物種間同源性關係分析
和
Newick 序列處理

國立高雄應用科技大學 資訊工程系

指導教授：鐘文鈺 教授

成員：1103108112 陳敏涵
1103108136 鄭文涵
1103108120 黃雅筠
1103108132 李少榮
1103108104 陳妍臻

內容

圖表目錄

摘要

生物科技研究的蓬勃發展，積累了各式龐大的資料與數據，例如 DNA 序列、胺基酸序列及其結構等。為了分析如此龐大卻仍舊爆炸性成長的分子資料，近年來衍生出了一門新興科學—生物資訊學。生物資訊學是一門結合生物學、計算機科學、統計學等學科所形成的新研究領域，其中包括分析大量生物資料的統計方法及演算法設計。

承上所述，現今的生物資訊分析也有兩點特色：處理資料量大且重複性質的步驟多、針對不同分析方法需運用不同軟體，也就是說生物學家們在做一項研究時，會需要分析多種物種，而這些物種可能都需要做同樣的步驟分析、處理，才能看出彼此之間的差異性。生物學家們在做序列比對時可能需要用到 A 軟體、在做結構域分析時要用 B 軟體、再畫同源關係樹時可能又必須用到 C 軟體。一項研究的過程都有高重複性工作且必須運用多種套裝軟體。

所以，鑒於現今物種同源性分析的方式過於繁雜，使得研究效率不彰，本專題目標：建構一個完整的方法，透過程式整合所有的流程，讓使用者可以藉由此系統，預測特定基因，分析物種演化的親疏關係。另外，我們將最後的結果轉換成 Newick 格式，並創新相關的讀取、分析功能，加強工具的實用性。

藉由不同領域的結合，程式設計不僅能解決反覆的操作，更能為生物研究提供大數據分析，以提高研究的效率，透過生物多樣性分析，進而推測物種習性，說明各物種間的差異，進行分類並判斷物種種群及數量，以進行養殖、棲地保育、保護，進而達到物種復甦、經濟養殖等目標。

研究動機與研究問題

近年來，有關於物種及其演化的基因體研究突飛猛進，部份甚至已經應用在相關領域上，例如美國能源部和國家衛生研究院所主導的國際人類基因體計畫，便對醫學、生物學，乃至整個生命科學產生無可估量的影響，其中包括了解病人完整的基因序列有助於醫療人員提供適當的治療、核醣體及器官的出現有利於生物進化和演變的研究等。基因體不僅僅只是基因的簡單排列，其經過長期的演化產生特有的結構，這些結構與基因所發揮的功能有著不可或缺的關係，因此，生物基因體的研究是解釋生命遺傳語言的關鍵。

隨著各項生物研究計畫的進行，越來越多的生物序列資料被建構成資料庫，並置於網際網路上免費供人使用，V2R(vomeronasal-type2 odorant receptor gene)基因分析[1]作為啟發我們研究的論文，即在 NCBI 上使用相關的 Blast 軟件進行 DNA 和蛋白質序列的對庫檢索，並針對相似性高的序列做兩端延伸以進行基因預測，通過 WISE2 和 HMMER 確定是否為基因的片段後，接著使用 TMHMM 塞選出屬於 V2R 基因的部份，其中 ClastalW 和 MAFFT 幫助將用於基因預測的序列做對齊，最後利用 MEGA 建構系統發育樹。

由上述的步驟可知，分析物種之間同源性關係所需要採用的網路資源和套裝

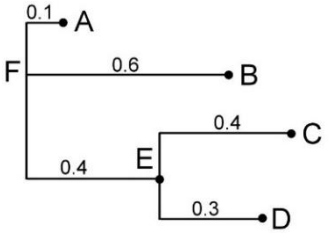
軟體多達 7 種。雖然在過去電腦尚未普及的年代，生物學家得純粹靠眼睛進行序列資料的比對和搜尋，過程繁複且耗時長久，現今應用軟體比對分析，縮短不少時間，卻還是要經過 7 種步驟才能實現，更不用說大量的數據分析了。而新物種的發現、DNA 的重新定序…等，每一條新資訊的出現或改變都會牽動著人們去重新推論關於演化關係的結果，若每條基因序列的分析都須經過這 7 道手續，那 10 種、100 種物種呢？假使每個物種有 5、6 條基因，又需要操作多少次？若能將這些重複性質高的操作，藉由程式來達成目的，甚至濃縮為只需要一條指令，是否會方便許多呢？

可想而知，這不僅能大大提高生物學家在作業上的效率，加速生物科技研究的發展，也同時降低了應用上的難度，甚至能惠及更多的使用者。試想一下，生物學家在針對規劃好的流程選定適用的軟體前，首先得熟悉各項資源的操作，這些資源大多零散分布在網路上，若不是經常做序列分析，或是只用到序列分析中的某幾種功能，使用網路資源的確會是一個很好的選擇，但對於需要做大量分析的人而言，有系統完整規劃過的程式顯然更優於網路上那些零碎的資源。因為套裝軟體可藉由自製程式將分析自動化，而網路資源卻為了防範駭客將自身作為跳板進行惡意攻擊，設計了相關的保護機制封鎖外掛程式，以致於只能手動一個接著一個步驟執行。舉例來說，一個方式只要透過本機端輸入指令即可取得相似性高的序列資料，另一個方式卻只能在網頁端一筆筆的載下來，更不用說將整個流程簡化，省去一條一條輸入資料和指令的麻煩，效率上及使用門檻上的差距是顯而易見的。

