行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　※
※　生物與資訊工程整合教育課程規劃及研究(3/3)　※
※　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　■整合型計畫

計畫編號：NSC92-2516-S-216-003

執行期間：92 年 10 月 01 日至 93 年 09 月 30 日

計畫主持人：鄭藏勝 教授

共同主持人：鄭芳炫 教授　李文權 研究員

林道通 副教授　黃三元 副研究員

許文龍 副教授　林恩仲 副研究員

周智勳 副教授　吳哲賢 副教授

劉志俊 助理教授 侯玉松 助理教授

羅　琪 副教授　俞征武 副教授

張慧玫 助理教授 劉世華 助理教授

執行單位：中華大學資訊工程學系

中　華　民　國　93　年　12　月　15　日

行政院國家科學委員會專題研究計畫成果報告

生物與資訊工程整合教育課程規劃及研究(3/3)

計畫編號：NSC 92-2516-S-216-003
執行期限：92 年 10 月 1 日至 93 年 9 月 30 日

主持人：中華大學機械工程學系 教授鄭藏勝
共同主持人：中華大學資訊工程學系 教授鄭芳炫
台灣動物科技研究所 研究員李文權
E-mail: tscheng@chu.edu.tw

一. 中文摘要

生物資訊已是二十一世紀全世界科技研究最重要的課題，生物科技與資訊科技的結合已是未來科技不可避免之發展趨勢。「人類基因體解讀計畫(Human Genome Project)」預計 2003 年以前將把人類的 3×10^9 DNA 序列全部解碼並決定基因的生物功能，隨著人類基因圖譜即將定序完成，科學家的下一個課題便是了解數萬個基因所代表的意義和其相互關係，以及蛋白質組功能研究。為了對這龐大資料加以處理以期達到完全了解基因，我們極需尋求計算的工具及技術，以便協助生物學家發覺序列中重要的特徵，並洞悉其功能及結構。目前全世界先進國家均積極在發展生物科技技術，而我國亦將生物科技列為國家之重點發展項目。然而我國面臨最大之瓶頸乃在於生物科技人才嚴重不足之問題，因此本計畫之研究重點及目標即是以生物科技人才之培育為主，希望為國內生物科技盡一已之力。

目前在國內生物科技乃是以有生化科系之學校為主導,如台大，陽明等學校，然而生物科技研究需借重大量之資訊處理技術及能力才能事半功倍，而專業之生物科技人才卻往往無法有效率的運用資訊科技來協助其研究。另一方面專業之資訊人才，卻因沒有生物知識之訓練而來協助生物科技之發展。因此若能結合生物與資訊工程之整合教育而培養出具生物資訊之專業人才將對我國生物科技之發展大有助益。目前在國內已有少數學校規劃出生物資訊之相關學程，不過仍以生化科系為主，如陽明醫學院之生物資訊學程，而以資訊為主而規劃之生物資訊學程則很少。因此本計畫以資訊工程為主體結合生物科技之訓練規劃出一系列之生物資訊學程，讓國內廣大之工程人才只要經計畫中課程之訓練即可成為生物科技產業之專門人才。本總計畫共規劃出四個專門學程即生物資訊學程，生物計算法學程，隨機分析學程和生物資料庫學程，且分別由四個子計畫來執行完成。本計畫前後共三年，第一年完成了大學部高年級十門課程，第二年亦完成了大學部高年級另十門課程，而第三年則完成了研究所四門課程。計畫之完成已為國內生物科技界培養出一批生力軍，而本計畫之課程規劃亦可移植至其他大學以培養更多之生物科技人才。

Abstract

Bioinformatics has becoming one of the major research topics in the 21th century. The constitution of computer technology and molecular biology technology is evidently essentially in the future. Due to the effort of the Human Genome Project, the full sequence of all 3-billion DNA bases is set to be completed by the end of 2003 and to determine the biological functions of the genes. The next challenge will be to fully realize the value of the data and gain a full understanding of the genome, the relation between genes and function of proteins. Powerful computational

tools and technologies are critical for scientists to discover the biological features to provide insight into their structures and functions. Most of the progressive countries pay more attention on the research of Bio-technology including Taiwan. However, the bottleneck of development of Bio-technology in Taiwan is lack of bio-engineers. Therefore, the objective of this project is to educate and train a sufficient member of bio-engineers to contribute on the development of Bio-technology for Taiwan.

The project is partitioned into four parts; Bio-informatics, Bio-computational algorithms, Stochastic analyses and Bio-database. Each part concerns on the course planning and research on its related topics. The undergraduate and graduate course are also included in the course planning. This project lasts for three years. In the first year, we had opened 10 courses for undergraduate students and another 10 courses for the second years. For the last years of the project, 4 courses for graduate students are opened.

二. 緣由與目的

目前在國內生物科技乃是以有生化科系之學校為主導,如台大, 陽明等學校, 然而生物科技研究需借重大量之資訊處理技術及能力才能事半功倍, 而專業之生物科技人才卻往往無法有效率的運用資訊科技來協助其研究。另一方面專業之資訊人才, 卻因沒有生物知識之訓練而來協助生物科技之發展。因此若能結合生物與資訊工程之整合教育而培養出具生物資訊之專業人才將對我國生物科技之發展大有助益。目前在國內已有少數學校規劃出生物資訊之相關學程，不過仍以生化科系為主，如陽明醫學院之生物資訊學程，而以資訊為主而規劃之生物資訊學程則很少。

本校目前已成立生物資訊學系，且已經和學校附近之台灣動物科技研究所建立合作關係，本校將協助台灣動物科技研究所規劃電腦中心、育成中心及網路系統，台灣動物科技研究所則支援本校生物相關課程之師資及實驗設備，因此在計畫中順利規劃了大學部及研究所之生物資訊學程，並已完全實行三年之課程，明年度將再規劃生物資訊之推廣教育，進一步服務專校和職校學生。

本校目前亦正積極籌設生物資訊研究中心，其目標即是以生物科技為基礎, 應用資訊科技之技術以完成生物資訊系統之研究及其應用技術之發展。因為生物資訊專業人材非常稀少，現在已有十五位以上之博士級資訊專長教師轉型投入此研究領域，本校生物資訊研究中心最重要的四大研究主題，（一）生物資訊計算法：由平行分散專長教師負責（二）生物資訊知識庫：由資料庫專長教師負責（三）3D 視覺化計算法及醫學影像：由人機界面專長教師負責（四）生物晶片設計及生物電腦技術：由積體電路專長教師負責。

本計畫的目的是以資訊工程為主體結合生物科技之訓練規劃出一系列之生物資訊學程,讓國內廣大之工程人才只要經計畫中課程之訓練即可成為生物科技產業之專門人才。

三. 計畫內容

目前在國內生物科技乃是以有生化科系之學校為主導,如台大, 陽明等學校, 然而生物科技研究需借重大量之資訊處理技術及能力才能事半功倍, 而專業之生物科技人才卻往往無法有效率的運用資訊科技來協助其研究。另一方面專業之資訊人才, 卻因沒有生物知識之訓練而來協助生物科技之發展。因此若能結合生物與資訊工程之整合教育而培養出具生物資訊之專業人才將對我國生物科技之發展大有助益。目前在國內已有少數學校規劃出生物資訊之相關學程，不過仍以生化科系為主，如陽明醫學院之生物資訊學程，而以資訊為主而規劃之生物資訊學程則很少。因此本計畫以資訊工程為主體結合了生物科技之訓練規劃出一系列之生物資訊學程,讓國內廣大之工程人才只要經計畫中課程之訓練即可成為生物科技產業之專門人才。本計畫共規劃出四個專門學程即生物資

訊學程，生物計算方法學程，隨機分析學程和生物資料庫學程，且分別由四個子計畫來執行完成。所有學程之規劃課程涵蓋大學部高年級及研究所。整體之學程規劃及各學程之關係如下圖所示。本計畫前後共三年，第一年完成了大學部高年級十門課程，第二年亦完成了大學部高年級另十門課程，而第三年則完成了研究所四門課程。經三年計畫執行已為國內生物科技界培養出一批生力軍，而本計畫之課程規劃亦可移植至其他大學以培養更多之生物科技人才。

四. 計畫成果自評

根據下圖之計畫內容規劃，本計畫第一年之課程主要以四大學程即生物資訊學程，計算方法學程，隨機分析學程及資料庫學程之基礎課程為主。在第一年度之計畫中我們已完全開立了計畫規劃之十門課程即基礎生物學，分子細胞生物學，生物資訊學導論，生物資訊學實務，資料結構，電腦演算法，機率，隨機過程，資料庫系統及進階資料庫，並完成第一年之期中報告及課程講義。第二年之計畫中，主要以四大學程即生物資訊學程，計算方法學程，隨機分析學程及資料庫學程之理論課程為主。在第二年度的計畫中我們也完全開立了計畫規劃之十門課程即分子遺傳學，生物化學，生物資訊學，進階生物資訊學，平行處理，作業研究，類神經網路，機器學習，生物資料庫設計及平行分散資料庫。計畫前二年的計畫期中報告及詳細課程講義均如期繳交完成。本年度依計畫內容在研究所共開立了四門進階之生物資訊課程，即生物電腦，生物知識庫，分子序列及演化樹分析及蛋白質結構及功能分析。課程詳細之成果詳列於以下各子計畫之成果報告中。詳細之課程講義亦整理於附件一併繳交。

<table>
<tr><td></td><td>第一學期 第二學期 第三學期 第四學期</td><td>第五學期 第六學期</td></tr>
<tr><td></td><td>基礎課程</td><td>研究課程</td></tr>
<tr><td>生物資訊課程
子計畫一</td><td>基礎生物學 分子細胞生物學 分子遺傳學 生物化學
生物資訊學導論 生物資訊學實務 生物資訊學 進階生物資訊學</td><td rowspan="4">生物資訊學專題(一)(二)：
生物電腦
生物資訊學專題(一)(二)：
生物知識庫
生物資訊學專題(一)(二)：
分子序列及演化樹分析
生物資訊學專題(一)(二)：
蛋白質結構及功能分析</td></tr>
<tr><td>計算方法課程
子計畫二</td><td>資料結構 電腦演算法 平行處理 作業研究</td></tr>
<tr><td>隨機分析課程
子計畫三</td><td>機率與統計學 隨機過程 類神經網路 機器學習</td></tr>
<tr><td>資料庫課程
子計畫四</td><td>資料庫系統 進階資料庫 生物資料庫設計 平行分散資料庫</td></tr>
</table>

參考文獻：

1. J. Setubal and J. Meidanis, ***Introduction to Computational Molecular Biology***," PWS Publishing Company, 1997.
2. M. Ercegovac, T. Lang, and J. H. Moreno, ***Introduction to Digital Systems***, John Wiley & Sons, Inc., 1999
3. J. P. Uyemura, ***A First Course in Digital Systems Design: An Integrated Approach***, Brooks/Cole publishing Company, 2000.
4. J. F. Wakerly, ***Digital Design: Principles & Practices***, 3rd Edition, Prentice Hall International, Inc, 2000.
5. K. C. Chang, ***Digital Systems Design with VHDL and Synthesis: An integrated Approach***, *IEEE Computer Society*, 1999.
6. A. Dewey, ***Analysis and Design of Digital Systems with VHDL***, PWS Publishing Company, 1997.
7. M. J. S. Smith, ***Application-Specific Integrated Circuits***, Addison-Wesley Longman, Inc., 1997.
8. P. Mazumder and E. M. Rudnick, ***Genetic Algorithms for VLSI Design, Layout & Test Automation***, Prentice Hall International, Inc, 1999.
9. D. E. Goldberg, ***Genetic Algorithms in Search, Optimization & Machine Learning***, Addison-Wesley Longman, Inc., 1989
10. Z. Michalewicz, ***Genetic Algorithms + Data Structures = Evolution Programs***, Springer-Verlag Berlin Heidelberg, 1992.
11. D. A. Pattern and J. L. Hennessy, ***Computer Organization & Design: the Hardware/Software Interface***, *Morgan Kaufmann Publishers, Inc.*, 1998.
12. J. P. Hayes, ***Computer Architecture and Organizatio***n, 3rd Edition, *McGraw-Hill Book Company*, 1998.
13. V. P. Heuring and H. F. Jordan, ***Computer Systems Design and Architecture***, Addison-Wesley Longman, Inc., 1997
14. D. A. Pattern and J. L. Hennessy, ***Computer Architecture: A Quantitative Approach***, 2nd Edition, *Morgan Kaufmann Publishers, Inc.*, 1996.
15. K. Hwang and F. A. Briggs, ***Computer Architecture and Parallel Processing***, *McGraw-Hill Book Company*, 1984.
16. M. J. Murdocca and V. P. Heuring, ***Principles of Computer Architecture***, Prentice Hall International, Inc, 2000.
17. J. W. Valvano, ***Embedded Microcomputer Systems: Real Time Interfacing***, Brooks/Cole, Inc, 2000.
18. K. L. Short, ***Embedded Microprocessor Systems Desogn***, Prentice Hall International, Inc, 1998.
19. M. J. Daley and M. G. Eramian, ***Models of DNA Computation***, private communication,1998.
20. S.F. Altschul *et al.*, "Basic local alignment search tool," *J. Mol. Biol.*, Vol. 215, pp. 403-410, 1990.
21. D.J. Lipman and W.R. Pearson, "Rapid and sensitive protein similarity searches," *Science*, Vol. 227, pp. 1435-1441, 1985.
22. S.F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST," *Nucleic Acids Res.*, Vol. 25, pp. 3389-3402, 1997.
23. FURY: fuzzy unification and resolution based on edit distance Gilbert, D.; Schroeder, M. Bio-Informatics and Biomedical Engineering, 2000. Proc. IEEE Intern. Symp. on , Pg. 330 -336.
24. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028) Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on , Volume: 1 , 1999.
25. Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000 Fuzzy Systems, IEEE Transactions on , Volume: 8 Issue: 5 , Oct. 2000.
26. Hybrid Protein Model (HPM): a method to compact protein 3D-structure information and physicochemical properties De Brevern, A.G.; Hazout, S.A. String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on , 2000, Page(s): 49 -54.
27. A necessary condition for self-reproduction in the Semar core

Suzuki, H. Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE '99. 1999 IEEE International , Volume: 1 , 1999 Page(s): 123 -128 vol.1.

28. A soft computing approach to the metabolic modeling Yen, J.; Lee, B.; Liao, J.C. Fuzzy Information Processing Society, 1996. NAFIPS., 1996 Biennial Conference of the North American , 1996 Page(s): 343 –347
29. A neural-fuzzy system for the protein folding problem Daugherity, W.C. Industrial Fuzzy Control and Intelligent Systems, 1993., IFIS '93., Third International Conference on , 1993 Page(s): 47 -49.
30. A computational neural approach to support the discovery of gene function and classes of cancer Azuaje, F. Biomedical Engineering, IEEE Transactions on , Volume: 48 Issue: 3 , March 2001 Page(s): 332 -339.
31. Making genome expression data meaningful: prediction and discovery of classes of cancer through a connectionist learning approach Azuaje, F. Bio-Informatics and Biomedical Engineering, 2000. Proceedings. IEEE International Symposium on , 2000 Page(s): 208 -213.
32. Fuzzy molecular modeling Ress, D.A. Intelligent Processing and Manufacturing of Materials, 1999. IPMM '99. Proceedings of the Second International Conference on , Volume: 2 , 1999 Page(s): 1225 -1233 vol.2.
33. Another logical molecular NAND gate system Mulawka, J.J.; Wasiewicz, P.; Pluciennicczak, A. Microelectronics for Neural, Fuzzy and Bio-Inspired Systems, 1999. MicroNeuro '99. Proceedings of the Seventh International Conference on , 1999 Page(s): 340 -345.
34. Fuzzy system identification for composite operation and fuzzy relation by genetic algorithms Ohtani, S.; Kikuchi, H.; Yager, R.R.; Nakanishi, S. Knowledge-Based Intelligent Electronic Systems, 1997. KES '97. Proceedings., 1997 First International Conference on , Volume: 1 , 1997 Page(s): 289 -295 vol.1.
35. Stratification structures on a kind of completely distributive lattices and their applications in theory of topological molecular lattices Cui Hongbin; Zheng Chongyon Fuzzy Systems Symposium, 1996. Soft Computing in Intelligent Systems and Information Processing., Proceedings of the 1996 Asian , 1996 Page(s): 484 -489.
36. Systematic passive shimming of a permanent magnet for P-31 NMR spectroscopy of bone mineral Battocletti, J.H.; Kamal, H.A.; Myers, T.J.; Knox, T.A. Magnetics, IEEE Transactions on , Volume: 29 Issue: 4 , July 1993 Page(s): 2139 -2151.
37. Quantitative measures of molecular similarity (drug design) Arteca, G.A.; Mezey, P.G. Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century., Proceedings of the Annual International Conference of the IEEE Engineering in , 1989 Page(s): 1907 -1908 vol.6.
38. Mining residue contacts in proteins using local structure predictions Zaki, M.J.; Shan Jin; Bystroff, C. Bio-Informatics and Biomedical Engineering, 2000. Proceedings. IEEE International Symposium on , 2000 Page(s): 168 -175.
39. Coding model for translation in E. coli K-12 May, E.E.; Vouk, M.A.; Bitzer, D.L.; Rosnick, D.I. BMES/EMBS Conference, 1999. Proceedings of the First Joint , Volume: 2 , 1999 Page(s): 1178 vol.2.
40. Identification of hidden Markov models for ion channel currents. I. Colored background noise Venkataramanan, L.; Walsh, J.L.; Kuc, R.; Sigworth, F.J. Signal Processing, IEEE Transactions on , Volume: 46 Issue: 7 , July 1998 Page(s): 1901 -1915.
41. Identification of hidden Markov models for ion channel currents. II. State-dependent excess noise Venkataramanan, L.; Kuc, R.; Sigworth, F.J. Signal Processing, IEEE Transactions on , Volume: 46 Issue: 7 , July 1998 Page(s): 1916 -1929.
42. Hidden Markov models in biomedical signal processing Cohen, A.

Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE , Volume: 3 , 1998 Page(s): 1145 -1150 vol.3.

43. Gene family identification network design Wu, C.H.; Shivakumar, S. Intelligence and Systems, 1998. Proceedings., IEEE International Joint Symposia on , 1998 Page(s): 103 -110.
44. A combination of a functional motif model and a structural motif model for a database validation Asogawa, M.; Fujiwara, Y.; Konagaya, A. System Sciences, 1995. Vol.1 Proceedings of the Twenty-Eighth Hawaii International Conference on , 1995 Page(s): 174 -183 vol.5.
45. Parameterization studies of hidden Markov models representing highly divergent protein sequences McClure, M.A.; Raman, R. System Sciences, 1995. Vol.1 Proceedings of the Twenty-Eighth Hawaii International Conference on , 1995 Page(s): 184 -194 vol.5.
46. A three-dimensional animation system for protein folding simulation Akahoshi, M.; Onizuka, K.; Ishikawa, M.; Asai, K. System Sciences, 1994. Vol.V: Biotechnology Computing, Proc. of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 173 -182.
47. Application of hidden Markov modeling to the characterization of transcription factor binding sites Raman, R.; Overton, G.C. System Sciences, 1994. Vol.V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 275 -283.
48. Stochastic context-free grammars for modeling RNA Sakakibara, Y.; Brown, M.; Underwood, R.C.; Mian, I.S.; Haussler, D. System Sciences, 1994. Vol.V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 284 -293.
49. HMM with protein structure grammar Asai, K.; Hayamizu, S.; Onizuka, K. System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on , 1993 Page(s): 783 -791 vol.1.
50. Protein modeling using hidden Markov models: analysis of globins Haussler, D.; Krogh, A.; Mian, I.S.; Sjolander, K. System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on , 1993 Page(s): 792 -802 vol.1.
51. Mining residue contacts in proteins using local structure predictions Zaki, M.J.; Shan Jin; Bystroff, C. Bio-Informatics and Biomedical Engineering, 2000. Proceedings. IEEE International Symposium on , 2000 Page(s): 168 -175.
52. Identification of hidden Markov models for ion channel currents. I. Colored background noise Venkataramanan, L.; Walsh, J.L.; Kuc, R.; Sigworth, F.J. Signal Processing, IEEE Transactions on , Volume: 46 Issue: 7 , July 1998 Page(s): 1901 -1915.
53. Identification of hidden Markov models for ion channel currents. II. State-dependent excess noise Venkataramanan, L.; Kuc, R.; Sigworth, F.J. Signal Processing, IEEE Transactions on , Volume: 46 Issue: 7 , July 1998 Page(s): 1916 -1929.
54. Gene family identification network design Wu, C.H.; Shivakumar, S. Intelligence and Systems, 1998. Proceedings., IEEE International Joint Symposia on , 1998 Page(s): 103 -110.
55. Parameterization studies of hidden Markov models representing highly divergent protein sequences McClure, M.A.; Raman, R. System Sciences, 1995. Vol.1 Proceedings of the Twenty-Eighth Hawaii International Conference on , 1995 Page(s): 184 -194 vol.5.
56. A three-dimensional animation system for protein folding simulation Akahoshi, M.; Onizuka, K.; Ishikawa, M.; Asai, K. System Sciences, 1994. Vol.V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 173 -182.
57. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on

Systems, Man, and Cybernetics (Cat. No.99CH37028) Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on , Volume: 1 , 1999.
58. Parameterizing genetic algorithms for protein folding simulation Schulze-Kremer, S.; Tiedemann, U. Molecular Bioinformatics, IEE Colloquium on , 1994 Page(s): 8/1 -8/7.
59. IEE Colloquium on 'Molecular Bioinformatics' (Digest No.1994/029) Molecular Bioinformatics, IEE Colloquium on , 1994.
60. A hybrid genetic algorithm application to a genetics sequencing problem Walker, J.D.; File, P.E.; Miller, C.J.; Samson, W.B. Molecular Bioinformatics, IEE Colloquium on , 1994 Page(s): 7/1 -712.
61. Genetic algorithm driven clustering for toxicity prediction Devogelaere, D.; Van Bael, P.; Rijckaert, M. Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International Conference on , Volume: 1 , 2000. Page(s): 173 -176 vol.1.
62. A hybrid genetic algorithm for finding stable conformations of small molecules Barbosa, H.J.C.; Raupp, F.M.P.; Lavor, C.; Lima, H.; Maculan, N. Neural Networks, 2000. Proceedings. Sixth Brazilian Symposium on , 2000 Page(s): 90 -94.
63. Application of neural networks: a molecular geometry optimization study Lemes, M.R.; Zacharias, C.R.; Dal Pino, A., Jr. Neural Networks, 2000. Proceedings. Sixth Brazilian Symposium on , 2000. Page(s): 288
64. Prediction of protein structures using a Hopfield network Scott, L.P.B.; Chahine, J.; Ruggiero, J.R. Neural Networks, 2000. Proceedings. Sixth Brazilian Symposium on , 2000 Page(s): 284
65. Memetic algorithms and the molecular geometry optimization problem Hodgson, R.J.W. Evolutionary Computation, 2000. Proceedings of the 2000 Congress on , Volume: 1 , 2000 Page(s): 625 -632 vol.1.
66. Convergence analysis of a segmentation algorithm for the evolutionary training of neural networks Huning, H. Combinations of Evolutionary Computation and Neural Networks, 2000 IEEE Symposium on , 2000 Page(s): 70 -81.
67. A technique of genetic algorithm and sequence synthesis for multiple molecular sequence alignment Ching Zhang; Wong, A.K.C. Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on , Volume: 3 , 1998, Page(s): 2442 -2447 vol.3.
68. Molecular binding in structure-based drug design: a case study of the population-based annealing genetic algorithms Chien-Cheng Chen; Leuo-Hong Wang; Cheng-Yan Kao; Ming Ouhyoung; Wen-Chin Chen Tools with Artificial Intelligence, 1998. Proceedings. Tenth IEEE International Conference on , 1998 Page(s): 328 -335.
69. Coal molecular structure construction by genetic algorithm Ibayashi, S.; Ohkawa, T.; Komoda, N. Intelligence and Systems, 1998. Proceedings., IEEE International Joint Symposia on , 1998. Page(s): 111 -115.
70. Toward efficient multiple molecular sequence alignment: a system of genetic algorithm and dynamic programming Ching Zhang; Wong, A.K.C. Systems, Man and Cybernetics, Part B, IEEE Transactions on , Volume: 27 Issue: 6 , Dec. 1997 Page(s): 918 -932.
71. Stalk: an interactive system for virtual molecular docking Levine, D.; Facello, M.; Hallstrom, P.; Reeder, G.; Walenz, B.; Stevens, F. IEEE Computational Science and Engineering , Volume: 4 Issue: 2 , April-June 1997 Page(s): 55 -65.
72. The circular schema theorem for genetic algorithms and two-point crossover Neubauer, A. Genetic Algorithms in Engineering Systems: Innovations and Applications, 1997. GALESIA 97. Second International Conference On (Conf. Publ. No. 446) , 1997 Page(s): 209 -214.

73. Search for native conformations of organic molecules by genetic algorithms Beiersdorfer, S.; Hesser, J.; Schmitt, J.; Manner, R.; Schulz, A.; Wolfrum, J. EUROMICRO 97. New Frontiers of Information Technology., Proceedings of the 23rd EUROMICRO Conference , 1997 Page(s): 624 -630.
74. Fuzzy system identification for composite operation and fuzzy relation by genetic algorithms Ohtani, S.; Kikuchi, H.; Yager, R.R.; Nakanishi, S. Knowledge-Based Intelligent Electronic Systems, 1997. KES '97. Proceedings., 1997 First International Conference on , Volume: 1 , 1997. Page(s): 289 -295 vol.1.
75. Exogenous parameter selection in a real-valued genetic algorithm Kaiser, C.E., Jr.; Lamont, G.B.; Merkle, L.D.; Gates, G.H., Jr.; Pachter, R. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 569 -574.
76. Ligation experiments in computing with DNA Jonoska, N.; Karl, S.A. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 261 -266.
77. New parallel hybrid genetic algorithm based on molecular dynamics approach for energy minimization of atomistic systems Celino, M.; Palazzari, P.; Pucello, N.; Rosati, M.; Rosato, V. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 115 -119.
78. Parallel genetic algorithms on PARAM for conformation of biopolymers Sundararajan, V.; Kolaskar, A.S. High Performance Computing, 1996. Proceedings. 3rd International Conference on , 1996 Page(s): 22 -26.
79. Hybrid genetic algorithms for minimization of a polypeptide specific energy model Merkle, L.D.; Lamont, G.B.; Gates, G.H., Jr.; Pachter, R. Evolutionary Computation, 1996., Proceedings of IEEE International Conference on , 1996 Page(s): 396 -400.
80. Four problems for which a computer program evolved by genetic programming is competitive with human performance Koza, J.R.; Bennett, F.H., III; Andre, D.; Keane, M.A. Evolutionary Computation, 1996., Proceedings of IEEE International Conference on , 1996 Page(s): 1 -10.
81. Molecular binding: a case study of the population-based annealing genetic algorithms Leuo-hong Wang; Cheng-yan Kao; Ming Ouh-young; Wen-chin Chen Evolutionary Computation, 1995., IEEE International Conference on , Volume: 1 , 1995 Page(s): 50.
82. Simple genetic algorithm parameter selection for protein structure prediction Gates, G.H., Jr.; Merkle, L.D.; Lamont, G.B.; Pachter, R. Evolutionary Computation, 1995., IEEE International Conference on , Volume: 2 , 1995 Page(s): 620 -624 vol.2.
83. Proceedings First International Symposium on Intelligence in Neural and Biological Systems. INBS'95 Intelligence in Neural and Biological Systems, 1995. INBS'95, Proceedings., First International Symposium on , 1995.
84. Parameterizing genetic algorithms for protein folding simulation Schulze-Kremer, S.; Tiedemann, U. Molecular Bioinformatics, IEE Colloquium on , 1994 Page(s): 8/1 -8/7.
85. IEE Colloquium on 'Molecular Bioinformatics' (Digest No.1994/029) Molecular Bioinformatics, IEE Colloquium on , 1994.
86. A hybrid genetic algorithm application to a genetics sequencing problem Walker, J.D.; File, P.E.; Miller, C.J.; Samson, W.B. Molecular Bioinformatics, IEE Colloquium on , 1994 Page(s): 7/1 -712.
87. A genetic algorithm for molecular sequence comparison Ching Zhang Systems, Man, and Cybernetics, 1994. Humans, Information and Technology., 1994 IEEE International Conference on , Volume: 2 , 1994 Page(s): 1926 -1931 vol.2.
88. A comparison of GA and RSNR docking Xiao, Y.L.; Williams, D.E. Evolutionary Computation, 1994. IEEE World Congress on Computational

Intelligence., Proceedings of the First IEEE Conference on , 1994 Page(s): 802 -806 vol.2.
89. Using an annealing genetic algorithm to solve global energy minimization problem in molecular binding Leuo-Hong Wang; Cheng-Yan Kao; Ming Ouh-Young; Wen-Chin Cheu Tools with Artificial Intelligence, 1994. Proceedings., Sixth International Conference on , 1994 Page(s): 404 -410.
90. Parameterizing genetic algorithms for protein folding simulation Schulze-Kremer, S.; Tiedemann, U. System Sciences, 1994. Vol.V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 345 -354.
91. Classification and function estimation of protein by using data compression and genetic algorithms Chiba, S.; Sugawara, K.; Watanabe, T. Evolutionary Computation, 2001. Proceedings of the 2001 Congress on , Volume: 2 , 2001 Page(s): 839 -844.
92. imensionality reduction using genetic algorithms Raymer, M.L.; Punch, W.F.; Goodman, E.D.; Kuhn, L.A.; Jain, A.K. Evolutionary Computation, IEEE Transactions on , Volume: 4 Issue: 2 , July 2000 Page(s): 164 -171.
93. A family competition evolutionary algorithm for automated docking of flexible ligands to proteins Jinn-Moon Yang; Cheng-Yan Kao Information Technology in Biomedicine, IEEE Transactions on , Volume: 4 Issue: 3 , Sept. 2000 Page(s): 225 -237.
94. Prediction of protein structures using a Hopfield network Scott, L.P.B.; Chahine, J.; Ruggiero, J.R. Neural Networks, 2000. Proceedings. Sixth Brazilian Symposium on , 2000 Page(s): 284.
95. Application of genetic algorithm for predicting tertiary structures of peptide chains Yap, A.; Cosic, I. BMES/EMBS Conference, 1999. Proceedings of the First Joint , Volume: 2 , 1999 Page(s): 1214 vol.2.
96. Molecular binding in structure-based drug design: a case study of the population-based annealing genetic algorithms Chien-Cheng Chen; Leuo-Hong Wang; Cheng-Yan Kao; Ming Ouhyoung; Wen-Chin Chen Tools with Artificial Intelligence, 1998. Proceedings. Tenth IEEE International Conference on , 1998 Page(s): 328 -335.
97. Revisiting the GEMGA: scalable evolutionary optimization through linkage learning Bandyopadhyay, S.; Kargupta, H.; Gang Wang Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on , 1998 Page(s): 603 -608.
98. Genetic algorithm for artificial neurogenesis Clergue, M.; Collard, P. Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on , 1998 Page(s): 410 -415.
99. Exogenous parameter selection in a real-valued genetic algorithm Kaiser, C.E., Jr.; Lamont, G.B.; Merkle, L.D.; Gates, G.H., Jr.; Pachter, R. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 569 -574.
100. A comparative study of GA and orthogonal experimental design Tanaka, H. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 143 -146.
101. Ligation experiments in computing with DNA Jonoska, N.; Karl, S.A. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 261 -266.
102. Identifying genetically spliced languages Kim, S.M. Evolutionary Computation, 1997., IEEE International Conference on , 1997 Page(s): 231 -235.
103. Hybrid genetic algorithms for minimization of a polypeptide specific energy model Merkle, L.D.; Lamont, G.B.; Gates, G.H., Jr.; Pachter, R. Evolutionary Computation, 1996., Proceedings of IEEE International Conference on , 1996 Page(s): 396 -400.
104. A soft computing approach to the metabolic modeling Yen, J.; Lee, B.;

Liao, J.C. Fuzzy Information Processing Society, 1996. NAFIPS., 1996 Biennial Conference of the North American , 1996 Page(s): 343 -347.

105. Simple genetic algorithm parameter selection for protein structure prediction Gates, G.H., Jr.; Merkle, L.D.; Lamont, G.B.; Pachter, R. Evolutionary Computation, 1995., IEEE International Conference on , Volume: 2 , 1995 Page(s): 620 -624 vol.2.

106. Parameterizing genetic algorithms for protein folding simulation Schulze-Kremer, S.; Tiedemann, U. Molecular Bioinformatics, IEE Colloquium on , 1994 Page(s): 8/1 -8/7.

107. IEE Colloquium on 'Molecular Bioinformatics' (Digest No.1994/029) Molecular Bioinformatics, IEE Colloquium on , 1994.

108. IEE Colloquium on 'Applications of Genetic Algorithms' (Digest No.1994/067) Applications of Genetic Algorithms, IEE Colloquium on , 1994.

109. Genetic algorithms for protein tertiary structure prediction Schulze-Kremer, S. Applications of Genetic Algorithms, IEE Colloquium on , 1994 Page(s): 6/1 -6/5.

110. Development needs for diverse genetic algorithm design Kingdon, J.; Dekker, L. Applications of Genetic Algorithms, IEE Colloquium on , 1994 Page(s): 3/1 -311.

111. Recognizing patterns in protein sequences using iteration-performing calculations in genetic programming Koza, J.R. Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proc. of the First IEEE Conf. on , Page(s): 244 -249 vol.1.

112. Automated learning of a detector for the cores of /spl alpha/-helices in protein sequences via genetic programming Handley, S. Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proc. of the First IEEE Conference on , Page(s): 474 -479 vol.1.

113. Incremental prediction of the side-chain conformation of proteins by a genetic algorithm Iijima, H.; Naito, Y. Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on , 1994 Page(s): 362 -367 vol.1.

114. Using an annealing genetic algorithm to solve global energy minimization problem in molecular binding Leuo-Hong Wang; Cheng-Yan Kao; Ming Ouh-Young; Wen-Chin Cheu Tools with Artificial Intelligence, 1994. Proceedings., Sixth International Conference on , 1994 Page(s): 404 -410.

115. Automated discovery of detectors and iteration-performing calculations to recognize patterns in protein sequences using genetic programming Koza, J.R. Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on , 1994 Page(s): 684 -689.

116. Protein structure prediction using hybrid AI methods Guan, X.; Mural, R.J.; Uberbacher, E.C. Artificial Intelligence for Applications, 1994., Proceedings of the Tenth Conference on , 1994 Page(s): 471 -473.

117. The comparison of structures and sequences: alignment, searching and the detection of common folds Johnson, M.S.; Overington, J.P.; Edwards, Y.; May, A.C.W.; Rodionov, M.A. System Sciences, 1994. Vol.V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 296 -305.

118. Parameterizing genetic algorithms for protein folding simulation Schulze-Kremer, S.; Tiedemann, U. System Sciences, 1994. Vol.V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on , 1994 Page(s): 345 -354.

119. On the applicability of genetic algorithms to protein folding Unger, R.; Moult, J. System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on , 1993 Page(s): 715 -725 vol.1.

120. Stochastic motif extraction using a genetic algorithm with the MDL principle Konagaya, A.; Kondou, H.

System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on , 1993 Page(s): 746 -755 vol.1.

121. IEE Colloquium on 'Symbols Versus Neurons' (Digest No.123) Symbols Versus Neurons, IEE Colloquium on , 1990.

122. Asai, K., K., Ueno, Y. & Yada, T. Recognition of human genes by stochastic parsing. Pac Symp Biocomput, pp. 228-239, 1998

123. Uberbacher, E. C., Xu, Y. & Mural, R. J. Discovering and understanding genes in human DNA sequence using GRAIL, Methods Enzymol 266, pp. 259-281, 1996

124. Casidio, R., Compiani, M., Fariselli, P. & Vivarelli, F. Predicting free energy contributions to the computational stability of folded proteins from the residue sequence with radial basis function networks. Intelligent Systems for Molecular Biology 3, pp.81-88, 1995

125. Hanke, J. & Reich, J. G., Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. Comput Applic Biosci 6, pp447-454, 1996

126. Brunak, S., Engelbrecht, J. & Knudsen, S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J. Mol Biol 220, pp.49-65, 1991

127. Gonnet, G., Korostensky, C. and Benner, S.(1999) Evaluation measures of multiple sequence alignments. J. Comp. Biol..

128. S. Gupta, J. Kececioglu, and A. Schaffer. Making the shortest-paths approach to sum-of-pairs multiple sequence alignment more space efficient in practice. *Proc. 6th Symp. on Combinatorial Pattern Matching*, pages 128 – 43, 1995.

129. S. K. Gupta, J. Kececioglu, and A. A. Schaffer. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. In *J. Computational Biology*, 1996.

130. FARACH, M., and S. KANNAN. 1999. Efficient algorithms for inverting evolution. J. Assoc. Comput. Mach. 46:437–449.

131. NIKAIDO, M., A. P. ROONEY, and N. OKADA. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotomuses are the closest extant relatives of whales. Proc. Natl.

