模糊搜尋在基因序列資料庫之應用

周良懋, 林佑錫 and 鄭錦聰
國立虎尾科技大學資訊工程系
g243114@moom.nfu.edu.tw g243116@moon.nfu.edu.tw

論文摘要

Fuzzy 這門科學經過了一段時間的發展已經到達成熟的地步了，日常生活中充斥著 Fuzzy，如自動洗衣機、自動感應冷氣機…等，已經在許多地方均可見到模糊的影子了。搜尋的技巧千變萬化，因時間、地點而不同，本論文利用模糊搜尋的技巧來進行容錯的搜尋，我們更可以利用此等技巧來進行基因序列資料庫之模糊搜尋，藉由模糊搜尋之特性達成基因序列資料庫之強健性。

關鍵詞：基因序列資料庫、模糊搜尋、容錯的搜尋、資料表關聯

Abstract

The science, fuzzy, had ripe stage already during a long time development, such as auto-wash machine, auto-detection air conditioner…etc. Now, you can see many application with fuzzy. The skill of search is changeable, but different because of time, place. In this paper, we use the fuzzy technology to make search RNA sequence and keyword term in order to reaching fault-tolerant. And we can use this technology to observe the gene has taken extensive of sudden changes.

Keywords: Fuzzy, Search, UCSC & GeneBank.

1. 前言

在現今的時代，生物科技已經是一門熱門的科技，但是做這門學問有一個很大的問題，往往需要費時一段不算少的時間，所以這種方式在現在已不適用了。以前的方法叫做在筆記本上做生物的研究，不可能在短時間和少量的成本得到想要的結果，再來有一種演進的方法就是在實驗室裡做一些實驗將基因的表現或是一些遺傳的問題在實驗室和試管中得到結果，這種叫做在試管中做生物科技的研究。所花的時間較少，所得到的結果也相對的較多，但是隨著時間的過去，資料量也愈來愈多，如何有效的保留和分析這些資訊則是一門學問。

「生物資訊學」(Bioinformatics)於是油然而生，它是一門運用電腦計算能力與分析方法來解決生物學問題的技術。現在有了電腦，電腦的運算速度，儲存容量大，佔用空間小，且可以實現一些複雜的問題，如比對序列的資訊和預測基因的表現…等。而這些序列就存在資料庫中。

每天都有許多生物學資料產生，像著名的計畫：基因定序計畫(Genome sequencing projects)，就是有大量的資料須要分析，而這些龐大的資料都存放在三個共有的資料庫，它們分別在美國、歐洲和日本。儘管人類基因定序計畫已經完成，但是資料量仍然以令人咋舌的速度在快速成長中。

Fuzzy 理論的誕生是在 1965 年，美國加州柏克萊大學 L.A. Zadeh(札德)教授在「資訊與控制」(Information and Control)學術雜誌上，發表「Fuzzy 集合」的論文，Fuzzy 理論於是正式誕生。而大部份的人都認為"模糊"不是好事，但是在人們的日常生活中，"模糊"卻隨時隨地、如影隨形般與人們生活在一起，如我們常說的：六十幾分、大約二點半、接近兩仟元…等。但有時候卻是不得不如此，和科學家愛因斯坦在 1921 年曾講過一句話：「數學定律若要盡量的逼近"真實"，則它們必然無法很"精確"；而它們要盡量"精確"，則必然無法"真實"」。這個偉大的科學家為了 Fuzzy 理論下了一個重要的詮釋。

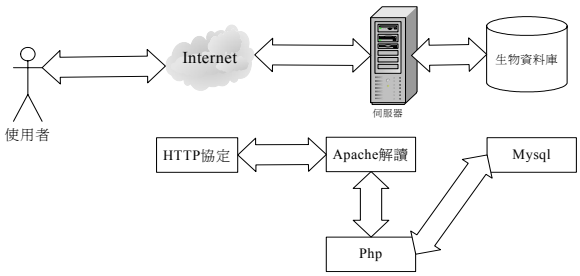
在傳統的資料庫中所儲存的資料必須是明確的，當我們以查詢語言從資料庫抓取所要的資料集合時，這個查詢式也必須是一個明確的描述，如此資料庫系統才有辦法根據我們所要求的條件，明確的把資料找出。

