測微核醣核酸轉錄起始位

支援向量學習機是一種用於分類的機器學習，主要是嘗試在空間中，尋找一個超平面 (Hyperplane) 能將不同類別的資料完美的分開，同時此超平面能與不同的類別的資料距離愈大愈好。為了找到 TSSseq 與 CAGE 於微核醣核酸的上游有高表現的區域，我們需要訓練的資料也必須具有類似的資訊，DBTSS 的資料庫中有許多老鼠的表現蛋白質基因的轉錄起始位資料，而 TSSseq 和 CAGE 的實驗資料在這些老鼠表現蛋白質基因的轉錄起始位置上，大部分都呈現高的表現，所以從 DBTSS 裏篩選出獨立的已知蛋白質編碼基因的轉錄起始位置，特別選用獨立的轉錄起始位，是因為獨立的已知蛋白質編碼基因轉錄起始位由於沒有相近的轉錄起始位置，所以 TSSseg 和 CAGE 的才會統一在轉錄起始位置有大量表現，可以由圖 4 老

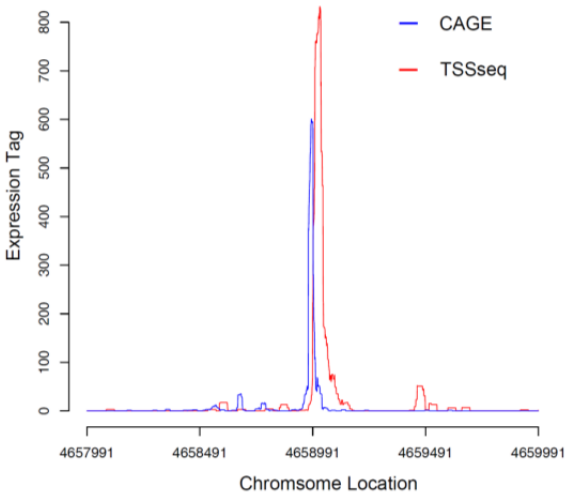

圖 4. 獨立轉錄起始位蛋白質編碼基因 (NM_172894)

鼠 NM_172894 基因的例子看出。而圖 5 可以看出含有叢集 (cluster) 轉錄起始位置的蛋白質編碼基因 NM_172894 會因其轉錄起始位相距太近使得 CAGE 和 TSSseq 的表現不會只集中在轉錄起始位的位置。而共有 11415 組的資料從 DBTSS 中篩選出配合 TSSseq 和 CAGE 的表現作為支援向量學習機訓練的資料。在支援向量學習機的預測下，老鼠的 178 個微核醣核酸可以在其中 138 個微核醣核酸找到被判斷為具有轉錄起始位之正向資訊的區域。

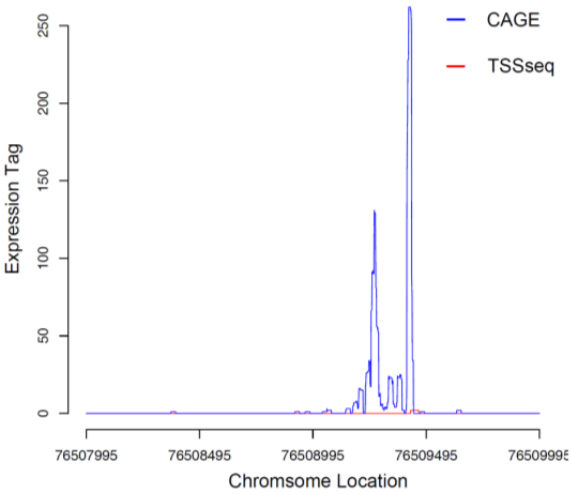

圖 5. 叢集轉錄起始位蛋白質編碼基因 (NM_198894)

