第一章 緒論

把著作表列當作基因序列來處理似乎是件有趣的事。人類的基因體(genome)大約由 30 億個核甘酸(nucleotide)組成，而這些鹼基對(base pairs)在每個人身上都有不盡相同的排列。任何兩個人間，大約一千個鹼基對就會有一個不同，因此我們可以用 DNA 來辨認一個人。在這個系統中，我們使用蛋白質序列(protein sequence)來辨認書目資料。首先將書目資料轉換成蛋白質序列，然後根據蛋白質序列的表現，在樣板資料庫(template database)中找到最相近的樣板，根據此樣板來解析書目資料以得到後設資料(metadata)。以前我們都是以資訊的觀點來解決生物的問題，反過來現在我們卻是以生物的觀點來解決資訊的問題。在這個系統上我們將每一筆著作表列當成一個小的生物看待，曾試想試著去研究它們的演化過程，建立起一棵演化樹(evolution tree)。雖然後來失敗了，但是我們相信還是有很多生物的規則可以套到這上面來。如果有一天可以在書目資料上發現了什麼新的規則，說不定還可以給我們一些啟發，將它套回人類身上，造福大眾。

1.1 研究動機

引用文獻可以彰顯自己文章的新穎性、也可以用書目資料來反駁他人的著作、甚至是沿用其著作上的方法、或者是引用文獻來加強說服力、…等等[1]。這些都是引用文獻的動機，但是引用時機對不對、有沒有道理，那就見仁見智了。不過不可否認的，書目資料存在確實有其重要性，之於上述理由研究書目資料的後設資料也變得格外重要。因此在實驗室網站中，我們也想要將實驗室成員的著作資料也存入網站當中。不過資料輸入一直都是一項繁瑣的事，往往需要透過人工一筆一筆鍵入。同時為了方便資料庫管理，我們通常會將一筆資料

分成好幾個欄位，希望使用者依照規定存入適當的位置。通常使用者會得到一個輸入介面如圖 1.1 所示。首先輸入作者，然後輸入論文名稱，期刊名稱，冊別，號碼別，頁別，發表月份、發表年份，…等等。但是以人工作這樣的事情實在是非常的費時。而且有些文件資料早就已經有電子檔案存在，只是沒有將欄位分開存放。如果可以藉由電腦自動判讀這些文件，擷取需要的部分並且存入資料庫中正確的位置，就可以省下不少時間。由於原始資料中的著作表列，往往依每個人習慣使用的格式而不盡相同。想要寫個通用的程式來判讀所有格式實為不易。為了解決這個問題，一開始必須蒐集一些著作表列的格式，然後撰寫程式去分辨每一種格式的後設資料。但是要作這樣的事並沒有想像中簡單，因為這會隨著蒐集增多，而出現越來越多問題。首先每當增加一種格式，整個程式幾乎就要重新改寫一遍，才能讓每一種格式都可以被辨識出來而不會混淆。第二，程式越變越大，只要加入一種新格式，就要加入好幾個判斷式。第三，花的時間越來越多，因為程式碼越來越大，想要加入新的判斷式就越來越複雜，考慮的事項就越來越多。最後會造成無從著手，甚致於以人工處理還來的比較快。因此必須要有一個架構，讓程式設計師，甚至連使用者都可以輕易的加入新的格式。而且在加入新的格式之後，還可以不影響之前的結果，這就是本文想要作的事。

1.2 問題敘述

BibTeX 是一個標準的程式及檔案格式，如圖 1.2。在 1985 年由 Oren Patashnik 和 Leslie Lamport 所制定。當初制定是為了 LaTeX 的需求。格式上完全以文字為基礎，所以可以套用在任何的程式。而它是以欄位為基礎，對於沒定義的欄位 BibTeX program 會將它忽視。它是現今 Internet 上最常見的目錄學格式，所以決定將它所使用的欄位

引用到我們的實驗室網站中。要將它的功能實作出來並不困難，困難的是要讓大家習慣使用它。因為大部分的實驗室成員在自己的網頁上都有一份著作表列，如圖 1.3。現在基於實驗室網站的需求，要求大家將自己的著作表列移植到實驗室網站上時，卻要將每一筆著作拆成好幾個欄位，然後再一個欄位接著一個欄位的填入適當的位置。由於 BibTeX 所制定的格式相當複雜，往往填入一筆資料就要耗費相當多的時間，如果沒有參考使用手冊經常會將資料填入錯誤的欄位。因為不同種類的著作需要將資料填入不同的欄位中。例如 conference 和 journal 直覺上是屬於同一種類，但是在 BibTeX 中它們所使用的欄位卻不盡相同，對於第一次使用的人會有很大的挫折感。因為有這樣的不方便性，會造成實驗成員不想花時間將資料移植到實驗室網站上，讓實驗室網站形同虛設。如能使用自動化的方法，將著作表列的後設資料解析出來，一筆一筆存入實驗室網站，最後再由實驗室成員各自去確認以及修改自己的部分，就可以省下不少時間，增加使用的意願。不過著作表列格式實在是千奇百怪，每個人都有自己習慣使用的格式，如果只是一昧往程式中加入新的程式碼來解析新的格式，這將會使整個程式變的動彈不得。想要從書目資料中自動解析出欄位看似簡單，其實並沒有想像中容易[2]。