因此，我們將 V2R 基因分析[1]的步驟作為依據，設計出一個完整的工具，涵括所有的流程，為了提高工具應用的寬廣度，有別於論文中只是針對 V2R 基因做物種的同源性分析，我們的系統讓使用者能運用此工具輸入任意一條要比對的蛋白質序列資料，獲取相關物種的同源性關係。

該論文中也提到，嗅覺在大多數物種的行為上扮演著極其重要的角色，因此我們採用 V2R 基因作為判定物種之間同源性關係的依據。由於此篇論文刊載於十多年前，不僅序列資料庫隨著相關的研究有所更新，分析工具也由於前人的修改變得更加精確，因此，本專題的研究有助於提高資訊的正確性與準確度。

普通的圖像輸出雖然能直觀的理解演化關係，但卻難以用程式設計做進一步的分析，對電腦來說，文字遠比圖片能做出更準確且有效率的操作，因此我們將結果輸出為 Newick 格式和樹狀圖兩種，以適應更多情況下的數據分析要求，其中 Newick 格式便於保存，當未來想運用相似物種進行分析時，許多軟體都可直接匯入。



圖表 1 演化樹示意圖

文獻回顧與探討

為了達成預期的目標，除了啟發研究的 V2R 基因分析[1]外，我們還閱讀了相關的論文及書籍，藉此熟悉軟體的操作，並幫助我們完善所需的功能。

提高效率

為何生物學需要統計呢？我們知道，許多資料經由彙整，成為有用的資訊，而透過分析資訊，可歸納出知識。同理，我們將 DNA 和蛋白質序列進行比對並分析兩者的異同處，以期能夠推測出它們的結構、功能及進化上的聯繫，其中如何尋找資訊的相似性，即屬於統計問題。但生物資訊所涉及的資料量相當龐大，想要設計出能在最短時間內完成搜尋及比對的工具，關鍵就在於演算法。

生物資訊演算法[2]這本書便簡略的提到幾個常見的演算法，並講解如何計算時間複雜度來衡量一個演算法是否有效率。該書內容也簡介了有關生物學方面的知識，藉由物種演化和序列的搜索比對延伸說明生物資訊學的起源與發展，這本書主要側重在描述解決生物問題時會應用到哪些基礎的算法與原理，幫助大家獲取所需的資訊。我們透過這本書所學到的知識，熟悉演算法在生物資訊上的應用，並優化程式碼增加執行的效率，藉以提高工具的實用性。

序列搜尋程式

NCBI 使用手冊[3]內容主要說明如何使用 NCBI(National Center for Biotechnology Information)這個網站來進行相似基因的搜尋，並且逐步介紹過程中可能會使用的工具。從這篇文章中我們找到一些在研究初期可以幫助我們快速進行研究的工具，比方 NCBI 內建資料庫其中附帶的特定基因搜尋、網頁版的 blast 工具以及如何解讀網頁版 blast 工具所產生出來的結果。雖然這些工具並不會出現在我們最終的程式中，但它們卻是可以讓我們快速了解本機端工具的使用，及解讀出我們所撰寫程式需要的資訊。

基因預測程式

真核基因預測[4]內容主要說明如何將演算法運用在真核基因的預測上，並且提高結果的準確性。我們將網路上蒐集的資料與該論文做結合，從中認識 DNA 的結構及其與蛋白質的轉換關係，透過了解基因預測的概念和方法，進一步得知可能存在的例外，再導入論文中閱讀。該論文所述基因預測的方法主要有兩種，其中 Wise2 屬於 expression based predictors 預測，根據已有的實驗證據(如 cDNA)、EST 和蛋白質序列進行蛋白質編碼基因的注釋；而 HMMER 則為 ab initio 的基因預測，根據基因體的 DNA 序列對蛋白質編碼基因進行預測。以下兩篇論文分別對這兩種方法做相關的介紹。

GeneWise and GenomeWise[5]提出兩種基因預測的演算法，能在使用正確數據的情況下得到準確且完整的基因結構，內容相較於上篇論文，更著重在算法及原理上。

HMMER[6]詳細解說了工具的原理及應用，該內容偏向測試取得相對準確的資料，於專題關聯性不高，但其中也有幾段論述值得我們參考，比如模擬研究，將

大量的序列資料重複測試，取得出現機率最高的模型，我們將此概念應用在基因預測上，藉以提高結果的準確度。

TMHMM[7]則說明如何運用工具將未確定的基因與已知的基因序列進行比對，比較兩者跨膜螺旋的相似性進而確認是否為相同的基因。我們從論文中了解TMHMM的原理及應用過程，並將此作法運用在篩選V2R基因上，驗證基因預測的結果。

演化樹分析

演化樹是運用不同資訊或數學演算法建構而成的，以節及分支來表示各物種間演化的親緣關係，其中每個節點都代表著重要的演化事件。

建樹演算法[8]便是針對當時基因體研究尚處於基礎階段，設計而成的一種系統發育樹的重建方法，滿足物種快速鑒定、系統發育樹自動增長以及結果準確等要求，該算法雖然無法完成大數據的構建，但其相關功能與本專題有所雷同，給我們帶來了不小的助力與啟發。

直系同源關係以及旁系同源關係對於基因之間功能性的關聯佔有很重要的地位。基因樹演算法[9]即針對上述兩種同源關係，提供了一個推斷基因樹上基因重複事件和物種分化事件的算法，該論文也透過簡介相關工具延伸系統發生學，說明在多數情況下，同源基因的鑑定不足以進行特定的功能預測，因為並非所有的同源基因都具有相同的功能，點明系統發生學的重要性。