模糊理論是為解決模糊現象而發展的學問。用來表現無法精確定義之模糊概念，尤其在人類語言的語意模糊現象，例如"好像"、"一點"…等。它利用歸屬函數描述一個概念特質，用 0 和 1 之間的數值來表示屬於某一概念程度，此值稱為元素對集合的歸屬度。

在修過一門"生物資訊導論"的課程，和修過一門"資料庫理論和應用"的課程，決定利用修課時所學到的技術，建立一個可用的生物資料庫，可供他人使用模糊搜尋相對應的序列。

2 系統架構和.研究方法

- **系統架構**



圖一

使用者可以透過網頁介面，連線到我們的伺服器，透過伺服器再對我們的資料庫下查詢，然後資料再透過伺服器傳結果到網頁給使用者。

● **研究方法與步驟**

1.在 linux 上架設 Apache、Php 和 Mysql

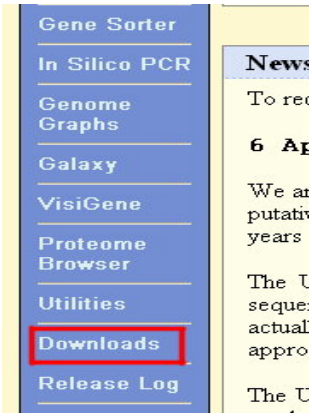
2.再利用 Matlab 從 UCSC 和 GeneBank 大量地下載回生物資料庫

3.讀懂和分析 UCSC 和 GeneBank 的檔案格式，再利用 mysql 語法寫到我們的資料庫中

4.撰寫符合我們 Fuzzy 需求的網頁

● **取得資料庫**

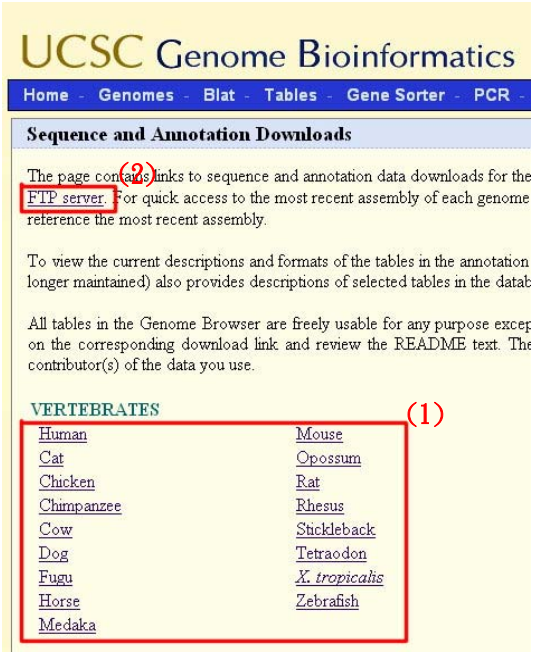
首先先到 UCSC 的網站，網址：http://genome.ucsc.edu/取得需要的資料庫。



圖二

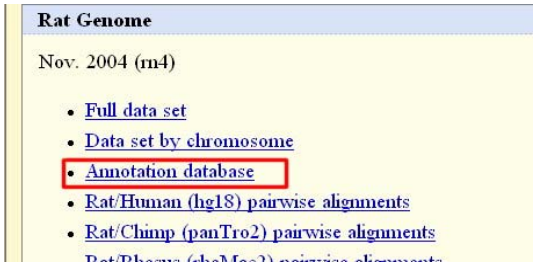
在左下角的 Downloads 的地方下載

進入之後選擇自己需要的物種的序列(1)或是使用 FTP 下載(2)



圖三

進入後請選擇 Annotation database



圖四

進入之後只選擇自己需要的 TABLE，因為在資料庫裡有太多資料但是有很多並不是很清楚內容是哪些所以只取自己所需要的就好。

在此我們所需要的東西 TABLE 是：gbCdnaInfo、cell、mrnaClone、library、sex、source、description、cds、productName、geneName、organism、keyword、author、development、tissue。每個 TABLE 有兩個檔案*.sql&*.txt.gz 匯入欄位名稱到主機上

```
LOCUS       A15H9FIB              1228 bp    DNA     linear   VRL 18-APR-2005
```