132. PAGEL, M. 1999. Inferring the historical patterns of biological evolution. Nature **401**:877–884.

133. SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: Long branch repulsion by likelihood in the Farris Zone. Cladistics **14**:209–220.

134. STEEL, M. 1999. Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences. Research Report NI 98025-BFG. Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

135. BRUNO, W. J., N. D. SOCCI, and A. L. HALPERN. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. **17**:189–197.

136. TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

137. BRUNO, W. J., and A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564–566.

138. FLYNN,M.J.,AND K.W.RUDD(1996),"Parallel Architectures, "ACM Computing Surveys,Vol.28,No.1(March),pp.67-70.

139. MARTIN,R.P.,A.M.VAHDAT,D.E.CULLER,AND T.E.ANDERSON(1997),"Effects of Communication Latency, Overhead, and Bandwidth in a Cluster Architecture, "Proc 24th Ann. Int. Symp. Comput. Arch.,ACM,pp.85-97.

140. QUINN,M.J.(1994),Parallel Computing

Theory and Practice, McGraw-Hill, New York.
141. FOSTER,I.(1995)Designing and Building Parallel Programs, Addison-Wesley, Reading Massachusetts.
142. WILSON,G.V.(1995),practical Parallel programming, MIT Press, Cambridge, Massachusetts.
143. BLELLOCH,G.E.(1996),"programming Parallel Algorithms," Comm. ACM, Vol.39, No.3,pp.85-97.
S. B. Needleman, C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology, vol. 48, pp. 443-453, 1970.

行政院國家科學委員會專題研究計畫成果報告
生物與資訊工程整合教育課程規劃及研究(3/3)
生物電腦

計畫編號：NSC 92-2516-S-216-003
執行期限：92 年 10 月 1 日至 93 年 9 月 30 日
主持人：中華大學機械工程學系教授鄭藏勝
共同主持人：中華大學資訊工程系助理教授侯玉松
中華大學資訊工程系副教授許文龍
E-mail: yshou@chu.edu.tw

一、中文摘要

在生物晶片與生物電腦這門課程中，教學目標在於介紹生物晶片的運作原理、資料分析及應用，與生物電腦的概念與未來發展。

所以在本子計畫中，我們將重點放在生物晶片構造、統計分析、群聚分析、調控網路建構、分子分類器、基因型晶片、相關分析軟體、奈米電腦的研究與探討，教學方式以專書研讀與論文研討等方式進行，以期增進同學在此領域的研究能力。

關鍵詞：生物晶片、生物電腦、生物晶片資料分析、奈米電腦

Abstract

In the course of biochips and biocomputers, teaching purposes were introducing the following terms: the principles of biochips, data analysis, applications, concepts of biocomputers and future developments.

In this subproject, the focus were structures of biochips, statistic analysis, clustering analysis, construction of regulatory networks, molecular classifiers, genotyping chips, analysis software, and nanocomputers. The teaching methods combined the study of books and papers to improve research power of students in the field.

Keywords: biochip, biocomputer, biochip data analysis, nanocomputer

二、緣由與目的

生物晶片是生物科技研究的一項利器，在過去，研究 DNA 或蛋白質特性，可能需要做多次實驗，才能獲得結論，相當耗費時間、人力與金錢。利用生物晶片，可以在短暫的時間內，同時完成多項實驗，所以在許多生物實驗室中，研究人員已採用生物晶片進行檢驗工作，並利用相關分析軟體，分析晶片資料，而完成許多學術論文。

在生物電腦方面，隨著奈米技術的逐漸發展，利用生物細胞內的物資，如 DNA，製作奈米設備(nanodevice)，如電晶體等基本硬體元件，已不再是夢想。在未來發展中，利用細胞製作輕薄短小又節源的奈米電腦，也將成為重要的研究課題。

為了使學生早先具備上述二大領域的研究能力，在生物與資訊工程整合教育課程研究計畫的第三年，我們特別安排這門課程，希望學生能趕上目前生物晶片的研究熱潮，及感受未來生物電腦的發展脈動，這是本子計畫的緣起與目的。

三、結果與討論

為了完整介紹生物晶片與生物電腦二領域，本課程實施重點分別為：

(1) 生物晶片概論：
介紹生物晶片的種類、運作原理及與

其他基因表現的實驗方法(如：SAGE)之優缺點比較。

(2) 基本晶片資料分析：

介紹晶片資料分析的統計方法，如 t 檢定與變異數分析等。

(3) PCA 分析：

PCA(principle component analysis)是常用的基因晶片資料分析工具，可將多維度資料投射到二維平面上，在此介紹其運作原理與使用方法。

(4) 群聚分析：

群聚分析(clustering analysis)可將基因表現行為相似的基因做分群，是建構基因調控網路的先期工作。

(5) 調控網路的逆向工程：

調控網路(regulatory network)的推測是基因晶片的一項重要應用。

(6) 分子分類器：

基因晶片另一項應用是疾病檢測，例如癌症預測，此處介紹如何使用生物晶片做疾病分類。

(7) 基因型晶片：

基因型(genotype)研究也是 DNA 突變研究的重要課題，此處介紹如何利用生物晶片，有效測量各種基因突變的類型。

(8) 相關分析軟體：

介紹統計軟體 package R 的使用方法。

(9) 奈米設備與細胞分子：

介紹奈米設備的電子學特性，與細胞分子的關係。

(10) 奈米生物電腦：

介紹利用奈米設備建構奈米生物電腦的構想。

(11) 論文研討：

與學生共同研討相關重要文獻。

本課程實施對象為碩士班二年級學生，課號：M02845A，學分數是 3 學分，一學期課程，修課人數 10 人。課程要求為期中考、期末考與論文報告，教學投影片如附件。

課程實施結果良好，大部分同學反應考題難易適中，對於生物晶片原理及分析軟體的操作，以及生物電腦概念，都能有所心得，對於日後研究生涯，已經奠定深厚的基礎。

事後檢討本課程，教學重點放在基因晶片，是優點也是缺點。好處是基因晶片是生物晶片中最常為人使用的工具，其精確度與分析方法較為健全，相對而言，蛋白質晶片技術較不成熟，實用性較差。

壞處是容易忽略其他生物資訊相關研究課題，求精就難求廣。

另外，受限於本校設備之不足，無法讓學生實際操作生物晶片的製作、顯象與分析，所以學生認為實務練習不夠，課程難與實務相結合。

在生物電腦方面，學生反映內容太難，牽涉到微電子學與奈米技術，所以不易完全了解。

解決之道應該需借重外界力量，例如邀請學界或業界在生物晶片及奈米技術有實務操作經驗者，到課堂演講；或是安排相關生物實驗室，進行生物晶片操作的實務練習，這樣會讓學生對課程較有具體的認識與體驗，對於日後的研發工作，會大有幫助。

四、計畫成果自評

本課程完成生物晶片與生物電腦的十個重要課題的介紹與探討，並針對相關論文，指導學生研讀，並在研討中互相交換觀點，課程內容充實，對於學生的未來研究，有很大幫助。

本課程教材與研討論文，都是採用最新的書籍與期刊論文，可以掌握最新資訊，是一項優點。但是相對地，準備課程必須耗費更多的時間，每個學年教材內容都會更新，無法完全沿用舊教材，備課較為辛苦，但是較能符合生物資訊科學的日新又新的情況。

本課程最大缺點是實驗設備不足，因受限於經費，無法構置生物晶片製片機器，以致同學無法實際接觸生物晶片實

驗，是相當可惜的，希望未來能添購相關設備，以增強學生的實務經驗與操作能力，對於未來研究或就業，都會很有幫助。

五、參考文獻

[1] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470; 1995.

[2] Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 19:442–447; 1995.

[3] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 270:484–487; 1995.

[4] Emili AQ, Cagney G. Large-scale functional analysis using peptide or protein arrays. Nat Biotechnol 18:393–397; 2000.

[5] Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4:844–847; 1998.

[6] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. Nat Genet 29:365–371; 2001.

[7] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258:818–821; 1992.

[8] Ried T, Liyanage M, du Manoir S, Heselmeyer K, Auer G, Macville M, Schröck E. Tumor cytogenetics revisited: comparative genomic hybridization and spectral karyotyping. J Mol Med 75:801–814; 1997.

[9] J.C. Ellenbogen et al, MITRE review articles, 1996, 1997, 1998, 1999 http://www.mitre.org/technology/nanotech/list_of_articles.html

[10] *Molecular Electronics*, edited by A. Aviram and M. Ratner, (The New York Academy of Science, 1998)

[11] *Molecular Electronics,* edited by J. Jortner and M. Ratner (Backwell Science, 1996)

[12] EE867 삼성반도체소자특강 *2001Sprin g* http://inca.kaist.ac.kr/home/home.htm

[13] Susmita Datta and Somnath Datta （2003）, Comparisons and validation of statistical clustering techniques for microarray gene expression data , BIOINFORMATICS ,Vol. 19, no. 4, pp. 459–466

[14] Markus Varsta （2002）, SELF-ORGANIZING MAPS IN SEQUENCE PROCESSING , Helsinki University of Technology Laboratory of Computational Engineering Publications.

[15] Juha Vesanto and Esa Alhoniemi （2000）, Clustering of the Self-Organizing Map , IEEE Transactions on Neural Networks, 11: pp.586-600.

[16] Susmita Datta and Somnath Datta , Comparisons and validation of statistical clustering techniques for microarray gene expression data , BIOINFORMATICS Vol. 19 no. 4 2003, pages 459–466

[17] Knudsen, S. A Biologist's Guide to Analysis of DNA Microarray Data, Wiley-Interscience, 2002.

[18] Kerr,M.K.,Martin,M.,and Churchill,G.A. Analysis of variance for gene expression microarray data. *J.Comput. Biol.* (2000). 7:819-37.

[19] Kerr,M.K.,and Churchill,G.A. Statistical design and the analysis of gene expression microarray data. *Genet Res.*(2001). 77:123-8. Review.

[20] Alter,O.,Brown,P.O.,and Botstein,D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* (2000) 97:10101-6.

[21] Wall, M. E., Dyck, P. A., and Brettin, T. S. SVDMAN－singular value decomposition analysis of microarray data. *Bioinformatics* (2001) 17:566-568

[22] Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray

experiments: Application to sporulation time series. *Pac. Symp. Biocomput.* (2000) 2000:455-66.

[23] Jarmer, H., Friis, C., Saxild, H. H., Berka, R., Brunak, S., and Knudsen, S. Inferring parsimonious regulatory network in *B. Subtilis. Pacific Symposium on Biocomputing.* (2002) 2002. Poster presentation.

[24] Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. Rich probabilistic models for gene expression. *Bioinformatics.* (2001) 17(Suppl 1):S243-S252.

[25] Tanay, A., and Shamir, R. Expansion on existing biological knowledge of the network: Computational expansion of genetic networks. *Bioinformatics.* (2001) 17(Suppl 1):S270-S278.

[26] Yeang, C. H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J., and Golub, T. Molecular classification of multiple tumor types. *Bioinformatics.* (2001) 17(Suppl 1):S316-S322

[27] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P.S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Genetics.* (2001) 7:673-679.

行政院國家科學委員會專題研究計畫成果報告
生物與資訊工程整合教育課程規劃及研究(3/3)
生物知識庫

計畫編號：NSC 92-2516-S-216-003
執行期限：92 年 10 月 1 日至 93 年 9 月 30 日
主持人：中華大學機械工程學系教授鄭藏勝
共同主持人：中華大學資訊工程系助理教授劉志俊
中華大學資訊工程系副教授周智勳
E-mail: ccliu@chu.edu.tw

一、中文摘要

在人類基因組計畫（Human Genome Project, 簡稱 HGP），完成人類基因組定序的草圖工作後，下一階段的主要研究工作在於瞭解這些基因組序列的意義。由於基因組的資料是如此的龐大，此項工作有賴於生物學家、生化學家與資訊研究人員的共同努力才能解決。由於人類基因的研究在醫學、農業、工業、製藥方面影響深遠且具有龐大的商機，故生物資訊被視為本世紀最重要的研究領域。

生物資訊是一項跨領域的研究學科。傳統的資訊研究人員對生物學的基本知識並不熟悉，必須補充分子遺傳學、分子生物學、生物化學與蛋白質結構等生物學專業知識，了解生物學專門語彙才能開展生物資訊的研究工作。在本計畫中，我們結合生物學、資料庫與資料探勘三個領域的主要內容，設計出生物知識庫課程，來奠定生物資訊在資料庫方面的研究基礎，並為國家培養專業的生物資訊人才，提昇我國生物資訊的整體競爭力。

關鍵詞：生物知識庫、核酸資料庫、蛋白質資料庫

二、緣由與目的

生物資訊是目前各先進國家積極投入的重點研究領域，而建立生物資料庫是從事生物資訊相關研究的首要基礎工作。而國外雖然現有許多蛋白質與基因資料庫可透過 WWW 來進行查詢，但要深入進行相關研究，必定要能完全直接掌握這些生物資料庫。所以，設計一支援生物資訊相關研究的生物資料庫十分重要。生物資料庫包含 DNA 資料庫與蛋白質資料庫兩大類，熟悉此兩類生物資料庫的操作是生物資訊相關研究的基本入門需求。

基因的主要資料型態是很長的字串(人類染色體每條有三千萬至三億對鹽基)，而蛋白質更有三維立體結構。所以基因資料庫與蛋白質資料庫和傳統的文數字為主的關連式資料庫相當不同。此外，各種基因與蛋白質分析工具皆需要與資料庫儲存與查詢系統互相配合，才能發揮各種生物資訊演算法的效能。

生物知識庫本課程介紹基本資訊探勘技術及其在生物資訊方面的應用。基本資訊探勘技術包含關聯規則分析、分類分析、叢集分析與序列分析。生物資訊探勘技術包含基因與蛋白質重複序列分析、基因表現與蛋白質摺疊結構關聯法則分析、基因表現與 DNA 序列關聯法則分析、基因與蛋白質重複序列與基因表現的關聯法則分析，訓練生物資料分析專業人才，奠定本土化生物資訊研究的基礎建設。

三、結果與討論

3.1 課程內容

本課程介紹基本資訊探勘技術及其在生物資訊方面的應用。基本資訊探勘技術包含關聯規則分析、分類分析、叢集分析與序列分析。生物資訊方面的應用包含重複序列探勘、蛋白質分類、蛋白質分類、基因表現資料探勘、蛋白質摺疊結構分類。

3.2 課程特色

- 強調資訊探勘技術在生物資訊方面的應用
- 建立基因表現資料庫
- 分析重複序列、基因表現、蛋白質摺疊結構間之關聯
- 強調實作: 修課學生必須實際完成至少一種資訊探勘技術

3.2 培養學生具體之能力

- 了解資訊探勘技術
- 了解基因表現與蛋白質摺疊結構
- 熟悉 Linux 作業系統與 PC Clustering 平台的操作
- MySQL 基因表現資料庫的建立
- 培養學生實際應用資訊探勘技術能力與經驗

3.3 課程之學術或應用之價值

A. 在學術方面: 資訊探勘與生物資訊皆是目前十分重要的研究主題。傳統的資訊探勘技術，如關聯規則分析、分類分析、叢集分析與序列分析是以關連式資料庫為分析對象, 而基因與蛋白質則是。

B. 在應用方面: 生物科技將是台灣未來的重要產業。而兼具生物與資料庫背景的人力十分稀少，本課程希望能培養生物資訊的專業人才，建立本土化的生物資料庫。

C. 在教材方面: 本科目的教材投影片可自由自 http://140.126.5.39/ 下載。

生物知識庫(研究所)：教材下載

檔案名稱	檔案大小	教材說明
Chap00.ppt	1176 KB	Chap 0 Preface 本章說明本課程的教材與課程大綱
Chap01.ppt	71 KB	Chap 1 Classification Analysis 本章介紹資料探勘中的分類分析的主要技術
Chap02.ppt	2,210 KB	Chap 2 Decision Tree 本章介紹決策樹的主要分析技術以及 ID3, C4.5 決策樹分析演算法
Chap03.ppt	3,029 KB	Chap 3 Data Mining Supports in SQL Server 本章介紹 SQL Server 資料探勘實際技術
Chap04.ppt	410 KB	Chap 4 Weka 本章介紹 Weka 資料探勘系統的安裝與使用
Chap05.ppt	471 KB	Chap 5 Weka 與資料庫 本章介紹 Weka 與生物資料庫連接技術
Chap06.ppt	631 KB	Chap 6 HMM 本章介紹 HMM 的基本觀念
Chap07.ppt	1,632 KB	Chap 7 HMM Applications 本章介紹 HMM 在蛋白質分類、蛋白質二級結構預測、基因預測、多重序列比對、非編碼 RNA 預測的應用
Chap08.ppt	505 KB	Chap 8 Protein Motif Databases 本章介紹主要的蛋白質分類資料庫與其分類方法
Chap09.ppt	181 KB	Chap 9 HMM Tools 本章介紹主要的 HMM 工具

以下檔案為研究所【生物知識庫】課程的投影片，為 PowerPoint XP 格式，歡迎下載。

四、計畫成果自評

91 學年度第 2 學期以及 92 學年度第 1

學期於中華大學資訊工程研究所開立“生物知識庫”課程，課程資料如下:

課程編號: M02824A (資訊工程研究所)
課程名稱: 生物知識庫
上課時間: 週三 234
選課人數: 11

本年度一共有 11 位研究生選修“生物知識庫”課程，學生對於有關資料探勘的基本技術如 HMM 與決策樹等知識在課程結束後都能建立良好的觀念。此外對於實作的要求亦使學生在生物知識庫實務方面的操作獲得實際的經驗。相信此門課程所培養的生物資料分析與資料探勘專門人才，在未來一定能夠為生物資訊的研究與開發貢獻心力。

行政院國家科學委員會專題研究計畫成果報告
生物與資訊工程整合教育課程規劃及研究(3/3)
生子序列及演化樹分析

計畫編號：NSC 92-2516-S-216-003
執行期限：92 年 10 月 1 日至 93 年 9 月 30 日

主持人：中華大學機械工程學系教授鄭藏勝
共同主持人：中華大學資訊工程系副教授吳哲賢
中華大學資訊工程系副教授俞征武
台灣大學畜產學系助理教授林恩仲
E-mail: jswu@chu.edu.tw

一、中文摘要

在生物資訊的研究領域中，我們時常討論的課題是，生物的基因或者組織是如何演化的。而研究這方面的生物學家們嘗試著去建構這些生存著的組織，牠們演化的歷史。這些演化歷史的分支可以被描繪成一棵樹型的結構，當這些子代的個數增加，建構出來可能的樹型會隨著快速的增加。

多重序列分析在生物資訊中同樣扮演著重要的角色，由於龐大的基因組計畫，在公開的資料庫裡有著相當多的序列資料，藉著這些資料我們可以搜尋相似的序列，因此就需要一個有效率的 MSA 程式來協助搜尋。

在本計畫中，我們採用 Fundamentals of Molecular Evolution 與 Molecular Evolution and Phylogenetics 兩本書做為教科書，側重在 MSA，NJ，MP 與 ML 的探討，介紹網路工具的運用技巧，並帶領同學運用這些工具處理問題，以達到理論與實務合一的成效。

關鍵詞： MSA、MP、ML、NJ

Abstract

How a group of genes or organisms evolved is a fundamental question in biology. Biologists who study such evolution try to reconstruct the evolutionary history of all living organisms. The divergence over evolutionary history can be described with a tree-like structure termed a phylogeny. As the number of descendants increases, the number of possible tree topologies increases very quickly.

Multiple sequence alignments (MSA) play a crucial role in molecular biology. Large genome projects result in an explosion of sequence data in public databases. For many genes a database search will reveal a whole number of homologous sequences. One wishes to learn about the evolution and the sequence conservation in such a group requires efficient MSA programs.

In this project, we use two books that Fundamentals of Molecular Evolution and Molecular Evolution and Phylogenetics be textbook, the following important topics were teach: MSA , Neighbor-join(NJ), Maximum parsimony (MP) and Maximum likelihood methods (ML), the using techniques for web tools.

Keywords: MSA、MP、ML、NJ

二、緣由與目的

在生物資訊的研究領域中，當要研究物種演化的歷史，我們會需要建構演化樹來協助我們[1-19]。例如：計算序列間的距離矩陣(Distance matrix on Sequence)、比較多條序列演算法(Multiple Sequence

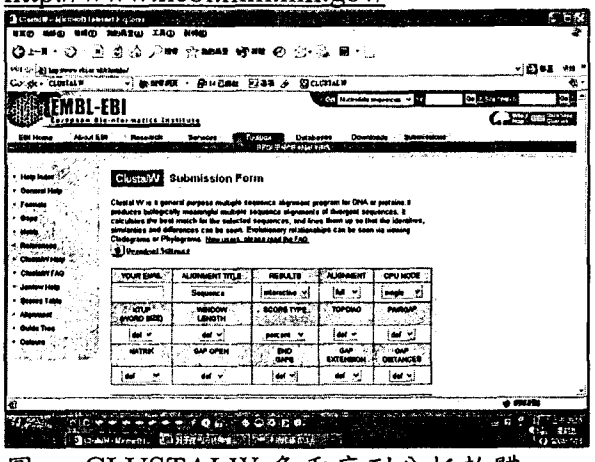
Alignment)、距離進行鄰接法(Neighbor-join)、最大簡約法(Maximum parsimony)...等問題。

這些問題大都有相對應的網路工具可以解決，例如：CLUSTALW、Phylip。

本課程的主旨，即在教導同學建構演化樹以及分子序列分析的演算法及相關工具，以解決建構演化樹及分析序列的相關問題。

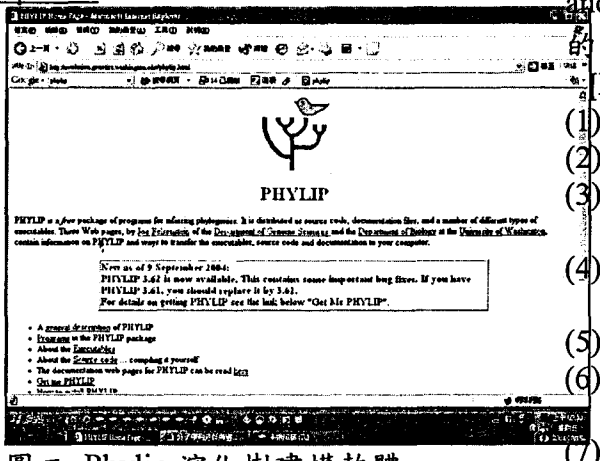
本課程所使用到的工具大部分都可以在底下的網頁找到

http://www.ncbi.nlm.nih.gov/



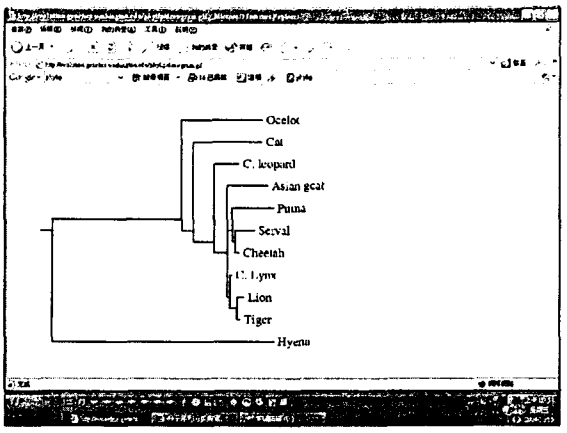
圖一 CLUSTALW 多重序列分析軟體

http://evolution.genetics.washington.edu/phylip.html

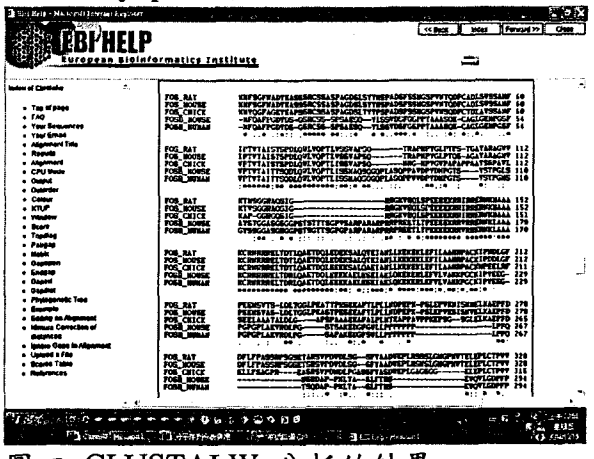


圖二 Phylip 演化樹建構軟體

底下是演化樹建構與多重序列分析的幾個圖示



圖三 Phylip 建構出的演化樹



圖四 CLUSTALW 分析的結果

在教學內容上，採取 Fundamentals of Molecular Evolution 與 Molecular Evolution and Phylogenetics 兩本教科書，主要有底下的章節：

[Fundamentals of Molecular Evolution]

(1) Genes, Genetic Codes, and Mutation
(2) Dynamics of Genes in Populations
(3) Evolutionary Change in Nucleotide Sequences
(4) Rates and Patterns of Nucleotide Substitution
(5) Molecular Phylogenetics
(6) Gene Duplication, Exon Shuffling, and Concerted Evolution
(7) Evolution by Transposition
(8) Genome Evolution

[Molecular Evolution and Phylogenetics]

(1) Molecular Basis of Evolution
(2) Evolutionary Change of Amino Acid Sequences
(3) Evolutionary Change of DNA sequences
(4) Synonymous and Nonsynonymous Nucleotide Substitutions
(5) Phylogenetic Trees
(6) Phylogenetic inference: Maximum Parsimony Methods

(7) Phylogenetic inference: Maximum Parsimony Methods
(8) Phylogenetic inference: Maximum Likelihood Methods
(9) Accuracies and Statistical Tests of Phylogenetic Trees
(10)Molecular Clocks and Linearized Trees
(11)Ancestral Nucleotide and Amino Acid Sequences
(12)Genetic Polymorphism and Evolution
(13)Population Trees from Genetic Markers
(14)Perspectives

在教材製作方面，課堂上採用投影片報告討論(附件)。投影片中都有詳細的課程報告。

三、結果與討論

本課程為一學期 3 學分課程，實施對象為資訊工程系碩一學生，共實作兩個程式和一個報告。

實施結果良好，大部分同學反應對於建構演化樹工具的使用方法及分子序列的分析程式實作等方面，都有所心得。

事後檢討本課程，在分子序列分析方面，我們可以去了解其分析理論基礎，來增加分析結果之正確性及可信度。但是由於計算量非常龐大，所以需要介由高效能的 PC 來計算才可以有較好的結果以及等待計算的時間也才能縮短；演化樹建構方面，我們了解可以藉由不同建構演化樹的方法，建構出生物學家再在研究物種歷史上所需不同的演化樹；網路工具應用方面，通常一個問題都需要多個工具一並使用才可以解決，不過如果運用得宜的話，將可以再短時間內把問題解決。在過去，沒有這些網路上的工具，那麼解決基因序列比對的問題，將會花上許多的時間，來撰寫程式，而且找出來的結果也不會有統一的格式。

四、計畫成果自評

在這 36 週的課程中，本人與學生探討了建構演化樹 14 個章節，與分子序列分析 8 個章節，並且讓學生輪流練習上台報告，在其中學習新知，一起研究討論問題，內容尚算豐富。

因為兩部分的課程內容豐富，如要兼顧實作部份，在一學期 18 週的時間中，僅能教完理論部分與工具運用，難以讓學生對工具不足之處作加以改善的動作。

所以後續的建構演化樹及分子序列分析課程，本人認為應當與生物資訊專題的課程內容相互配合，讓學生能運用目前現有的工具並且可以進而發現目前工具的缺點並且加以改善之。

若經費支援充足，如能購置多部更高效能的 PC，對於學生運用基因序列工具和撰寫這方面的程式將有更大的幫助。

最後，課程的終極未來發展，希望學生可以寫出一套介面友善，整合多功能，並且可以得到不錯的結果的工具。這是長遠的課程發展方向，需要一點一滴的努力，經驗累積，才能達成。

五、參考文獻

[1] OXFORD, "Molecular Evolution and Phylogenetics", Published by Oxford University Press, Inc., 2000.
[2] Dan Graur and Wen-Hsiung Li, "Fundamentals of Molecular Evolution", Sinauer Associates, Inc., Publishers Sunderland, Massachusetts, 2000.
[3] Adachi, J. and Hasegawa, M. "MOLPHY: programs for molecular phylogenetics", Version 2.2. Institute of Statistical Mathematics, Tokyo, 1994.
[4] Agarwala, R., Applegate, D. L., Maglott, D., Schuler, G. D., Schäffer, A. A., "A Fast and Scalable Radiation Hybrid Map Construction and Integration Strategy", *Genome Research*, vol. 10,pp. 350-364, 2000.
[5] Bäck, T., "Evolutionary algorithm in theory and practice", Oxford University Press, New York, USA, 1996.
[6] Bäck, T., Hammel, U., and Schwefel, H. P. (1997) Evolutionary computation: Comments on the history and current state. IEEE Transaction on Evolutionary Computation, vol. 1, pp. 3-17.
[7] Barash, Y. and Friedman, R. (2002) Context-Specific Bayesian Clustering for

Gene Expression Data. Journal of Computational Biology, vol. 9, pp. 169-191.

[8] Berger, M. P. and Munson, P. J. (1991) A novel randomized iterative strategy for aligning multiple protein sequences. Computer Applications in the Biosciences (CABIOS), vol. 7, pp. 479-84.

[9] Beyer, H. G. (1995) Toward a theory of evolution strategies: On the benefit of sex- the ( ) , / (λ μ μ theory. Evolutionary Computation, vol. 3, pp. 81-111.

[10] Biedl, T., Brejova, B., Demaine, E. D., Hamel, A. M., and Vinar, T., "Optimal Arrangement of leaves in the tree representing hierarchical clustering of gene expression data", Technical report CS-2001-14, University of Waterloo, April 2001.

[11] Blanchette, M., Bourque, G., and Sankoff, D. (1997) Breakpoint phylogenies. Genome Informatics, vol. 8, pp. 25-34.

[12] W. I.Chang and E. L. Lawler, "Approximate string matching in sublinear expected time" In Proceedings of the 31st Annual IEEE Symposium on Foundations Computer Science, 116-124. IEEE,1990.

[13] K. M. Chao, W. R. Pearson and W. Miller, "Aligning two sequences within a specified diagonal band", Computer Applications in the Biosciences 7:347-352, 1992.

[14] M. O. Dayhoff, R. M. Schwartz and B. C. Orcutt, "A model of evolutionary change in proteins", In M. O. Dayhoff, ed., Atlas of Protein Sequence and Structure, volume 5, supplement 3. National Biomedical Research Foundation, Washington D.C. pp. 345-352, 1978.

[15] A. Dembo, and S. Karlin, "Strong limit theorems of empirical functional for large exceedances of partial sums of i.id. variables", Annals of Probability 19:1737-1755, 1991.

[16] W. Feller, "An Introduction to Probability Theory and its Applications", Vol II. John Wiley and Sons, 1971.

[17] A. P. Gultyaev, "The Computer simulation of RNA folding involing pseudoknot formation", Nucleic Acids Research 19:2489-2494, 1991.

[18] F. Jacob, "Evolution and tinkering", Science 196:1161-1166, 1977.

[19] W. H. Jefferys, and J. O. Berger, "Ockham's razor and Bayesian analysis", American Scientist 80:64-72, 1992.

行政院國家科學委員會專題研究計畫成果報告

生物與資訊工程整合教育課程規劃及研究(3/3)

蛋白質的結構與功能之分析

計畫編號：NSC 92-2516-S-216-003
執行期限：92 年 10 月 1 日至 93 年 9 月 30 日

主持人：中華大學機械工程學系教授鄭藏勝
共同主持人：中華大學生物資訊系助理教授張慧玫
中華大學資訊工程系副教授林道通
台灣動物科技研究所副研究員黃三元
E-mail: yshou@chu.edu.tw

1.1.1 Introduction:

已過的 20 年，生物界經歷了兩大變革，徹底地翻轉生物學的思維和研究方法。

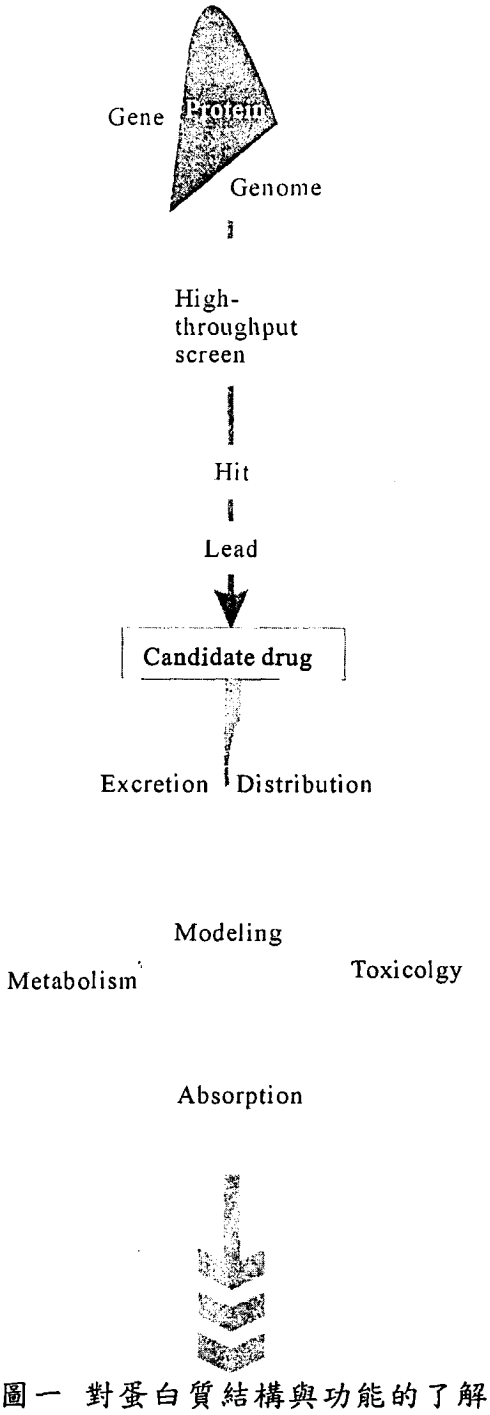
第一大變革，就是結構生物學，利用 3D 巨分子的電腦模型視訊化，提供了過去做不到的實驗結果的預測、實驗設計、和高複雜度資料的整合與解釋。尤其是填補了過去對蛋白質酵素動力學中間產物了解上的空白，這方面的貢獻已具體的推動著目前結構導向的藥物設計(structure-aided drug design)的發展(見圖一)。

第二大變革，就是基因體生物學，各物種基因體序列的公開，加上高產量檢測(high-throughput screen)技術的提升，提供了我們對細胞中各基因整體運作的了解，對細胞從產生至凋亡的生命史、細胞刺激後的訊號調節反應、個體中細胞間的互動等方面，擴大了解答生物學蛋白質功能問題的在時間與空間上的深度。

本課程以大部分的重點在強調變革中所帶來的結果，就是這些蛋白質結構與功能的生物意義；而以較少部分在強調變革的實驗方法。課程中的介紹包括基因序列、到胺基酸序列、到蛋白質各級結構與功能、到蛋白質生化和細胞功能、到蛋白質調節上的彈性。並加上相關網站與資料庫，與課本中的實例，相信對資訊背景的研究所學生，可提供扎實而具體、且多樣性的蛋白質資訊與所需的生物相關性之訓練。

課程目標

本課程的設計，提供資訊背景的研究所學生，可利用課程中所學資料，幫助他們以電腦在解生物方面訊號處理、影像處理、資料結構處理、電腦模型建置、視訊化與模擬等問題時，作為最好的橋樑。目前較多的資訊工具偏重於解蛋白質結構分類與預測等問題 (如圖二)，我們深信如此的課程訓練，可儲備學生解蛋白質功能的新視野(如圖三)，對將來在研究功能基因體(functional genomics)、蛋白體(proteomics) 、或系統生物學(systems biology) 的實作上會有很大的貢獻。



圖一 對蛋白質結構與功能的了解可推動結構導向藥物設計的發展

為了避免學生高門檻的適應感到困擾，本課程首先以簡易生物化學開始，課程如此設計的理由是對不論是否修過生物化學的學生，都能有起點的前置準備。課程主體在於對蛋白質已知結構與功能的生物資料，並達到學生可以在修過本課程後，能獨立看懂蛋白質結構與功能的相關生物文獻，如此對學生才有長遠的幫助與影響。本課程特色乃是：第一，起點的前置準備，適合學生銜接課本內容章節的安排。第二， 此課程章節設計依目前蛋白質結構與功能的相關研究以知的詳盡程度不同為出場序。 第三，課程中所選課本為章章相扣，息息相關，方便進入更深的生物內涵。

課程大綱與教學時數分配:

	Topic	Text book	Time
1	Fundamentals of protein structure	C	3h
2	Protein tertiary and quaternary structure	B;C	3
3	Protein stability and flexibility	A	6
4	Protein function recognition	A	6
5	Catalysis function	A	3
6	Ligand regulation	A	3
7	GTPase regulation	A	3
8	Degradation, phosphorylation, proteolysis, splicing and other PTM regulation	A	3
9	Olfactory proteins	A;B; C	0.5
10	Experimental tools for probing protein function	A	1.5
11	Sequence alignment, homology modeling, profile-based threading and rosetta	A;B	3
12	Identification of binding sites and catalytic residues	A	3
13	One structure with diverse functions and diverse structures with one function, protein with more than one function	A	1.5
14	Prion, amyloid and serpins	A	0.5
15	Structure determination	A	1.5
16	Proteins as drug targets	A;B	0.5
17	Final exam	-	