1.3 論文的結構

第一章做完概略介紹之後，第二章將針對相關研究作探討。首先是書目自動索引，因為大部分與本研究有直接相關的著作都是集中在書目自動索引這個領域，由於相關研究不是很多，所以可作較詳細的介紹。因為本系統是以 BLAST 作為蛋白質序列比對的工具，所以針對序列比對工具的演進歷史在第二章中稍作說明。以及說明為何挑選 BLAST 當作基因字串比對的工具。同時也對 BLAST 的演算方法稍作介紹。第三章的主題是系統架構之設計。這一章進入本論文研究的主

題，介紹本系統如何將一筆書目資料轉成基因格式，以及如何建立樣板資料庫。最後再藉由 BLAST 從資料庫中找出最相似的樣板，然後根據這個樣板解析出正確的後設資料。第四章是實驗與結果，在這一章中我們設計了幾個實驗來探討系統準確度。根據實驗的結果作些比較，了解本系統跟現有系統的差異。最後，第五章則是結論。

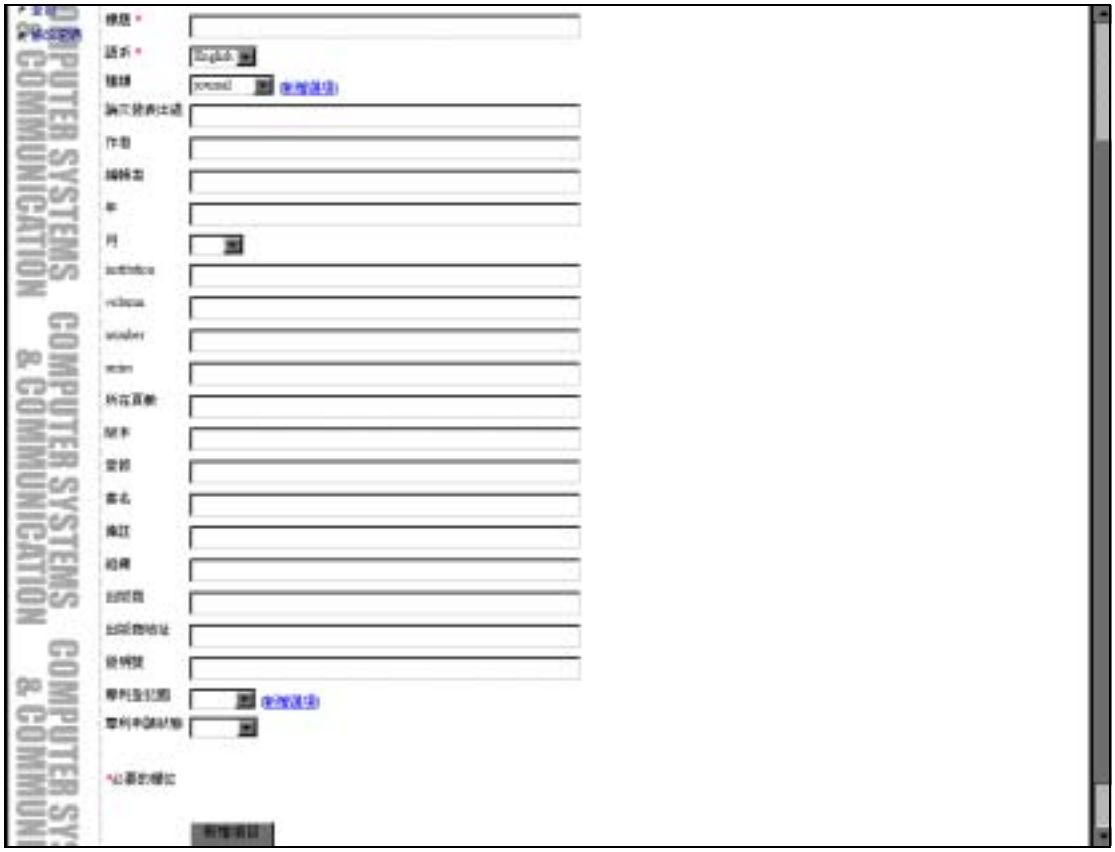

圖 1.1 著作輸入範例

```
@Article{SaZi94,
   author =          "Bruno Salvy and Paul Zimmermann",
   title =    "Gfun: a {M}aple package for the manipulation of generating
                      and holonomic functions in one variable",
   journal =         "ACM Transactions on Mathematical Software",
   year =    1994,
   volume =          20,
   number =          2,
   pages =           "163--177"
}
```

圖 1.2 BibTeX 格式範例

Publications

Refereed Papers:

1. Yieh-Ran Huang and Ian-Ming Ho, "Distributed Call Admission Control for a Heterogeneous PCS Network", to appear IEEE Trans. On Computers, vol. 51, no. 11, Nov. 2002.
2. Ray-I Chang, Meng-Chang Chen, Ming-Tat Ko and Ian-Ming Ho, "Schedulable Region for VBR Media Transmission with Optimal Resource Allocation and Utilization," Information Sciences, Vo. 141, Issue 1-2, pp. 61-79, 2002. (SCI)
3. Ray-I Chang, Meng-Chang Chen, Ming-Tat Ko and Ian-Ming Ho, "VBR Traffic Shaping for Streaming of Multimedia Transmission," Multimedia Networking: Technology, Management and Applications, Mahbubur Rahman Syed (Eds.), to appear. [Book Chapter]
4. Ray-I Chang, Meng-Chang Chen, Ian-Ming Ho, and Ming-Tat Ko, "Schedulable Region for VBR Media Transmission with Optimal Resource Allocation and Utilization," Information Science: an International Journal, special issue on "Intelligent Multimedia Computing and Networking", accepted for publication.
5. Ray-I Chang, Meng-Chang Chen, Ian-Ming Ho & Ming-Tat Ko, "Online Traffic Smoothing for Delivery of VBR Media Streams", Circuits, Systems, Signal Processing - Special Issue on Multimedia Communication, Vol. 20, No. 1, 2001.
6. Ian-Ming Ho, Shih-Kan Huang, Tyng-Ruey Chuang and D. T. Lee, "On creation and management of digital libraries: system environment, human-computer interface and research issues," J. Library & Information Science, (26,2), 38-48, Oct. 2000.
7. Y.-R Haung, Y.-B Lin, and I.-M Ho, "Performance Analysis for Voice/Data Integration on a Finite-Buffer Mobile System," IEEE Trans. Veh. Technol., 49(2), March 2000.
8. Shian-Hua Lin, Meng Chang Chen, Ian-Ming Ho, and Yueh-Min Huang, "ACIRD: Intelligent Internet Documents Organization and Retrieval", appear to IEEE Transactions on Knowledge and Data Engineering.
9. R.I. Chang, M.C. Chen, I.M. Ho and M.T. Ko, "Optimal bandwidth-buffer tradeoff for VBR media transmission over multiple relay-servers," IEEE Multimedia System (also in Proc. IEEE ICMCS'99), vol. 2, pp.31-35, 1999.
10. Cheng-Hsing Yang, Sao-Jie Chen, Ian-Ming Ho and Chia-Chun Tsai, "Efficient routability check algorithms for segmented channel routing"; ACM Trans. Des. Autom. Electron. Syst. 5, 3, Pages 735 - 747, Jul. 2000.
11. Shiao-Li Tsao, Meng Chang Chen, Ian-Ming Ho, Ming-Tat Ko, and Yueh-Min Huang, "Data Allocation and Dynamic Load Balancing for Distributed Video Storage Server" Journal of Visual Communication and Image Representation, vol. 10, no. 2, 1999, p197-218.
12. Der-Jen Lu, Yu-Chung Wang, Ian-Ming Ho, Meng-Chang Chen, and Ming-Tat Ko, "Experience in Designing A Using TCP as Transport Protocol VOD System over A Dedicated Network", IEEE Trans. on Consumer Electronics, Vol. 44, No. 4, Nov. 1998.
13. I.-S. Cheng, S.-J. Chen, and I.-M. Ho, "Efficient bipartitioning algorithm for size-constrained circuits", IEE Proc.-Comput. Digit. Tech., Vol. 145, No. 1, January 1998.
14. Ian-Ming Ho and M.T. Ko, "Bounded fan-out m-center problem", Information Processing Letters, vol. 63, 1997, pp. 103-108.
15. Cheng-Hsing Yang, Sao-Jie Chen, Ian-Ming Ho, and Chia-Chun Tsai, "Hmap: A Fast Mapper for EPOAs Using Extended OBDD Hash Tables", ACM Transactions on Design Automation of Electronic Systems, vol. 2, no. 2, Apr. 1997.
16. Meng Chang Chen , Ian-Ming Ho, Ming-Tat Ko and Shie-Yuan Wang, "A SCSI Disk Model for Multimedia Storage Systems", o appear in Computer System Science & Engineering, June 1999.
17. M.-H. Lee, C.-H. Chang, M. C. Chen, I.-M. Ho, M.-T. Ko, Y.-J. Oyang, K.-H. Tsai, and S.-Y. Wang, "Design and Implementation of a Predictable High-Throughput Video Conference Recorder", IEEE Trans. on Consumer Electronics, Feb. 1996 (also in Proceedings 1995 International Workshop on HDTV and the Evolution of Television, Nov. 1995).
18. Ian-Ming Ho, M. Sarrafzadeh and A. Suzuki, "An Exact Algorithm for Single-Layer Wire-Length Minimization", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, Jan. 1993.

圖 1.3 著作表列範例