3.3 微核醣核酸轉錄起始位預測結果

將 11935 個獨立的蛋白質編碼基因轉錄起始位取前後 1100 位置，分別 200bp 為單位計算每個區塊中 TSSseq 和 CAGE 表現的密度，以此做為支援向量學習機訓練模型的資料，再來從 mirBase 收集 178 個老鼠基因間微核醣核酸，以每個基因間微核醣核酸序列起始位置到上游 50000 的範圍內，每隔 100 單位取前後各 1100 的範圍內以分別 200bp 為單位延伸出 11 格的區塊，分別計算每個區塊中 TSSseq 標籤和 CAGE 標籤的密度。將這些基因間微核醣核酸上游中若有被判斷為正向資訊的區域，這些區域為可能有轉錄起始位的區域，選出這些區域中具有 TSSseq 和 CAGE 高表現的位置，做為可能的轉錄起始位，除了 TSSseq 和 CAGE 的高表現外，表現序列標籤和序列保守性這兩種特徵可以判斷序列在基因中是否表現以及演化的重要性，故用作判斷轉錄起始位的依據，UCSC 網站有老鼠相關的表現序列標籤和序列保守性資料，利用這兩種特徵來看是否與具有 TSSseq 和 CAGE 高表現位置可能的轉錄起始位重疊，同時這些序列不能與蛋白質編碼基因序列重疊。若以上條件符合，則此位置很有可能為轉錄起始位。分析結果從 178 個基因間微核醣核酸中找到了 138 個符合分析區域不與蛋白質重疊的微核醣核酸，再利用支援向量學習機方法結合以上的特徵分析找到 79 個基因間微核醣核酸的轉錄起始位置（請見補充資料）。

3.4 預測準確度

有關於實驗的準確度，可以從敏感性 (sensitive)、特異性 (specific)、準確性 (accuracy)、精確性 (precision) 4 種參數判斷，將用來製作支援

向量學習機模型的資料放入支援向量學習機模型進行預測 (表 1) 。

表 1. 預測準確度

Sensitive	Specific	Accuracy	Precision
98.7%	92.5%	94.4%	84.7%

4. 討論

4.1 預測方法的優勢

微核醣核酸調控蛋白質基因,使得它在生物學上扮演重要的調控角色。但想了解整個微核醣核酸調控網路，就需要了解上游調控微核醣核酸的轉錄因子 (transcription factors),所以一個好的預測微核醣核酸轉錄起始位置方法用於定義啟動子區域 (promoter region)是很重要的。本篇藉由兩種與轉錄起始位置有關的高通量定序的實驗,來預測 mirBase 中的基因間微核醣核酸轉錄起始位置，實際上此方法在人類微核醣核酸轉錄起始位置上已有成功研究的範例[17]。而此方法用於老鼠上，經由之前的分析，也可以觀察到 TSSseq 和 CAGE 的實驗資料在轉錄起始位置的表現雷同，故此預測方法在老鼠上也是可行的。但本實驗中沒有用到組蛋白修飾作用的相關實驗資料，原因在於現在的網路資料庫並沒有現成的有關於老鼠組蛋白修飾作用 Chip-seq 序列定位的資料，相較之下缺乏一個特徵可能會影響到微核醣核酸轉錄起始位置的預測。

4.2 預測的困難及改善

微核醣核酸與轉錄起始位置的距離通常很遠，可以由本實驗的結果統計微核醣核酸與轉錄起始的位置，在 50kb 的範圍內都可以找到轉錄的起始位置，但這種特性增加了微核醣核酸轉錄起始位置預測的困難，結果中有部分的微核醣核酸沒有找到轉錄起始位置，表示這些微核醣核酸的轉錄起始位無法被本實驗的特徵所辨認，其原因可能為特徵不足，使得部分微核醣核酸沒有訊號而被辨認，另外也有可能部分微核醣核酸只能利用核醣核酸聚合酶 II 所辨認，再來是這些起位置超出了 50kb 之外。

為了能找出未辨認微核醣核酸轉錄起始位置，方法的進一步改善是必須的，可以加入組蛋白修飾作用 特徵來為微核醣核酸做預測，或是加入 RNA-seq 的資料，藉由與完整的基因轉錄體互相比對找出微核醣核酸的起始位置。