成績評定:

平時表現 10 % ＋專題報告 20% ＋作業 30% ＋期末考 40%。

課程內容:

課程內容包括講義、Powerpoint 簡報、相關網站輔助參考資料。教授方式以黑板說明、Powerpoint 簡報、學生 15 分鐘上台報告文獻閱讀心得。課程教授內容如附件。

相關網站:

1. For protein prediction and classification:
   http://scop.berkeley.edu/
   http://bioinformatics.ljcrf.edu/pdb_blast/PB_help.html
   http://www.uark.edu/chemistry/facultystaff/faculty/sakon/homepage/expasy.html
   http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html
   http://www.sdsc.edu/pb/edu/pharm207/8/8.html#goals
   http://www.protonet.cs.huji.ac.il/
   http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
2. For protein docking tools:
   http://zlab.bu.edu/zdock/benchmark.shtml
   http://www.scripps.edu/pub/olson-web/doc/autodock/
   http://www.techfak.uni-bielefeld.de/~sneumann/agaiprot/
   http://dock.compbio.ucsf.edu/
   http://www.sdsc.edu/CCMS/Papers/DOT_sc95.html
   http://www.bmm.icnet.uk/docking/
   http://www.csd.abdn.ac.uk/~dritchie/hex/
   http://www.bioinfo.de/isb/gcb99/poster/zimmermann/
   http://abagyan.scripps.edu/lab/web/man/frames.htm
   http://pc-gamba.math.tau.ac.il/
3. For protein structure alignment:
   http://zlab.bu.edu/zlab/protein.shtml
   http://lore.came.sbg.ac.at:8080/CAME/CAME_EXTERN/PROSUP/
   http://www.embl-heidelberg.de/predictprotein/predictprotein.html
   http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
   http://hmmer.wustl.edu/
   http://www.cse.ucsc.edu/research/compbio/sam.html
   BMERC PSA Server: http://bmerc-www.bu.edu/psa/
   STRIDE and PREDATOR: http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/argos/stride/stride_info.html
   DSSP: http://www.cmbi.kun.nl/swift/dssp/
   FSSP: http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html
   HSSP: http://www.cmbi.kun.nl/swift/hssp/
   PDBselect: http://www.cmbi.kun.nl/swift/pdbsel/
   PDBFinder: http://www.cmbi.kun.nl/swift/pdbfinder/
   PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/
   PSI-pred: http://www.psipred.net
   DSC : http://www.aber.ac.uk/~phiwww/prof/
4. Kyoto Encyclopedia of Genes and Genomes:
   http://www.genome.ad.jp/kegg/
   http://genome.cse.ucsc.edu/
5. For proteins in cell biology:
   http://www.biocarta.com/
   http://www.cbs.dtu.dk/services/SignalP/
   http://www.cbs.dtu.dk/services/TargetP/
6. For Signal Transduction Knowledge Environment:
   http://stke.sciencemag.org/
   http://www.grt.kyushu-u.ac.jp/spad/
   http://geo.nihs.go.jp/csndb/
   http://www.indstate.edu/thcme/mwking/oxidative-phosphorylation.html
7. For biomolecular interaction network database:
   http://www.blueprint.org/bind/bind.php
   http://dip.doe-mbi.ucla.edu/
8. For metabolic pathway / genome databases:
   http://biocyc.org/
   http://www.ncgr.org/pathdb/
9. For protein function prediction:
   http://predictome.bu.edu/
   http://dunbrack.fccc.edu/SCWRL3.php

http://fold.doe-mbi.ucla.edu/
http://www.cbs.dtu.dk/services/ProtFun/
Module finding:
http://www.bork.embl-heidelberg.de/Modules/
http://jura.ebi.ac.uk:8765/holm/ddd2.cgi
10. For signaling pathway and gene regulation:
http://193.175.244.148/
http://www.amaze.ulb.ac.be/
11. For structural genomics:
http://cubic.bioc.columbia.edu/genomes/RES/2002_bioinformatics/
12. For bioinformatics tools:
http://www.cse.ucsc.edu/~karplus/compbio_pages.html/
http://www.public.iastate.edu/~pedro/research_tools.html
Alignment:
http://www.c2b2.columbia.edu/research/papers_topics.html
http://www.ch.embnet.org/software/TCoffee.html
http://www.drive5.com/muscle/
http://www.fccc.edu/research/labs/dunbrack/pisces/
Hydrophobicity:
http://bioinformatics.weizmann.ac.il/hydroph/
13. For particular protein family:
http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/
http://pfam.wustl.edu/
Channel proteins:
http://www.neuro.wustl.edu/neuromuscular/mother/chan.html
G-protein: http://www.cmbi.kun.nl/7tm/
14. For protein structure visualization:
http://www.umass.edu/microbio/rasmol/
http://www3.ebi.ac.uk/tops/
Protein Explorer:
www.umass.edu/microbio/chime/explorer
PDB Lite:
www.umass.edu/microbio/rasmol/pdblite.htm
DRuMS Standard Color Schemes for Macromolecules:
www.umass.edu/molvis/drums
NonCovalent Bond Finder:
www.umass.edu/microbio/chime/find-nc
b
Biochemistry in 3D:
www.worthpublishers.com/lehninger3d
www.umass.edu/molvis/freichsman
http://www.biochem.ucl.ac.uk/bsm/pdbsum/
15. Evaluation of protein structure prediction methods:
LiveBench
http://bioinfo.pl/LiveBench/
16. Databases:
http://www.rcsb.org/pdb/
ftp://ftp.ebi.ac.uk/pub/databases/
http://tw.expasy.org/sprot/
NMR: http://biotech.ebi.ac.uk:8400/

課程實施成效

壹、預期效益之達成：如圖 4，本課程內容約教授達預期的 80%，約達 A 項指定課本的 70%。修課學生人數為 8 人較預期多，上課出席率約達 70%。本課程的授課對象是非生物背景的學生，為達到學生可以在修過本課程後，能獨立看懂蛋白質結構與功能的相關生物文獻，所以課本是使用生物學原文書，學生必須相當用功，大部分的學生都還算努力。其中一位外系生(土木系環工組)，起初程度落差較大課程後段也漸漸跟得上。至於學生專題簡報，學生多能理解所選的論文主題，能有清楚的表達能力，不過學生的英文程度，如何檢選、消化、整理資料並合成自己的創見，並力圖能在自己專業上融會貫通而應用，尚需加油。

貳、非同步教學教材之編製：如圖 4，本課程教學教材之編製秉持，第一，為使本課程更適合學生之需求，本課程每堂課以三本課本為主，依材料合適程度編製簡單講義，學生再依課堂黑板的講解，加註在講義中，經過少許的抄寫，加強學生的印象和參與。學生發現經過講義才是自己的東西，在應付考試時，也很有幫助。由於考試是 open-book 的方式，平常閱讀的心

得或其他資料蒐集到的資訊都可彙整在講義中。除了幫助課堂教材內容同步學習，也適合未來發展非同步教學教材之參考。第二，本課程也以各相關網站作為教學的輔助資源 (如前項課程內容所示)。

參、教學實作環境之配合：如圖 4，本課程為配合學校學程課程間，相關系所實習資源規劃整合，本課程屬資訊工程系，與生物資訊系生化虛擬實驗室相關教材相容，學生可到生物資訊系虛擬實驗室，適合學生體驗教學內容實作環境。對非生物背景系所的學生的課程與實作有很好的聯結。另外學程規劃整合運用校內相關資源，本課程可改善並考慮配用數小時電腦教室，學生課堂直接上網操作蛋白質範例或適時做分析應用，如此將更有助於課程的推展。

肆、涵蓋面的平衡之調整：如圖 4，本課程在一學期時間，要囊括目前蛋白質結構與功能的各種發展，常常需要有取捨。本課程採取「質」重於「量」，涵蓋主題專而少但內容深具探索性。本課程在美國許多大學均為研究所本科生選修的課程設計，得到相當肯定的教學效果；然而在施行層面上，本課程就時間控制上，在起初基礎扎根，用掉部分時間(15%)。另外章節切割與上課時間的搭配上，課本在蛋白質功能調節方面介紹相當詳細，有部分的主題如資訊工具介紹較弱未能發揮盡致，涵蓋面的平衡尚有改進空間。

伍、與其他課程之相關性：隨著人類基因體計劃的完成，與數個真核生物基因體之讀序完成，"分子遺傳學"就變得更加重要，因為分子遺傳學主要是研究基因結構與遺傳功能之間的關係(與本課程蛋白質從第 1 主題到第 8 主題有功能變化與調控的相關性)，以期對未來人類疾病防治鋪路。然而後基因時代核心議題想探討乃是基因功能，而基因功能的執行者正是蛋白質，因此完整基因組的蛋白質結構與功能成為下一紀元努力的目標。蛋白質結構根基在基因序列，基因突變可造成蛋白質結構與功能也改變，如許多遺傳疾病的例子。而"分子細胞生物學" 主要是研究蛋白質功能執行的舞台，在動態的調節中(與本課程蛋白質從第 3 主題到第 16 主題有細胞空間的相關性)表現出多樣性的生命現象。上學年"巨分子基因體與蛋白體學" 主要是研究基因體統整的分析，當個別蛋白質研究速度不足應付所需時，統整的分析提供全面蛋白體相關性(與本課程蛋白質從第 1 主題到第 16 主題有分析規模擴大的相關性)。

陸、學生後續發展相關性：根據最近數個月的就業市場，顯示一般生技公司在抗體定製研發工程師、藥廠蛋白製藥研發工程師，均要求具備巨觀的蛋白體知識與資訊工具研發的能力；由於生物資訊關系所陸續成立，相信本課程對學生不論畢業後就業或繼續升學之後續發展方向，都具有相當正面影響。

柒、激發不同專業的系際研究風氣：如圖 4，本課程的內容，對非生物背景的學生，同時兼具在基礎扎根以及與未來進階課程銜接的任務。在本學期的課程實行中，我們有兩位電機系的教授與副教授全程參與課程旁聽，使本課程增加許多相當有深度的討論，減少單向教學可能有的盲點。如此跨領域的交流實在可貴，並激發跨領域的可能的嶄新研究思路，其中共同提出申請研究計劃案有，93 年國科會提升私校院研發能量計劃，總計劃名稱: 生物資訊於系統生理之研究及應用，子計劃名稱: 訊號傳遞蛋白體之建構與模擬。兩位電機系的教授在加入清大團隊 94 年基因體計劃，本課程也支援生物資源部分，促使電機系的教授子計劃申請案撰寫完成。如圖 4，本課

程有助於不同專業的系際研究風氣刺激，未來可預期的研究發展，將是不可限量的。

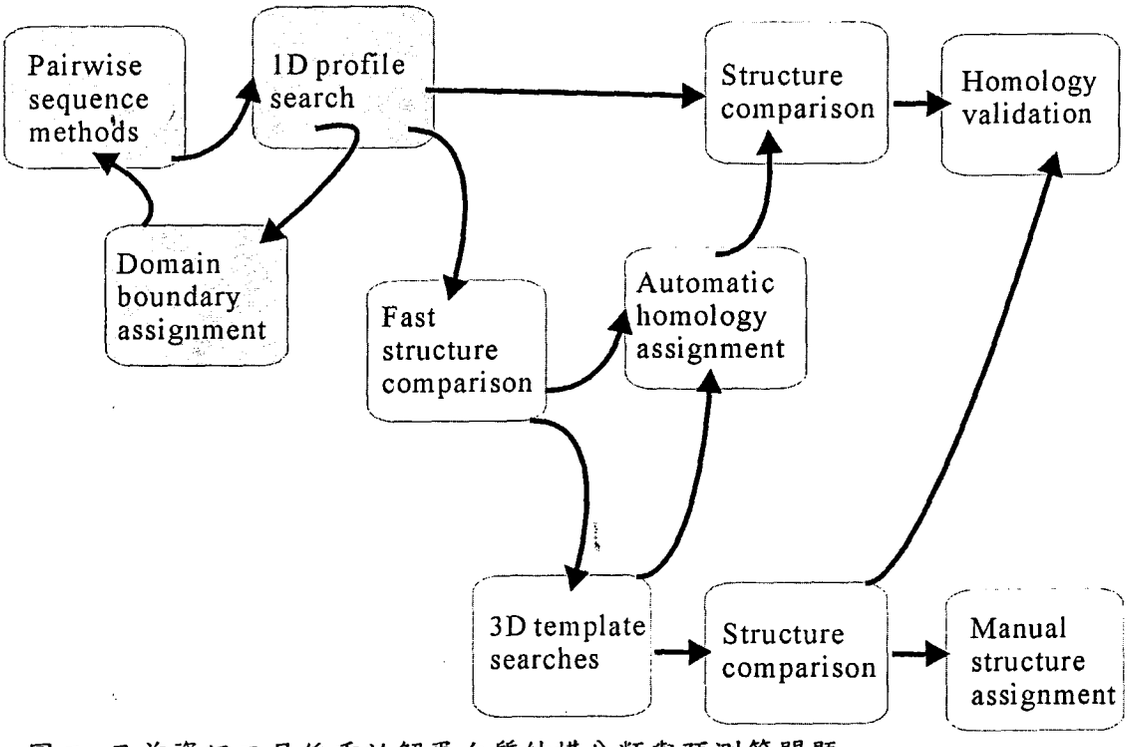
捌、計畫經費需求與學校行政與教學資源的配合：另外在課程管理上，利用本校計算中心的資訊服務，經由 e-mail 提醒未到課學生作業或考試的範圍及時間，以及講義的寄送等，增加教學互動，學生也可隨時查看自己學習成績的累積狀況。 如圖 4，本課程學校對本學程的支持程度相當高，使本課程得以順利得著各方面配合，特此感謝學校行政與教學資源的支援與國科會計畫經費的支援。

Reference:

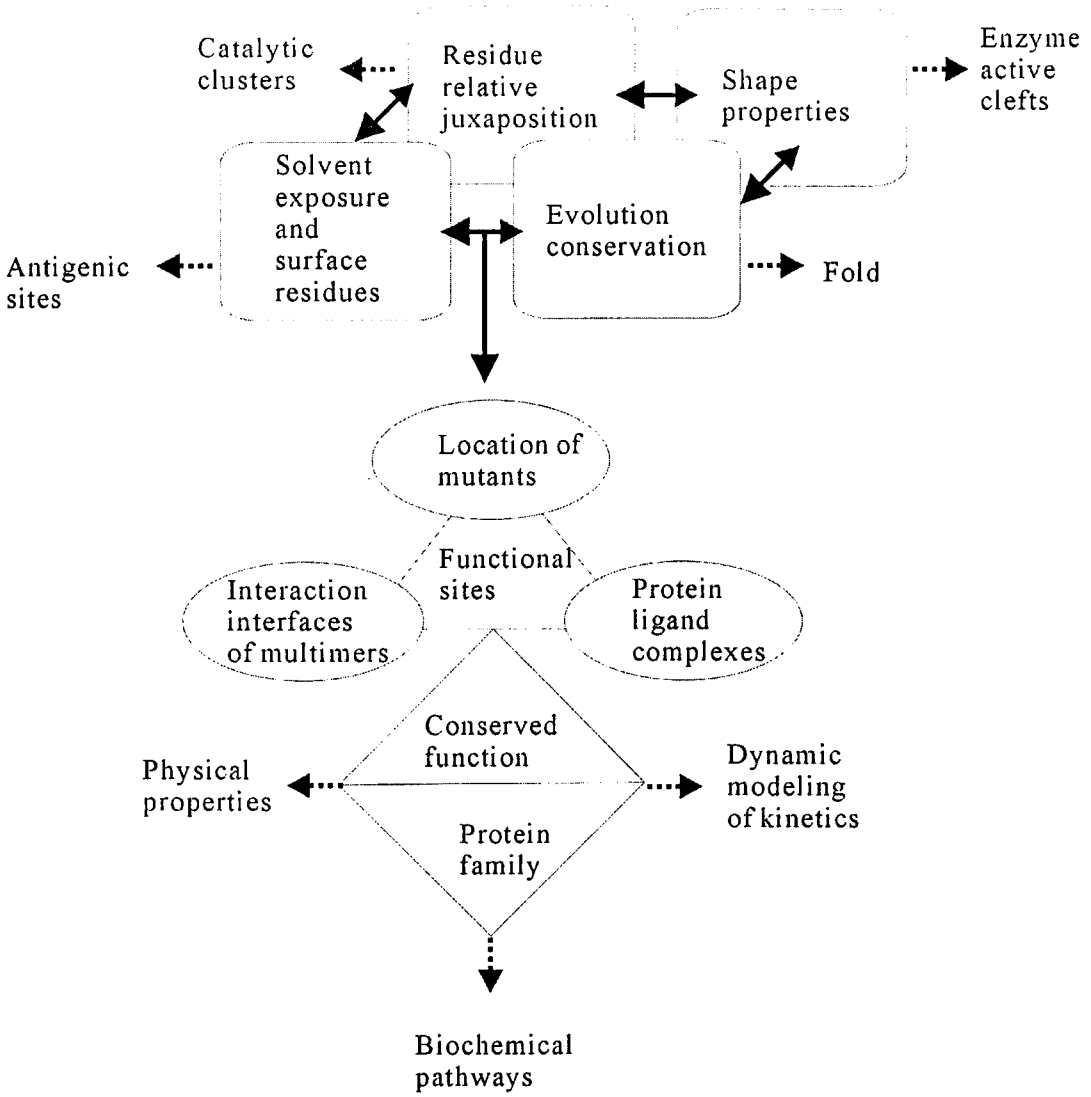
(A) "Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

(B) "Structure bioinformatics ", 2003. by Philip E. Bourne and Helge Weissig, by Wiley-Liss publishers, Inc.，藝軒書局.

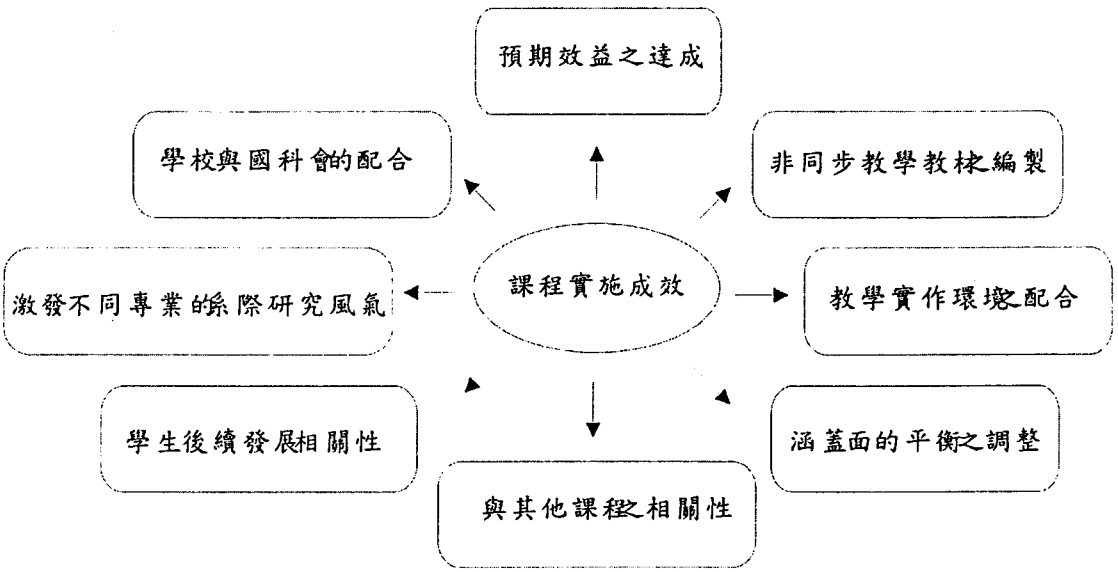
(C) "Biochemistry" 3rd Edition, 2001. Mary K. Campbell 原著， 林順富、陳師瑩、顏瑞鴻、蕭慧美編譯，偉明圖書出版.



圖二 目前資訊工具偏重於解蛋白質結構分類與預測等問題



圖三 蛋白質功能的新視野亟需資訊工具的新開發



圖四 值得肯定的"蛋白質結構與功能"課程實施成效

生物電腦教材內容

Section 1 Introduction

Textbook: A Biologist's Guide to Analysis of DNA Microarray Data, Steen Knudsen, Wiley-Liss, 2002

Hybridization(1)

- 2 DNA strands hybridize if they are complementary to each other.
- One or two strands can be replaced by RNA.
- Techniques using hybridization: Southern blotting, Northern blotting and DNA microarrays.

Hybridization(2)

- DNA arrays:
  - Oligonucleotide probes have been immobilized on a surface at μm distances.
  - The sample is labeled with a fluorescent dye that can be detected by a light scanner.
  - Tens of thousands of hybridizations are run at the same time. It offers a systemic view of how cells react in response to certain stimuli…

Hybridization(3)

- DNA arrays:
  - A probe matches a mRNA in cells. A mRNA is a result of gene expression, so it can apply to expression analysis. (⇨an expression profile)
  - Another application: to detect mutation in specific genes. (i.e. genotyping, see Section 11)
  - For expression analysis, 2 major technologies are available: Affymetrix chips and spotted arrays.

Affymetrix Genechip(1)

- Affymetrix uses equipments similar to be used for making silicon chips.
- Photolithographic process: Affymetrix uses masks to control synthesis of oligonucleotides on a chip. (Figure 1)
- Length of oligos <= 25.

Figure 1: photolithographic construction of microarrays

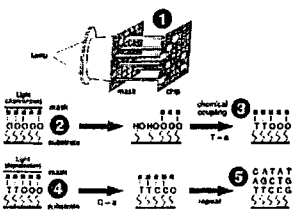


Figure 3.3
The photolithographic construction of microarrays. Synthesized high-density oligonucleotide microarray manufacturing with photolithography. Using selective masks, photolabile protecting groups are light activated for DNA synthesis (1, 4); photoprotected DNA bases are added and coupled to the intended coordinates (2). This cycle is repeated (3) with the appropriate masks to allow for controlled parallel synthesis of oligonucleotide chains in all coordinates on the array (5). (Derived from Lipshutz et al. [121].)

Affymetrix Genechip(2)

- Up to 40 oligos are used for detection of each gene.
- Affymatrix selected a region of each gene that has least similarity to others.
- From this region, 11-20 oligos are chosen as perfect matches (PM) and 11-20 oligos are mismatch oligos (MM), which are identical to PM except for central position 13. (Figure 1.2)

Figure 1.2 Affymetrix GeneChip

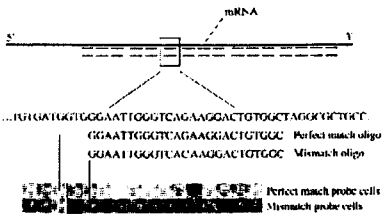
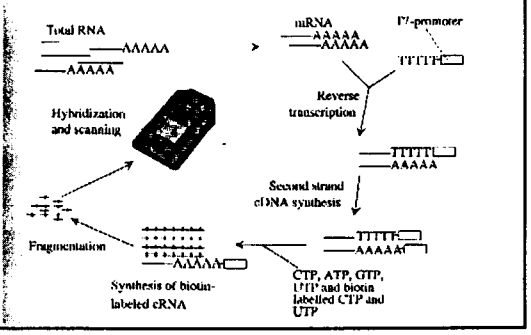


Fig. 1.2 The Affymetrix GeneChip technology. The presence of messenger RNA is detected by a series of probe pairs that differ in only one nucleotide. Hybridization of fluorescent messenger RNA to these probe pairs on the chip is detected by laser scanning of the chip surface. (Figure by Christoffer Bro.)

Affymetrix Genechip(3)

- PM and MM:
  - MM: detect nonspecific, background hybridization.
  - However, for weakly expressed mRNA, subtracting MM from PM add considerably to the noise in the data (Schadt, et al., 2000).
  - Average difference between probe pairs is calculated as signal intensity. (Section 3.1)
- Target mRNA is labeled with fluorochrome. The steps from cell to chip is shown in Fig.1.3.

Figure 1.3 Preparation of sample for GeneChip arays



Performance of GeneChip Tech.

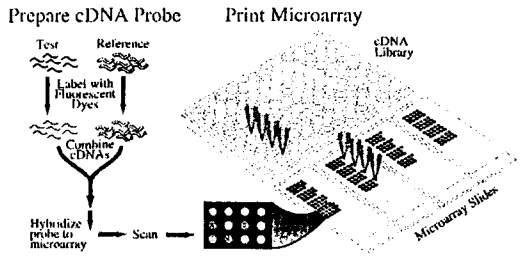
Table 1.1 Performance of the Affymetrix GeneChip technology. Numbers refer to chips in routine use and the current limit of the technology (Lipshutz, et al., 1999; Baugh, et al., 2001).

	Routine use	Current limit
Starting material	5 μg total RNA	2 ng total RNA
Detection specificity	$1:10^5$	$1:10^6$
Difference detection	twofold changes	10% changes
Discrimination of related genes	70-80% identity	93% identity
Dynamic range (linear detection)	3 orders of magn.	4 orders of magn.
Probe pairs per gene	20	4
Number of genes per array	12,000	40,000

Spotted Arrays(1)

- A robot spotter is used to move small quantities of probe from a microtiter (滴定) plate to the surface of a glass plate.
- The probe can consist of cDNA, PCR product, or oligonucleotides.
- Each probe is complementary to a unique gene.
- Probes are fixed to the surface. (Fig. 1.4)

Figure 1.4: Spotted Array Tech.



Spotted Arrays(2)

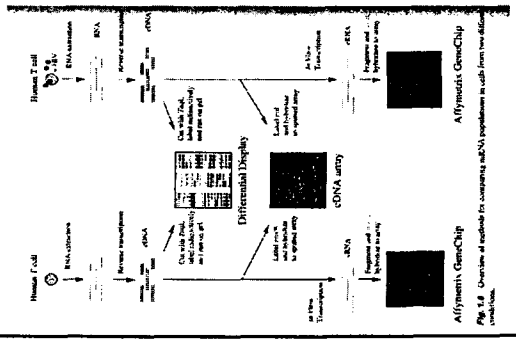
- Advantage for spotted arrays:
  - We can design any probe for spotting on the array
  - The sample and the control are hybridized to the same chip, whereas 2 Affymetrix chips are required to compare a sample and a control.
- Disadvantage:
  - Spotting will not be nearly as uniform as Affymetrix chips.
  - The cost of oligos becomes high.

Table 2.1

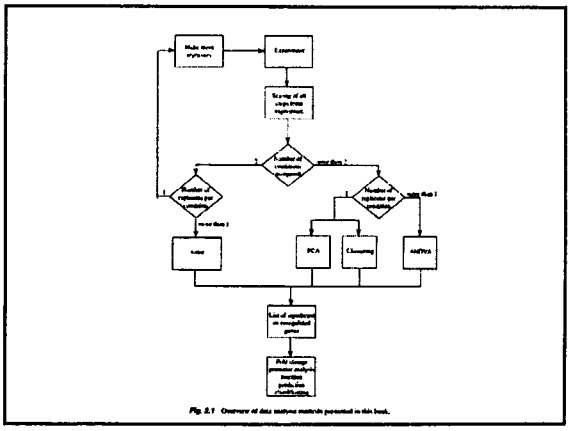
Table 1.2 Performance of the spotted array technology (Schena, 2000).

	Routine use
Starting material	10-20 μg total RNA
Dynamic range (linear detection)	3 orders of magnitude
Number of probes per gene	1
Number of genes or ESTs per array	$\approx$10,000

1.5 Example (Figure 1.8)



Section 2 Overview of Data Analysis



Section 3 Basic Data Analysis

Absolute Measurements(1)

- Image processing software:
  - Identifies the probe cells
  - Calculates signal intensities
  - Subtracts background
  - Measures noise level
- Intensity level:
  - cDNA array: int. no.s for red/green channel
  - Affymetrix: Average Difference is calculated

Absolute Measurements(2)

- Average difference and weighted average difference (Li and Wong, 2001a, b) :

$$AvgDiff = \frac{\sum_N PM - MM}{N}$$

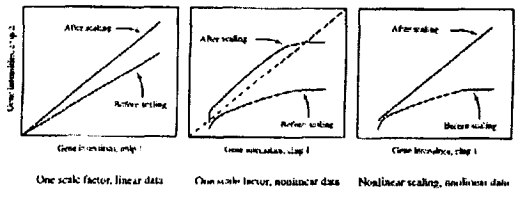
$$\tilde{\theta} = \frac{\sum_N (PM_n - MM_n)\phi_n}{N}$$

Scaling(1)

- Scaling: ensure the expression levels in sample are comparable to exp. levels in control.
  - Maintenance genes: expressed at a constant level. And then we check these genes are expressed at the same level in sample/control.
  - Assume the total amount of mRNA measured from each cell is constant. We then multiply all exp. level by a scaling factor until the sum of all exp. level in sample/control is identical.

Scaling(2)

- Nonlinear scaling:
  - Better method: scale weakly and highly exp. genes separately. (Schadt, et al. 2000, Li and Wong 2001b. Workman, et al. 2001)



One scale factor, linear data | One scale factor, nonlinear data | Nonlinear scaling, nonlinear data

Detection of Outliers

- Outliers in chip:
  - Entire chip is bad.
  - An individual gene on a chip that deviates from the same gene on other chips from the same sample.
- Detection of outliers:
  - The simplest model is equality among replicates. If one replicate deviates from mean, it is an outlier.
  - A replicate with a low significance t-test measure.
  - More advanced models (Li and Wong 2001a, b)

Fold Change

- Compare expression level in sample & control.
  - Sample/control : >1 ⇨ up-regulated
  - AffyFold: fold change is symmetric
  - Logfold: log (sample/control), symmetric and continuous.

$$AffyFold = \frac{Sample - Control}{min(Sample, Control)} + \begin{cases} 1 & \text{if Sample} > \text{Control} \\ -1 & \text{if Sample} < \text{Control} \end{cases}$$

Significance (t-test)

- Sample / control = 2. Is it an experimental error? Is it significance ?
- If you repeat control and sample, you can use a **t-test** to determine expression is significantly different between control and sample.
- T-test: 當兩常態母體之變異數未知但相等, 或樣本大小n1 = n2, 則可利用t-test檢測兩母體的期望值是否相等. (p.25 Ex, t-test.xls)

T-test

亦可爲t檢定之雙尾檢定，即假設的建立爲:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

檢定統計量t爲:

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

而$S_p^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$

式中n_1、n_2爲兩組獨立樣本的樣本大小，$\overline{Y}_1$、$\overline{Y}_2$爲兩組樣本的平均數，

Table 3.2 *t*-test on difference between patient categories *A* and *B*.

Gene	Patient A_1	A_2	B_1	B_2	*P*-value
a	190	210	290	310	0.019
b	390	410	590	610	0.005
c	110	90	120	80	1.000
d	400	90	600	200	0.606

190	290	
210	310	
200	300	
200	200	200
7.071068		
0.019419		

Significance (ANOVA)

- If you have more than two conditions, the method analysis of variance (ANOVA) will, using F distribution, calculate the probability that all come from the same distribution.
- ANOVA: 變異數分析, 是 F 統計量的重要應用, 用來檢定多個常態母體期望值是否全等.
- Example: p.26, Table 3.3

ANOVA

$$SST = SSR + SSE$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

$$- \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2 + \sum_{i=1}^{k}(n_i - 1)\hat{S}_i^2$$

- MSR = SSR / (k – 1)
- MSE = SSE / [k (n – 1)]
- F = MSR / MSE, 第一自由度=k-1, 第二自由度=k(n-1)
- k: 條件數, n: 每個條件的樣本數

Table 3.3 ANOVA on difference between patient categories N, A and B.

Gene	N_1	Patient N_2	A_1	A_2	B_1	B_2	P-value
a	90	110	190	210	290	310	0.0018
b	190	210	390	410	590	610	0.0002
c	90	110	110	90	120	80	1.0000
d	200	100	400	90	600	200	0.5560

200	400	600		
100	90	200		
150	245	400	265	
5000	48050	80000	133050	
63700				
31850				
44350				
0.718151				
0.556097				

Significance

- Nonparametric tests (無母數檢定):
  - Both t-test and ANOVA assume data follow normal distribution.
  - Wilcoxon/Mann-Whitney rank sum test will assess significance without assuming normality.
- P-value: 母體期望值相等的機率, 即基因表現未改變的機率.

Example 2: Number of Replicates

- Number of Replicates 即 t-test 的 n_i 或 ANOVA的 n 值, n 值越小, T 或 F 值越大, 檢定結論較傾向拒絕H_0, 母體期望值不相等, 即基因表現有顯著差異... 所以table 3.4與3.5, 當replicates數較小時, False negative值較大.
- 隨著replicates數增加, true positives遞增.
- How many replicates need? This depends on how large is variance btwn repl. and how small a fold change do you wish to detect.

Section 4 Visualization by Reduction of Dimensionality

Principal Component Analysis (PCA)

- If we measured 6000 genes in 15 patients, the data constitute a 15x6000 matrix.
- Imagine 6000 genes as points in a 15-dim hyperspace. A cloud of 6000 points in hyperspace. Exist one direction where the cloud will be extended. This is the axis of first principle component (PC).
- 2nd PC: is orthogonal to first PC and captured maximum variation in data.

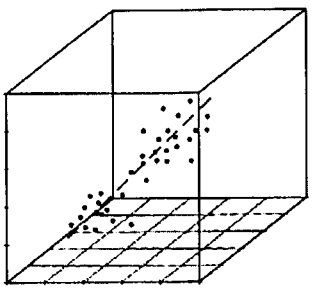
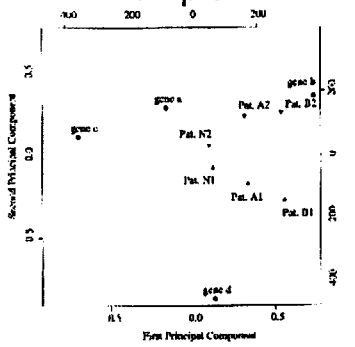
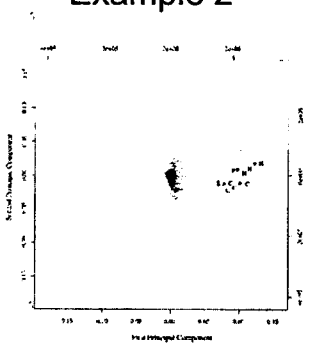


Fig. 4.1 A cloud of points in three-dimensional space. The cloud is not regular. It extends more in one direction than in all other directions. This direction is the first principal component (dashed line).

Example 1



Example 2



Section 5 Cluster Analysis

Hierarchical Clustering (H Clustering)

- Once you have more experiments under different conditions, it makes sense to group significantly changed genes into clusters that behave similarly under different conditions.
- H clustering method:
  - Each gene is a N-dimension vector.
  - Calculate the distance between two genes.
  - Group those genes together that are closest. (nearest neighbor or centroid)

Table 5.1 Expression readings of five genes in two patients.

Gene	Patient N_1	A_1
a	90	190
b	190	390
c	90	110
d	200	400
e	150	200

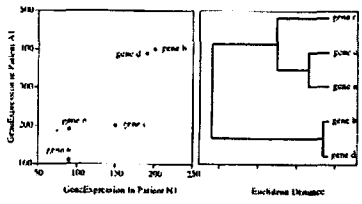


Fig. 5.1 Hierarchical clustering of genes based on their Euclidean distance.

K-Means Clustering

- H clustering only fails when you have a large number of genes (several thousand).
- K-means clustering is faster.
  - Do not calculate distances between all genes.
  - Number of clusters is decided.
  - Computer randomly assigns each gene to a cluster.
  - Calculate the distance between each gene and centroid.
  - If a gene is closer to center of another cluster then it is reassigned to the closer cluster.
  - Stop algo until centroids will no longer change.

A Example: K-means clustering

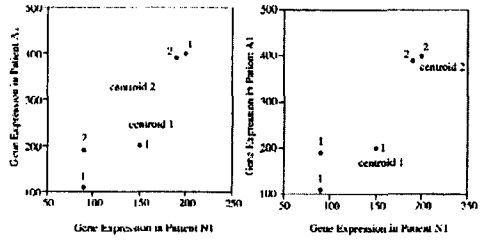
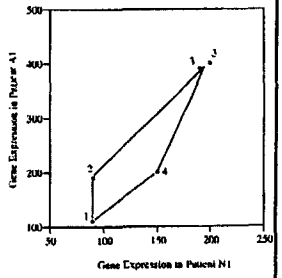


Fig. 5.3 K means clustering of genes based on their Euclidean distance. First, genes are randomly assigned to one of the two clusters in K: 1 or 2 (Left). The centroids of each cluster are calculated. Genes are then reassigned to another cluster if they are closer to the centroid of that cluster (Right). After just one iteration, the final solution is obtained (Right).

Self-Organizing Maps (SOM)

- SOM is similar to K-means, but instead of allowing centroids to move freely in space, they are constrained to a 2D grid.
- Fig. 5.4: centroids are at each corner of the grid.