Newick 格式是一種用文字替代圖形表示演化樹的方式，使用圓括號和逗號表示物種親緣關係，數字表示其演化的距離，我們將分析結果輸出成Newick格式並參考Newick format [10]提到的功能，新增幾種分析演化樹的方式，例如：分群、尋找最接近的物種關係、尋找距離內的物種…等，加強工具針對各種問題的實用性。

Newick格式可用以下八種方式解釋圖一：

(,,(,));	沒有節點被命名。
(A,B,(C,D));	葉節點被命名。
(A,B,(C,D)E)F;	全部的節點皆被命名。
(:0.1,:0.6,(:0.4,:0.3):0.4);	除了根節點之外的所有節點都有到父節點的距離。
(:0.1,:0.6,(:0.4,:0.3):0.4):0.0;	所有節點都有到父節點的距離。
(A:0.1,B:0.6,(C:0.4,D:0.3):0.4);	除了根節點之外的所有節點都有到父節點的距離，且所有葉節點皆被命名。(常用)
(A:0.1,B:0.6,(C:0.4,D:0.3)E:0.4)F;	除了根節點之外的所有節點都有到父節點的距離，且所有節點皆被命名。
((B:0.6,(C:0.4,D:0.3)E:0.4)F:0.1)A;	基於葉節點的樹。(很少見)

圖表 2 Newick格式說明表

研究方法及步驟

以下依序為我們的研究步驟，研究時間上的進度分配可參考圖二。

蒐集資料

初步蒐集資料，主要蒐集的資料為「物種同源性分析的方式」&「現今生物資訊分析所用的套裝軟體」，我們首先要先了解生物學家們做物種間同源性分析的流程、使用的工具。

以啟發我們研究這項專題的論文 V2R 基因分析[1]為依據，先初步蒐集論文中所提到工具的資料，清楚熟悉每項工具的使用方式。

目前我們已蒐集的資料已在參考文獻的部分提及，但在整理資料的過程中我們還遇到了許多問題。V2R 基因分析[1]中，只提到了分析的大致流程卻沒有每個步驟的詳細規劃，例如：

- 基因分析軟體所需用到的指令？
- 建立基因預測用的 HMM 模型需要調整到哪種程度才能視為穩定？
- HMM 基因預測模型的建立是使用 DNA 序列還是蛋白質序列較為適當？
- 如何評估我們每項步驟得出的結果是正確的？那些資料對於人來說只是很難理解其義的英文字母
- 現今生物學家在得出物種同源性關係圖時，是否還需要運用到那些分析方式，是適合整合進系統中的？

以上是我們目前蒐集資料所遇到的相關問題，還望在開始研究此專題時，能蒐集更多的資料，以幫助克服難題。

評估十年前物種同源性分析流程所使用的程式

因為現今所尋到的主要參考資料為 2006 年的著作，有些套裝軟體可能已不再更新，曾經的方式可能也不適合現今的情況，所以我們需要找出已被淘汰掉的套裝軟體，評估此項流程在現況下的可行性，並尋找現今最恰當的替代方案。

程式改進和功能新增

評估完以前所使用軟體在現今是否可繼續使用後，我們必須更精進現在所使用的功能。現今生物資訊分析的需求越來越大，以前的軟體也無法完全應付現今的需求，如果沒有軟體適合做為替代方案，將由我們寫出新的程式以完善整個系統，增進生物學家們在做物種同源性分析時的方便性。

目前，我們所思考必須增加的功能，為生物資訊上 Newick format 的資訊分析應用，在我們目前所閱讀的文獻中，並沒有可以完整應用 Newick format 做各項分析的程式，我們想把它實做出來，整合到整個系統中。

<u>初步預想中 Newick 分析的功能：</u>

- 分群：計算每個物種和根的距離，用 Stack 尋找括號，用括號後的距離做計算。

  以圖一為例，若將此樹按照距離分為兩群則：A 距離根節點 F0.1、B 距離根節點 F0.6、C 距離根節點 F(0.4 + 0.4) = 0.8、D 距離根節點 F(0.4

+ 0.3) = 0.7，可得此結論 AB 一群、CD 一群。

- 找最近的親緣關係：同上一項使用 stack，配合 Breadth-first search 和 Depth-first search 找出最短距離。

  以圖一為例，若要找出與 C 物種有最相近的親緣關係則：A 物種與 C 物種的距離為(0.1 + 0.4 +0.4) = 0.9、B 物種與 C 物種的距離為(0.6 + 0.4 +0.4) = 1.4、D 物種與 C 物種的距離為(0.3 + 0.4) = 0.7，可得此結論 D 物種與 C 物種親緣關係最為相近。

- 找特定距離內的親緣關係：如上一項的方法，但需尋找更多的物種的最短路徑，一旦超出範圍即停止，跳下一個物種繼續。

  以圖一為例，若要找出與 C 物種距離 1 以內的物種則：A 物種與 C 物種的距離為(0.1 + 0.4 +0.4) = 0.9<1，A 物種符合；B 物種與 C 物種的距離為(0.6 + 0.4 +0.4) = 1.4>1，B 物種不符合；D 物種與 C 物種的距離為(0.3 + 0.4) = 0.7<1，D 物種符合，可得此結論 A 物種及 D 物種與 C 物種距離 1 以內。