● **Genebank 文件說明**

Locus：基本上這個只是一個開頭標記，並沒有什麼特別的意義。

A15H9FIB：這個是 Locus 的名稱，長度不超過 10 個字，第一個字為英文，從第二個以後為數字或字母，英文都是大寫字母，在此是一個唯一的值並無重複，其實這是一個舊東西，原本是要移除的，但是有很多軟體都依賴於這個獨一無二的名稱所以並沒有被移除。

1228 dp：這個是序列的長度，長度從 1 到 350000bp，實際上 Genebank 和其他的資料庫並不接受 50bp 以下的序列。

DNA：分子類型通常是 DNA 和 RNA 但是也有少量的其他類型出現但是都會表明單鏈

VRL：這個是 Genebank 的分類碼，由三個英文字母所組成，這具有物種分類的意義，或者出於其他的分類目的，此分類碼的存在是為了保相容性，它可以追朔到以前 Genebank 為了保持可管理文件大小而將整個資料庫照物件分類而分割成幾個文件的時候。

18-APR-2005：這個是數據最後被公開的日期，在許多情況下，也是第一次公開的日期，但是這個日期並無法律保證，所以並不可靠，這個日期並無被拿來做為申請專利的依據。

```
DEFINITION  Adenovirus type 15H9 (Morrison) fibre gene, nonenveloped DNA.
```

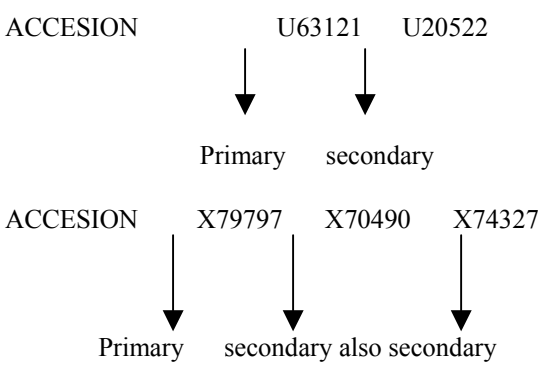
DEFINITION 行在 GeneBank 紀錄中用以記錄生物意義的。這行東西也會出現在 NCBI 的 FASTA 文件中這樣任何人進行 BLAST 相似性搜索時都會看到這些訊息。

生成這一行時要特別小心，因為許多記錄生成工作可以部份的自動進行。所以管理資料庫的工作人員

必須小心的檢查這一行必須保證訊息的一致性和有效性。但是用一行文字來說明生物的背景並不總是可行的。對於不同的資料庫採用了各自的解決方法。其中有一些共識並且每個資料庫也都了解其他資料庫的解決方法。下面是 DEFINITION 行結構標準的一個小節對於 mrna，可以是以上的類型：
屬種　產物名稱 (基因符號)　mrna, complete cds
或者對於基因組的紀錄：
屬種　產物名稱 (基因符號)　gene, complete cds
當然各個資料庫採用的解決方法也考慮到了其他類型的紀錄下列這些規則應用於細胞的序列，以保證用戶及資料庫工作人員明了 DNA 的來源和生物背景
DEFINTION　屬種 蛋白質 X　(xxx) gene, (下列選一) complete cds.
,編碼粒線體蛋白質的核基因
,編碼葉綠體蛋白質的核基因
,編碼粒線體蛋白質的粒線體基因
,編碼葉綠體蛋白質的葉綠體基因
或者
DEFINTION　屬種　XXS　核糖體 RNA　gene, (下列選一)　complete sequence
,編碼粒線體 RNA 的粒線體基因
,編碼葉綠體 RNA 的葉綠體基因
在一項合作資料庫之間達成的一項協議在 DEFINTION 行中給出屬和種的全名而不使用通用名或是屬名的縮寫，只有愛滋病是例外它將在裡面用 HIV1，HIV2 表示。

```
ACCESSION   X76706
```

這個是在查詢資料庫的一個紀錄的一個重要的關鍵詞。各個號碼將在參考文件被引用並始終和序列在一起。當序列被更新時這個號碼是不會變的。這個號碼採取下列兩種方式之一：1+5 或是 2+6 的方式 1+5 的格式是指一個大寫的英文字母後面跟著五個數字的格式;；2+6 的格式是指兩個大寫的英文字母後面跟著六個數字的格式。通常 ACCESION 都是單獨存在於這一行其餘出現的都為二級檢索碼




```
VERSION     X76706.1  GI:436054
```