Distance Measures

- Euclidean distance: sqrt((a1-b1)^2+... +(aN-bN)^2)
- Vector angle: Fig. 5.5. (角度越小, 向量越有正比關係)
- Pearson correlation coefficient.

$$\cos\theta = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}$$

$$\frac{\sum_{i=1}^{N}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i-\bar{b})^2}}$$

Example: Comparison of Distance Measures

Table 5.2 Expression readings of four genes in six patients

Gene	N_1	N_2	A_1	A_2	B_1	B_2
			Patient			
a	90	110	190	210	290	310
b	190	210	390	410	590	610
c	90	110	110	90	120	80
d	200	100	400	90	600	200

Table 5.3 Euclidean distance matrix between four genes

Gene	a	b	c	d
		Gene		
a	0.00	5.29	3.20	4.23
b	5.29	0.00	8.38	5.32
c	3.20	8.38	0.00	5.84
d	4.23	5.32	5.84	0.00

Table 5.4 Vector angle distance matrix between four genes.

Gene	a	b	c	d
		Gene		
a	0.00	0.02	0.42	0.52
b	0.02	0.00	0.41	0.50
c	0.42	0.41	0.00	0.51
d	0.52	0.50	0.51	0.00

Table 5.5 Pearson distance matrix between four genes.

Gene	a	b	c	d
		Gene		
a	0.00	0.06	1.45	1.03
b	0.06	0.00	1.43	0.98
c	1.45	1.43	0.00	0.83
d	1.03	0.98	0.83	0.00

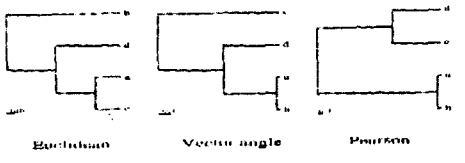


Fig. 5.6 Hierarchical clustering of distances (with three different distance measures) between genes in the example.

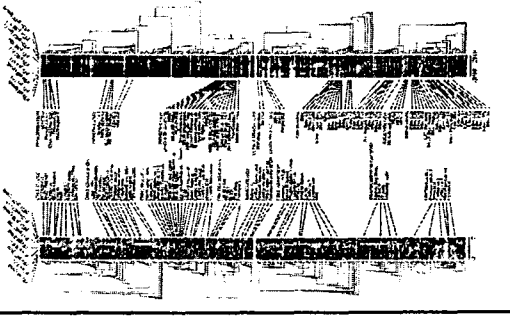
Comparison

- Vector angle distance is the best way to represent gene expression responses.
- Genes a, b and d all have increasing expression over the 3 patient categories. The vector angle clustering has captured this trend, grouping a and b close together and d nearby.

Normalization

- 3 common ways of normalizing expression of a gene:
  - Means: subtract the mean from all numbers.
  - Length: divide all numbers by that length.
  - S.D.: divide all numbers by standard deviation.
- Remember, you have already scaled data, so expression should be comparable.
- Author suggest: use vector angle distance on non-normalized expression for clustering.

Visualization of Clusters



Section 6 Beyond Cluster Analysis

Function Prediction

- Genes that appear in same cluster have similar transcription response to different conditions. It is likely caused by some commonality in function. You can infer function of orphan genes in same cluster.
- When no genes with known function, a number of properties can be used to predict a likely function class. (proteins with similar function also share some similarities in AA length, posttranslational modification, cellular destination signal…)

Discovery of Regulatory Elements in Promoter Region (1)

- If some genes share a regulatory response to some stimuli, we assume they share a binding site for a transcription factor in their promoter.
  - ClustArray allows you to select a cluster and search their upstream promoter region for regulatory elements. (promoter region are known)
  - Saco-patterns: searches for patterns are fully conserved (for example, AGCTTAGG). But transcription factor binding sites may degenerate.

Discovery of Regulatory Elements in Promoter Region (2)

  - Ann-spec: searches degenerate patterns, but it is sensitive to choice of parameters.
- Example 1: discovery of proteasomal element
  - Take 6269 annotated yeast ORFs and extract 200 bp starting 300 bp upstream of ORF to cover most promoter regions.
  - Divide 6269 promoter regions into proteasome (31) and those have not (6238). (positive /bground set)
  - Run saco-patterns, GGTGGCAAA present 25 of positive set and 26 of background (false positive).

Discovery of Regulatory Elements in Promoter Region (3)

- Example 2: Rediscovery of Mlu cell cycle box (MCB)
  - Sort yeast promoter regions by expression in a cell cycle experiment.
  - We look for patterns are more frequent in up-regulated genes than in nonregulated genes.
  - Fig. 6.1 shows the distribution of genes that contain a MCB pattern ACGCGT, was discovered to be significant.

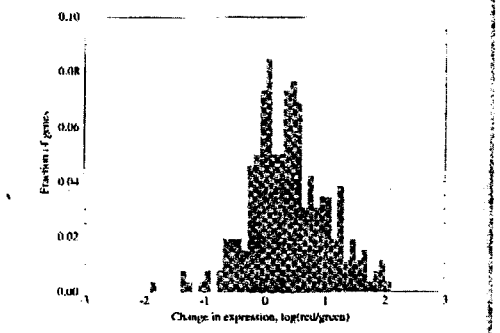


Fig. 6.1 Logfold distribution of all yeast genes (open bars) in a cell cycle experiment. Log fold distribution of genes containing Mlu cell cycle box (shaded bars) in the same experiment. Drawing by Lars Juhl Jensen based on data from Jensen and Knudsen (2000).

Integration of Data (1)

- Data analysis from an expression experiment becomes more powerful if information about function of genes and knowledge of promoter elements controlling expression are included.
- Such an integration of data can be done manually on a small scale.
- If you have a group of genes with significant different expression btwn 2 conditions, you can subject them to further analysis.

Integration of Data (2)

- 2 further analysis:
  - Look at functional annotation (Medline abstracts) of genes to discover clues of a pathway they may all be a part of.
  - Run a promoter analysis in Section 6.2. (problem: 從up-regulated genes的promoter regions的序列中, 找出共同特徵序列, 當作regulatory element)
- Such analysis takes a long time, and you only do it manually for a small number of genes.

Section 7 Reverse Engineering of Regulatory Networks

Introduction

- A gene can affect expression of another gene by binding of the gene product to promoter region of another gene.
- Between more than 2 genes, the regulatory network is referred as regulatory interactions.
- Reverse engineering problem can be solved in 2 ways: using time-series data & steady-state data.

The Time-Series Approach (1)

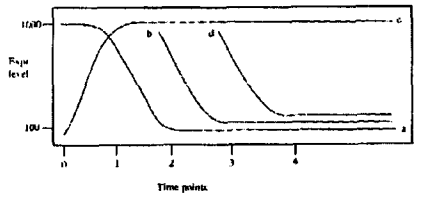
- Expression level of a certain gene at a certain time point can modeled as some function of expression levels of all other genes at all previous time point.
- Linear modeling approach:
  - First, remove genes don't show significant change.
  - Cluster genes that behave the same way.
  - Build a linear model of remaining gene clusters.

The Time-Series Approach (2)

$$x_j(t) = \sum_{i=1}^{N} r_{i,j} x_i(t-1)$$

- Xj(t): expression level of gene j at time t.
  rij: a weight factor showing how gene i affects gene j.
- Another solution: Holter et al. (2000) and Bayesian networks (Friedman, et al. 2000).

Example 4 (Section 7.7)



- Expression of each gene at each time as log10 of fold change relative to time 0. gene a is expressed 0, 0, -1, -1, -1 at time points 0, 1, 2, 3, 4.

Example 4: expression matrix and linear equation model

Table 7.2 Expression matrix for four genes.

Gene	Time 0	1	2	3	4
a	0	0	−1	1	1
b	0	0	0	−1	−1
c	0	1	1	1	1
d	0	0	0	0	1

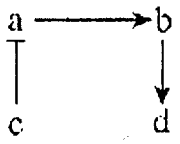
$$-1(t-3) = r_{b,a}0 + r_{c,a}1 + r_{d,a}0$$

$$-1(t-2) = r_{b,a}0 + r_{c,a}1 + r_{d,a}0$$

$$0(t-1) = r_{b,a}0 + r_{c,a}0 + r_{d,a}0$$

Example 4: interaction matrix and regulatory network

Gene	Gene: a	b	c	d
a		+		
b				+
c	−			
d				



The Steady-State Approach

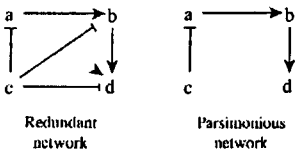
- The effect of deleting a gene on the expression of other genes is measured.
  - If exp of gene b increases after deletion of gene a, then gene a repressed the exp of gene b.
  - If exp of gene b decreases after deletion of gene a, then gene a enhanced the exp of gene b.
  - Reference: Ideker, et al. 2000 and Hughes, et al. 2000.

Limitation of Network Modeling

- Genetic network approached ignore regulatory interactions take place at protein-protein level. In future, regulatory network models must include protein-protein interaction maps...
- We need a way to combine prior biological knowledge of regulatory networks, information deduced from time-series / steady-state experiments. (1, 3 => 2)

Example 1: Steady-State Model

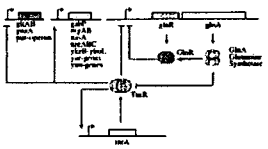
Gene	Gene: a	b	c	d
a		+		−
b				+
c				−
d				



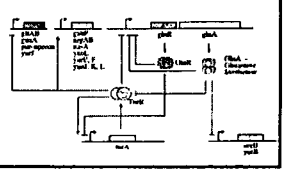
Redundant network => parsimonious network: delete those paths for each pairs of genes excluding the longest path.

Example 2: Steady-State Model on Real Data

- Figure 7.2: Known regulatory network in Bacillus subtilis. (Jarmer, et al., 2002)



- Figure 7.3: Regulatory network reverse engineered from real steady-state data.



Section 8 Molecular Classifiers

Introduction

- Problem: What if you have several cancer specimens from one subtype and several specimens from other subtypes?
- You need to look for genes that all specimens from one subtype have in common and are absent in all specimens from other subtype.
- You have just selected genes to fit your data. No general method for classification.

Basic Rules to Build a General Method

- Avoid overfitting data. Use fewer estimated parameters than the number of specimens.
- Validate your method by testing it on an independent data set.
- Cross validation: When data set is small, you subdivide data set into test and training several time. (tenfold cross-validation for 10 datum)

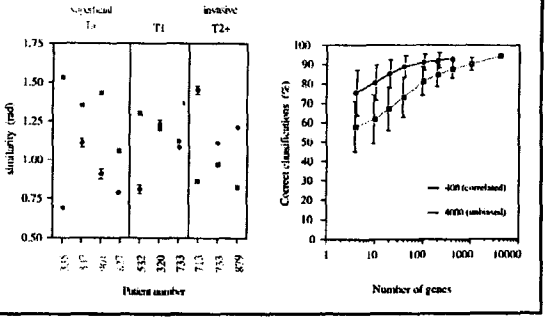
Classification Schemes

- Nearest Neighbor:
  - For each patient, find k nearest neighbors.
  - Predict most common class among the k neighbors by majority vote.
  - It works well if the classes are well separated in PCA or clustering.
- Neural Networks:
  - It is suitable to classify large data set (50 – 100).
- Support Vector Machine: machine learning.

Example 1: Classification of Cancer Subtypes

- 2 subtypes of bladder cancer: superficial and invasive.
- Only had biopsies from 10 patients.
- The Method: to measure the angle between the vector of all gene expression levels for each patient and the vectors of 2 reference pools of superficial and invasive cancer. (Figure 8.1)
- To test 4 new patients, they were correctly classified.

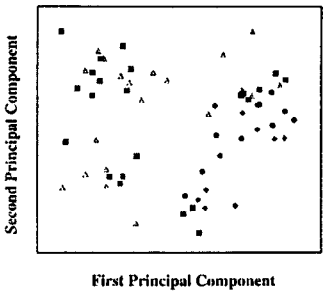
Figure 8.1



Example 2

- Khan, et al. (2001) have classified small, round blue cell tumors into 4 classes.
- A k nearest neighbor classifier.
  - Data set: 2000 genes, 63 tumor samples
  - K = 3, Euclidean distance, and leave-one-out cross-validation
  - Correct rate: 61/63
- A neural network classifier: pp. 78-79.

Figure 8.2



Section 9 Selection of Genes for Spotting on Arrays

Introduction

- You may have so much knowledge about the genes that you spot on a custom array.
- If you are not sure that all genes, you can do a homology search (Psi-Blast) / Medline search at NCBI website.
- Another way of selecting genes for a spotted array is to use a commercial Affymetrix array to identify genes about a particular problem.

Gene Finding (1)

- No matter what organism, a large fraction of genes is not functionally characterized.
- They have been predicted by the existence of a cDNA or EST clone with matching sequence, by a match to a homologous gene in another organism, or by gene finding in genomic seq.
- Gene finding uses software to predict structure of genes based on DNA sequence alone.

Gene Finding (2)

- Quality judgment of gene finding:
  - Expression analysis may be a good method for experimental verification of predicted gene.
  - Skovgaard, et al. 2001: the predicted number of 4300 genes for E. coli probably contains about 500 false positive predictions. All ORF's longer than 100 triplets were annotated as genes for the Archaea Aeropyrum pernix, half are probably false.

Gene Finding (3)

- Better performance is achieved when including codon usage statistics. These frequencies are to some degree specific to the organism.
- Even better performance is including models for specific signals like splice sites, promoters, and start codons. Such signals are best combined within HMM.

Selection of Regions with Genes

- How can you prevent spotting probes that are complementary to more than one gene?
- There is software available to help search for regions that have least similarity to other genes.
- ProbeWiz: takes a list of gene identifiers and uses Blast to find regions in those genes that are the least homologous to other genes. It uses a DB of the genome from the organism...

Section 10 Limitations of Expression Analysis

Introduction (1)

- Expression analysis measures only the transcriptome.
  - Important regulation takes place at translation and enzyme activity. Those regulations are ignored.
- The effect of alternative splicing is ignored.
  - To what extent are changes in observed signal from a particular mRNA due to alternative splicing rather than due to a changes in cross-hybridisation?
  - Changes in relative probe intensity within a gene might reveal alternative splicing.

Introduction (2)

- mRNA is an unstable molecule.
  - Messengers are programmed for enzymatic degradation and half-lives of messengers vary considerably.
  - Messengers with very short half-lives may be difficult to extract. Thus, it is impossible to detect with statistical significance.

Relative vs. Absolute RNA Quantification

- Absolute quantification is a much harder task.
  - We need to know how well each probe hybridizes to its target for each different mRNAs.

Section 11 Genotyping Chips

Introduction

- Genotyping chips are available.
  - They measure DNA.
  - p53 chip is available from Affymetrix for detecting mutations in DNA of human p53 tumor supressor.
  - It does so with overlapping oligos that each are complementary to 20 bps of TP53 gene (Figure 11.1). (SNP)
  - There are still limitations of accuracy of this determination. Develop a NN software improve it. (Figure 11.2)

Figure 11.1

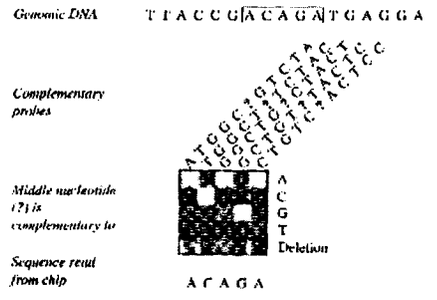
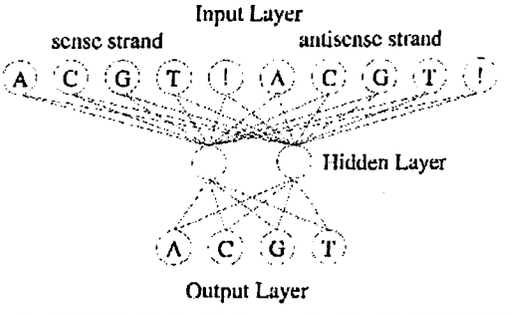


Figure 11.2



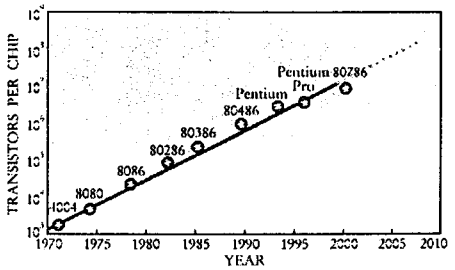
Nanodevice I

1. Solid state nanoelectonics
2. Molecular electronics

References

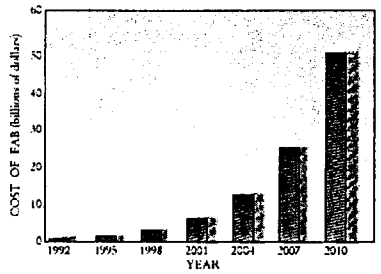
- J.C. Ellenbogen et al, *MITRE review articles*, 1996, 1997, 1998, 1999 http://www.mitre.org/technology/nanotech/list_of_articles.html
- *Molecular Electronics*, edited by A. Aviram and M. Ratner, (The New York Academy of Science, 1998)
- *Molecular Electronics*, edited by J. Jortner and M. Ratner (Backwell Science, 1996)
- EE867 삼성 반도체소자 특강 *2001 Spring* http://inca.kaist.ac.kr/home/home.htm

Moore's 1st Law



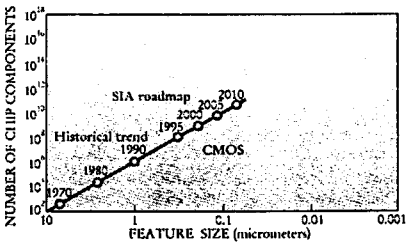
- Chip functionality increases a factor of 4 every 3.4 years

Moore's 2nd Law



- Manufacturing cost increases by a factor of 2 every 3 years

Physical Limitations of CMOS



- Scaling of electron devices expects that only 8 electrons are required for on/off by 2010

Limitations of CMOS in ~ 10 years

- Fundamental physical limit
  - 8 electron per bit (today 1000 e/bit)
- Manufacturing cost
  - $50 Billion/FAB

Moore's law scaling to an end ??

Transistor Technologies — Sep. 2000

Why Scaling?

- Higher Density
  - Integration of more transistors onto a smaller chip
- Higher Performance
  - Higher current drives
  - Smaller capacitances
  - Reduced gate delay
- Lower Voltage, Lower Power

We Need...

Novel Technology Development and Optimization
- Various Process Conditions
- Structures
- Materials

Exploring Scaling Limit with Conventional Technologies

EE867 삼성 반도체소자 특강 *2001 Spring* http://inca.kaist.ac.kr/home/home.htm

Nano-computers

- Electronic Nanocomputers
  Quantum mechanical tunneling effect
- Chemical Nanocomputers
  Stores information in the chemical bonds: DNA etc.
- Mechanical Nanocomputers
  Moving molecular scale parts: how to assemble ?
- Quantum computer
  Each bit of information as a quantum state: spins etc

Nanoelectronic Devices

Solid state nanoelectronic devices

- Quantum dots "artificial atoms" (QD): *0-D*
- Resonant Tunneling Devices (RTD, RTT): *1- or 2-D*
- Single Electron Transistor (SET): *3-D*
- Quantum Cellular Automata (QCA)
- Magnetic Random Access Memory (MRAM)

} Islands confine electrons

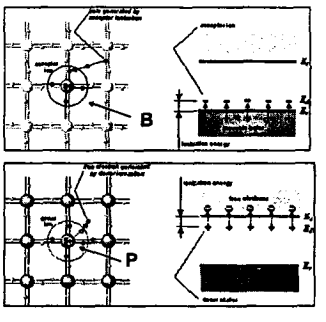
Molecular electronic devices

chemically assembled configurations rather than artificially drawn structures

Microelectronic Transistor: structure, operation, obstacles to miniaturization

- Function of transistor
  - Two state device or switch
  - Amplification

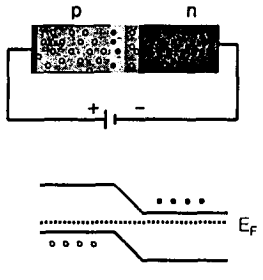
n- and p-type Silicon Semiconductor



- p-type
- B doped
- Hole carrier

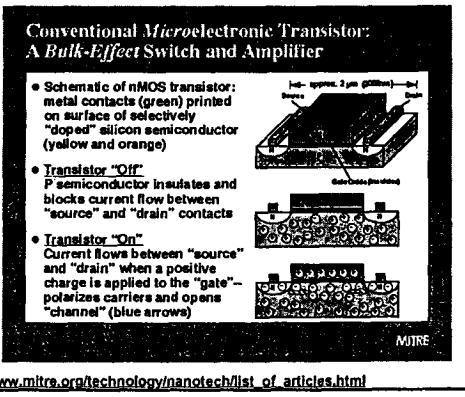
- n-type
- P, As doped
- Electron carrier

Depletion region



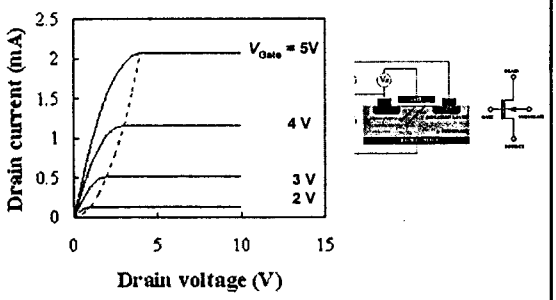
Some of the free electrons in the n-region diffuse across the junction and combine with holes to form negative ions. In so doing they leave behind positive ions at the donor impurity sites.

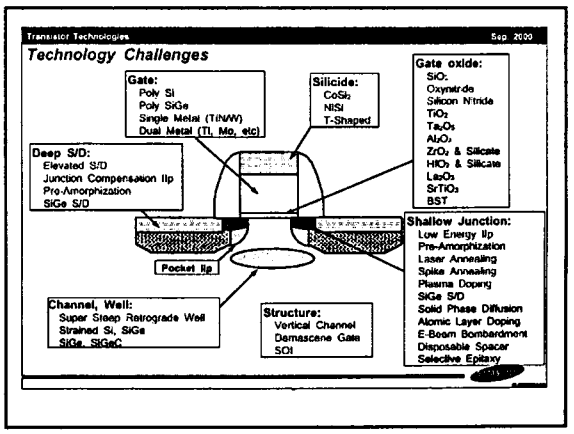
Structure and Operation of MOSFET

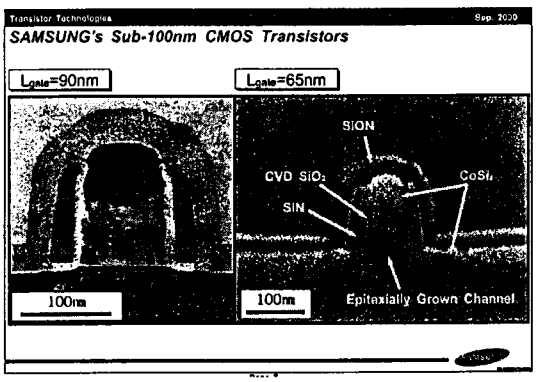


http://www.mitre.org/technology/nanotech/list_of_articles.html

I-V Charateristics of n-type MOSFET







Obstacles to Further Scaling of FET

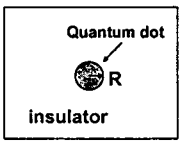
- *High electric field due to bias voltage to short distance*
  - avalanche break down due to high kinetic energy electrons
  - tunneling through insulating layers
- *Heat Dissipation of transistors due to limited thermodynamic efficiency*
  - molecular scaled device: as much heat as gunpowder
- *Vanishing bulk properties and nonuniform dopant concentration*
  - *few dopant atoms in nanoscale: functioning as transistor?*
  - quantum mechanical effect
- *Shrinkage of depletion regions*(<0.1μm)
  - current leakage
- *Shrinkage and uneveness of the thin oxide layer*
  - tunneling

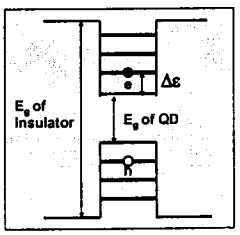
Electron Nature in Smaller Sizes

- *Energy quantization*
  d ~ Fermi wave length of electron in a metal (λ_F) or exciton diameter in a semionductor
- *Charge quantization*
  Charging energy (E_c) >> Thermal energy (kT)
- *Ballistic*
  d<mean free path (λ)

Free electron case: see Kittel
$\Psi = \exp(ikr)$, $k = 2\pi n/L$
$E = \hbar^2k^2/2m$
$N = 2x\ (4\pi k_F^3/3)/(2\pi/L)^3 = Vk_F^3/3\pi^2$
Let electron concentration N = N/V
$E_F = (\hbar^2/2m)\ k_F^2 = (\hbar^2/2m)\ (3\pi^2\ N)^{2/3}$
$k_F = (3\pi^2\ N)^{1/3}$
$\lambda_F = 2\pi/k_F = 2\pi\ (3\pi^2\ N)^{-1/3}$

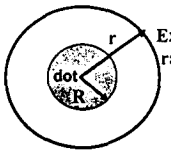
Energy Quantization of Electron in a QD





- *Particle in a box problem*
- Energy levels of *electron and hole are quantized due to the finite size of the dot*
  $E(R) = \hbar^2\pi^2/2\mu R^2$
- R ↓⇒ Energy spacing Δε ↑
- For energy quantization:
  Au: λ_F=0.52 nm, Si: $r_{exciton}$ = 4.3 nm

Quantum Confinement



Exciton radius

Energy for the lowest excited state relative to E_{gap}

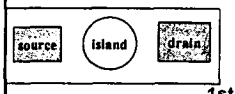
$E(R) = \hbar^2\pi^2/2\mu R^2 - 1.8e^2/2\varepsilon R$ ...

Particle in a box problem

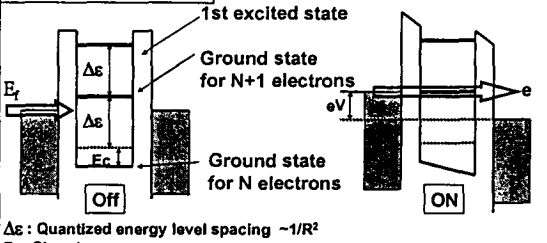
- R<< r: Strong Confinement
  - 1st term (localization) dominant
  - *Electron and hole are quantized*
  - *Energy gap ~1/R²*
  - eg) Si<4.3 nm, Ge<11.5 nm, GaAs<12.4
- R>> r: Weak confinement
  - 2nd term (coulomb attraction) dominant
  - Exciton confinement character

L.E. Brus, J. Chem. Phys. 80, 4403(1984)

Quantum well for RTD

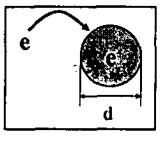


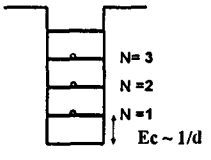
- Quantum effects:
  - Energy quantization
  - Tunneling
- Multiple on and off states



Δε : Quantized energy level spacing ~1/R²
Ec: Charging energy

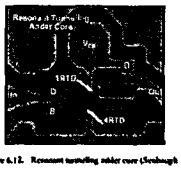
Charge Quantization





- Charging energy: $Ec = e^2/2C >> kT$
  - At T =300K
  - kT = 26 meV
  - $C << 3.1x10^{-19}$ F
  - $C = 4\pi\varepsilon d$
  - $4\pi\varepsilon = 1.1x10^{-10}$ J $^{-1}$ C^2m^{-1}
  - $Ec = e^2/8\pi\varepsilon d \sim 1/d$
- The condition for charge quantization: d<< 28 nm

Resonant Tunneling Device (RTD)



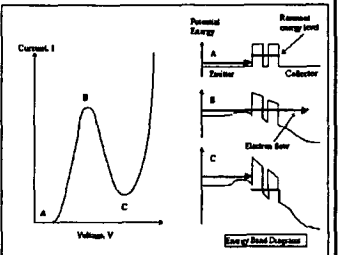
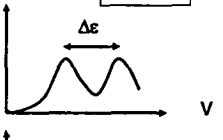


Figure 6.11. Resonant tunneling device (Moffat 1999).

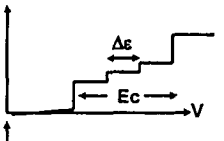
Comparison: QD, RTD and SET

$\Delta\varepsilon \sim 1/d^2$
$Ec \sim 1/d$

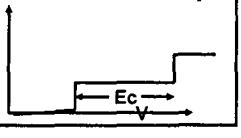
RTD: $\Delta\varepsilon >> Ec$
- 1-D Narrow islands
- Short dimension (5-10nm): $\Delta\varepsilon$ large
- Long dimension: Ec small



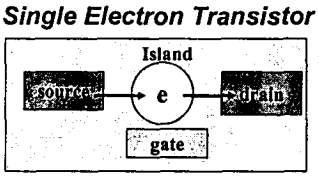
QD : $\Delta\varepsilon < Ec$
- 0D- islands: short all in 3 D
- Metal or semiconductor
- $\Delta\varepsilon$ large and Ec large

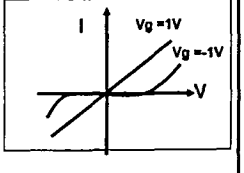


SET: $\Delta\varepsilon << Ec$
- No very short and no very long dim.
- For energy quantization: d(metal) <<d(Semic.)
- Ec is much less sensitive than $\Delta\varepsilon$ to choice of materials as islands
- Metal islands emphasize Ec over $\Delta\varepsilon$
- Coulomb blockade



Single Electron Transistor





Single Electron Transistor
- Island potential is capacitively controlled by the gate.
- Coulomb blockade is overcome by changing the gate voltage

Advantage
- ultra low power operation
- fast

History

1969 Lambe and Jaklevic
Charge Quantization in a small box

1987 Averin and Likharev(SUNY)
Single Charge Transfer
Single Electron Transistor at very low temperature

1991 Fulton(Bell Lab)
Single Charge Sensing Structure

1993 Nakazato and Ahmed, Dresselhaus(MIT) and Likharev
Single Electron Memory Prototype

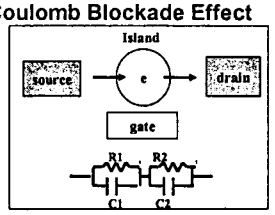
1994 Many Groups from Hitachi, IBM, Minnesota, etc
Single Electron Memory at Room Temperature using Quantum Dots

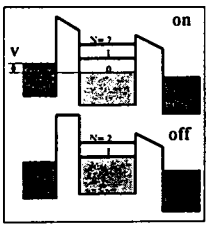
1996 Yano(Hitachi Centural Lab)
First Single Electron Memory Array. 64bits

1998 Yano(Hitachi Centural Lab)
First ULSI Single Electron Memory Prototype. 128Mbits

1999 Nakazato(Hitachi Cambridge) and Ahmed(Cavendishi Lab)
Announce Manufacturable Device for Next Generation Memory

Coulomb Blockade Effect



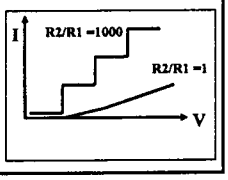


Coulomb Blockade Effect

Quantum tunneling of electron between source and drain can be blocked if the charging energy $Ec = e^2/2C >> kT$

$\Delta E = eV - Ec < 0$: blocked

$2Ec = e^2/(C_1+C_2)$



Coulomb Staircase

- Asymmetric junction (R2>>R1)
- Current steps at $e/2C_2 + n(e/C_2)$

S.Y. Chou et al, Appl. Phys. Lett 67, 938 (199

Example: SET at 100K

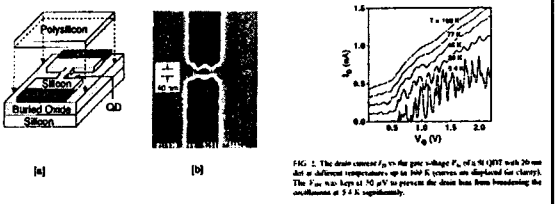
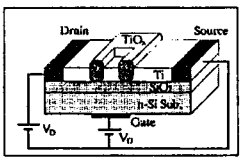
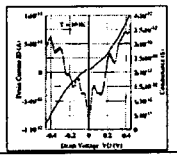


FIG. 1. The drain current I_D vs the gate voltage V_G of a Si QDT with 20-nm dot at different temperatures up to 300 K (curves are displaced for clarity). The V_{DS} was kept at 50 μV to prevent the drain bias from broadening the oscillations at 4.2 K significantly.

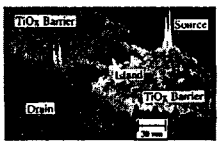
- Si dot with 20nm diameter : energy spacing = 40meV
- Current oscillation due to the interference between different modes of quantum waves in a cavity

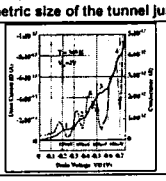
Example: *SET operating at room temperature*





Unclear coulomb staircase due to the symmetric size of the tunnel juncton

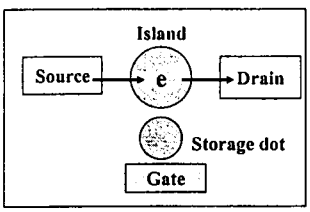


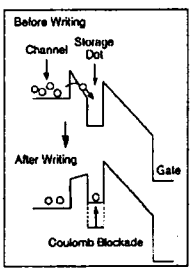


K. Matsumoto et al, Appl. Phys. Lett. 68, 34 (1996)

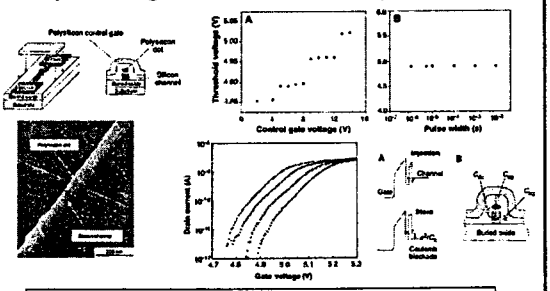
Coulomb staircase with periods of 150mV

Single Electron Memory





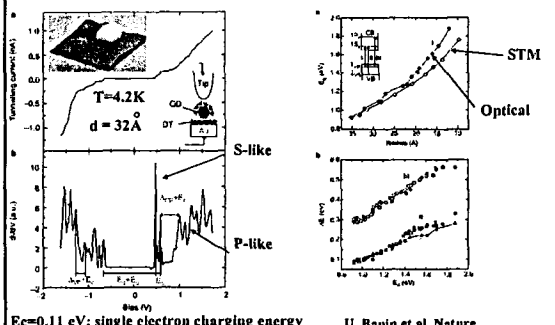
Example: Single Electron Memory



- Discrete shift in the threshold voltage
- Staircase relation between the charging voltage and the shift
- Self-limiting charging process

L. Guo, E. Leobandung, S.Y. Chu, Science 275, 649 (1997)

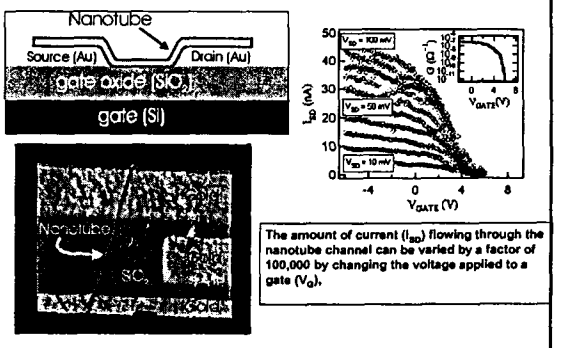
Tunneling Spectroscopy of InAs QD



Ec=0.11 eV: single electron charging energy
Eg=1.02 eV: nanocrystal band gap

U. Banin et al, Nature, 400, 926 (2000)

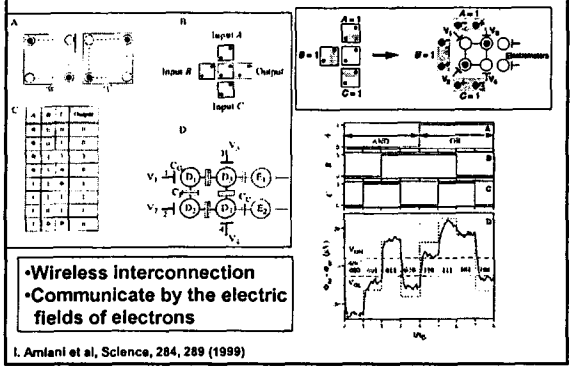
Nanotube Single Electron Transistor



The amount of current (I_{SD}) flowing through the nanotube channel can be varied by a factor of 100,000 by changing the voltage applied to a gate (V_G).