以上初步構想為我們目前的藍圖，更詳細的程式流程設計和功能規劃，在專題進行時，會讓其變得更加完善。

第一次規劃系統

做完上述的需求分析後，我們需要規劃出整個系統所需的功能。系統功能需求要符合以下我們最初訂定的目的。

- 讓同源性分析更方便
- 輸出結果可依條件搜尋，藉以方便生物學家分析

而因為我們團隊的背景為資訊工程，在對於生物資訊分析的流程可能並不是那麼清楚，所以我們第一次規劃出系統後，要再將實際資料投入我們規劃的步驟做分析，判斷我們所列出的步驟是否為正確的，再做整體系統步驟的微調和整理。

重新蒐集資料

第一次規劃完成，我們將選用 zebrafish 作為我們測試分析所用的主要 DNA，因為 zebrafish 有悠久的研究歷史，相對於其他物種有較完善的 DNA 資料庫，具有爭議的地方也較少。另一方面，也希望我們在測試我們整體系統的過程中，能更新 V2R 基因分析[1]的分析結果。畢竟，在這十年間，zebrafish 的 DNA 資料庫已更新的更為完善。

確定最終系統步驟

透過前面步驟的初步試探，我們可以設定出最符合需求且輸出結果最正確的系統。後續的程式設計，都將遵循此步驟規劃出的結果再做整合。在這個環節，每一個規劃出來的步驟，因為都需分項做程式設計，所以每步驟的輸入輸出需在此階段做統一，以便後續的系統整合。

分項規劃程式流程圖

整體系統規劃出來後，需再分項規劃每個環節中所需程式的程式設計流程圖，

例如此步驟需要的詳細功能和細部的程式設計問題。

程式撰寫

照著上一環節規劃出的流程圖，開始撰寫程式。

測試除錯

做程式功能性的測試和除錯。

系統整合

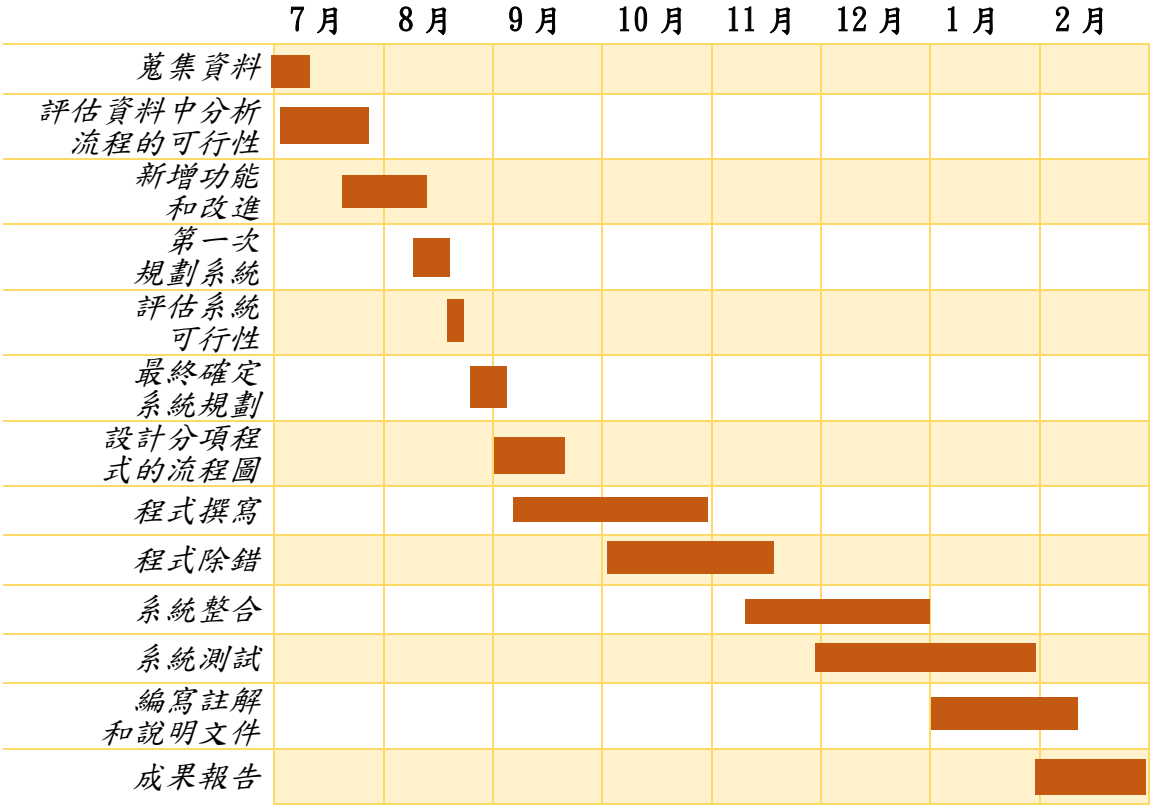
將分向的程式整合為一個完整的系統。也就是前面分項規劃的程式可能都需要手動輸入及輸出，整合為系統後，我們的系統將能直接從第一步驟執行成功執行到最後步驟的輸出，而輸出的結果在我們的系統中可做搜尋分析。