這行指明了 X76706 的版本為第一版，GI 為 geninfo identifer。

```
KEYWORDS    fiber gene; fiber protein.
```

KEYWORDS 是一個歷史遺物，並且不幸地在很多情況下被誤用了給一個紀錄加上關鍵詞通常並不是十分有效因為在過去的年月中有許多作者選用了不在受控制詞表中的詞，並且在整個資料庫中的用法也不一致。因此 NCBI 不鼓勵用關鍵詞，但在查詢時加入關鍵詞是可以的，特別是那些在其他紀錄中未出現過的詞，或以一種受控的方式來使用的詞(例如對於 EST，STS，GSS，HTG 紀錄)。

```
SOURCE      Human adenovirus type 15
  ORGANISM  Human adenovirus type 15
            Viruses; dsDNA viruses, no RNA stage; Adenoviridae; Mastadenovirus.
```

SOURCE 行中有生物通用名或科學名稱。有些情況下也有其他來源的信息。現在正一致努力以保證來源特性中包含所有必須的訊息，並且所有關於分類的訊息可以從來源特性以及 NCBI 分類 SERVER 中獲得。

```
REFERENCE   1  (bases 1 to 1228)
  AUTHORS   Pring-Akerblom,P. and Adrian,T.
  TITLE     Characterization of adenovirus subgenus D fiber genes
  JOURNAL   Virology 206 (1), 564-571 (1995)
   PUBMED   7831811
REFERENCE   2  (bases 1 to 1228)
  AUTHORS   Pring-Akerblom,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-DEC-1993) P. Pring-Akerblom, Medizinische Hochschule
            Hannover, Nationales Referenzzentrum f.Adenoviren, Institut f.
            Virologie & Seuchenhygiene, Konstanty-Gutschow-Str. 8, 30625
            Hannover, FRG
```

每一個 GeneBank 記錄至少要有一篇參考文獻。許多情況下有兩篇。參考文獻提供了科學證據以及一個背景來解釋這個特定的序列為何這樣確定。

```
FEATURES             Location/Qualifiers
     source          1..1228
                     /organism="Human adenovirus type 15"
                     /mol_type="genomic DNA"
                     /strain="intermediate"
                     /isolate="morrison"
                     /db_xref="taxon:28276"
                     /map="0.88-0.92 units"
```

來源特性是唯一一個必須在所有 GeneBank 紀錄中出現的特性所有的特性都有一系列合法的限定詞有些是強制性的。所有的 DNA 序列紀錄都有出處，即使是合成序列這樣極端的特例一樣。大多數的情況下，下一個紀錄只能有一個來源特性，並帶有/organism 的限定詞。

限定詞 organism 包含屬種的科學名稱，有些情況下，還可以在亞種水平描述。對於來源，一系列限定詞包括關於 BioSource 的所有材料，這可能包含圖譜染色體或組織和克隆標示以及其他資料庫訊息。

```
gene            50..1138
                /gene="fiber gene"
CDS             50..1138
                /gene="fiber gene"
                /codon_start=1
                /product="fiber protein"
                /protein_id="CAA54127.1"
                /db_xref="GI:436055"
                /db_xref="GOA:P68983"
                /db_xref="InterPro:IPR000931"
                /db_xref="InterPro:IPR000939"
                /db_xref="InterPro:IPR000978"
                /db_xref="InterPro:IPR008982"
                /db_xref="InterPro:IPR009013"
                /db_xref="UniProtKB/Swiss-Prot:P68983"
                /translation="MSKRLRVEDDFNPVYPYGYARNQNIPFLTPPFVSSDGFQNFPPG
                VLSLKLADPIAIVNGNVSLKVGGGLTLQDGTGKLTVNADPPLQLTNNKLGIALDAPFD
                VIDNKLTLLAGHGLSIITKETSTLPLGLRNTLVVLTGKGIGTESTDNGGTVCVRVGEGG
                GLSFNNDGDLVAFNKKEDKRTLWTTPDTSPNCKIDQDKDSKLTLVLTKCGSQILANVS
                LIVVDGKYKIINNNTQPALKGFTIKLLFDENGVLMESSNLGKSYWNFRNENSIMSTAY
                EKAIGFMPNLVAYPKPTAGSKKYARDIVYGNIYLGGKPDQPVTIKTTFNQETGCEYSI
                TFDFSWAKTYVNVEFETTSFTFSYIAQE"
misc_feature    596..1135
                /gene="fiber gene"
                /note="hyper variable region"
```