P. Avouris et al ,Appl. Phys. Lett, 73, 2447 (1998)

Quantum Dot Cellular Automata

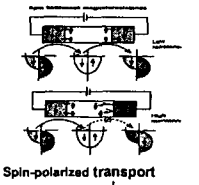


- Wireless interconnection
- Communicate by the electric fields of electrons

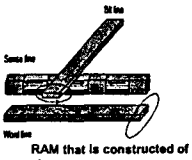
I. Amlani et al, Science, 284, 289 (1999)

Magnetic Random Access Memory (MRAM)

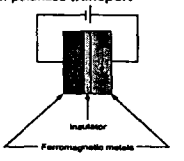
G.A. Prinz, *Science* 282, 1660 (1998)



Spin-polarized transport



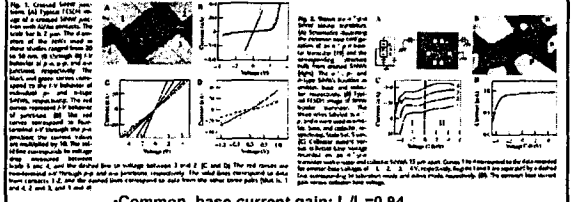
RAM that is constructed of GMR elements



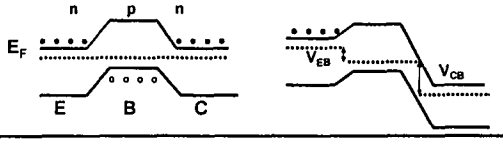
Magnetic Tunnel Junction(MTJ)

RAM that is constructed of magnetic tunnel junction

Nanowire Transistor Y. Cui and C.M. Lieber, Science 291, 851 (2001)



- Common base current gain: I_C/I_E=0.94
- Common emitter current gain: I_C/I_B=16



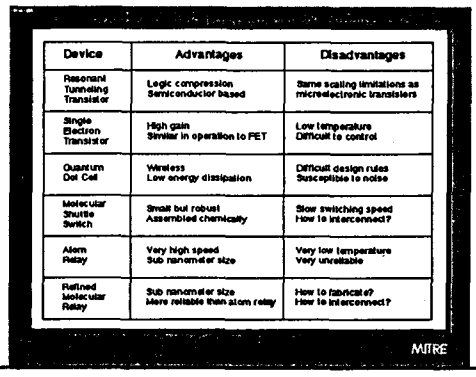
Comparison of Nanodevices

Comparison of Nanotransistor Technologies

Device	Operating principle	Status
Resonant Tunneling Transistor	Quantum resonance in double barrier potential wells	Capable of large scale fabrication
Single Electron Transistor	Coulomb Blockade in small quantum dots	Experimental only at very low temperatures
Quantum Dot Cell	Single electron confinement in arrays of quantum dots	Quantum dots can be fabricated in the laboratory. Quantum dot cells are still theoretical
Molecular Shuttle Switch	Movement of a molecular "bead" between two stations on a switch molecule	Experimental, can only be switched chemically
Atom Relay	Vibrational movement of a single atom in and out of an atom wire	Theoretical
Refined Molecular Relay	Rotational movement of a group in and out of an atom wire	Theoretical

MITRE

Comparison of Nanodevices



Device	Advantages	Disadvantages
Resonant Tunneling Transistor	Logic compression Semiconductor based	Same scaling limitations as microelectronic transistors
Single Electron Transistor	High gain Similar in operation to FET	Low temperature Difficult to control
Quantum Dot Cell	Wireless Low energy dissipation	Difficult design rules Susceptible to noise
Molecular Shuttle Switch	Small but robust Assembled chemically	Slow switching speed How to interconnect?
Atom Relay	Very high speed Sub nanometer size	Very low temperature Very unreliable
Refined Molecular Relay	Sub nanometer size More reliable than atom relay	How to fabricate? How to interconnect?

MITRE

Drawbacks and Obstacles to Solid- State Nanoelectronic Devices

- *Valley current in RTD*
  - not clear on and off
- *Sensitivity to input and current fluctuations*
  - accidental on and off
- *Cryogenic operation: 30nm Si SET at 150 K*
  - reducing size: Ec and $\Delta\varepsilon$ increase.
- *Materials: Si is still preferred*
  - SiO_2 is still best insulator
  - GaAs is far inferior to electric fields
- *Background charge problems*
- *Extreme sensitivity of the tunneling current to width of potential barriers*
  - SOI technology or nonpolarizable organic compounds
- *Extreme difficulty of making islands and tunnel barriers precisely and uniformly*

세미나 공고

- 제목: *"Polymeric Nanoparticles for Drug Delivery"*
- 연사: 조종수 교수(서울대)
- 장소: 응용공학동 2층 2 세미나실
- 시간: 9월 13일(목) 17:00-17:30

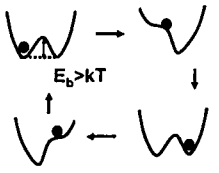
2. *Molecular Electronics*

- Molecular-scale electronics
- Molecular materials for electronics

- Molecular wire
- Diode, rectifier
- Molecular switches
- Molecular memory
- Sensors
- Optics and optical switches
- Displays
- Electrochemical devices
- Molecular heterostructure and quantum well devices

Molecular Electronic Switching Devices

- *Electric-field* controlled molecular switching devices including *quantum-effect* molecular electronic devices
- *Electromechanical* molecular electronic devices
- *Photoactive/photochromic* molecular switching devices
- *Electrochemical* molecular devices



Bistability

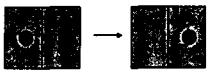
강의 순서

- Information transfer
- Interconnections
- Electron transport
- Molecular wire
- Molecular rectifier
- Molecular switches
- Molecular RTD
- Molecular memory
- Atom relay

Computational Limits

- Rolf Landauer
  Nonreversible computer performing Boolean logic operation requires minimum energy for a bit operation,

$$P = nk_BT\ ln2$$



Ex) energy required to add two 10 digit decimal numbers ?
P ~ 100kT ln2 ~ 3×10^{-18} joule per additions

10^{18} additions per joule

This is roughly the equivalent of 10^9 Pentiums !

Why Molecular electronics ?

ENIAC, 1947

17,468 vacuum tubes
60,000 pounds
16,200 cubic feets
174 kilowatts (233 horsepower)

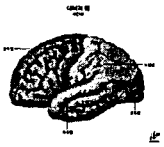
HP Jornada, 2000

- *Popular Mechanics* (1949, March issue) predict that someday ENIAC would contain only 1500 vacuum tubes
- Now, Improve power efficiency by 10^8 and shrink by 10^8
- Their prediction was based wrong foundation(vacuum-tube technology)

Technology and Biology

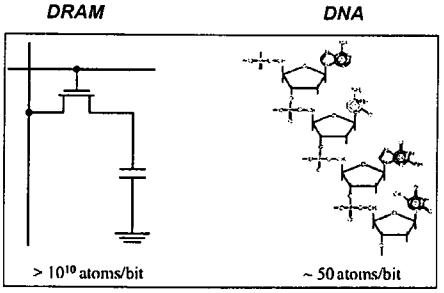
Technology	*Biology*
• Artificial	Natural
• Si-based	C-based
• Manufactured	Self-assembled
• Short history	Long history
• *Function:* logical numerical operation	Replication, adaptation...
• Room temp	Room temp

Microprocessor vs Brain



MPU	*Brain*
• # of MOSFET ~ 10^7	# of cells ~ 10^{12}
• # of switches ~ 10^7	# of switches ~ 10^{11} neurons
• # of connections ~10^7	# of connections ~ 10^{15} synapses
• *Wiring*: fixed	flexible
• *Architecture*: serial	parallel

DRAM vs DNA



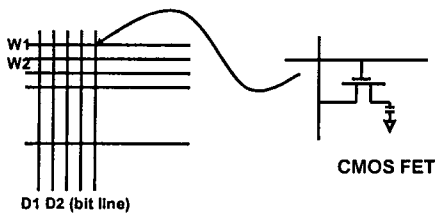
Advantages of Molecular Electronics

- Nano-scaled structures with *identical size*
- *Ultra High density:* 10^6 times denser than Si logic circuits
- *Very cheap*

Critical issues on Molecular Electronics

- What device types can provide *bistable* operation ?
- How can these devices be *organized* into high density 2D and 3D arrays ?
- How can these devices be *connected* in large number to input/output lines?

Dynamic Random Access Memory (DRAM)



Interconnection Dilemma

- Today's chip densities are such that the wires consume some 70% of the real estate
  → they cause some 70% of the defects that lower chip yield
- The rate of defects in a chemically fabricated nanocircuit : ppb level
  → million defects in a system containing 10^{15} components

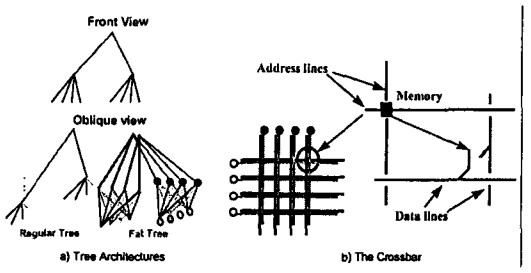
It is impossible to tolerate such a level of defects!!

Several Fascinating Possibilities

Future computing architecture should be highly tolerant defects!!

- **Defect-tolerant computing architecture:**
  Heath, and Stoddart at UCLA, teamed up with Williams at HP
- **Nanocell:**
  Tour at Rice, Reed at Yale teamed up with Penn State in Nov. 1999
- **Switching with nanotubes:**
  Lieber at Harvard University
- **DNA assembly, computation:**
  Seeman at New York University

Defect-Tolerant Computing Architecture

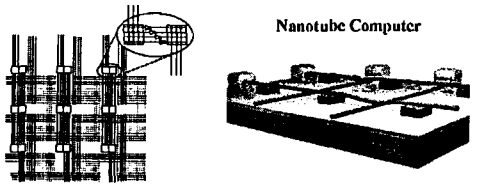


•Fat tree architecture enable one to route around and avoid the defect.
•Manufacturing by chemical assembly is feasible.

HPL Teramac
1THz multi-architecture computer

- ***Tera:*** 10^{12} operations per sec
- ***Mac:*** Multiple architecture computer
- 10^6 gates operating at 10^8 cycle/sec
- Largest defect-tolerant computer
- Contains 256 effective processors
- Computes with look-up tables
- 220,000 (3%) defective components

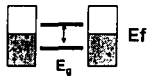
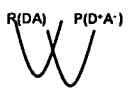
Nanotube Interconnects



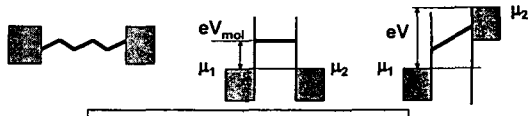
•Molecular scale wire with atomic perfection
•Interconnect at high density
•High thermal conductivity
•Very stiff materials

Lieber et al

Wire conductance vs. Electron transfer

	Molecular wire	*Intramolecular electron transfer*
Potential Energy		
Obsevable	current	rate constant
Continuum	electrode	vibronic levels
Process	electron tunneling	electron tunneling
Theory	$I = (2\pi e/h) \int dE\, T^2(E,V) [f_D(E)(1-f_A(E+eV))$	$k = (2\pi/h)T^2_{DA}(FC)$

Conductance : Landauer formula

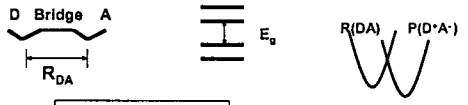


$$I = (2e/h) \int dE\, T(E,V) [f(E-\mu_1)-f(E-\mu_2)]$$

•F(E): Fermi function
•T(E,V): transmission function
⇒ *molecular energy levels* and *their coupling to the metallic contacts*
μ_1, μ_2: electrochemical potentials
μ_1= Ef- eV_{mol}
μ_2=Ef+eV- eV_{mol}

W. Tian et al, J. Chem. Phys. 109, 2874(1998)

Intramolecular Electron Transfer



$k = (2\pi/h)V^2_{DA}(FC)$

$V_{DA} = a\exp(-\beta R_{DA})$, $\beta = -(1/a)\ln(2t/E_g)$

V_{DA} : electronic coupling through direct and superexchange

FC: Frank-Condon factor

R_{DA} : DA distance

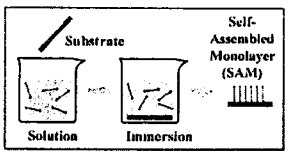
a: length between electronic basis function in the bridge structure

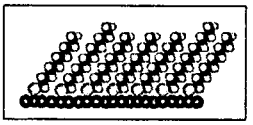
T: matrix elements between those bridge structure

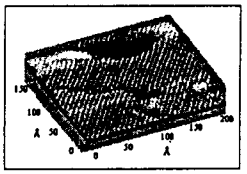
E_g : energy gap

Chap1 and 2 in *Molecular Electronics* ed. J. Jortner and M. Ratner

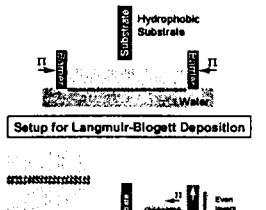
Fabrication: *Self-Assembled Monolayer (SAM)*

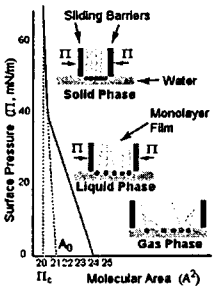






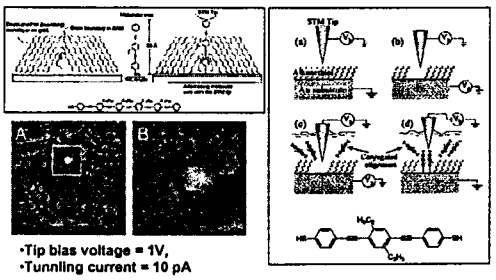
Fabrication: *Langmuir-Blodgett Technique*





http://www.public.iastate.edu/~miller/nmg/lbfilms.html

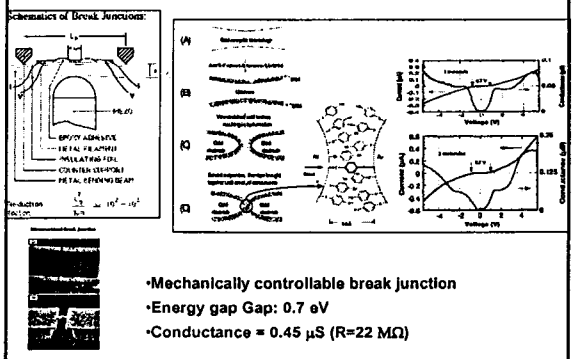
Conductance of molecular wire: STM



•Tip bias voltage = 1V,
•Tunnling current = 10 pA

L. A. Bumm et al , *Science* 271, 1705 (1996).

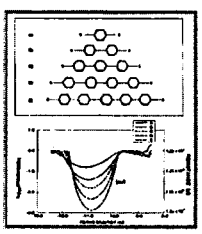
Conductance of Molecular Wire: MBJ

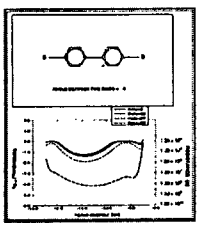


•Mechanically controllable break junction
•Energy gap Gap: 0.7 eV
•Conductance = 0.45 μS (R=22 MΩ)

MA REED,C Zhou, CJ Muller, TP Burgin, JM Tour, Science 278, 252 (1997)

Structural Effects on Conductance:Theory

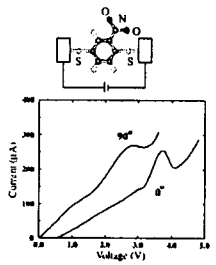


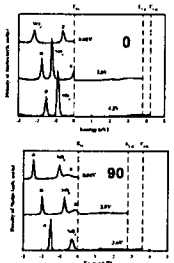


•Resistance increases exponentially with the # of the rings
•Relative orientation of the rings
•The bonding between them.

MP Samanta et al, Phys. Rev. B53, R7626 (1996)

Temperature Effects : Theory





- Unusual temperature induced large shift (~1eV) in is due to:
  - The rotation property of the NO_2 group
  - Different symmetry of the states localized on the NO_2 group with respect to the orbitals of the carbon ring

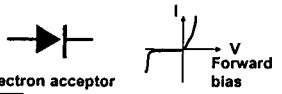
M Di Ventra, SG Kim, ST Pantelides, ND Lang, PRL86, 288(2001)

Comparison of Conductivity

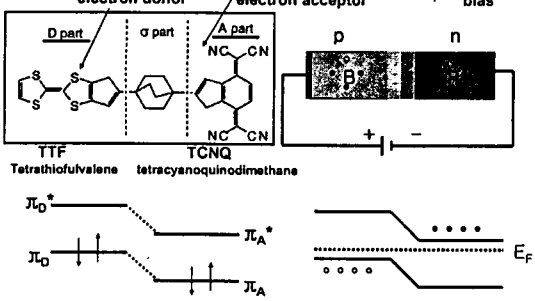
	1,4 benzene ditiol	*polyphenylene wire (3ring)*	*carbon nanotube*	*copper wire*
App. Vol	1	1	1	$2x10^{-3}$ (10cm)
Current (A)	$2x10^{-6}$	$3.2x10^{-5}$	$1x10^{-7}$	1
Cross section (nm^2)	~0.05	~0.05	~3.1 r=1nm	$\sim 3.1x10^{12}$ r=1mm
Current density (e/sec-nm^2)	$2x10^{12}$	$4x10^{12}$	$2x10^{11}$	$2x10^{6}$

Molecular Rectifier

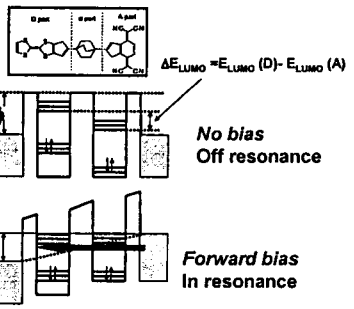
A. Aviram and M.A. Ratner, Chem. Phys. Lett. 29, 277 (1974)



electron donor — electron acceptor

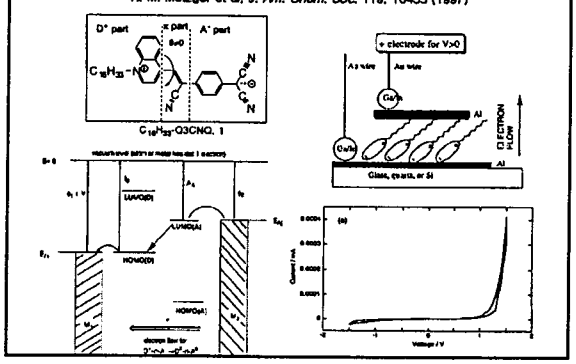


Energy Levels of Molecular Orbitals

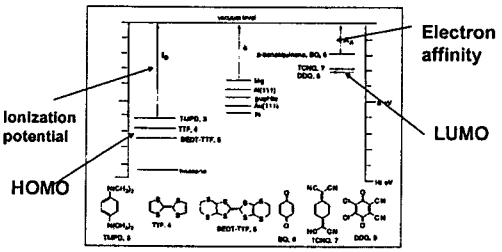


Electrical Rectification of LB films

R. M. Metzger et al, *J. Am. Chem. Soc.* 119, 10455 (1997)

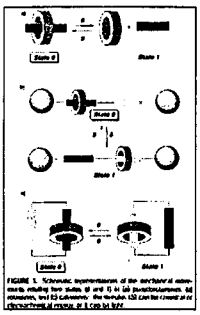


Energy levels

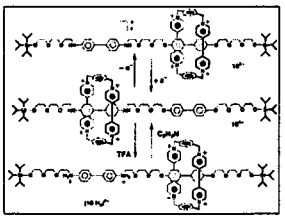


- Ionization potetials I_D for D end must be *small* and match as closely as possible work function ($\phi 1$) of metal layer (M1).
  - If I_D is too low, the molecule would oxidize in air
- Electron affinity A_A for A end must be as *large* as possible, match with the work function metal layer M2($\phi 2$): this is not easy !

Molecular Switches

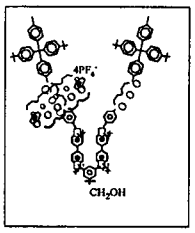


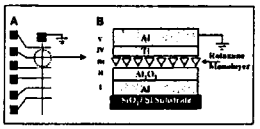
1. Chemical switching
2. Electrochemical switching
3. Photochemical switching

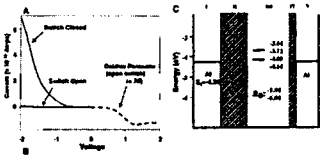


V.Balzani et al, *Acc. Chem. Res. 31*, 405 (1998)

Molecular Switches and Gates: Rotaxane



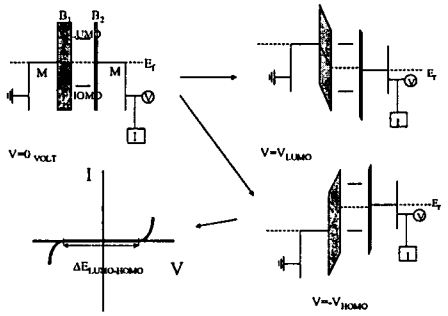




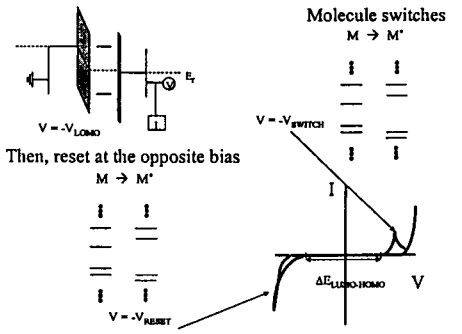
•Closed at reducing voltage(-2V): current flow due to resonant tunneling
•Open at oxidizing voltage(>0.7V): irreversible

C.P. collier et al, Science 285, 391(1999)

Electron Transport in Single molecule

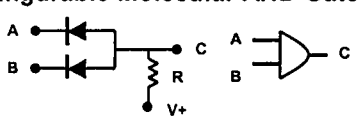


Electron Transport in Single molecule



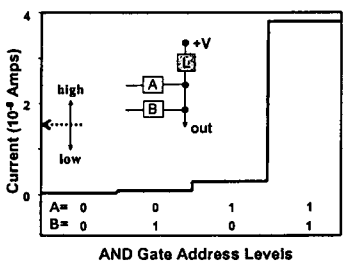
Then, reset at the opposite bias

Configurable Molecular AND Gate



A	B	C
0	0	0
1	0	0
0	1	0
1	1	1

•Difference between high and low current levels: 15 ~30

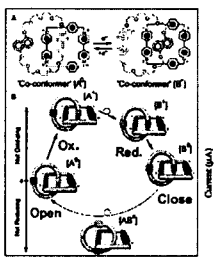


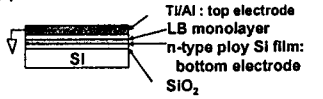
C.P. Collier, E.W. Wong et al.

AND Gate Address Levels

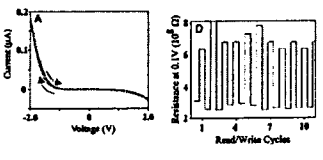
Reversible Molecular Switches: [2]Catenane

Collier et al, Science, 289,1172 (2000)





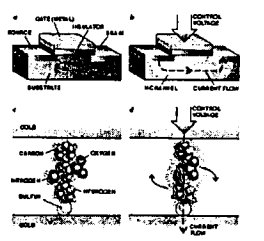
Ti/Al : top electrode
LB monolayer
n-type ploy Si film: bottom electrode
SiO_2

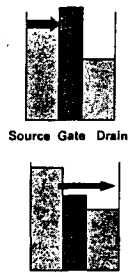


•Upon oxidation, the TTF become positively charged, the Coulombic repulsion between TTF⁺ and the tetracationic cyclophane causes to circumrotate.
•Reversible switching :Opened >+2V, closed <-1.5V

Molecular Field Effect Transistor

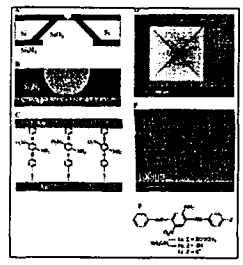
R.A. Reed and J.M. Tour, Sci. Amer. 282,86 (2000)

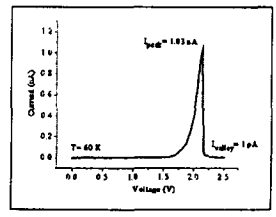




CONVENTIONAL MICROTRANSISTOR (a) has three terminals, known as the source, gate and drain. A positive voltage applied to the gate draws electrons to the insulator (S), enabling current to flow from the source to the drain. A molecule based on three benzene rings (c) was also used to switch an electric current. The center ring had asymmetric fragments, enabling it to be treated by an electrical field (E). With a specific voltage applied, the electrical field twisted the molecule and permitted current to flow.

Reversible Molecular Switch with NDR Effect

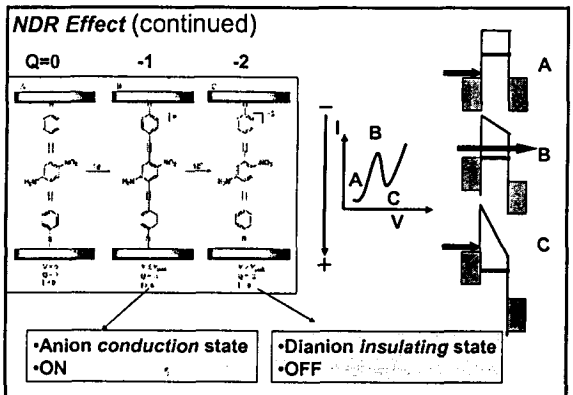




- Negative Differential Resistance ~ 400 MΩcm²
- Peak current denisty: 50A/cm²
- Peak to valley ratio = 1030:1 (typical device=30:1)
- Temperature induced shift : rotation of ligand (JACS,122,3015 (2001))

J Chen, MA Reed, AM Rawlett, JM Tour, Science 286, 1550 (1999)

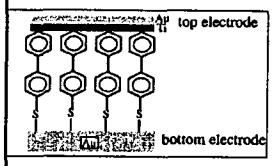
NDR Effect (continued)

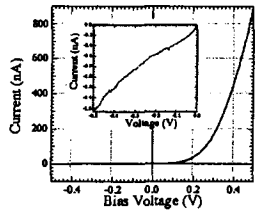


- Anion *conduction* state
- ON

- Dianion *insulating* state
- OFF

Molecular Diode

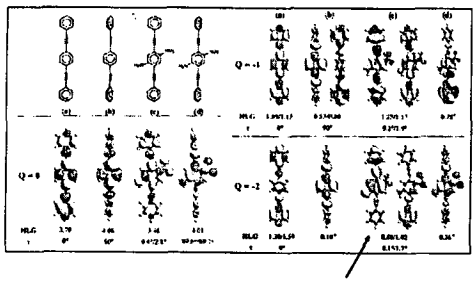




- The prominent rectifying behavior is due to the *asymmetry* of the molecular heterostructure.
- The barrier from the bottom electrode is higher than the barrier for electrons from the top T electrode

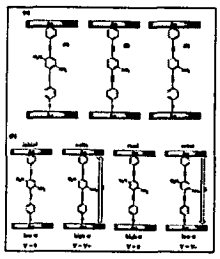
C. Zhou, M. R. Deshpande, M. A. Reed, L. Jones II, and J. M. Tour, Appl. Phys. Lett., 71, 611 (1997).

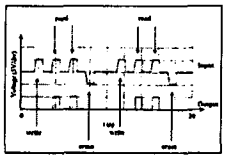
Lowest Unoccupied Molecular Orbital (LUMO)

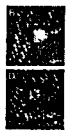


No LUMO states on the ring

Molecular RAM



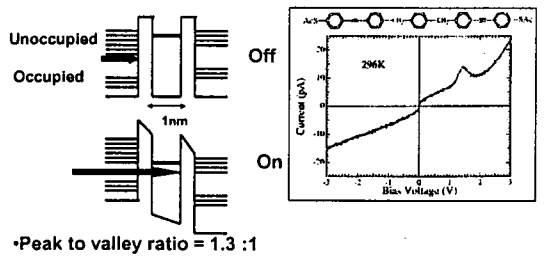




- 15min hold time DRAM at room temperature
- Reversible molecular memory
- Over one billion cycles and counting with no degradation

M.A. Reed et al, Appl. Phys. Lett. 78, 3735 (2001),
Z.J. Donahuer et al, Science 292, 2305 (2001)

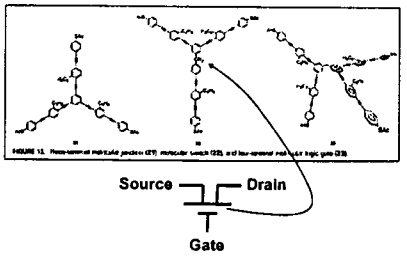
Molecular Resonant Tunneling Diode (MRTD)



- Peak to valley ratio = 1.3 :1
- Electrically active device by molecular orbital engineering

M.A. Reed, Proc. IEEE , Volume: 87, 652 (1999)

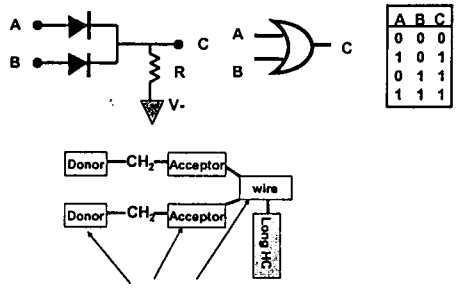
Synthesis of Molecular Devices



- Nano-scaled structures with *identical size and shape*
- *High density and low power*

J.A Tour, Acc. Chem. Res. 33, 391 (2000)

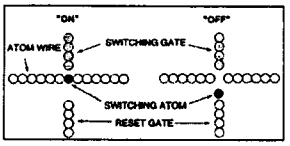
Logic Gate : OR



Conjugated aromatic organic molecules

Electromechanical Molecular Electronics

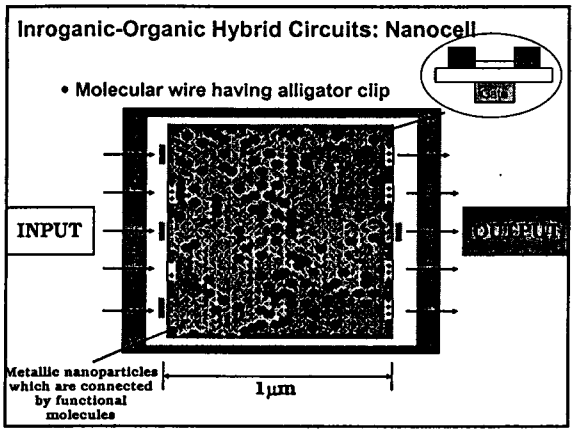
1. Single molecule electromechanical amplifier:
   Joachim and Gimzewski Chem Phys. Lett. 265, 353 (1997)
   - conductance modulation due to electromechnical deformation of C_{60} cage
2. Atom relay transistor
   Y. Wada, JVST A17, 1399 (1999)



- Upon charging a gate, a mobile switching atom move into a line of atomic wire

Inroganic-Organic Hybrid Circuits: Nanocell

- Molecular wire having alligator clip



Further challenges

- Combining individual devices
- Mechanisms of conductance
- Nonlinear I/V behavior
- Energy dissipation
- Necessity of gain in molecular electronic circuits
- Slow speed
- New computer architecture
- Synthesis of new molecules

숙제 1.

Read Feyman's lecture "***There is plenty of room at the bottom***" listed at www.zyvex.com/nanotech/feynman.html

공고

- 금주 20일 (목요일): 휴강
- 스케쥴 변경

week 7 : 10.16-10.18 Macromolecular nanostructures (김상윤 교수)
week 8 : 10.20-26 중간고사

switch

○ 강의명 :

○ 참고문헌 *Nanotechnology Research Directions: IWGN Workshop Report(1999)*
http://dot.loyola.edu/nano/IWGN.Research.Directions/

○ 강의내용 기능성 나노구조의 합성, 제조, 물리 화학적 성질과 나노구조에 대한 특성분석 방법 등을 다루고 나노구조를 응용한 나노센서 및 나노소자의 개념을 소개함

week 1 : 9 4 Introduction (김세훈 교수)
week 1/2 : 9.6-9.11 Nano-characterization (김세훈 교수)
week 2/3 : 9.13-9.18 Nano device I (김세훈 교수)
week 3/4 : 9.20-9.25 Template based nanostructures (유동 교수)
week 4/5 : 9.27-10.4 Template based nanostructures (유동 교수)
week 6 : 10.9-10.11 Macromolecular nanostructures (김상윤 교수)
week 7 : 10.16-10.18 휴강
week 8 : 10.20-26 Macromolecular nanostructures (김상윤 교수)
week 9 10.30-11.1 Nano-fabrication and nano-lithography (김진백 교수)
week 10 : 11.6-11.8 Nano-quantum chemistry (이충섭 교수)
week 11 : 11.13-11.15 Nano-thermodynamics (이억근 교수)
week 12 11.20-11.22 Nano-sensor and nano-device II (박주현 교수)
week 13 : 11.27-11.29 Nano-sensor and nano-device II (박주현 교수)
week 14 : 12.4-12.6 Nanoparticles and nanowires (전진우 교수)
week 15. 12.11-12.13 Nanoparticles and nanowires (전진우 교수)
week 16 : 12.15-12.21 기말고사

生物知識庫教材內容

Chap 0 Preface
生物知識庫課程大綱

劉 志 俊 (Chih-Chin Liu)

中華大學 資訊工程系

February 2003

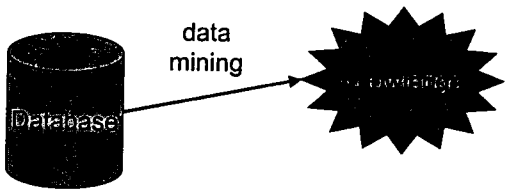
Research Issues in Bio-Databases

- Data Modeling
  - How to store/represent biological data
- Data Retrieval
  - To retrieve similar biological objects
- Data Mining
  - How to find rules behind biological data
- Simulation
  - Pathway Simulation, Virtual Cell, Virtual Life

Data Mining

Area	Term
Artificial Intelligent	Machine Learning
Signal Processing	Pattern Recognition
Database	Data Mining

Data Mining



Data Mining Techniques

- Classification Analysis
- Clustering Analysis
- Association Rule Analysis
- Generalization Analysis

Traditional Data Types

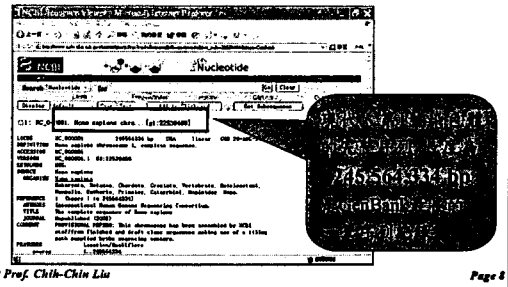
- Strings
  - char
  - varchar
- Numeric
  - integer
  - float/double
  - decimal
  - money
- Date

New Data Types in Bio-Databases

- Large Strings
  - DNA Sequences, Protein Sequences
- Biological Images
  - 2D Gels, Microarray Images
- 3D Structures
  - Proteins, Compounds
- Network
  - Pathways

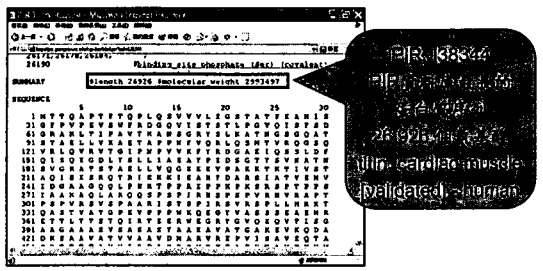
New Data Types in Bio-Databases

- Large Strings: DNA Sequences



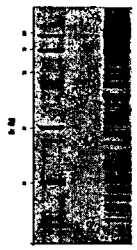
New Data Types in Bio-Databases

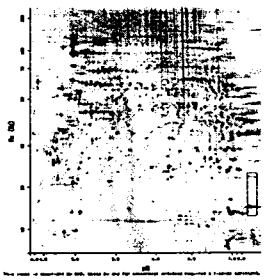
- Large Strings: Protein Sequences



New Data Types in Bio-Databases

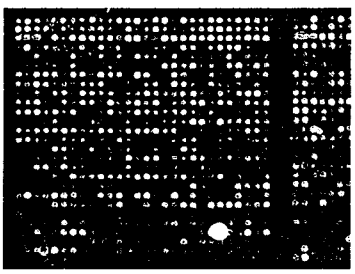
- Images: 1D-Gel, 2D-Gel





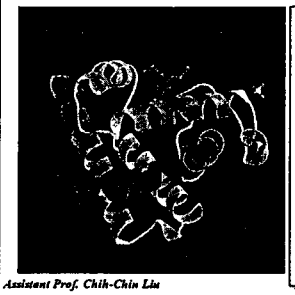
New Data Types in Bio-Databases

- Images: Microarray (Stanford Microarray Database)



New Data Types in Bio-Databases

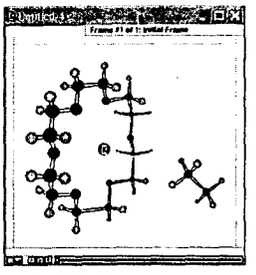
- 3D Structures: Protein



```
ATOM      1  N   VAL     1      -4.004
15.224  18.836  1.00 32.64           N
ANISOU    1  N   VAL     1     4512   9448
4441  -935  -2875   320       N
ATOM      2  CA  VAL     1      -3.520
15.758  14.000  1.00 18.42           C
ANISOU    2  CA  VAL     1     1478   2235
3280  -280  -487   555       C
ATOM      3  C   VAL     1      -2.862
14.738  15.628  1.00 17.00           C
ANISOU    3  C   VAL     1     1803   1981
2980  -152  -486   234       C
ATOM      4  O   VAL     1      -3.858
15.560  15.714  1.00 16.61           O
ANISOU    4  O   VAL     1     1758   2150
3163  -489  -394   501       O
```

New Data Types in Bio-Databases

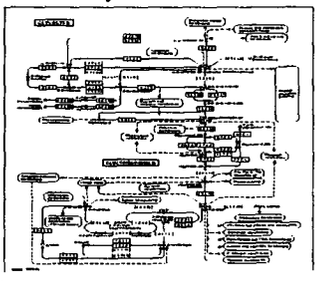
- 3D Structures: Chemical Compound



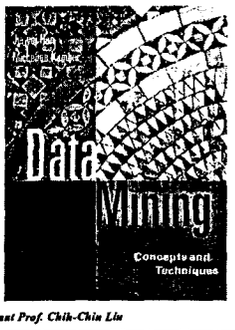
Atom	X	Y	Z
C(1)	-2.571	3.253	0.987
O(2)	-2.662	1.984	1.576
O(3)	-1.012	2.425	-0.556
C(4)	-1.175	3.397	0.442
C(5)	-3.964	1.866	2.084
C(6)	-4.072	0.512	2.733
O(7)	-3.894	-0.450	1.728
C(8)	-4.087	-1.782	2.327
C(9)	-4.002	-2.739	1.240
O(10)	-2.704	-2.688	0.712
O(11)	0.355	0.367	-1.720
C(12)	0.495	1.661	-2.258
C(13)	0.257	2.630	-1.116
O(14)	-0.807	-2.103	-1.651
C(15)	0.457	-1.918	-2.220
C(16)	0.494	-0.524	-2.793
C(17)	-2.637	-3.662	-0.294

New Data Types in Bio-Databases

- Network: Pathways



Textbook



Data Mining: Concepts and Techniques

Jiawei Han and Micheline Kamber

Publisher: Morgan Kaufmann; 1st edition (August 2000)

ISBN: 1558604898

Journals

- *Nucleic Acid Research*
  http://nar.oupjournals.org/
- *Bioinformatics*
  http://bioinformatics.oupjournals.org/

KDD Conferences and Journal

- KDD Workshops 1989, 1991, 1993, 1994
- KDD Conference annually since 1995
- KDD Journal since 1997
- ACM SIGKDD http://www.acm.org/sigkdd

Chap 1 Classification Analysis

劉志俊 (Chih-Chin Liu)

中華大學 資訊工程系

February 2003

Outline

- Introduction to Classification
- Classification vs. Clustering
- Classification Approaches
- Criteria for Comparing Classification Methods

Introduction to Classification

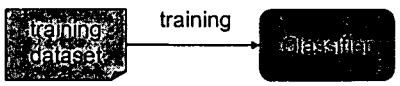
- ***Classification*** and ***prediction*** are two forms of data analysis that can be used
  - to extract models describing important data classes
  - to predict future data trends
- Many classification methods have been proposed in machine learning, pattern recognition, artificial intelligence, expert systems, and statistics.