系統程式測試除錯

整體系統的測試和除錯。

編寫註解說明文件

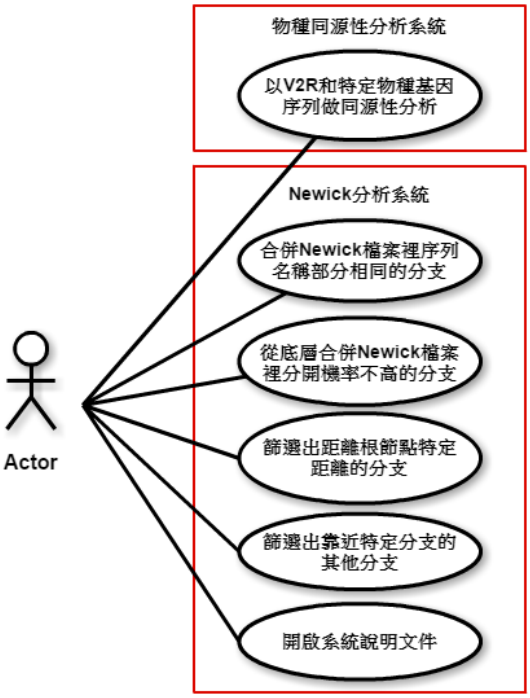
將在我們的程式碼中增加註解，以利後續的系統維護和程式功能新增。也要撰寫說明文件，讓使用者對我們的系統操作方式能輕鬆上手。



圖表 3 研究步驟甘特圖

系統開發成果

- **系統範圍**

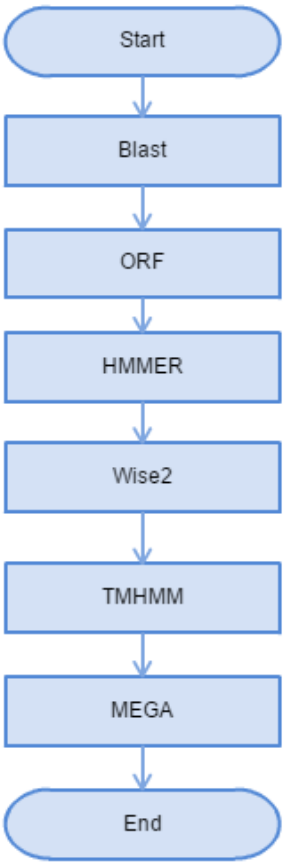


圖表 4 Usercase 示意圖

圖以 usercase 的方式說明系統可以支援的各種分析功能。

系統分為兩部分物種同源性分析系統顧名思義可用來做物種間的同源性分析，輸出的檔案可送至 Newick 分析系統，以圖示中的其中四種方式做進一步的分類，以便使用者能透過這些輸出資料做進一步的分析。