4.3 老鼠與人類微核糖核酸轉錄起始位的比對

我們將老鼠跟人類的微核醣核酸基因轉錄起始位做比對，其中老鼠微核醣核酸的轉錄起始位置在預測時只選出為帶有表現序列標籤或序列保守性的結果，在算出轉錄起始位置與微核醣核酸的距離。再以相同的預測方法用在人類微核醣核酸上（數據沒有呈現），不過人類在預測上不一定具有表現序列標籤或序列保守性的驗證，比對後可以得到表 3 與表 4 的結果，其中表 4 為距離相近且共用轉錄起始位置的叢集微核醣核酸，這些資料可以觀察到除了一組微核醣核酸叢集外（mir-106a, mir-20b, mir-363 與 mir-92a-2），大部分的人類微核醣核酸與其轉錄起始位置的距離是大於老鼠的，這表示在演化的過程中，人類微核醣核酸與其轉錄起始位置之間多出了許多序列，而這些多出的序列是否有意義或帶有功能且有利於人類是個有趣的問題。

表 2. 微核醣核酸與轉錄起始位置的距離

miRNA	Species	Distance between TSS and miRNA
mir-451a	human	36318
	mouse	35253
mir-193a	human	10294
	mouse	7328
mir-150	human	46188
	mouse	17934
mir-192	human	44646
	mouse	28601
mir-298	human	21384
	mouse	16705
mir-296	human	22023
	mouse	16932
mir-30b	human	32103
	mouse	25668
mir-30d	human	27765
	mouse	21780

表 3. 微核醣核酸叢集與轉錄起始位置的距離

miRNA cluster	Species	Distance between TSS and miRNA cluster
mir-450, mir-542	human	8412
	mouse	7703
mir-221, mir-222	human	20845
	mouse	20262
mir-106a, mir-20b, mir-363, mir-92a-2	human	3613
	mouse	30193

從 Ensembl 的資料統計關於蛋白質編碼基因的編碼區序列與內含子的平均總長，從表 5 看出人

類的編碼區序列 (Coding Domain Sequence; CDS) 與老鼠的編碼區序列的平均總長度極為接近，而內含子的平均總長度變化卻很大。而其中可能的原因是編碼區序列中的外顯子是蛋白質編碼基因中轉譯蛋白質的中心區域，只要外顯子減短或增長都會對生物造成不可預期的影響，所以演化上外顯子的序列大部分保留下來，長度也幾乎不變。而在微核糖核酸的部份，由於微核糖核酸基因也具有內含子，所以微核醣核酸與其轉錄起始位置演化上的距離的變化很有可能主要是受到內含子長度變化的影響。但由於微核糖核酸基因是非編碼基因，它的轉錄主要來自微核糖核酸的結構性，所以演化中其非結構性限制部分之外顯子序列長度的改變很有可能不會影響微核糖核酸的功能，這也可能造成人類微核糖核酸與老鼠微核糖核酸與轉錄起始位置之間的距離改變。

表 4. 老鼠與人類每個蛋白質編碼基因外顯子與內含子平均總長

Species	Average CDS size	Average combined intron size
mouse	1369.50	39011.00
human	1229.37	47038.29

誌謝

本論文承蒙國家科學委員會之計畫經費贊助，計畫編號為 NSC 100-2621-B-005-001，僅此誌謝。

參考文獻

[1] I. Alvarez-Garcia, E.A. Miska, "MicroRNA functions in animal development and human disease", Development, vol. 132. 2005, pp. 4653-4662.

[2] N. Bushati, S.M. Cohen, "microRNA functions", Annual review of cell and developmental biology, vol. 23. 2007, pp. 175-205.

[3] G.A. Calin, C.M. Croce, "MicroRNA signatures in human cancers", Nature reviews. Cancer, vol. 6. 2006, pp. 857-866.

[4] S.D. Hsu, F.M. Lin, W.Y. Wu, C. Liang, W.C. Huang, et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions", Nucleic acids research, vol. 39. 2011, pp. D163-169.

[5] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease", Nucleic acids research, vol. 37. 2009, pp. D98-104.

[6] G.L. Papadopoulos, M. Reczko, V.A. Simossis, P. Sethupathy, A.G. Hatzigeorgiou, "The database of experimentally supported targets: a functional update of TarBase", Nucleic acids research, vol. 37. 2009, pp. D155-158.