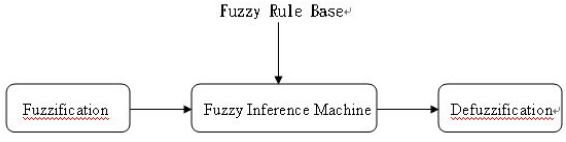
圖五

CDS 指示讀者如何將兩個序列連接在一起，或如何根據核苷酸序列以及基因編碼得到胺基酸序列。在分析這些序列時，我們必須從 DNA 座標推導出胺基酸位置，並且我們對於編碼蛋白質的了解也將僅限於對從 DNA 特性的描述中獲得。這一限制可被 Sequin 克服這一例子也顯示了資料庫交叉索引(db-xref)的使用。這一受限制詞允許資料庫將另一部資料庫中使用的標識符號交叉索引。允許 db-xref 的資料庫都是合作資料庫所維護的。

正如上面提到的，NCBI 給每個序列給一個 gi 號碼。這意味著翻譯產物蛋白質序列(不是簡單附屬於 DNA 紀錄，如同在 GeneBank 紀錄中顯示的)，也有自己的 gi 號碼。一個特定的 gi 號碼當且僅當序列更改時才更改。蛋白質 gi 號碼現在作為 PID db-xref 或蛋白質號碼出現。但是 PID 已經被取消掉了，Protein-id(或核苷酸序列產生成的蛋白質檢索號)將由 3 英文個字母加五位數字構成，後跟著一個句號和另一個整數，顯示這個蛋白質的版本。當序列更新時這一數字也跟著增加。

- **模糊理論做法**

我們的做法並不真的改變資料庫結構，也就是說它仍然假設資料模式為標準的關聯式結構，所以嚴格說來它並沒有儲存模糊資料的能力。但它在資料庫的查詢語言之上另外附加一層的前處理器，也就因為這前處理器的緣故，使用者可用模糊的述詞或形容詞來表達它們所要查詢的條件，而這些模糊查詢就由前處理器加以轉換成標準的資料庫 SQL 查詢式子，然後再由 DBMS 來加以處理與執行。而前置處理器的部分我們是用 apache+php+mysql 來實現，我們利用網頁具有高度親和力的特性來實現簡單的模糊搜尋輸入，可圖示如下



圖六

(1)模糊集合(fuzzy set)與歸屬函數(membership function)

$$u_A(x)=\begin{cases}1, & x\in A\\ 0, & x\notin A\end{cases}$$

傳統的普通集合其推論是基於二值邏輯，即對於每個元素，若屬於這個集合，其特徵函數為 1；若不屬於這個集合，則其特徵函數為 0，我們可以表示成：

1965 年 Zadeh 教授提出「模糊集合論」用來解決在現實生活中必須允許「是與否」之間擁有中界狀態的模糊現象，他仿照特徵函數表示普通集合的方法，建立歸屬函數來表示模糊集合。即歸屬函數表示出論域(universe of discourse)U 中至[0,1]上的一個映射 u

uA:U→[0,1]

對於任何 x 屬於 U，都有一個 uA (x)，其值介於 0 與 1 之間，來代表 x 屬於模糊集合 A 的歸屬度。

(2)模糊集合的表示方式

模糊集合 $\underset{\sim}{A}$ 具體表示方式有很多種，以下為兩種常用的表示法：

1. Zadeh 表示法

$$\underset{\sim}{A}=\frac{u_A(x_1)}{x_1}+\frac{u_A(x_2)}{x_2}+\frac{u_A(x_3)}{x_3}+\frac{u_A(x_4)}{x_4}+\cdots\cdots+\frac{u_A(x_n)}{x_n}$$