- **系統功能架構**

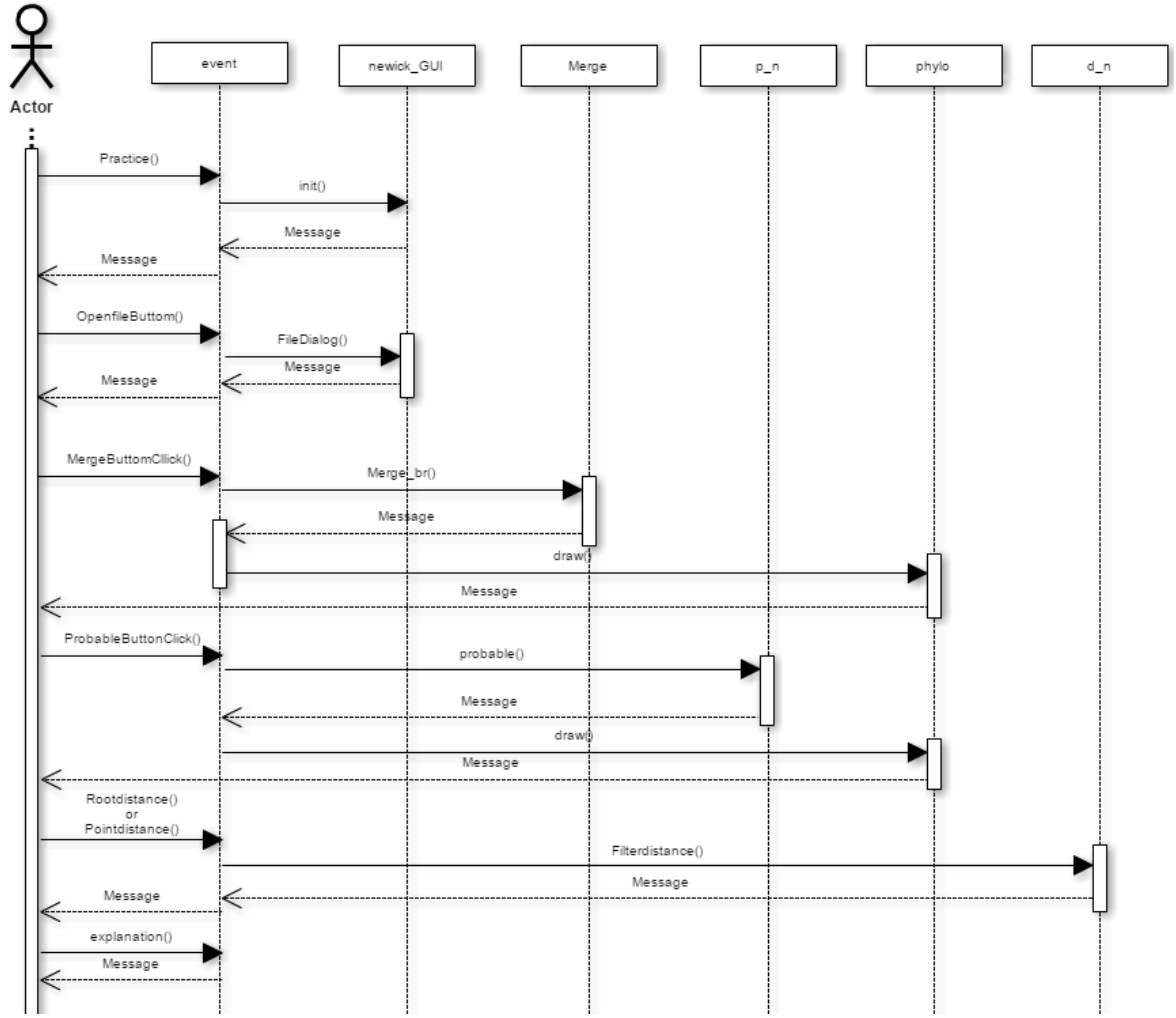


同源性分析系統最終完成的流程圖如左圖。使用者須先準備數條 V2R 蛋白質序列，及要比對的物種基因序列，作為系統的輸入。

最開始會由 Blast 找到特定物種中相似的每段序列，將序列起始延伸 3000 字元，末端延伸 5000 字元。之後使用 open reading frame 的方法大致預測出基因序列中的 Exon。再經由 HMMER 建構 HMM 模型進一步預測每段基因和 V2R 基因的同源性關係，此為第一次過濾。因 ORF 的方式預測出的 Exon 並不完全準確，我們將 HMMER 過濾出的序列挑出，用他們的 DNA 序列以 Wise2 重新預測每條基因的 Exon，將 Wise2 跑出的最後結果以密碼子的方式轉為蛋白質序列並丟到 TMHMM 中做結構域分析。以此篩選出的最終結果，視為和 V2R 同源關係相近的序列。

系統會將得到的多條序列一起放進 MEGA 中分析，以建構同源關係樹，此樹會由 Newick 格式表示，最終的輸出結果也是 Newick 檔，此檔可在我們之後的 Newick 分析系統中進行分析。

圖表 5 同源性分析性統簡易流程圖



圖表 6 系統序列圖

如圖表六，以系統序列圖來表示 Newick 系統分析工具的功能，此工具依功能可分為四種，合併分支(Merge)、機率分群(p_n)、以根節點距離篩選(d_n)、以相近距離節點篩選(d_n)。使用的介面如括號內所述。

- **設計限制**

  運行平台:LINUX,其中同源性分析所用到的工具多在 LINUX 平台運行，所以此系統僅限用於 LINUX 平台。

  檔案類型：同源性分析系統僅限於輸入 fasta 檔，Newick 分析工具僅限於輸入 Newick 類型檔案。

  導入的外部系統及模組：系統依賴多種分析工具和模塊運行，使用者使用前須先安裝下列系統及模組。系統：BLAST、WISE2、HMMER、TMHMM、MEGA；模組：wxPython、bioPython、Matplotlib、numpy。

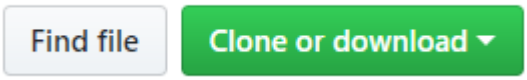
系統操作說明

一、 同源性分析 pipiline

第一步驟：

在 GitHub 網站上下載程式

https://github.com/joyce850722/V2R-biological-information



圖表 7 github 畫面截圖

第二步驟：

在 Linux 系統 commandline 上下載所需工具的指令

1. sudo apt-get update
2. 此為執行 BLAST 所需的工具
   - sudo apt-get instal ncbi-blast+
3. 此為執行 WISE2 所需的工具
   - sudo apt-get install wise
4. 此為執行 HMMER 所需的工具
   - sudo apt-get install hmmer
5. 此為執行 PYTHON 所需的工具
   - sudo apt-get install python3
6. 此為執行 TMHMM 所需的工具

   Step1. 升級 CPAN

   → wegt http://www.megasoftware.net/do_force_download/megacc_ 7.0.26-1_amd64.deb

   Step2. 進入 CPAN

   → sudo perl -MCPAN -e shell

   Step3. 檢查 Perl 模組 版本需在 5 以上

   → perl -v

   Step4. 安裝 Perl

   → sudo apt-get updatesudo apt-get install perl

   安裝 TMHMM

   去官方申請下載，使用學術性單位信箱

   Step5. http://www.cbs.dtu.dk/services/TMHMM/

   Step6. 把解壓縮後 decodeeanhmm 文件放到 usr/bin 底下

   Step7. 修改 tmhmm 和 tmhmmformat.pl 中 perl 路徑

   #!/usr/bin/perl;bin/tmhmm and

bin/tmhmmformat.pl

（if not /usr/local/bin/perl）

Step8. **用 uname -s 和 uname -m 兩個命令獲得機器型號，更改 decodeanhmm**

例：decodeanhmm.Linux_X86_64

7. 此為執行 MEGA 所需的工具

Step1.