[7] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, et al., "miRecords: an integrated resource for microRNA-target interactions", Nucleic acids research, vol. 37. 2009, pp. D105-110.

[8] V. Tarasov, P. Jung, B. Verdoodt, D. Lodygin, A. Epanchintse, et al., "Differential Regulation of microRNAs by p53 Revealed by Massively Parallel Sequencing", Cell Cycle, vol. 6:13,. 2007, pp. 1586-1593.

[9] E.R. Mardis, "Next-generation DNA sequencing methods", Annual review of genomics and human genetics, vol. 9. 2008, pp. 387-402.

[10] E.R. Mardis, "The impact of next-generation sequencing technology on genetics", Trends in genetics : TIG, vol. 24. 2008, pp. 133-141.

[11] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq", Nature methods, vol. 5. 2008, pp. 621-628.

[12] L.J. Zhu, C. Gazin, N.D. Lawson, H. Pages, S.M. Lin, et al., "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data", BMC bioinformatics, vol. 11. 2010, pp. 237.

[13] D.L. Corcoran, K.V. Pandit, B. Gordon, A. Bhattacharjee, N. Kaminski, et al., "Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data", PloS one, vol. 4. 2009, pp. e5279.

[14] Y.W. Guohua Wang, Changyu Shen, Yi-wen Huang, Kun Huang, Tim H. M., K.P.N. Huang, Lang Li, Yunlong Liu, "RNA Polymerase II Binding Patterns Reveal Genomic Regions Involved in MicroRNA Gene Regulation", PloS one, vol. 2010, pp.

[15] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, et al., "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage", Proceedings of the National Academy of Sciences of the United States of America, vol. 100. 2003, pp. 15776-15781.

[16] S. Dambacher, M. Hahn, G. Schotta, "Epigenetic regulation of development by histone lysine methylation", Heredity, vol. 105. 2010, pp. 24-37.

[17] C.H. Chien, Y.M. Sun, W.C. Chang, P.Y. Chiang-Hsieh, T.Y. Lee, et al., "Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data", Nucleic acids research, vol. 39. 2011, pp. 9345-9356.

[18] R. Yamashita, H. Wakaguri, S. Sugano, Y. Suzuki, K. Nakai, "DBTSS provides a tissue specific dynamic view of Transcription Start Sites", Nucleic acids research, vol. 38. 2010, pp. D98-104.

[19] H. Kawaji, J. Severin, M. Lizio, A.R. Forrest, E. van Nimwegen, et al., "Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation", Nucleic acids research, vol. 39. 2011, pp. D856-860.

[20] H. Kawaji, J. Severin, M. Lizio, A. Waterhouse, S. Katayama, et al., "The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation", Genome biology, vol. 10. 2009, pp. R40.

[21] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, et al., "Ensembl 2012", Nucleic acids research, vol. 40. 2012, pp. D84-90.

[22] L.R. Meyer, A.S. Zweig, A.S. Hinrichs, D. Karolchik, R.M. Kuhn, et al., "The UCSC Genome Browser database: extensions and updates 2013", Nucleic acids research, vol. 41. 2013, pp. D64-69.

[23] T. Doerks, R.R. Copley, J. Schultz, C.P. Ponting, P. Bork, "Systematic identification of novel protein domain families associated with nuclear functions", Genome research, vol. 12. 2002, pp. 47-56.

[24] S.H. Nagaraj, R.B. Gasser, S. Ranganathan, "A hitchhiker's guide to expressed sequence tag (EST) analysis", Briefings in bioinformatics, vol. 8. 2007, pp. 6-21.

[25] A. Kozomara, S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data", Nucleic acids research, vol. 39. 2011, pp. D152-157.

[26] C.C. Chang, C.J. Lin, "LIBSVM: A Library for Support Vector Machines", Acm T Intel Syst Tec, vol. 2. 2011, pp.

[27] S. Baskerville, D.P. Bartel, "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes", Rna, vol. 11. 2005, pp. 241-247.