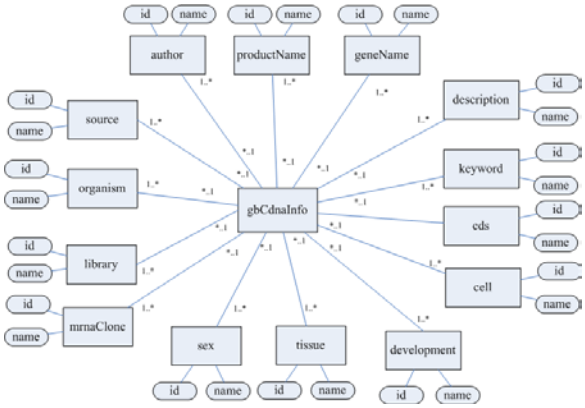
其中並非分數，而是表示論域中的元素 xi 及其歸屬度之間的對應關係，"+"也並不表示"求和"，而是表示模糊集合在論域上的整體。

2. 序偶表示法

$$\underset{\sim}{A}=\left\{ (x_1,u_{\underset{\sim}{A}}(x_1)),(x_2,u_{\underset{\sim}{A}}(x_2)),(x_3,u_{\underset{\sim}{A}}(x_3)),\ldots(x_n,u_{\underset{\sim}{A}}(x_n)) \right\}$$

其中(a,b)表示論域中的元素 x_i 及其歸屬度之間的對應關係

- **UCSC 的資料表關聯**

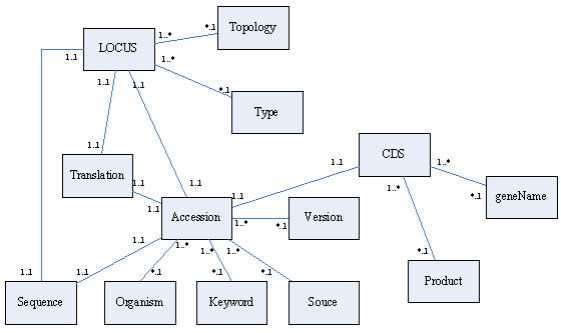


圖七

gbCdnaInfo 為最主要存著外圍的那些 table 的 id 要查詢序列的資訊時使用 gbCdnaInfo 裡的 acc 來作為唯一的鍵值來查詢然後得到其他 TABLE 的 id 就可以查到其他的資訊。

這裡的最主要是查詢序列的相關資訊裡面並無序列存在所以可以得到的就是關於序列的其他資訊。

- **我們的資料表關聯**



圖八

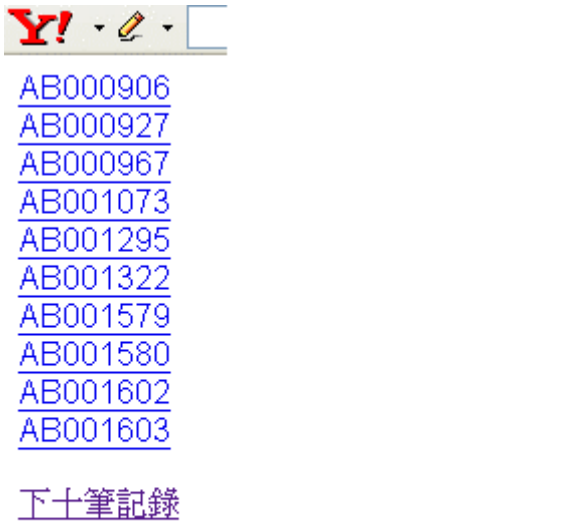
3. 實驗結果

圖九是輸入欲查詢的字串並選擇適當的模糊度，在此不慎輸入錯的單字(pratain)。圖十是利用 Fuzzy 的模糊搜尋來顯示可能的物件。圖十一是輸入欲查詢的 RNA 字串並選擇適當的模糊度。圖十二是利用 Fuzzy 的模糊搜尋來顯示可能的物種。圖十三是更近一步的了解物種。圖十四是了解此物種的蛋白質序列。圖十五是了解此物種的 RNA 序列。