http://www.megasoftware.net/do_force_download/megacc_7.0.26-1_amd64.deb

Step2. 在 commandline 輸入 **sudo dpkg -i megacc_7.0.26-1_amd64.deb**

第三步驟：

- 在此目錄下輸入 python pipl.py

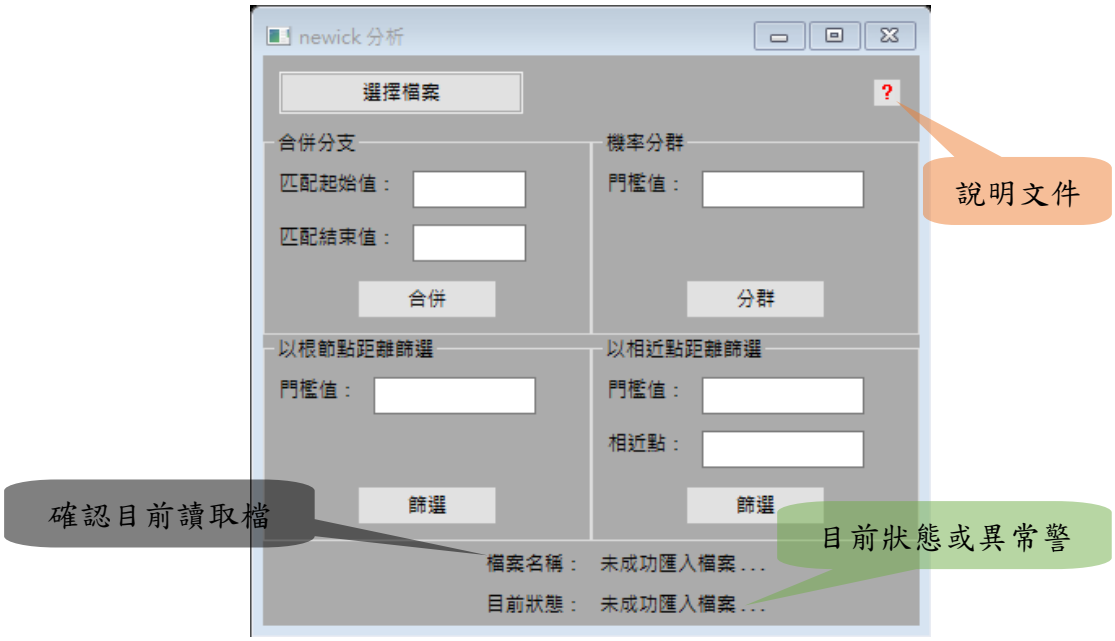
```
~/Downloads/pipl$ python pipl.py
```

第四步驟：

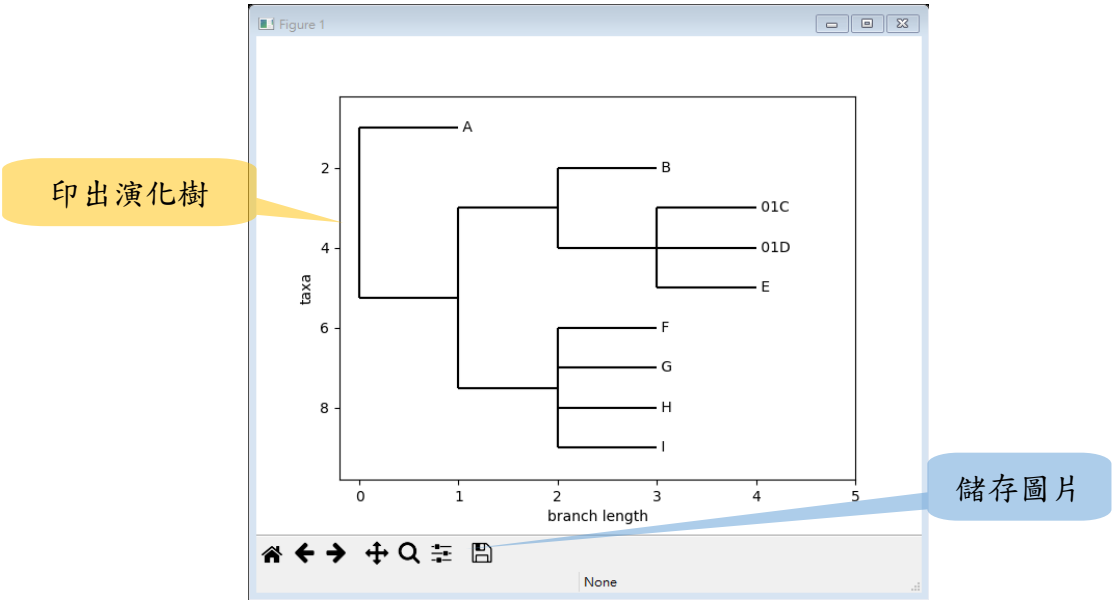
- 檔案輸入一：欲比對物種序列檔案（.fasta）
- 檔案輸入二：欲比對基因序列檔案（.fasta）

二、Newick 分析說明文件

- **介面展示**



圖表 8 Newick 程式介面



圖表 9 輸出演化樹程式畫面

- **檔案輸入：**Newick format (.nwk)
  ex. ('A' : 0.7, (('B' : 0.2, 'C' : 0.9) 0.5 : 0.1, 'D' : 0.01) 0.7 : 0.5) ;
- **檔案輸出：**
  1. **合併分支：**merge_output (.nwk)、figure (.png/ .jpg/ .pdf)。
  2. **機率分群:**probable_output (.nwk)、figure (.png/ .jpg/ .pdf)。
  3. **以根節點距離塞選：**distance_output (.nwk)、root_point (.txt)。
  4. **以相近點距離塞選：**distance_output (.nwk)、point_point (.txt)。
- **針對 Newick 部份本工具提供簡易的圖型化介面，並實現以下四種功能：**
  1. **合併分支**

     主要是將分支下同一物種的序列做合併，以便使用者能快速分辨出物種間的親疏關係。鑒於序列名稱有所不同，本工具設置匹配起始值與結束值獲取判斷相似物種的依據。其中匹配起始值與結束值僅能輸入大於 0 的數值，且前者不可大於後者。

     ex. ((01gene23 : 0.5, 01gene26 : 0.08) 0.8 : 0.2, 14gene31 : 0.12) 0.6 : 0.1);