Introduction

- Objectives of classification
  - To find *common properties* among objects in a class
  - To *predict the properties* of an unknown object
  - To *organize large amounts of data* into hierarchies
  - To *retrieve similar objects* in the same class for a given query example

Introduction

- Training Phase



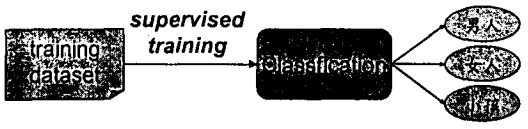
- Prediction Phase

Introduction

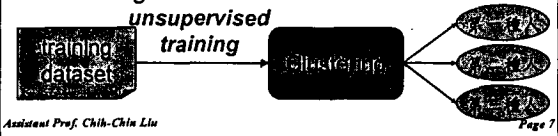
- Biological targets to classify/predict
  - Protein Functions
  - Protein Annotations
  - Protein Structures
  - Gene Functions
  - Gene/Intron/Exon Boundaries
  - Operons

Classification vs. Clustering

- Classification



- Clustering



Data Mining Techniques

- Decision Tree Approach
- Bayesian Classifier
- k-Nearest Neighbor Classifier
- HMM Classifier
- Neural Network Approach

Classification Techniques

- Association Rule Mining Approach
- Genetic Algorithm Approach
- Fuzzy Set Approach
- Rough Set Approach
- Case-Based Reasoning Approach

Data Mining Techniques

No classification method
is superior over all others
for all data types and domains!

Criteria for Comparing Classification Methods

- ***Predictive Accuracy***: how accurate is it ?
- ***Speed***: computation costs for training and prediction
- ***Robustness***: can it handle data with noises and missing values ?
- ***Scalability***: can it handle large amounts of data ?
- ***Interpretability***: any understanding or insight provided ?

References

- [Han00] Jiawei Han and Micheline Kamber, "Chap 7 Classification and Prediction," *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.
- [Duda73] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
- [Winston92] Winston,P.H., Artificial Intelligence, 3ed, Addison-Wesley, 1992.

Chap 2 Decision Tree-Based Classifiers

劉志俊 (Chih-Chin Liu)

中華大學 資訊工程系

February 2003

Outline

- Introduction to Decision Trees
- Decision Tree Classification Algorithms
  - CLS
  - ID3
  - C4.5
- C4.5 Tools
  - C4.5 by Quinlan
  - Other C4.5 Tools

Introduction to Decision Trees

- A ***decision tree*** is a flow-chart-like tree structure, where
  - each ***internal node*** denotes a test on an attribute
  - each ***branch*** represents an outcome of the test
  - each ***leaf node*** represents a class

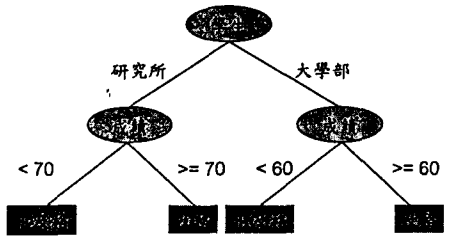
Introduction to Decision Trees

Non-Categorical Attributes *Categorical Attribute*

成績	學生	大學及格
80	研究生	及格
85	大學部	不及格
65	研究生	及格
68	大學部	及格
50	大學部	不及格
55	大學部	不及格
62	大學部	及格
40	研究生	不及格
90	研究生	及格
90	大學部	及格
40	大學部	不及格
50	大學部	不及格

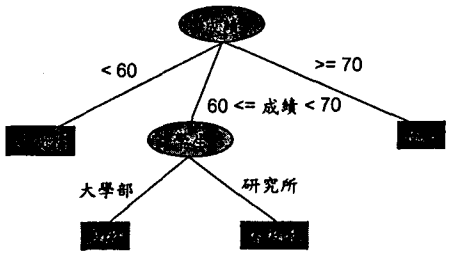
Introduction to Decision Trees

- Example: A Decision Tree



Introduction to Decision Trees

- Example: Another Decision Tree

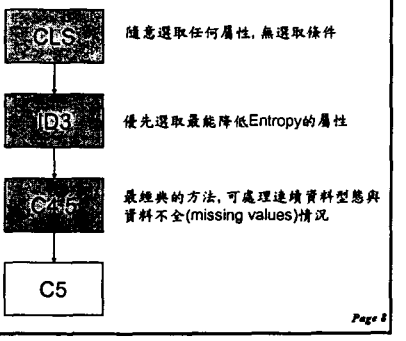


Introduction to Decision Trees

Problem:

- Input: Given a ***relational table*** with
  - ***N non-categorical attributes*** and
  - ***A categorical attribute***
- Output: A ***decision tree*** which can generate the categorical attribute value according to the corresponding N non-categorical attribute values

Decision Tree Classification Algorithms



CLS Algorithm

S1: T←the whole training set.
Create a T node.
S2: If all examples in T are positive, create a 'yes' node with T as its parent and stop.
S3: If all examples in T are negative, create a 'no' node with T as its parent and stop.
S4: Select an attribute X with values $v_1,\ldots,v_N$ and partition T into subsets $T_1,\ldots,T_N$ according to their values on X.
Create N T_i nodes ($i=1,\ldots,N$) with T as their parent and $X=v_i$ as the label of the branch from T to T_i.
S5: For each T_i do: $T \leftarrow T_i$ and go to S2.

ID3 Algorithm

- Definition: 分類資料 X 的***Entropy H(X)***

$$H(X)=\sum_{j=1}^{N} P_j \log_2(1/P_j)$$

- Example: 12 個學生, 6 個及格, 6 個不及格
  則此分類資料的 Entropy 為

$$\frac{6}{12}\log_2\left(\frac{1}{\frac{6}{12}}\right)+\frac{6}{12}\log_2\left(\frac{1}{\frac{6}{12}}\right)=\frac{1}{2}+\frac{1}{2}=1$$

ID3 Algorithm

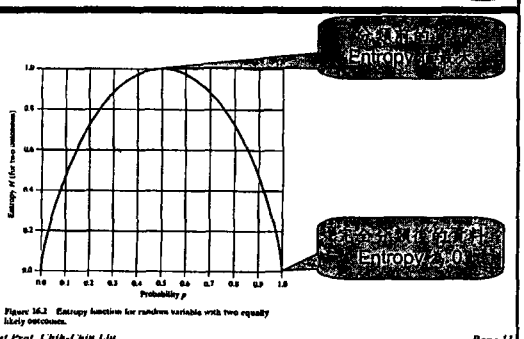
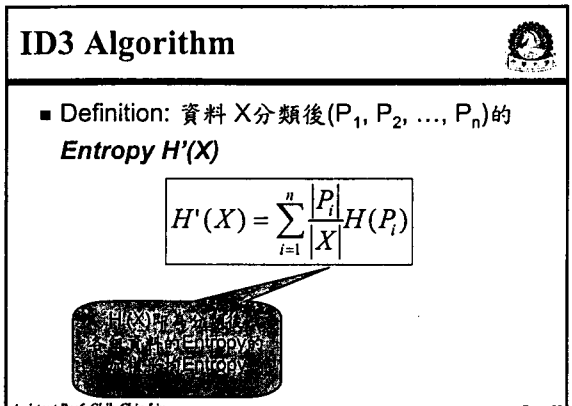


Figure 16.2 Entropy function for random variable with two equally likely outcomes.

ID3 Algorithm

- Decision Tree 中任一 ***leaf node 的 Entropy 必為 0***
- *Proof*:
  leaf node 中的所有 categorical attribute values 之值必相同, 令其值為 A
  $\therefore P_{cat\_attr=A} = 1, P_{cat\_attr\neq A} = 0$
  $\therefore$ Entropy = $1 \log_2(1) = 0$

ID3 Algorithm



- Definition: 資料 X分類後(P_1, P_2, ..., P_n)的 ***Entropy H'(X)***

$$H'(X) = \sum_{i=1}^{n} \frac{|P_i|}{|X|} H(P_i)$$

ID3 Algorithm

- Example: 12 個學生, 6 個及格, 6 個不及格; 分類後資料為 4 個研究生(3 個及格, 1個不及格), 8 個大學生(3 個及格, 5 個不及格)

$$H(\text{研究生}) = \frac{3}{4}\log_2\left(\frac{1}{\frac{3}{4}}\right) + \frac{1}{4}\log_2\left(\frac{1}{\frac{1}{4}}\right) = 0.811$$

$$H(\text{大學生}) = \frac{3}{8}\log_2\left(\frac{1}{\frac{3}{8}}\right) + \frac{5}{8}\log_2\left(\frac{1}{\frac{5}{8}}\right) = 0.954$$

$$H'(\text{學生}) = \frac{4}{12}0.811 + \frac{8}{12}0.954 = 0.907$$

ID3 Algorithm

- Definition: 分類後亂度增加量 Gain(X)

$$Gain = H(X) - H'(X)$$

ID3 Algorithm

- Example: 12 個學生, 6 個及格, 6 個不及格; 分類後資料為 4 個研究生(3 個及格, 1個不及格), 8 個大學生(3 個及格, 5個不及格), 則按學生分類的亂度增加量Gain(學生)為

$$\begin{aligned} Gain(\text{學生}) &= H(X) - H'(\text{學生}) \\ &= 1 - 0.907 \\ &= 0.093 \end{aligned}$$

ID3 Algorithm

```
function ID3 (R: a set of non-categorical attributes,
              C: the categorical attribute,
              S: a training set) returns a decision tree;
begin
    If S is empty, return a single node with value Failure;
    If S consists of records all with the same value for
       the categorical attribute,
       return a single node with that value;
    If R is empty, then return a single node with as value
       the most frequent of the values of the categorical attribute
       that are found in records of S; [note that then there
       will be errors, that is, records that will be improperly
       classified];
    Let D be the attribute with largest Gain(D,S)
       among attributes in R;
    Let {dj| j=1,2, ..., m} be the values of attribute D;
    Let {Sj| j=1,2, ..., m} be the subsets of S consisting
       respectively of records with value dj for attribute D;
    Return a tree with root labeled D and arcs labeled
       d1, d2, ..., dm going respectively to the trees

          ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ..., ID3(R-{D}, C, Sm);
end ID3;
```


ID3 Algorithm

ID3演算法在每個
決策樹的節點選取
都以分類後可以得到
Gain值最大的屬性
來建分類樹！

ID3 Algorithm

Non-Categorical Attributes | Categorical Attribute

Order	Outlook	Temperature	Humidity	Windy	Decision
1	rain	hot	high	true	*Don't Play*
2	rain	cool	normal	true	*Don't Play*
3	overcast	mild	high	true	*Play*
4	overcast	mild	normal	false	*Play*
5	rain	hot	high	false	*Play*
6	overcast	cool	normal	true	*Play*
7	sunny	hot	normal	true	*Don't Play*
8	sunny	mild	high	true	*Don't Play*
9	sunny	mild	normal	false	*Play*
10	rain	cool	normal	false	*Play*
11	rain	hot	high	false	*Play*
12	sunny	hot	high	false	*Don't Play*
13	sunny	cool	normal	false	*Don't Play*
14	rain	mild	normal	true	*Don't Play*

ID3 Algorithm

分類前資料的Entropy

$$I(7,7) = -\frac{7}{7+7}log_2\frac{7}{7+7} - \frac{7}{7+7}log_2\frac{7}{7+7} = 1$$

OUTLOOK欄位來分類可使資料的Entropy 降到最低

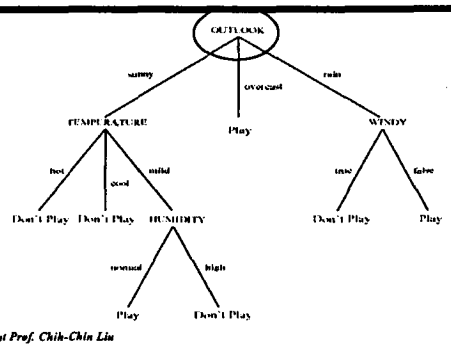
$$EI(OUTLOOK) = \frac{3+3}{14}I(3,3) + \frac{3+0}{14}I(3,0) + \frac{1+4}{14}I(1,4) = 0.69$$

$$EI(TEMPERATURE) = \frac{2+3}{14}I(2,3) + \frac{2+2}{14}I(2,2) + \frac{3+2}{14}I(3,2) = 1.21$$

$$EI(HUMIDITY) = \frac{3+3}{14}I(3,3) + \frac{4+4}{14}I(4,4) = 1$$

$$EI(WINDY) = \frac{2+5}{14}I(2,5) + \frac{5+2}{14}I(5,2) = 1.15$$

ID3 Algorithm



ID3 Algorithm

If *OUTLOOK* = *overcast* then *Play*.

If *OUTLOOK* = *rain* and *WINDY* = *true* then *Don't Play*.

If *OUTLOOK* = *rain* and *WINDY* = *false* then *Play*.

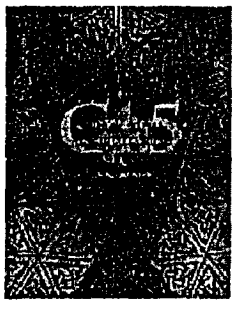
If *OUTLOOK* = *sunny* and *TEMPERATURE* = *hot* then *Don't Play*.

If *OUTLOOK* = *sunny* and *TEMPERATURE* = *cool* then *Don't Play*.

If *OUTLOOK* = *sunny* and *TEMPERATURE* = *mild* and *HUMIDITY* = *normal* then *Play*.

If *OUTLOOK* = *sunny* and *TEMPERATURE* = *mild* and *HUMIDITY* = *high* then *Don't Play*.

C4.5 Algorithm
Quinlan 經典著作



C4.5: Programs for Machine Learning

J. Ross Quinlan

ISBN 1558602380
Morgan Kaufmann

1993

Price: £ 42.95

C4.5 Algorithm

Non-Categorical Attributes | Categorical Attribute

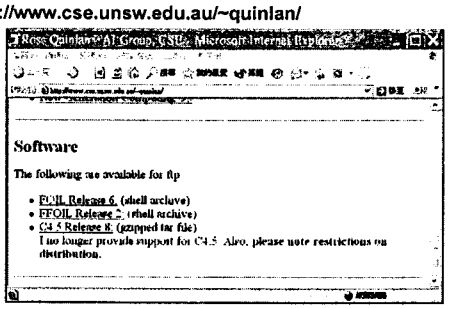
Outlook	Temperature	Humidity	Windy	Play (positive) / Don't Play (negative)
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

C4.5 Algorithm

- C4.5 演算法對 ID3 的改進之處
  - 可以處理連續值(continuous)的非分類屬性
  - 可以處理缺值(missing values)的非分類屬性

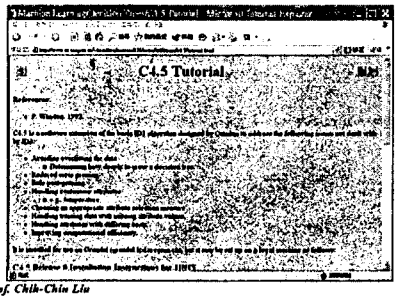
Tools: C4.5 by Quinlan
原始程式下載

http://www.cse.unsw.edu.au/~quinlan/

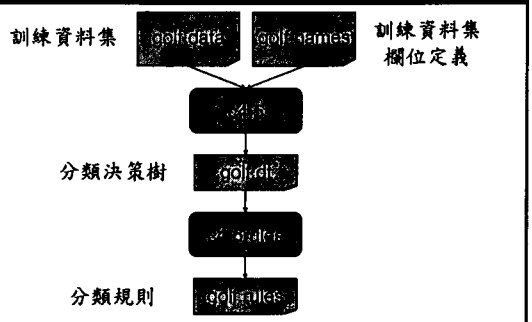


Tools: C4.5 by Quinlan
簡介與指令格式

http://www.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html

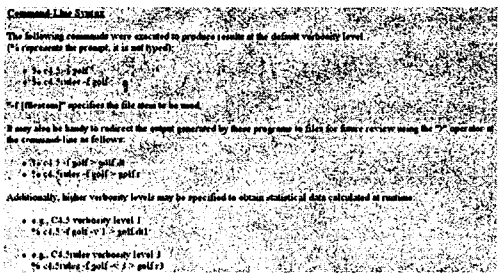


Tools: C4.5 by Quinlan
應用範例: Golf



Tools: C4.5 by Quinlan
應用範例: Golf

http://www.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/c4.5_prob1.html



Tools: C4.5 by Quinlan
應用範例: Golf

- Training Data Set: **golf.data**

```
sunny, 85, 85, false, Don't Play
sunny, 80, 90, true, Don't Play
overcast, 83, 78, false, Play
rain, 70, 96, false, Play
rain, 68, 80, false, Play
rain, 65, 70, true, Don't Play
overcast, 64, 65, true, Play
sunny, 72, 95, false, Don't Play
sunny, 69, 70, false, Play
rain, 75, 80, false, Play
sunny, 75, 70, true, Play
overcast, 72, 90, true, Play
overcast, 81, 75, false, Play
rain, 71, 80, true, Don't Play
```

Tools: C4.5 by Quinlan
應用範例: Golf

- Training Data Set: **golf.names**

```
Play, Don't Play.

outlook: sunny, overcast, rain.
temperature: continuous.
humidity: continuous.
windy: true, false.
```

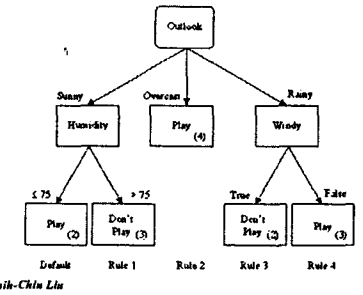

Tools: C4.5 by Quinlan
應用範例: Golf

- Result: Decision Tree **golf.data**

```
C4.5 [release 8] decision tree generator Thu Jun 15 09:15:50 2000
----------------------------------------

Options:
        File stem <golf>

Read 14 cases (4 attributes) from golf.data

Decision Tree:

outlook = overcast: Play (4.0)
outlook = sunny:
|   humidity <= 75 : Play (2.0)
|   humidity > 75 : Don't Play (3.0)
outlook = rain:
|   windy = true: Don't Play (2.0)
|   windy = false: Play (3.0)
```


Tools: C4.5 by Quinlan
應用範例: Golf

- Result: Decision Tree **golf.data**

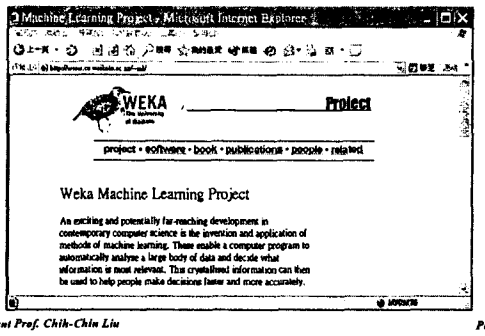


Tools: C4.5 by Quinlan
應用範例: Golf

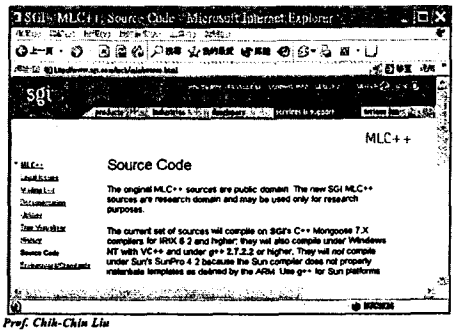
- Result: Rules **golf.rules**

```
Rule 2: outlook = overcast
        -> class Play  [70.7%]

Rule 4: outlook = rain
        windy = false
        -> class Play  [63.0%]

Rule 1: outlook = sunny
        humidity > 75
        -> class Don't Play  [63.0%]

Rule 3: outlook = rain
        windy = true
        -> class Don't Play  [50.0%]

Default class: Play
```

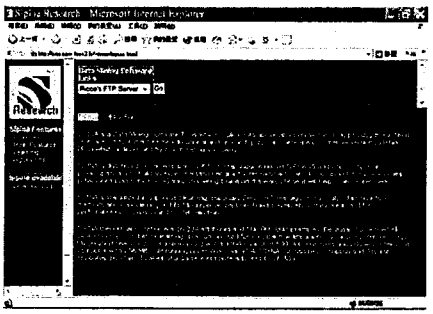

Tools: WEKA
http://www.cs.waikato.ac.nz/~ml/



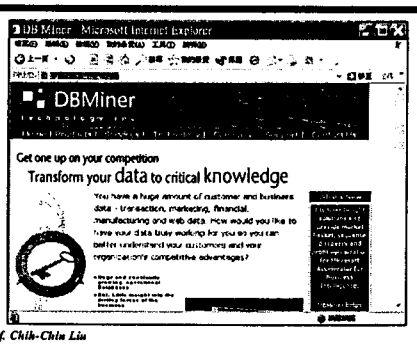
Tools: MLC++
http://www.sgi.com/tech/mlc/



Tools: SIPINA
http://eric.univ-lyon2.fr/~ricco/sipina.html

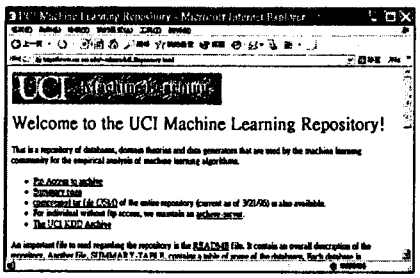


Tools: DBMiner
http://www.dbminer.com/



Data Set Repository
http://www.ics.uci.edu/~mlearn/

- Training Data Sets 集整之處



References

- [Han00] Jiawei Han and Micheline Kamber, "Chap 7 Classification and Prediction," *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000. (Decision Tree, ID3 簡介)
- [Quinlan93] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993. (C4.5 經典著作)
- [Winston92] Winston,P.H., Artificial Intelligence, 3ed, Addison-Wesley, 1992. (ID3 簡介)
- http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html (C4.5 簡介)
- http://www.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/c4.5_prob1.html (C4.5 實作範例)

Chap 3 Biological Data Classification Using Decision Tree Based Classifiers

劉志俊 (Chih-Chin Liu)

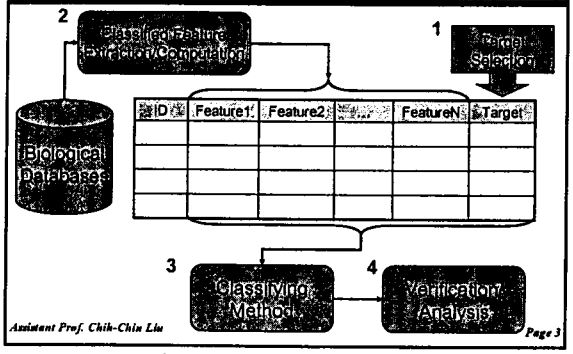
中華大學資訊工程系

March 2003

Outline

- Research Issues
- 蛋白質關鍵字自動註解: [Kretschmann01]
- 基因功能分類與預測: [King00]
- 蛋白質功能分類與預測: [King01]
- 插入子/表現子邊界預測: [Ting94]

Research Issues in Biological Data Classification

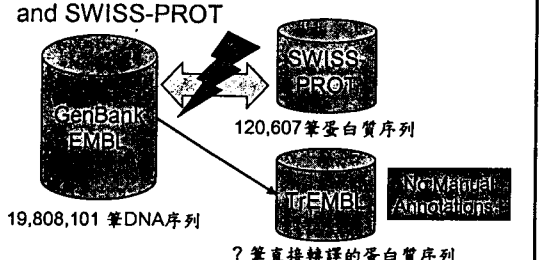


Research Issues in Biological Data Classification

- Target Selection: 分類標的為何？
- Feature Extraction: 如何由生物資料庫中擷取或計算分類用的特徵值？
- Classifying Method: 採用哪種分類演算法比較適當？
- Verification/Analysis: 如何驗證/分析所採用分類方法與分類結果的好壞？

[Kretschmann01]

- Motivation: ***Gap*** between GenBank/EMBL and SWISS-PROT



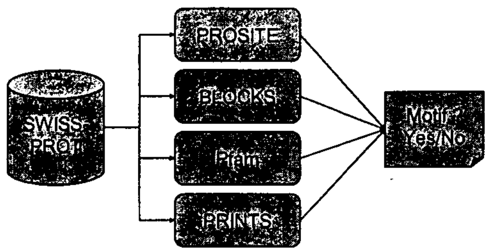
[Kretschmann01]

- Target: Protein Annotations (Keywords)

SWISS-PROT ID	Prosite pattern PS00487	Pfam pattern PF01493	Mammalia	FAD
Q9ZNZ7	-	yes	-	yes
Q9JI00	-	-	yes	-
Q9ZL14	yes	-	-	-
Q9NYQ3	-	-	yes	-
Q9UJM8	-	-	yes	-
Q9NYQ2	-	-	yes	-
Q43155	-	yes	-	yes
Q9TOP4	-	yes	-	yes
Q9WU19	-	-	yes	-

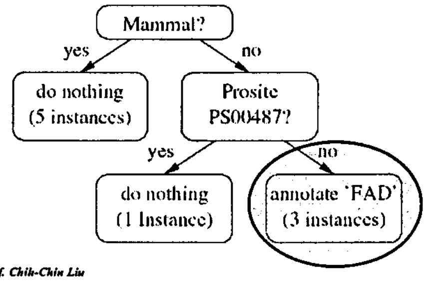
[Kretschmann01]

- Feature Extraction: ***Motif*** Identification



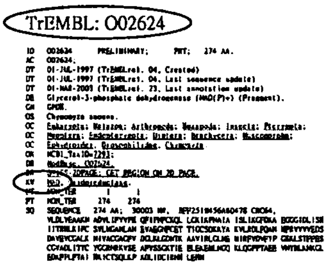
[Kretschmann01]

- Classification Method: C4.5 (Weka)



[Kretschmann01]

- Application: TrEMBL Keywords

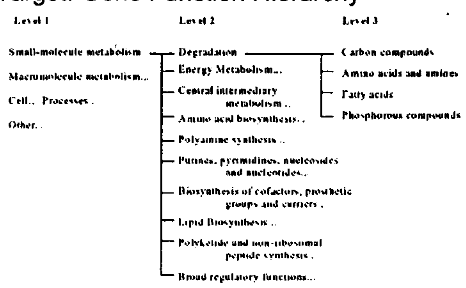


[King00]

- Motivation:
  - For the existing sequenced genomes function can be assigned to typically only between ***40-60% of the genes.***
  - The new science of functional genomics is dedicated to discovering the function of these genes, and to further detailing gene function.
  - A novel data-mining approach to predicting protein functional class from sequence is presented.

[King00]

- Target: Gene Function Hierarchy



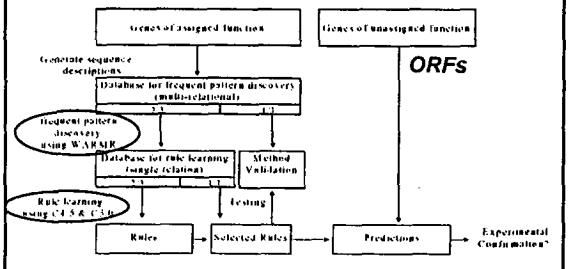
[King00]

- Feature Extraction:

Database argument	Description
hom(A)	refers to a homologous protein (*A*) found by PSI-BLAST
keyword(A, Word)	refers to a SwissProt keyword found in *A*
classification(A, Class)	refers to the phylogenic classification of the organism *A* came from, taken from SwissProt
species(A, Species)	refers to the species of *A*, taken from SwissProt.
mol_wt_rule(A, Weight)	refers to the molecular weight of *A*: 1 very low, 2 low, 3 medium, 4 high, and 5 very high.
amino_acid_ratio_rule(A, Residue, Weight)	refers to the percentage composition of the residue in the sequence.
e_val_rule(A, Weight)	refers to the PSI-Blast sequence similarity measure (note that a low value means a high sequence similarity).
e_val_gt(Value) e_val_lteq(Value)	refers to the PSI-Blast sequence similarity measure, greater than or less than/equal to a certain value
mol_wt_lteq(A, Weight) mol_wt_gt(A, Weight)	refers to the molecular weight of *A* being greater than or less than/equal to some value
amino_acid_pairs_wg()	and others similar, refers to the number of pairs of these two amino acids, in this case tryptophan and glycine
amino_acid_pair_ratio_qh (Ratio)	and others similar, refers to the ratio of one amino acid to another in the gene, in this case the ratio of glutamine(q) to histidine(h). This ratio is not a percentage, not out of a hundred, instead it's a ratio out of a thousand. So for example 2.8 means 0.28%.
amino_acid_ratio_g (Percentage)	and others similar, refers to the percentage composition of the residue in the sequence of the gene, in this case the percentage of glycine
psi_iter_gt(Number) psi_iter_lteq(Number)	refers to the number of iterations of the PSI_BLAST search (greater than or less than/equal to some number)

[King00]

- Classification Method: C4.5 and C5.0



[King00]

- Classification Method: C4.5 and C5.0

If
there exists a homologous protein in SwissProt with the keyword "membrane" and
there exists a homologous protein in *Bacillus subtilis* and there does not exist a homologous protein with very low molecular weight, a large percentage of glutamic acid, and medium sequence similarity
and there does not exist a homologous protein in SwissProt with good sequence similarity, low percentage of cysteine, the keyword "transmembrane" and a fairly high molecular weight
there does not exist a firmicutes sp. protein in SwissProt with the keyword "transmembrane", with medium molecular weight, and a very high amount of low entropy sequence
and there exists a homologous mammalian protein in SwissProt with the keyword "repeat" with very high molecular weight
Then
the ORF has the functional class "Degradation of macromolecules"

[King00]

- Verification/Analysis: 預測 65% 功能未知的ORFs, 正確率 60-80% (使用 *Mycobacterium tuberculosis*)

	Level 1	*Level 2*	*Level 3*	*Level 4*
Number of rules found	25	30	20	3
Rules predicting more than one homology class	19	18	8	1
Rules predicting a new homology class	14	15	1	0
Average test accuracy	62%	65%	62%	76%
Default test accuracy	48%	14%	6%	2%
New functions assigned	886 (58%)	507 (33%)	60 (4%)	19 (1%)

[King01]

- [King00]的改進版: 特徵值分為三類
  - SEQ: 與氨基酸組成成分相關統計特徵值, 933 個整數或浮點數屬性
  - SIM: 與氨基酸序列相似度(PSI-BLAST)相關特徵值, 13799 個布林值屬性
  - STR: 與蛋白質二次結構相關特徵值, 18342個布林值屬性

[King01]

- SEQ: 與氨基酸組成成分相關統計特徵值

Attribute	Description	No.	Type
amino_acid(P, R)	The number of residues of type R in the sequence	21	int
amino_acid_perc(P, R)	The percentage composition of residues of type R	21	real
amino_acid_pair(P, R1, R2)	The number of residue pairs of type R1, R2 in the sequence	441	int
amino_acid_pair_perc(P, R1, R2)	The percentage composition of residue pairs of type R1, R2	441	real
sequence_length	The number of residues in the sequence	1	int
molecular_weight	The computed molecular weight	1	int
aliphatic_index	The computed aliphatic index	1	real
hydro	The Grand average of hydropathicity (GRAVY), the value was discretized. 1 for low values, increasing up to 5 for high values.	1	int
pI	The theoretical isoelectric point (pI) for this ORF	1	real
atomic_comp(P, E)	The ORF's atomic composition of element E; where E is one of the following: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), or sulfur (S)	5	int

[King01]

- SIM: 與氨基酸序列相似度相關特徵值

Database Segment	Description
hom(P)	P is a homologous protein found by PSI-BLAST
e_val_sub(P, E)	P is a homologous protein found by PSI-BLAST with SIM measure E
e_val_hog(P, X)	P is a homologous protein found by PSI-BLAST with SIM measure less than X
e_val_gt(P, X)	P is a homologous protein found by PSI-BLAST with SIM measure greater than X
pid_e_val_sub(P, H)	P is a homologous protein found by PSI-BLAST on Iteration H
pid_hom_hog(P, X)	P is a homologous protein found by PSI-BLAST on Iteration less than X
pid_hom_gt(P, X)	P is a homologous protein found by PSI-BLAST on Iteration greater than X
species(P, Species)	The protein P comes from a species Species
classification(P, Class)	The protein P comes from a species with SwissProt phylogenic classification Class
mol_wt_sub(P, X)	The protein P has determined molecular weight X
mol_wt_hog(P, X)	The molecular weight of P is less than X
mol_wt_gt(P, X)	The molecular weight of P is greater than X
keyword(P, Word)	The SwissProt keyword Word describes protein P

[King01]

- STR: 與蛋白質二次結構相關特徵值

Datatype argument	Description
sec(S, T)	Position s in protein is predicted to be a secondary structure element of type T
sec(S1, S2, T)	Given the secondary structure at position S1, the length-using position S2, with S2 = S1 + 1, has a secondary structure prediction of type T
s_alpha(S, pc, X)	Position S is predicted to be an alpha-helix of length greater than X (similarity freq instead of pc)
s_beta(S, pc, X)	Position S is predicted to be a beta-strand of length greater than X (similarity freq instead of pc)
s_coil(S, pc, X)	Position S is predicted to be a coil of length greater than X (similarity freq instead of pc)
sec_alpha(S1, S2, pc, X)	Positions S1 and S2 (where S2 = S1 + 2) are predicted to be alpha-helices, S2 has length greater than X (similarity freq instead of pc)
sec_beta(S1, S2, pc, X)	Positions S1 and S2 (where S2 = S1 + 2) are predicted to be beta-strands, S2 has length greater than X (similarity freq instead of pc)
sec_coil(S1, S2, pc, X)	Positions S1 and S2 (where S2 = S1 + 2) are predicted to be coils, S2 has length greater than X (similarity freq instead of pc)

[King01]