圖九

圖十

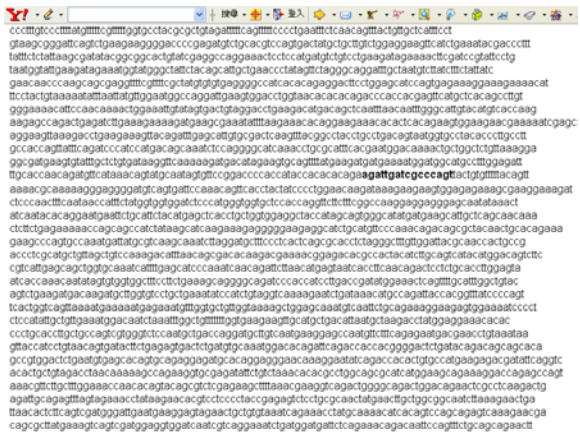
圖十一



圖十二

圖十三

圖十四



圖十五

5. 結論

本系統可對使用者不甚了解的關鍵字或是拼錯關鍵字時進行搜尋,在使用者不甚了解關鍵字的正確拼法時,此系統會自動地把在模糊度以上的關鍵字(可選擇)列表出來,讓使用者從中挑選合適的物種來進行觀察。可在一長串的 RNA 序列中進行比對,如此一來,就可以觀察出某一 RNA 序列在某一代出現了變化。總而言之,本系統的優點是:容錯、基因序列突變搜尋,本系統的缺點是:誤用模糊標準會找出比較多的資訊。

參考文獻

1.作者:Andreas D.Baxevanis、B.F.Francis Ouellette 著　李衍達、孫之榮等譯 出版年份:不可考 書名:生物信息學基因和蛋白質分析的實用指南　出版社:清華大學 (簡體書)

2.作者:艾特伍德、史密斯、陳進和、AttwoodTeresa K、SmithDavid J. Parry 出版時間:民 92[2003] 書名:生物資訊入門　出版社:藝軒圖書發行

3.作者:巴克西凡尼斯、歐利特、李衍達、孫之榮、BaxevanisAndreas D 、 OuelletteB. F. Francis 出版時間:2001　書名:生物訊息學:基因和蛋白質分析的實用指南　出版社:北京市 清華大學

4.作者:陳俊宏　出版時間:民 89[2000]　書名:PHP and MySQL 徹底研究 網頁資料庫設計　出版社:旗標

5.作者:卡斯塔內托、許鳴程、Castagnetto,Jesus 出版時間:民 89 [2000]　書名:專業 PHP 程式設計　出版社:碁峰資訊

6.作者:蓋壚　出版時間:民 92　書名:實用模糊數學　出版社:亞東

7.作者:Zadeh,L.A、陳國權　出版時間:民 71　書名:模糊集合.語言變量及模糊邏輯　出版社:北京市 科學出版社

8.作者:九章出版社編輯部　出版時間:民 81 [1992] 書名:模糊數學入門　出版社:九章出版臺北縣新店市

9.作者:中國生產力中心技術引進服務組　出版時間:民 83[1994]　書名:Fuzzy 實用化範例 用 C 語言　出版社:全華

10.作者:陳俊宏　出版時間:民 91　書名:Linux 8 進階系統安裝 DIY　出版社:旗標

11. 網址名稱:Apache 官方網站　網址:http://httpd.apache.org/

12. 網址名稱:Mysql 官方網站　網址:http://www.mysql.com/

13. 網址名稱:Php 官方網站　網址:http://www.php.net/

14.網址名稱:Phpmyadmin 官方網站　網址:http://www.phpmyadmin.net/home_page/index.php

15. 網址名稱:UCSC 官方網站　網址:http://genome.ucsc.edu/

16. 網址名稱:GeneBank 官方網站　網址:http://www.ncbi.nlm.nih.gov/Genbank/index.html

17.網址名稱:學習模糊系統(Fuzzy System)程式設計　網址:http://ccy.dd.ncu.edu.tw/~chen/resource/fuzzy/fuzzy.htm

18.網址名稱:中正大學資訊工程研究所　網址:http://caipc4.cs.ccu.edu.tw/course/fuzzysystem/

19.網址名稱:模糊理論筆記(Fuzzy Note)　網址:http://irw.ncit.edu.tw/peterju/fuzzy.html

20. 網址名稱:模糊數學　網址:http://www.math.tku.edu.tw/chinese/mathhall/mathinfo/lwymath/chaos.htm

21.網址名稱:模糊綜合評判應用於人力資源管理之研究　網址:http://www.management.fju.edu.tw/manadepat/review/paper.asp?no=0424#1

22. 網址名稱:模糊搜尋 Q&A　網址:http://163.14.136.86/c60/help/fuzzy_faq.htm