     ↓

     匹配起始值：1，匹配結束值：2

     ↓

     ((01gene : 0.02, 14gene31 : 0.12) 0.6 : 0.1);

2. **機率分群**

輸入門檻值，從根節點開始分群，若機率高於門檻值則分群，直至小於門檻值為止。其中門檻值僅能輸入大於 0，小於 1 的數值。

3. **以根節點距離塞選**

輸入門檻值，計算序列與根節點的距離，若距離小於門檻值則輸出。其中門檻值僅可輸入大於 0 的數值。

4. **以相近點距離塞選**

輸入門檻值與序列名稱，計算序列與相近點的距離，若距離小於門檻值則輸出。其中門檻值僅可輸入大於 0 的數值。

ex. ((01gene23 : 0.5, 01gene26 : 0.08) 0.8 : 0.2, 14gene31 : 0.12) 0.6 : 0.1);

↓

門檻值：0.6，相近點：01gene23

↓

01gene23 : 0

01gene26 : 0.58

系統效益

我們的系統將可用來作為物種同源性關係的分析，不論你是在尋找基因或在進行基因預測，每部分間的操作皆不需要使用者動手，而是由程式自行去處理。並不需要了解每套軟體是如何操作，能大幅縮短所需要花費的時間，相較於網路上的套裝軟體，我們的系統主要是針對 V2R 基因進行分析，在 V2R 基因上比其他軟體能達成更精準的預測，並在最後輸出系統發育樹的 newick 檔。

而關於物種演化分析部分，在網路上可找到的物種演化分析相關軟體絕大多數都是將系統發育樹進行視覺化處理，像是：可以完全訪問每個節點（修改節點標題，分支標題，分支寬度，分支長度）、將樹摺疊、展開或是編輯樹等，例如：TreeMe，都是對樹本身進行處理。而我們的軟體則可將樹上的數據進行統整，讓你不只是單純的看樹而是幫助你了解這棵樹進行整理及篩選，甚麼樣的生物同源關係最為相近、每個物種間的演化機率、將分支下同一物種的序列做合併，我們的軟體可以讓使用者快速分辨出物種間的親疏關係。

結論

目前的系統已完整的進行架設從基因搜尋、基因預測到物種演化分析所需的全部功能，使用者只需決定好要比對的物種及用來分析同源性的蛋白質則可進行分析，我們改進了十年前的分析方式，重新歸納了新的分析流程，以優化同源性分析時的運行速度。並可運用 newick 程式客製化的觀察你所關注的基因及基因間的同源性關係。

未來展望

我們利用不同分析物種同源性的工具串接寫出來的程式,不僅省去時間查詢指令，也將結果輸出為 Newick 格式和樹狀圖兩種，以適應更多情況下的數據分析要求。

希望功能性的完整：

1. 運行平台：在 Windows 作業系統上能執行。
2. 導入的外部系統及模組：打包所有下載的工具做成可執行檔，省去事前安裝的不便。
3. 程式部分：簡化程式，省去不必要的時間，讓執行更為流暢。

參考文獻

1. Yasuyuki Hashiguchi and Mutsumi Nishida, "Evolution and origin of vomeronasal-type odorant receptor gene repertoire in fishes," *BMC Evolutionary Biology*, vol. 6, pp. 76-88, Oct. 2006.
2. Neil Jones and Pavel Pevzner, "An Introduction to Bioinformatics Algorithms," *MIT Press*. Aug. 2004.
3. National Center for Biotechnology Information, "NCBI handbook," Retrieved fromhttps://www.ncbi.nlm.nih.gov/books/NBK143764/, 2013.
4. Michael Brent, "How does eukaryotic gene prediction work?" *Nature Biotechnology*, vol. 25, no. 8, pp. 883-885, Aug. 2007.
5. Ewan Birney, Michele Clamp and Richard Durbin, "GeneWise and genomeWise," *Genome Research*, vol. 14, no. 5, pp. 988-995, May. 2004.
6. Sean Eddy, "A probabilistic model of local sequence alignment that simplifies statistical significance estimation," *PLoS Computational Biology*, vol. 4, no. 5, e1000069, May. 2008.
7. Anders Krogh, B. Larsson, von Heijne and Erik Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567-580, Jan. 2001.
8. Christian and Sean Eddy, "A simple algorithm to infer gene duplication and speciation events on a gene tree," *Bioinformatics*, vol. 17, no. 9, pp. 821-828, Sep. 2001.
9. Ilan Wapinski, Avi Pfeffer, Nir Friedman and Aviv Regev, "Automatic genome-wide reconstruction of phylogenetic gene trees," *Bioinformatics*, vol. 23, no. 13, pp. 549-558, Jul. 2007.
10. Thomas Laubach, Arndt von Haeseler and Martin Lercher, "TreeSnatcher plus: capturing phylogenetic trees from images," *BMC Bioinformatics*, vol. 13, no. 1, pp. 110, 2012.
11. Christian Allende, ErikSohn and Cedric Little, "Treelink: data integration, clustering and visualization of phylogenetic trees," *BMC Evolutionary Biology*, vol. 16, pp. 414, Jul. 2015.
12. Michael Lynch and John Conery, "The Evolutionary Fate and Consequences of Duplicate Genes," *Science*, vol. 290, no. 5494, pp. 1151-1155, 2000.
13. Hansen A, Rolen SH, Anderson K, Morita Y, Caprio J, Finger TE, "Correlation between olfactory receptor cell type and function in the channel catfish," *J Neurosci*, vol.23, pp. 9328-9339, 2003.