- Verification/Analysis: 比較哪一種特徵分類效果較好

使用*E. Coli* 基因組

若限用一類特徵值來分類則序列相似度SIM結果最好

Attributes	Accuracy %			Coverage %			No. of Predictions		
	1	2	3	1	2	3	1	2	3
SEQ	64	63	41	20	18	4	259 (17)	345 (11)	63 (3)
SIM	75	74	69	29	26	16	250 (13)	288 (13)	152 (7)
STR	59	44	17	10	1	5	149 (7)	38 (2)	74 (3)
SEQ + SIM	84	71	60	23	26	16	231 (11)	272 (13)	115 (5)
SEQ + STR	69	64	50	20	22	3	317 (15)	401 (19)	37 (2)
SIM + STR	75	69	54	25	27	20	195 (9)	301 (14)	152 (7)
SEQ + SIM + STR	75	69	61	28	26	15	353 (16)	269 (12)	135 (6)
WTD_VOTE_ALL	60	54	42	53	38	36	863 (40)	818 (38)	475 (22)
VOTE_ALL	75	66	68	32	34	17	400 (18)	377 (17)	122 (6)
WTD_VOTE_SIM	64	66	52	41	34	22	626 (29)	462 (21)	296 (14)
VOTE_SIM	86	88	90	12	11	1	57 (3)	91 (4)	16 (1)

多數決定(兩票以上)投票制的正確率最高

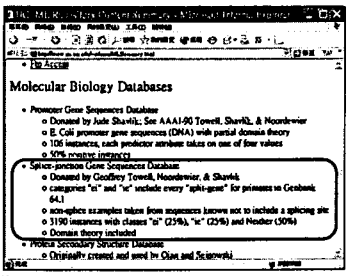
(依正確性)加權投票制的涵蓋率最高

[Ting94]

- Motivation:
  - DNA => Protein 的轉譯過程中, Intron/Exon 的決定很重要
  - 使用Data Mining 技術預測 Intron/Exon 以及 Exon/Intron 的邊界
  - C4.5 執行速度快, 但 ID2-of-3 的正確率較高
  - 設計稱為 MoN 的演算法, 希望能兼具C4.5 與 ID2-of-3 的優點

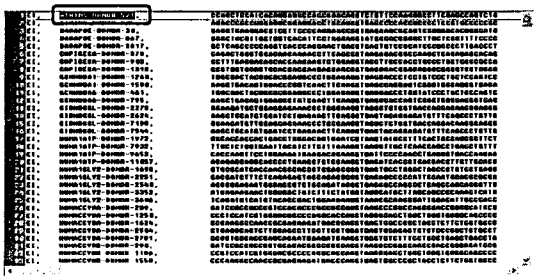
[Ting94]

- Data Set: *UCI Machine Learning Repository*



[Ting94]

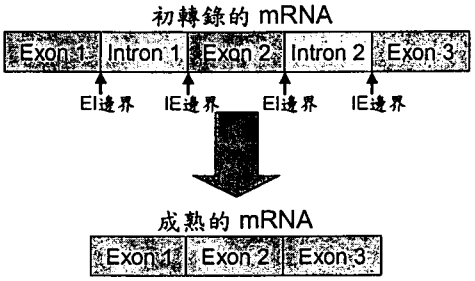
- Data Set: *UCI Machine Learning Repository*



[Ting94]

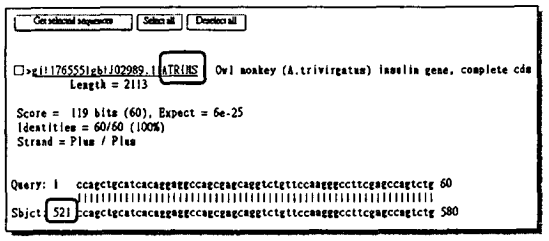
- Data Set: *UCI Machine Learning Repository*

初轉錄的 mRNA



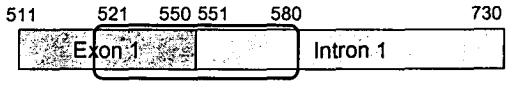
[Ting94]

- Source: *Genbank*



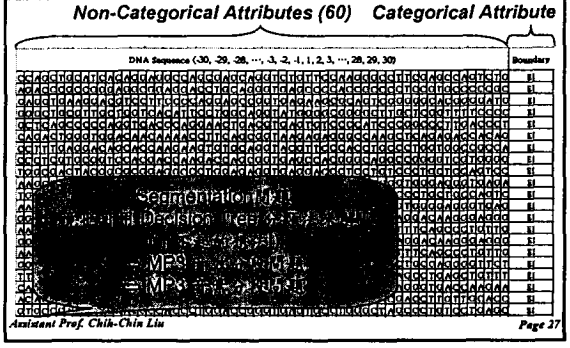
[Ting94]

- Source: *Genbank*




```
FEATURES             Location/Qualifiers
     source          1..2113
                     /organism="Aotus trivirgatus"
                     /db_xref="taxon:9505"
     prim_transcript 511..2008
                     /note="ins mRNA and introns"
     intron          551..730
                     /note="ins intron A"
```


[Ting94]



[Ting94]

- Result: 正確率 vs. 執行時間

	C4.5rules	*MoN*SA*	*MoN*G*	*ID2-of-3*
%Err	11.5 σ2.6	9.7 σ2.0	10.5 σ1.6	9.5 σ1.7
#Cnpt	32.5 σ2.8	20.0 σ3.5	26.7 σ4.6	12.4 σ0.7
Time	18.0	93.6	19.8	23546.7

Table 1: Result Summary for Splice-junction domain

演算法確有成效，但效果非常有限

C4.5rules # *rules*	*MoN*SA* [#m-n:#mv]	*MoN*G* [#m-n:#mv]	#*nodes*
328	200[100:93]	267[146:145]	118[50:9]

Table 2: Types of rules induced

References

- [Quinlan93] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993. (C4.5 經典著作)
- [Kretschmann01] E. Kreschmann, W. Fleischmann, and R. Apweiler,"Automatic Rule Generation for Protein Annotation with the C4.5 Data Mining Algorithm Applied on SWISS-PROT," *Bioinformatics*, Vol. 17, No. 10, pp. 920-926, 2001.
- [Ting94] K. M. Ting, "A M-of-N Rule Induction Algorithm and Its Application to DNA Domain," in Proc. of the Hawaii Intl. Conf. on System Sciences, pp. 133-140, 1994.

References

- [King01a] R. D. King, *et. al.*, "The Utility of Different Representations of Protein Sequence for Predicting Functional Class," *Bioinformatics*, Vol. 17, No. 5, pp. 445-454, 2001. ([King00] 加強版)
- [Dehaspe98] Dehaspe, L., Toivonen, H., & King, R.D. "Finding frequent substructures in chemical compounds," In: *The Fourth International Conference on Knowledge Discovery and Data Mining*. R. Agrawal, P. Stolorez, & G. Piatetsky (eds.) AAAI Press, Menlo Park. 30-36, 1998.
- [King00] King, R.D., Karwath, A., Clare, A., & Dehaspe, L, "Genome scale prediction of protein functional class from sequence using data mining," In *ACM Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 384-389, 2000.
- [King01b] King, R.D., Srinivasan, A., & Dehaspe, L. "Warmr: A Data Mining Tool for Chemical Data." *Journal of Computer-Aided Molecular Design*. 15, 173-181, 2001. ([King00][King01a][Dehaspe98] 等文所使用的方法)

Chap 3 Data Mining Supports in SQL Server

劉志俊 (Chih-Chin Liu)

中華大學 資訊工程系

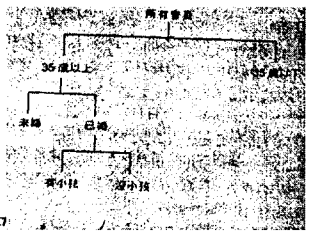
February 2003

SQL Data Mining

- Data Mining : found hidden knowledge from data base.
- Technique : using Data Mining Model to analysis.
- Model : DS for describing analysis object.
- Data Mining tech. :
  - neural net
  - decision tree
  - genetic algorithm
  - generalization

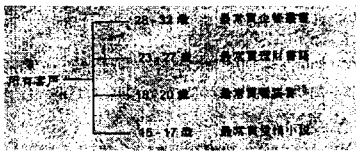
SQL Data Mining

- Data Mining Model in Analysis Services :
  - decision tree :



SQL Data Mining

- clustering :



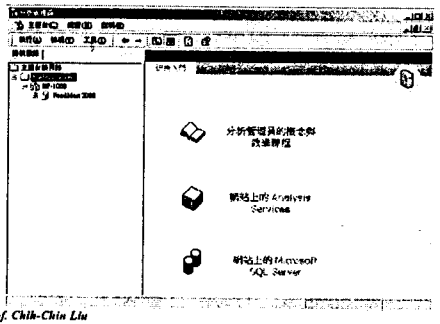
SQL Data Mining

- SQL data mining object
  - cube
  - table in data base

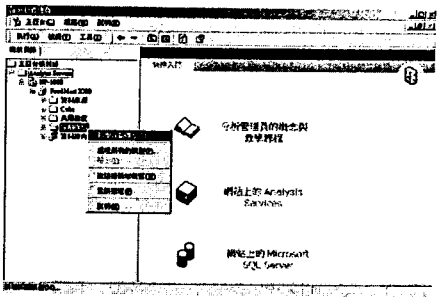
SQL Data Mining

- Analysis Services

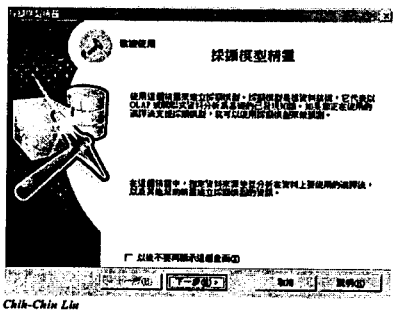
Analysis Services



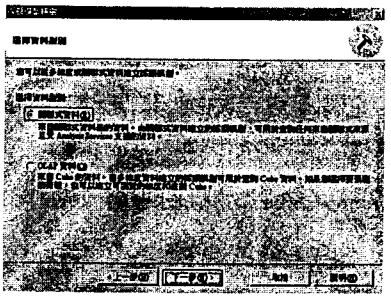
Build up model from table



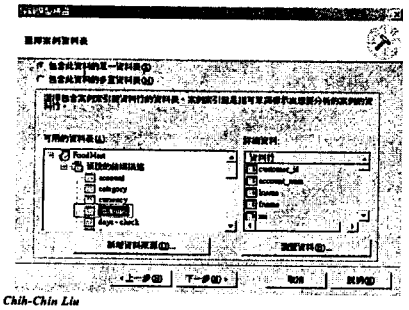
Build up model from table



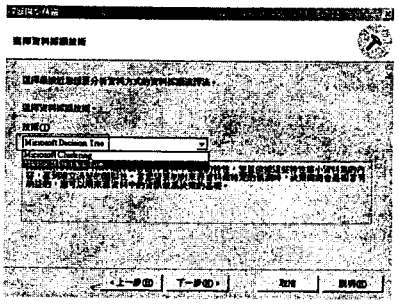
Build up model from table



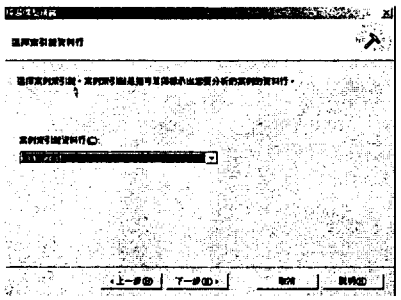
Build up model from table



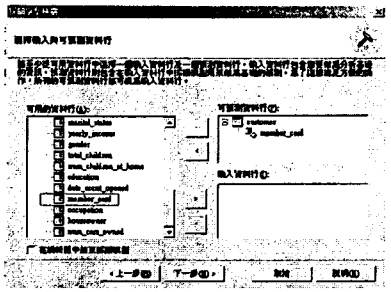
Build up model from table



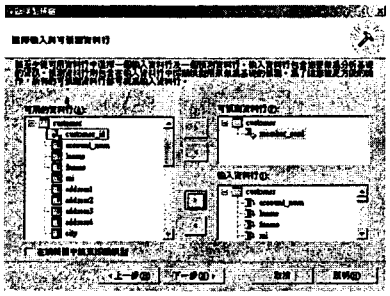
Build up model from table



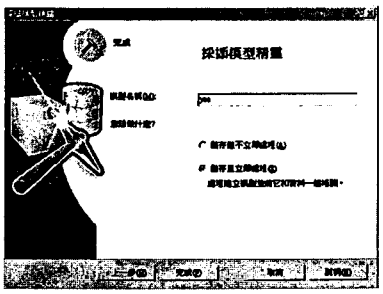
Build up model from table



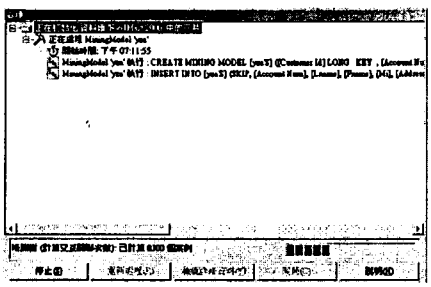
Build up model from table



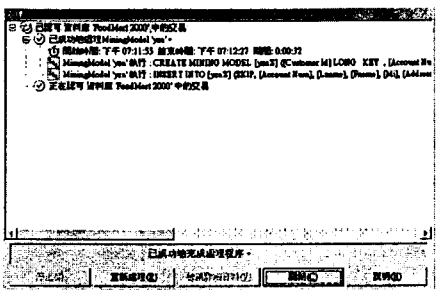
Build up model from table



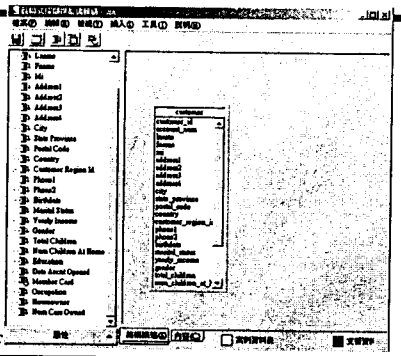
Build up model from table

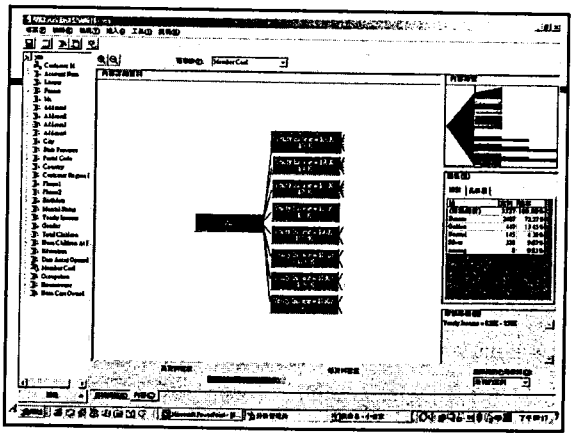


Build up model from table

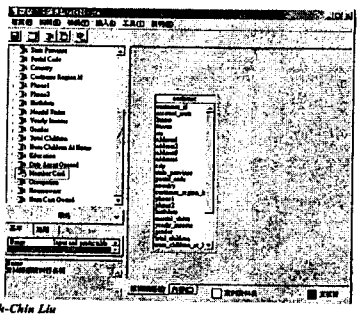


Build up model from table

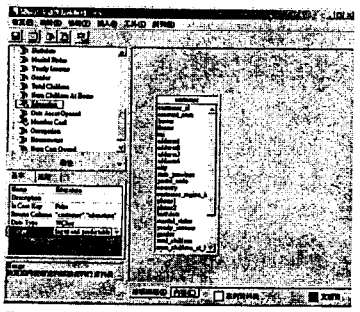




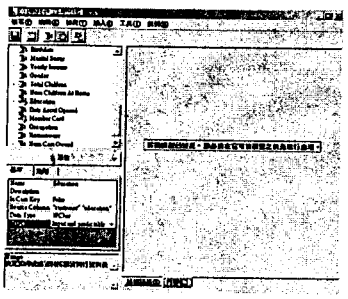
Build up model from table



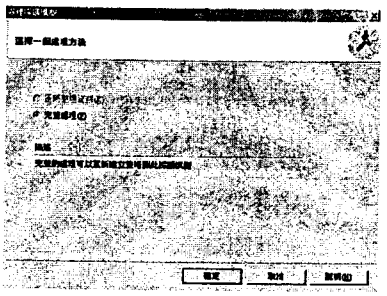
Build up model from table



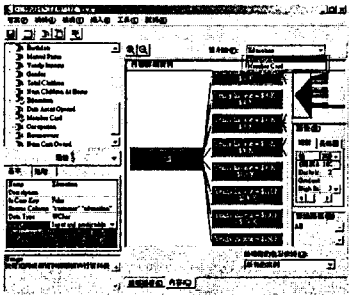
Build up model from table

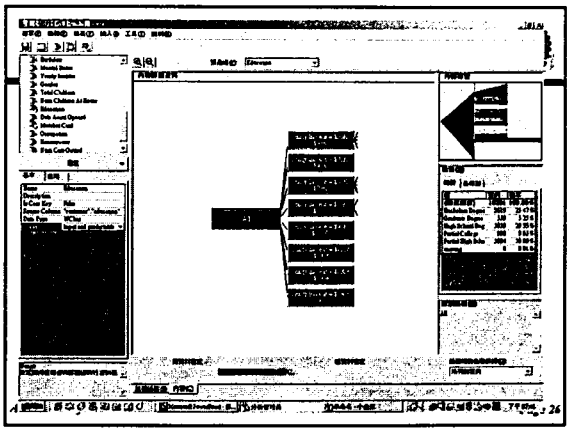


Build up model from table



Build up model from table





Chap 4 ＷＥＫＡ簡介

指導老師：劉志俊

秧雞生活介紹

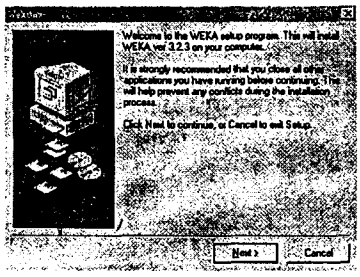
- 秧雞科的鳥類都喜歡在沼澤、稻田及茭白荀田等淺水地帶過生活
- 中國江南魚米之鄉，秧雞科的種類及數量都很豐富
- 在清大園裡，經常出現在園科區圍牆外的沼澤裡，大清晨偶爾會飛到原子爐前的睡蓮池裡覓食

白腹秧雞

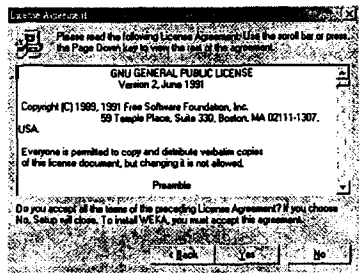
程式下載

- WEKA首 http://www.cs.waikato.ac.nz/~ml/
- 下載首頁 http://www.cs.waikato.ac.nz/~ml/weka/index.html
- 不含虛擬機器
  http://prdownloads.sourceforge.net/weka/weka-3-2-3.exe
- 含虛擬機器
  http://prdownloads.sourceforge.net/weka/weka-3-2-3jre.exe

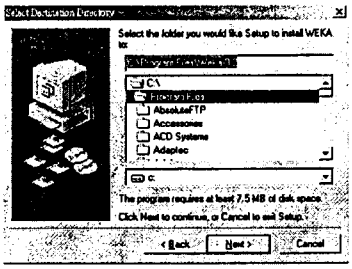
安裝流程



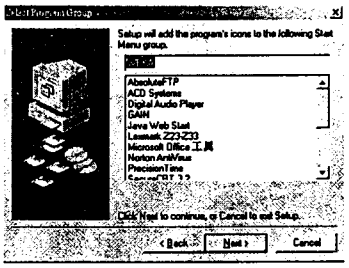
安裝流程



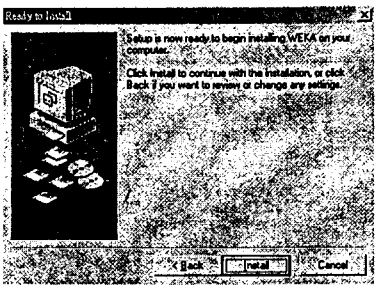
安裝流程



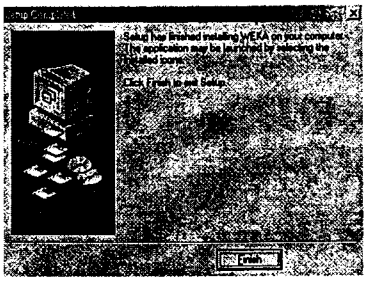
安裝流程



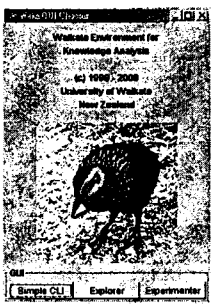
安裝流程



安裝流程



執行畫面



執行畫面

含有的演算法

- package **weka.classifiers.adtree**
- package **weka.classifiers.evaluation**
- package **weka.classifiers.j48**
- package **weka.classifiers.kstar**
- package **weka.classifiers.m5**
- package **weka.classifiers.neural**

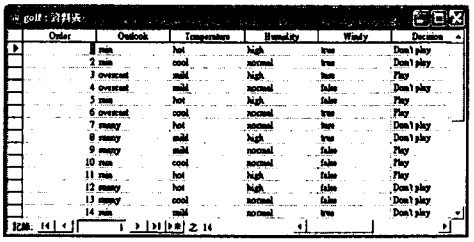
檔案格式

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@attribute test {y,n}

@data
sunny,85,85,FALSE,no,y
sunny,80,90,TRUE,no,n
overcast,83,86,FALSE,yes,y
rainy,70,96,FALSE,yes,y
rainy,68,80,FALSE,yes,y
rainy,65,70,TRUE,no,n
overcast,64,65,TRUE,yes,y
sunny,72,95,FALSE,no,n
sunny,69,70,FALSE,yes,y
rainy,75,80,FALSE,yes,y
sunny,75,70,TRUE,yes,y
overcast,72,90,TRUE,yes,y
overcast,81,75,FALSE,yes,y
rainy,71,91,TRUE,no,n
```

WEKA-與資料庫連結應用

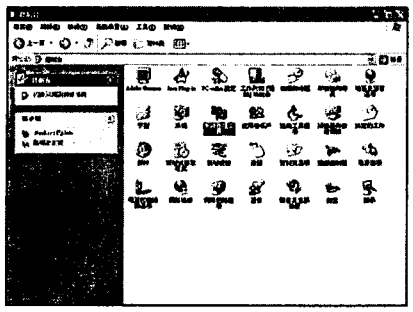
指導老師:劉志俊

Golf(使用access)



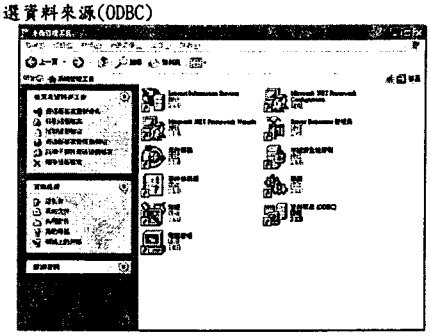
Step 1: Create a User DSN

- 打開控制台-系統管理工具



Step 1: Create a User DSN

- 點選資料來源(ODBC)

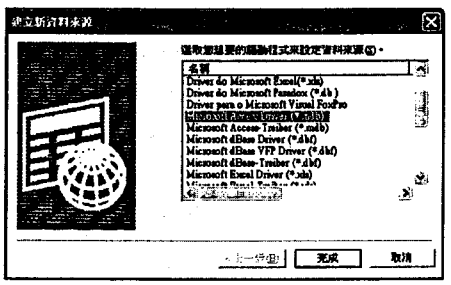


Step 1: Create a User DSN

- 選擇新增

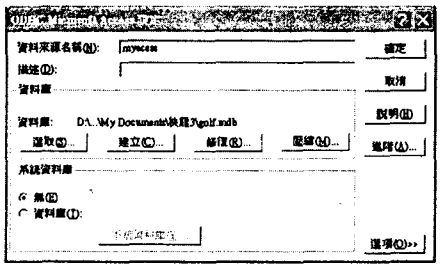
Step 1: Create a User DSN

- 選擇資料庫的驅動程式



Step 1: Create a User DSN

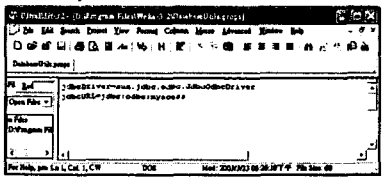
- 選擇資料庫



Step 2: Set up the DatabaseUtils.props file

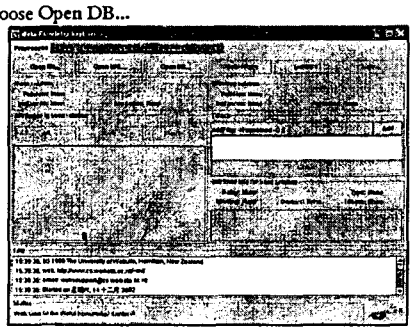
- The file is a text file that needs to contain the following lines:

jdbcDriver=sun.jdbc.odbc.JdbcOdbcDriver
jdbcURL=jdbc:odbc:*dbname*

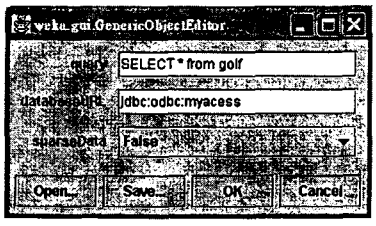


Step 3: Open the database

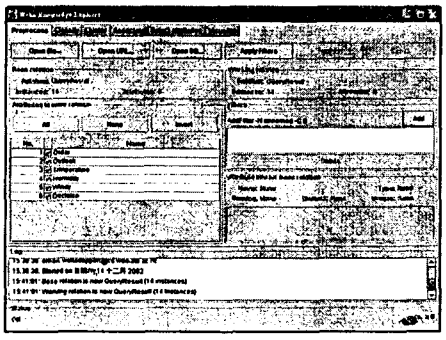
- Choose Open DB...



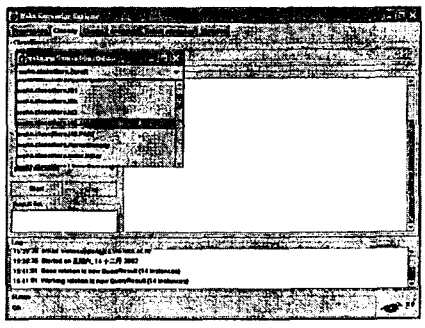
Step 3: Open the database



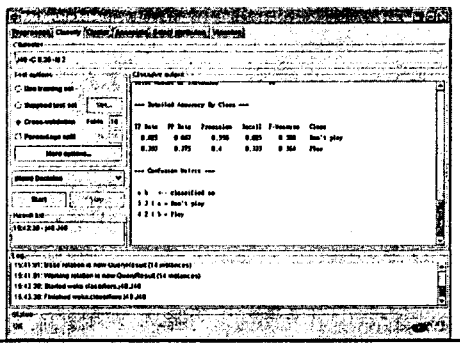
Step 3: Open the database



開始做分類-選擇C4.5



START!



Chap 6 Hidden Markov Model

劉志俊 (Chih-Chin Liu)

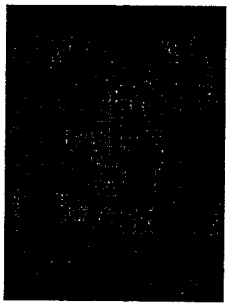
中華大學 資訊工程系

April 2003

Outline

- HMM Tutorial
- Observable Markov Model
- Hidden Markov Model
- HMM 三大問題
  - 比對問題
  - 解釋問題
  - 訓練問題

語音處理與 HMM 經典書籍



Fundamentals of Speech Recognition

Lawrence Rabiner, and Biing-Hwang Juang

Pearson Education POD

1 edition, 1993

ISBN: 0130151572

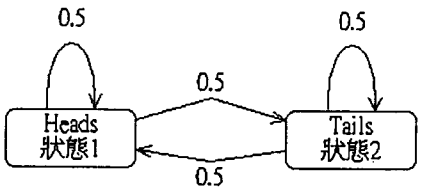
HMM Tutorial 經典論文

Proceedings of IEEE, Vol. 77, No. 2, pp. 257-286, Feb. 1989

A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

Observable Markov Model 範例一: The Coin-Toss Model



公平的銅板: 正反面出現機率一樣

Observable Markov Model 範例一: The Coin-Toss Model

- A Set of *N* Distinct ***States***

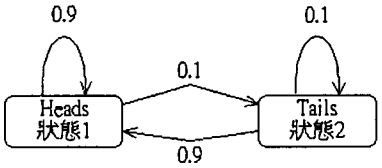
  {1, 2 } = {Heads, Tails}

- Matrix of ***State-Transition Probabilities***

$$A = \{a_{ij}\} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

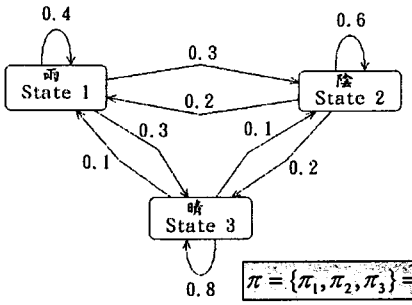
a_{ij} 表示由狀態 i 轉變為狀態 j 的機率

Observable Markov Model
範例一: The Coin-Toss Model



作弊用的銅板: 正面出現機率為0.9

Observable Markov Model
範例二: The Weather Model



$\pi = \{\pi_1, \pi_2, \pi_3\} = \{0,0,1\}$

Observable Markov Model
範例二: The Weather Model

- Set of ***States***

  {1, 2, 3 } = {雨, 陰, 晴}

- Matrix of ***State-Transition Probabilities***

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

Observable Markov Model
範例二: The Weather Model

- ***First Order*** Markov Chain

$$a_{ij} = P[q_t = j \mid q_{t-1} = i]$$
$$= P[q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots]$$

- Properties of State-Transition Probabilities

$$a_{ij} \geq 0 \quad \forall i, j$$
$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$$

Observable Markov Model
範例二: The Weather Model

- 問題1: 連續 8 天的天氣為"晴-晴-晴-雨-雨-晴-陰-晴"的機率有多大?

Sol: 令 O = (晴, 晴, 晴, 雨, 雨, 晴, 陰, 晴)

= (3, 3, 3, 1, 1, 3, 2, 3)

P(O | Model) = P((3, 3, 3, 1, 1, 3, 2, 3) | Model)

= P[3]P[3|3]P[3|3]P[1|3]P[1|1]P[3|1]P[2|3]P[3|2]

$= \pi_3 a_{33} a_{33} a_{13} a_{11} a_{31} a_{23} a_{32}$

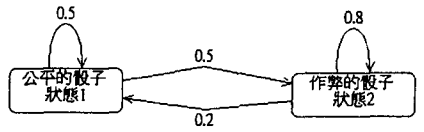
= (1.0)(0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2)

$= 1.536 \times 10^{-4}$

Hidden Markov Model

- ***Observable Markov Model***: The output of each state is not random. (沒有例外, 欠缺彈性)
- This model is too restrictive to be applicable to many problems.
- ***Hidden Markov Model***: The observation (output) is a probabilistic function of state.
  - Doubly embedded stochastic process
    - **State Transition: Hidden**
    - **Symbol Emission: Observable**

Hidden Markov Model
範例三: The Coin-Toss Model



P(H|公平的骰子) = 0.5
P(T|公平的骰子) = 0.5

P(H|作弊的骰子) = 0.9
P(T|作弊的骰子) = 0.1

使用兩顆骰子

Hidden Markov Model
範例三: The Coin-Toss Model

- Set of *States*
  S = {1, 2} = {公平的骰子, 作弊的骰子}
- Set of *Symbols*
  V = {H, T}
- Matrix of *State-Transition Probabilities*

$$A = \{a_{ij}\} = \begin{bmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix}$$

- Matrix of *Symbol Emission Probabilities*

P(H|公平的骰子) = 0.5	P(T|公平的骰子) = 0.5
P(H|作弊的骰子) = 0.9	P(T|作弊的骰子) = 0.1

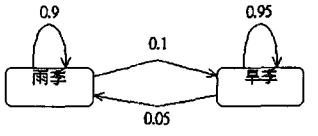
- The Initial state distribution

$$\pi = \{\pi_1, \pi_2\} = \{0.5, 0.5\}$$

Hidden Markov Model
常用Set of Symbols

- 銅板實驗: V = { H, T }
- 天氣實驗: V = { 晴, 陰, 雨 }
- 音樂實驗: V = { C, #C, D, #D, E, F, #F, G, #G, A, #A, B }
- DNA實驗: V = { A, T, C, G }
- 蛋白質序列實驗: V = { A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y }
- 蛋白質結構實驗: V = { α, β, turn }

Hidden Markov Model
範例四: The Weather Model



P(晴|雨季) = 0.01	P(晴|旱季) = 0.98
P(陰|雨季) = 0.09	P(陰|旱季) = 0.015
P(雨|雨季) = 0.9	P(雨|旱季) = 0.005

$\pi_{雨季} = 0.5$ $\pi_{旱季} = 0.5$

Hidden Markov Model
範例四: The Weather Model

- 問題2: 連續 8 天的天氣為"晴-晴-晴-雨-雨-晴-陰-晴"出現在雨季與旱季的機率各有多大?

Sol: 令 O = (晴, 晴, 晴, 雨, 雨, 晴, 陰, 晴)

Q_1 = (雨季, 雨季, 雨季, 雨季, 雨季, 雨季, 雨季, 雨季)

$P(O, Q_1|\text{Model}) = \pi_{雨季} P(晴|雨季)P(雨季|雨季)\ P(晴|雨季)$
$P(雨季|雨季)P(晴|雨季)P(雨季|雨季)P(雨|雨季)$
$P(雨季|雨季)P(雨|雨季)P(雨季|雨季)P(晴|雨季)$
$P(雨季|雨季)P(陰|雨季)P(雨季|雨季)P(晴|雨季)$

=(0.5)(0.01)(0.9)(0.01)(0.9)(0.01)(0.9)(0.9)(0.9)(0.9)
(0.9)(0.01)(0.9)(0.09)(0.9)(0.01) = $1.7433922005 \times 10^{-12}$

Hidden Markov Model
範例四: The Weather Model

Sol: 令 O = (晴, 晴, 晴, 雨, 雨, 晴, 陰, 晴)

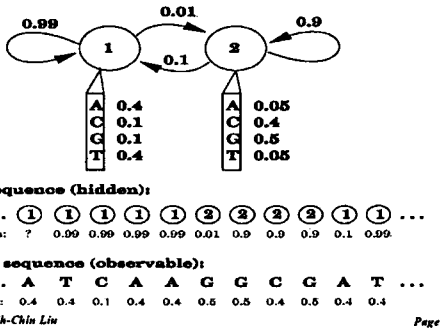
Q_2 = (旱季, 旱季, 旱季, 旱季, 旱季, 旱季, 旱季, 旱季)

$P(O, Q_2|\text{Model}) = \pi_{旱季} P(晴|旱季)P(旱季|旱季)\ P(晴|旱季)$
$P(旱季|旱季)P(晴|旱季)P(旱季|旱季)P(雨|旱季)$
$P(雨季|旱季)P(雨|旱季)P(旱季|旱季)P(晴|旱季)$
$P(雨季|旱季)P(陰|旱季)P(旱季|旱季)P(晴|旱季)$

=(0.5)(0.98)(0.95)(0.98)(0.95)(0.98)(0.95)(0.005)
(0.95)(0.005)(0.95)(0.98)(0.95)(0.015)(0.95)(0.005)

$=6.03866331428782 \times 10^{-10}$

Hidden Markov Model
範例五: The DNA Model



HMM 三大問題

Problem 1: 比對問題	Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O\|\lambda)$, the probability of the observation sequence, given the model?
Problem 2: 解釋問題	Given the observation sequence $O = O_1 O_2 \cdots O_T$, and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 \cdots q_T$ which is optimal in some meaningful sense (i.e., best "explains" the observations)?
Problem 3: 訓練問題	How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O\|\lambda)$?

HMM 三大問題
Problem 1: 比對問題

- 已知一 HMM λ 與一觀察序列 (observation sequence) $O = (o_1, o_2, \ldots, o_T)$, 則在此 HMM 模型下出現 O 的機率 $P(O \mid \lambda)$ 為何？
- 解法一: 暴力法
  - 1. 求在某一狀態序列 $q = (q_1, q_2, \ldots, q_T)$ 下出現O 的機率 $P(O \mid q, \lambda)$
  - 2. 求此 HMM 所有可能的狀態序列 q, 則$P(O \mid \lambda)$為所有狀態序列 q 出現 O 的機率的總和

HMM 三大問題
Problem 1: 比對問題

$$\begin{cases} P(O \mid q, \lambda) = \prod_{t=1}^{T} P(O \mid q_t, \lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \\ P(q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \\ P(O, q \mid \lambda) = P(O \mid q, \lambda) P(q \mid \lambda) \end{cases}$$

$$\Rightarrow P(O, q \mid \lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

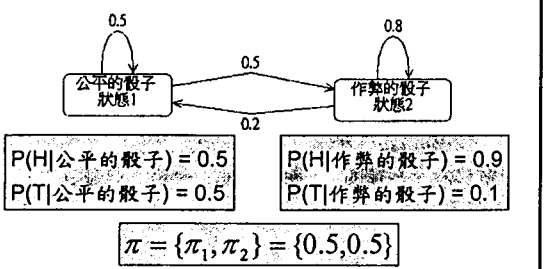
$$P(O \mid \lambda) = \sum_{all\ q} P(O \mid q, \lambda) P(q \mid \lambda) \quad O(2TN^T)$$

$$= \sum_{q_1 q_2 \cdots q_T} b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

$O(N^T)$ $O(2T \cdot N^T)$

HMM 三大問題
Problem 1: 比對問題

- 問題三: 出現 O = HHH 的機率為何？



$$\pi = \{\pi_1, \pi_2\} = \{0.5, 0.5\}$$

HMM 三大問題
Problem 1: 比對問題

Sol: $O(N^T) = 2^3 = 8$ 種 $O(2TN^T) = 6 \times 8 = 48$

$P(O,q_1q_1q_1|\lambda) = b_1(H)b_1(H)b_1(H)\pi_1 a_{11}a_{11} = (0.5)(0.5)(0.5)(0.5)(0.5)(0.5) = 0.015625$

$P(O,q_1q_1q_2|\lambda) = b_1(H)b_1(H)b_2(H)\pi_1 a_{11}a_{12} = (0.5)(0.5)(0.9)(0.5)(0.5)(0.5) = 0.028125$

$P(O,q_1q_2q_1|\lambda) = b_1(H)b_2(H)b_1(H)\pi_1 a_{12}a_{21} = (0.5)(0.9)(0.5)(0.5)(0.5)(0.2) = 0.01125$

$P(O,q_1q_2q_2|\lambda) = b_1(H)b_2(H)b_2(H)\pi_1 a_{12}a_{22} = (0.5)(0.9)(0.9)(0.5)(0.5)(0.8) = 0.081$

$P(O,q_2q_1q_1|\lambda) = b_2(H)b_1(H)b_1(H)\pi_2 a_{21}a_{11} = (0.9)(0.5)(0.5)(0.5)(0.2)(0.5) = 0.01125$

$P(O,q_2q_1q_2|\lambda) = b_2(H)b_1(H)b_2(H)\pi_2 a_{21}a_{12} = (0.9)(0.5)(0.9)(0.5)(0.2)(0.5) = 0.02025$

$P(O,q_2q_2q_1|\lambda) = b_2(H)b_1(H)b_1(H)\pi_2 a_{22}a_{21} = (0.9)(0.9)(0.5)(0.5)(0.8)(0.2) = 0.0324$

$P(O,q_2q_2q_2|\lambda) = b_2(H)b_2(H)b_2(H)\pi_2 a_{22}a_{22} = (0.9)(0.9)(0.9)(0.5)(0.8)(0.8) = 0.23328$

$$P(HHH \mid \lambda) = 0.015625 + 0.028125 + \cdots + 0.23328 = 0.43318$$

HMM 三大問題
Problem 1: 比對問題

- 解法二: 前向推導法(Forward Procedure)
  - 1. 定義在 t 時間 i 狀態出現部分觀察序列$o_1, o_2, \ldots, o_t$的機率 $\alpha_t(i)$
  - 2. 前向推導 $\alpha_{t+1}(i)$
  - 3. 求此 HMM 所有可能在 T 時間 i 狀態出現觀察序列 O 的機率 $\alpha_T(i)$, 則P(O | λ)為所有 $\alpha_T(i)$ 的總和

HMM 三大問題
Problem 1: 比對問題

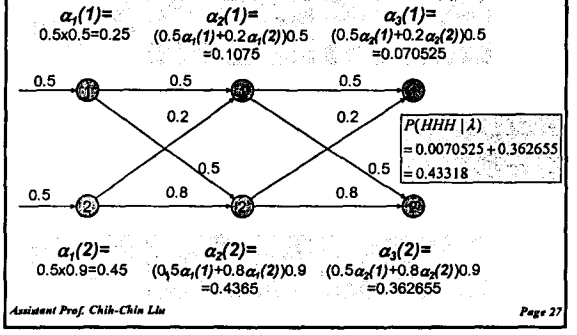
$$\alpha_t(i) = P(o_1 o_2 \cdots o_t, q_t = i \mid \lambda)$$

$$\Rightarrow \begin{cases} \alpha_1(i) = \pi_i b_i(o_1), \ 1 \le i \le N \\ \alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \ 1 \le t \le T-1 \text{ and } 1 \le j \le N \end{cases}$$

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

O(N²T)

HMM 三大問題
Problem 1: 比對問題



HMM 三大問題
Problem 1: 比對問題

- 解法三: 反向推導法(Backward Procedure)
  - 1. 定義在 t 時間 i 狀態出現部分觀察序列$o_{t+1}, o_{t+2}, \ldots, o_T$ 的機率 $\beta_t(i)$
  - 2. 末端 $\beta_T(i) = 1$
  - 3. 反向推導 $\beta_{t-1}(i)$
  - 4. 求此 HMM 所有可能在 *初始* ($t=1$) 時間 i 狀態出現觀察序列 O 的機率 $\beta_1(i)$, 則P(O | λ)為所有 $\pi_i \beta_1(i) b_i(o_1)$ 的總和

HMM 三大問題
Problem 1: 比對問題

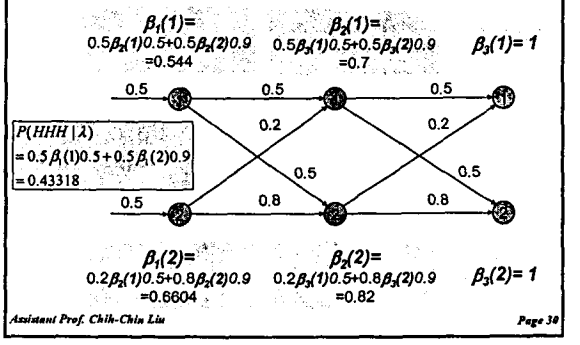
$$\beta_t(i) = P(o_{t+1} o_{t+2} \cdots o_T \mid q_t = i, \lambda)$$

$$\Rightarrow \begin{cases} \beta_T(i) = 1, \ 1 \le i \le N \\ \beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1}), \ 1 \le t \le T-1 \text{ and } 1 \le i \le N \end{cases}$$

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i \beta_1(i) b_i(o_1)$$

O(N²T)

HMM 三大問題
Problem 1: 比對問題



HMM 三大問題
Problem 2: 解釋問題

- 已知一 HMM λ 與一觀察序列 (observation sequence) $O = (o_1, o_2, \ldots, o_T)$, 則在此 HMM 模型下最有可能的狀態序列 $q = (q_1, q_2, \ldots, q_T)$ 為何？
- 解法一: 求在 t 時間最有可能的狀態 q_t
- 解法二: 求最有可能的狀態序列 $q = (q_1, q_2, \ldots, q_T)$ Viterbi Algorithm

HMM 三大問題
Problem 2: 解釋問題

- 解法一: 求在 t 時間最有可能的狀態 q_t
  - 1. 已知一 HMM λ 與一觀察序列 (observation sequence) $O = (o_1, o_2, \ldots, o_T)$, 定義在 t 時間狀態 $q_t = i$ 之機率

$$\gamma_t(i) = P(q_t = i \mid O, \lambda)$$

  - 2. 求使 $\gamma_t(i)$ 獲得最大值的 i 即為 q_t

HMM 三大問題
Problem 2: 解釋問題

- 解法一: 求在 t 時間最有可能的狀態 q_t
  - 1. 已知一 HMM λ 與一觀察序列 (observation sequence) $O = (o_1, o_2, \ldots, o_T)$, 定義在 t 時間狀態 $q_t = i$ 之機率

$$\gamma_t(i) = P(q_t = i \mid O, \lambda)$$

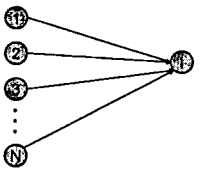
  - 2. 求使 $\gamma_t(i)$ 獲得最大值的 i 即為 q_t

HMM 三大問題
Problem 2: 解釋問題

$$\begin{aligned}\gamma_t(i) &= P(q_t = i \mid O, \lambda) \\ &= \frac{P(O, q_t = i \mid \lambda)}{P(O, \lambda)} \\ &= \frac{P(O, q_t = i \mid \lambda)}{\sum_{j=1}^{N} P(O, q_t = j \mid \lambda)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\end{aligned}$$

HMM 三大問題
Problem 2: 解釋問題

- 解法二: Viterbi Algorithm
- 類似前向推導法的作法
  - 前向推導法: 加總 $\alpha_t(i)$ 得到 $\alpha_{t+1}(i)$
  - Viterbi演算法: 取最大 $\delta_t(i)$ 得到 $\delta_{t+1}(i)$



HMM 三大問題
Problem 2: 解釋問題

- 解法二: Viterbi Algorithm
  - 1. 定義在 t 時間沿著路徑 $q_1, q_2, \ldots, q_t$ 到達 i 狀態使得出現部分觀察序列 $o_1, o_2, \ldots, o_t$ 的機率最大化的機率 $\delta_t(i)$
  - 2. 前向推導 $\delta_{t+1}(i)$
  - 3. 求此 HMM 所有可能沿著路徑 $q_1, q_2, \ldots, q_T$ 在 T 時間 i 狀態出現觀察序列 O 的機率 $\delta_T(i)$, 則 $q = (q_1, q_2, \ldots, q_T)$ 即為所求

HMM 三大問題
Problem 2: Viterbi Algorithm

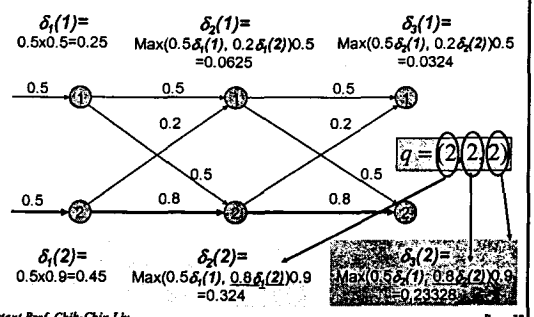
$$\delta_t(i) = \max_{q_1 q_2 \cdots q_{t-1}} P(q_1 q_2 \cdots q_{t-1}, q_t = i, o_1 o_2 \cdots o_t \mid \lambda)$$

$$\Rightarrow \begin{cases} \delta_1(i) = \pi_i b_i(o_1), \ 1 \le i \le N \\ \delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(j) a_{ij}] b_j(o_t), \ 2 \le t \le T \text{ and } 1 \le j \le N \end{cases}$$

- 與前向推導法的差別
  - 前向推導法: 加總 $\alpha_t(i)$
  - Viterbi演算法: 取最大值 $\delta_t(i)$

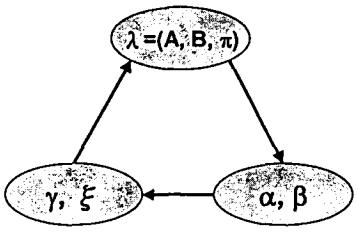
O(N²T)

HMM 三大問題
Problem 2: Viterbi Algorithm



HMM 三大問題
Problem 3: 比對問題

- 解法一: Baum-Welch Method (EM Method)



HMM 三大問題
Problem 3: 比對問題

- 解法一: Baum-Welch Method (EM Method)
  - 1. 定義在 t 時間狀態 $q_t = i$ 且 $t+1$時間狀態 $q_{t+1} = j$ 之機率 $\xi_t(i, j)$
  - 2. 由 A, B, $\alpha_t(i)$, $\beta_t(i)$, 求 $\xi_t(i, j)$
  - 3. 由 $\xi_t(i, j)$ 求 $\gamma_t(i)$
  - 4. 由 $\xi_t(i, j)$ 與 $\gamma_t(i)$ 更新 A, B, π
  - 重複 2, 3, 4 至收斂 (λ =(A, B, π) 穩定)

HMM 三大問題
Problem 3: 比對問題

- 解法一: Baum-Welch Method (EM Method)

$$\begin{aligned} \xi_t(i, j) &= P(q_t = i, q_{t+1} = j \mid O, \lambda) \\ &= \frac{P(q_t = i, q_{t+1} = j, O \mid \lambda)}{P(O \mid \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \end{aligned}$$

HMM 三大問題
Problem 3: 比對問題

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j)$$

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from state } i \text{ in O}$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{expected number of transitions from state } i \text{ to state } j \text{ in O}$$

HMM 三大問題
Problem 3: 比對問題

$$\overline{\pi}_i = \gamma_1(i), 1 \le i \le N$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, 1 \le i \le N, 1 \le i \le N$$

$$\overline{b}_j(k) = \frac{\sum_{t=1,\, o_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, 1 \le i \le N, 1 \le i \le N$$

References
Baum 的 HMM 古典論文

- [Baum66] Baum, L. E. and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Ann. Math. Stat.*, Vol. 37, pp. 1554-1563, 1966.
- [Baum67] Baum, L. E. and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Ann. Math. Stat.*, Vol. 37, pp. 1554-1563, 1966.

References
Rabiner 的 HMM 經典論文

- [Baum66] Baum, L. E. and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Ann. Math. Stat.*, Vol. 37, pp. 1554-1563, 1966.
- [Baum67] Baum, L. E. and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Ann. Math. Stat.*, Vol. 37, pp. 1554-1563, 1966.

Chap 6 HMM Applications

劉志俊 (Chih-Chin Liu)

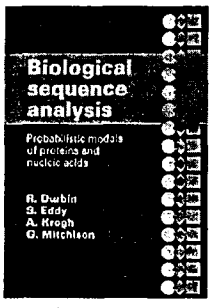
中華大學資訊工程系

April 2003

Outline

- Books and Tutorials About HMM
- Biological Applications of HMM
  - Profile HMM: [Eddy98][Haussler93]
  - Protein 2nd Structure Prediction: [Asai93][Karplus97]
  - Gene Prediction: [Krogh94][Kulp96]
  - Multiple Alignment: [Eddy95]
  - ncRNA: [Eddy02][Eddy01]

HMM在生物資訊的應用參考書籍



Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids

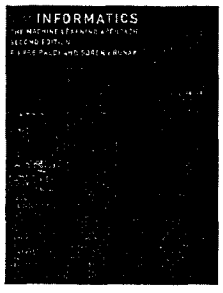
R. Durbin, S. Eddy, A. Krogh, and G. Mitchison

Cambridge University Press

1st edition, 1998

ISBN: 0521629713

HMM在生物資訊的應用參考書籍



Bioinformatics: The Machine Learning Approach

Pierre Baldi and Søren Brunak

The MIT Press

2nd edition, 2001

ISBN: 026202506X

Motivation

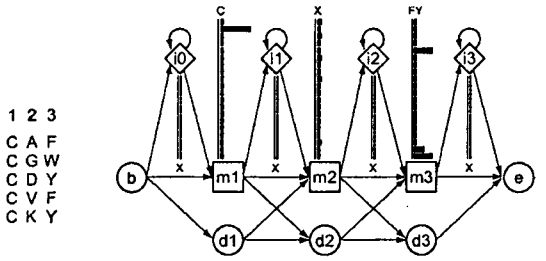
- Bioinformatics 在 Post-genome Era 扮演重要角色
- 胺基酸序列比對法(BLASTP, BLASTX, PSI-BLAST) 無法完全決定未知核酸/蛋白質的身分與功能
  - ***"Established sequence comparison algorithms detect significant similarities to 35-80% of new proteins, depending on the organism."*** [Eddy1996]
  - ***"In the 1990s, only roughly a third of the newly predicted protein sequences show convincing similarity to other known sequences, using pairwise comparisons."*** [Baldi2001]

Motivation

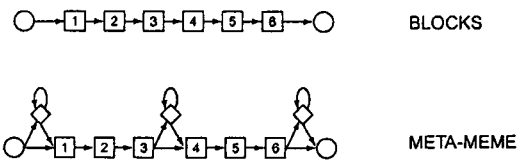
- David Haussler, Anders Krogh, and colleagues at UC Santa Cruz recognized that all the profile methods could be expressed as hidden Markov models (HMMs) [Haussler1994].

- Major Research Groups
  - UC Santa Cruz: Krogh, Haussler
  - WU St. Louis: Eddy
  - Baldi

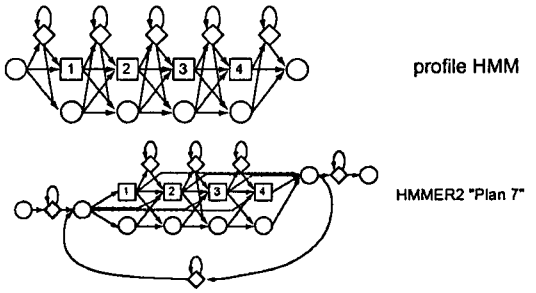
[Eddy98] Profile HMM Architecture



[Eddy98] Profile HMM Architecture



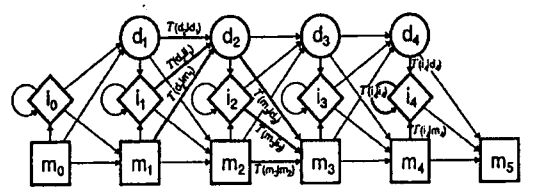
[Eddy98] Profile HMM Architecture



[Eddy98] Profile HMM HMM Software

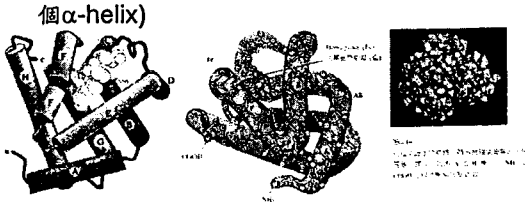
Software	URL
SAM	http://www.cse.ucsc.edu/research/compbio/sam.html
HMMER	http://genome.wustl.edu/eddy/hmmer.html
PFTOOLS	http://ulrec3.unil.ch:80/profile/
HMMpro	http://www.netid.com/
GENEWISE	http://www.sanger.ac.uk/Software/Wise2/
PROBE	ftp://ncbi.nlm.nih.gov/pub/neuwald/probe1.0/
META-MEME	http://www.cse.ucsd.edu/users/bgrundy/metameme.1.0.html
BLOCKS	http://www.blocks.fhcrc.org/
PSI-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/newblast.html

[Haussler93] Profile HMM Architecture



[Haussler93] Profile HMM Analysis of Globins

- 肌紅素(myoglobin)是第一個以X射線晶體分析法決定全部結構的蛋白質(153個胺基酸, 8個α-helix)



[Haussler93] Profile HMM Analysis of Globins

- Manual Aligned

Bachford以人工比對7個Globins的結構，得到右圖的多重比對結果。以此為基礎，完成對226個Globins序列的比對。

[Haussler93] Profile HMM Analysis of Globins

Figure 2: The alignment of seven representative globins from Bashford et al. [1]. Above the alignments are indicated which alpha helix each column belongs to. Bashford et al. identified 8 helices (A to H), but some regions are not well defined (especially CD, D, and FG). The sequences and their SWISS-PROT identifiers are from the top: Human α (HBA$HUMAN), human β (HBB$HUMAN), sperm whale myoglobin (MYG$PHYCA), larval *Chironomous thummi* globin (GLB3$CHITP), sea lamprey globin (GLB5$PETMA), *Lupinus luteus* leghemoglobin (LGB2$LUPLU), and bloodworm globin (GLB1$GLYDI).

Reference: [1] D Bashford, C Chothia and A M Lesk, *Journ Mol Biol* 196 (1987) 199-216.

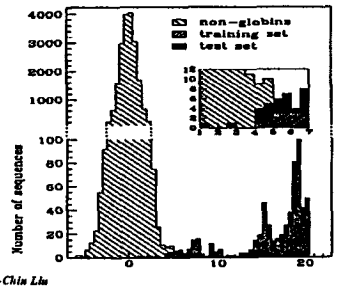
[Haussler93] Profile HMM Analysis of Globins

- HMM Trained

Haussler用400個Globins訓練出一個能分辨Globin的HMM，再拿Swiss-Prot中長度小於2000的蛋白質序列來測試結果：
Globin 625個錯2個
非Globin 19458個錯10個

[Haussler93] Profile HMM Analysis of Globins

- Globin 分類相似度統計



[Eddy02] ncRNA 預測 Motivation

- The number of known ***noncoding RNA genes*** is expanding rapidly.
- Computational analysis of genome sequences, which has been revolutionary for protein gene analysis, should also be able to address questions of the number and diversity of noncoding RNA genes.
- However, noncoding RNAs present computational genomics with a new set of challenges.

References
HMM 與生物資訊古典論文

- [Krogh94] Krogh, A., B. Brown, I.S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," *J. Mol. Bio.*, Vol. 235, pp. 1501-1531, 1994.
- [Baldi94] Baldi, P., *et al.*, "Hidden Markov Models of Biological Primary Sequence Information," *Proc. NAS USA*, Vol. 91, No. 2, pp. 1059-1063, 1994.
- [Eddy95] Eddy, S. R., G. Mitchison, and R. Durbin, "Maximum Discrimination Hidden Markov Models of Sequence Consensus," *J. Mol. Bio.*, pp. 9-23, 1995.

References
HMM 與生物資訊 Tutorial Papers

- [Eddy96] Eddy, S. R., "Hidden Markov Models," *Current Opinion in Structural Biology*, pp. 361-365, 1996.

References
HMM 在生物資訊各類應用

- [Eddy98] Eddy, S. R., "Profile Hidden Markov Models," *Bioinformatics*, Vol. 14, pp. 755-763, 1998. (使用 HMM 來建立蛋白質Profiles 的簡介)
- [Krogh94] Krogh, A., et al, "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," *J. Mol. Bio.*, Vol. 235, pp. 1501-1531, 1994. (最早使用 HMM 來建立蛋白質Profiles)
- [Asai93] Asai, K., S. Hayamizu, and K.I. Handa, "Prediction of Protein Secondary Structure by the Hidden Markov Model," *Comput. Applic. Biosci.*, Vol. 9, pp.141-146, 1993. (使用HMM作蛋白質二級結構預測)
- [Karplus97] Karplus, K. et al., "Prediction Protein Structure Using Hidden Markov Models," *Proteins*, Vol. 1, pp. 134-139, 1997.

References
HMM 在生物資訊各類應用

- [Kulp96] Kulp, D., *et al.* "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," *Nucleic Acids Research*, Vol. 22, pp. 4768-4778, 1994. (使用HMM對大腸桿菌DNA作基因預測)
- [Eddy95] Eddy, S. R., "Multiple Alignment Using Hidden Markov Models," in *Proc. Intl. Conf. Intelligent Systems for Molecular Biology*, AAAI Press, pp. 114-120, 1995. (使用HMM來作序列多重比對)
- [Eddy02] S.R. Eddy, "Computational genomics of noncoding RNA genes," *Cell*, 109:137-140, 2002. (Eddy 鼓吹使用計算基因體學對ncRNA 進行研究, 以進一步了解基因體)
- [Rivas01] E. Rivas and S.R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," *BMC Bioinformatics*, 2:8, 2001. (使用三個 HMMs 預測非編碼RNA)

Chap 4 Protein Motif Databases

劉 志 俊 (Chih-Chin Liu)

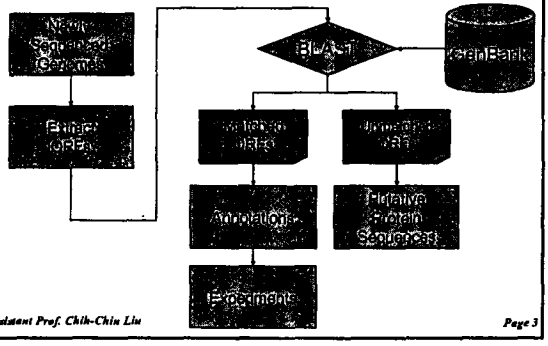
中華大學 資訊工程系

March 2003

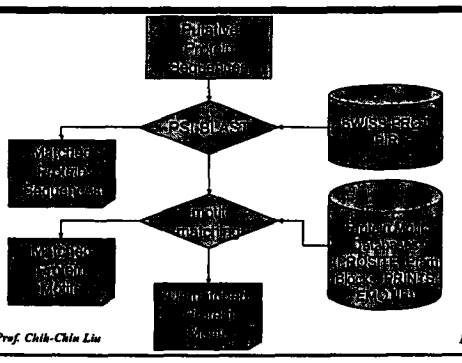
Outline

- Introduction
- PROSITE (RE, Profile)
- Pfam (HMM)
- PRINTS (Profile)
- Blocks (Profile)
- EMOTIF/IDENTIFY (RE)

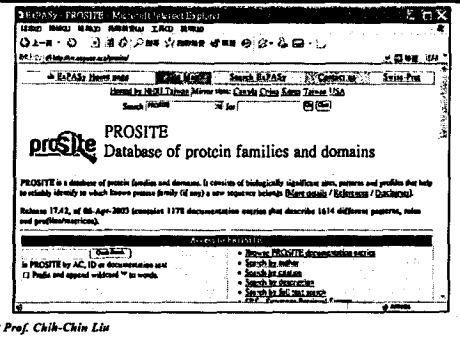
Gene Function Prediction



Gene Function Prediction



PROSITE: Protein Profile Database http://www.expasy.org/prosite/



PROSITE: Protein Profile Database http://www.expasy.org/prosite/

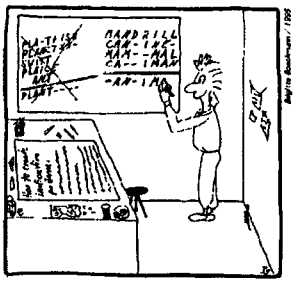
- 動機: 許多蛋白質功能相似, 但序列差異度很大
  - E.g.: globins, SH2 domains, SH3 domains
  - 傳統的 PSI-BLAST 序列相似性比對方法無法辨識
- 辨識機制: 與 Protein Profile 比對
- A ***profile*** is a table of position-specific animo acid weights and gap costs.
- Pattern = Motif = Signature = Fingerprint
  - 已定序的基因組(genomic)序列
  - 已定序的 cDNA 序列

PROSITE: Protein Profile Database

- PROSITE 為提供一種辨識特性未定的蛋白質(uncharacterized proteins)之功能的一個蛋白質 Profile 資料庫
- 特性未定的蛋白質來源
  - 已定序的基因組(genomic)序列
  - 已定序的 cDNA 序列
- Profile 建立方法:
  - 多重蛋白質序列比對
  - [Gribskov90][Luethy94][Thompson94]

PROSITE: Protein Profile Database
Profile 建立方法

How we develop Prosite patterns!



PROSITE: Protein Profile Database
五大設計原則

- **完整性(Completeness)**: 資料庫中的 patterns 與 profiles 涵蓋的資料愈完整愈好
- **高專一性(High Specificity)**: Profile 判斷 Motifs 的專一性高, 盡可能降低誤判
- **文件註解完整(Documentation)**: PROSITE 中每筆紀錄均有完整文件註解
- **定期檢討(Periodic Review)**: PROSITE 定期建討確保每筆紀錄均有效
- **與 SWISS-PROT 緊密結合**: 同一個開發單位, SWISS-PROT 與 PROSITE 同步更新

PROSITE: Protein Profile Database
下載資料檔案

- PROSITE 下載檔案 (ftp://ftp.expasy.org/databases/prosite/)
  - **PROSITE.DAT** : Profile 資料
  - **PROSITE.TXT** : Profile 語法描述
  - **PROSITE.DOC** : Profile 性質描述
- 相關軟體
  - ScanProsite
  - ProfileScan
  - FrameProfileScan
  - InterProScan

PROSITE: Protein Profile Database
下載資料檔案

Rel.	Date	Doc.	Entries	Notes
1.0	03/89	58	60	Only released in PC/Gene (Version 5.10)
2.0	05/89	129	132	Only released in PC/Gene (Version 6.00)
3.0	05/89	?	160	
4.0	10/89	?	202	Printed release (EMBL Bioconputing document)
5.0	04/90	296	338	
6.0	11/90	375	433	
7.0	05/91	441	508	
8.0	11/91	530	605	
9.0	06/91	580	689	
10.0	12/92	635	803	
11.0	10/93	715	927	
12.0	06/94	785	1029	First release to include profiles
13.0	11/95	889	1167	
14.0	12/97	997	1335	
15.0	06/98	1014	1352	
16.0	07/99	1034	1374	
17.0	11/01	1108	1501	

PROSITE資料格式
欄位說明

Motif編號	ID Identification (Begins each entry; 1 per entry)
PROSITE存取號	AC Accession number (1 per entry)
資料建立日期	DT Date (1 per entry)
簡述	DE Short description (1 per entry)
Pattern 資料	PA Pattern (1 or more per entry)
Profile資料	MA Matrix/profile (1 or more per entry)
Rule資料	RU Rule (1 or more per entry)
正確率等統計數值	NR Numerical results (1 or more per entry)
註解	CC Comments (1 or more per entry)
SWISS-PROT編號	DR Cross-references to SWISS-PROT (1 or more per entry)
PDB編號	3D Cross-references to PDB (1 or more per entry)
PROSITE文件編號	DO Pointer to the documentation (1 per entry)
單筆紀錄結束	// Termination line (Ends each entry; 1 per entry)

PROSITE資料樣本 Pattern

```
ID   T4_DEIODINASE; PATTERN.
AC   PS01205;
DT   NOV-1997 (CREATED); JUL-1999 (DATA UPDATE); JUL-1999 (INFO UPDATE).
DE   Iodothyronine deiodinases active site.
PA   R-P-L-[IV]-x-[NS]-F-G-S-[CA]-T-C-P-x-F.
NR   /RELEASE=40.7,103373;
NR   /TOTAL=16(16); /POSITIVE=16(16); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=0;
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC   /SITE=12,active_site;
DR   P49894, IOD1_CANFA, T; O42411, IOD1_CHICK, T; P49895, IOD1_HUMAN, T;
DR   Q61153, IOD1_MOUSE, T; O42449, IOD1_ORENI, T; P24389, IOD1_RAT  , T;
DR   P79747, IOD2_FUNHE, T; Q92813, IOD2_HUMAN, T; Q9Z1Y9, IOD2_MOUSE, T;
DR   P49896, IOD2_RANCA, T; P70551, IOD2_RAT  , T; O42412, IOD3_CHICK, T;
DR   P55073, IOD3_HUMAN, T; P49898, IOD3_RANCA, T; P49897, IOD3_RAT  , T;
DR   P49899, IOD3_XENLA, T;
DO   PDOC00925;
//
```


PROSITE資料樣本 Pattern Examples

[AC]-x-V-x(4)-{ED}.

This pattern is translated as:

[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

<A-x-[ST](2)-x(0,1)-V.

This pattern, which must be in the N-terminal of the sequence ('<'), is translated as:

Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

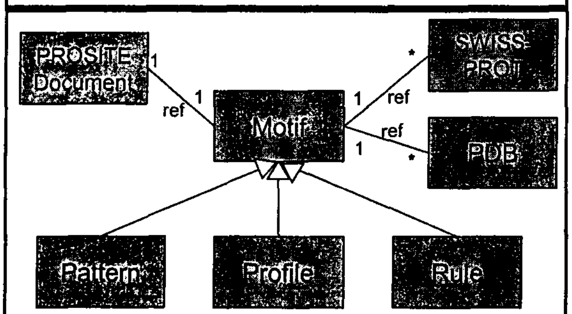
PROSITE資料樣本 Profile

```
ID   HSP20; MATRIX.
AC   PS01031;
DT   JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); NOV-1995 (INFO UPDATE).
DE   Heat shock hsp20 proteins family profile.
...  Some lines omitted..
```


PROSITE資料樣本 Profile Example

	FFYFFLF	KKPPKEK	LAIVVFL	LFVVLIL	SGGKASG	HQQEAEN	CTEAVCV	LMLIIIL	LFLLAIV	VQGKDQC
A	-18	-10	-1	-8	8	-3	[illegible]	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	[illegible]	[illegible]	-34	-31	0
E	[illegible]	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	[illegible]	[illegible]	-15	4	12	-29
G	[illegible]	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	[illegible]	14	-22	[illegible]	-23	-10
I	3	[illegible]	21	25	-29	-23	-8	33	19	-23
K	-26	25	[illegible]	[illegible]	-6	4	-15	[illegible]	[illegible]	0
L	14	[illegible]	19	27	-27	-20	-9	33	26	-21
M	3	-15	[illegible]	[illegible]	-17	-10	-9	[illegible]	[illegible]	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	[illegible]	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	[illegible]	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	[illegible]	[illegible]	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	[illegible]	[illegible]	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

PROSITE類別圖



PROSITE 蛋白質顧問群

Field of expertise　Name　Emai

If you have questions concerning one of the fields of listed below, you may click on the expert's Email add You will be given a form for your message to the expe

14-3-3 proteins
　Aitken A.: http://kr.expasy.org/cgi-bin

ABC transport membrane comp.
　Bauna E.: http://kr.expasy.org/cgi-bin

ADP-ribosylation factors
　Kahn R.A.: http://kr.expasy.org/cgi-bin

Alcohol dehydrogenases
　Joernvall H.: http://kr.expasy.org/cgi-bin
　Persson B.: http://kr.expasy.org/cgi-bin

Aldehyde dehydrogenases
　Joernvall H.: http://kr.expasy.org/cgi-bin
　Persson B.: http://kr.expasy.org/cgi-bin

"Wow, I am on the expert list !!!"

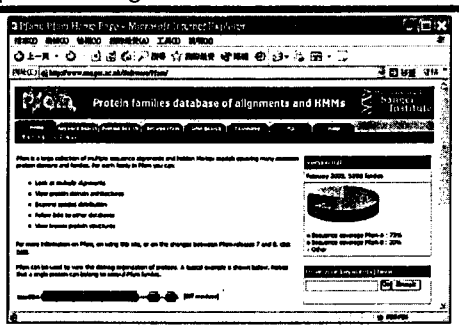
References
PROSITE資料庫簡介

- [Falquet02] Falquet, L., *et al.*, "The PROSITE Database, Its Status in 2002," *Nucleic Acids Research*, Vol. 30, No. 1, pp. 235-238, 2002. (PROSITE資料庫簡介)
- [Bucher94] Bucher, P. and A. Bairoch, "A Generalized Profile Syntax for Biomolecular Sequence Motifs and Its Function in Automatic Sequence Interpretation," in *Proc. of Intl. Conf. on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 53-61, 1994. (PROSITE資料庫中的profile語法格式)
- [Bucher96] Bucher, P., et al., "A Flexible Motif Search technique Based on Generalized Profiles," *Comput. Chem.*, Vol. 20, pp. 3-23, 1996. (PROSITE資料庫中的profile語法格式)

References
Profile 分析技術

- [Gribskov87] Gribskov, M., A. D. McLachlan, and D. Eisenberg, "Profile Analysis: Detection of Distantly Related Proteins," *Proc. of NAS*, Vol. 84, pp. 4355-4358, 1987. (Profile 分析技術)
- [Gribskov90] Gribskov, M., R. Luethy, and D. Eisenberg, "Profile Analysis," *Methods Enzymol.*, Vol. 183, pp. 146-159, 1990. (產生Profile之多重序列比對方法)
- [Luethy94] Luethy, R. , I. Xenarios, and P. Bucher, "Improving the Sensitivity of the Sequence Profile Method," *Protein Sci..*, Vol. 3, pp. 139-146, 1994. (Profile 分析技術改進)
- [Thompson94] Thompson, J. D., D. G. Higgins, and T. J. Gibson, "Improving the Sensitivity of Profile Searches Through the Use of Sequence Weights and Gap Excision," *Comput. Appl. Biosci.*, Vol. 10, pp. 19-29, 1994. (Profile 分析技術改進)

Pfam: Protein Families Database
http://www.sanger.ac.uk/Software/Pfam/



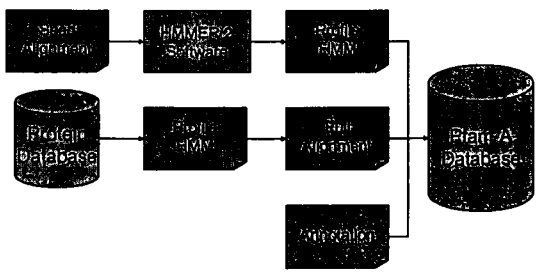
Pfam: Protein Families Database
http://www.sanger.ac.uk/Software/Pfam/

- 國際合作研發: 英國(Sanger中心)、瑞典(Karolinska研究院)、美國(Washinton大學)
- 3071 個蛋白質家族 (版本6.6, 2002)
- 涵蓋 69% 在 SWISS-PROT 39 與 TrEMBL 14 中的蛋白質序列
- Motif 的樣板稱為 ***Profile HMM***, 由 HMMER2 產生 (程式在 http://hmmer.wustl.edu/)
- 前 20 大蛋白質家族中, 每個家族的成員個數皆在2500個蛋白質序列以上

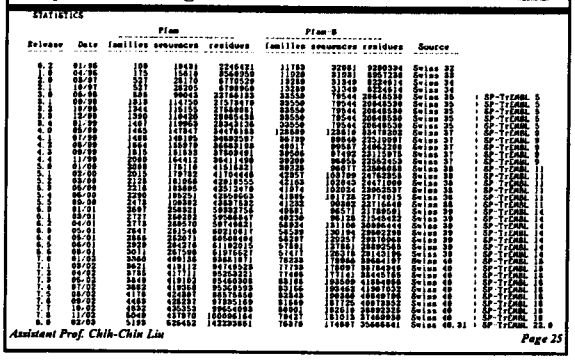
Pfam: Protein Families Database
http://www.sanger.ac.uk/Software/Pfam/

Wellcome Trust Sanger Institute (UK)	Karolinska Institute (Sweden)	St Louis (USA)
Alex Bateman - Group leader	Erik Sonnhammer	Sean Eddy
Richard Durbin - Head of Department		
Mhairi Marshall - Webmaster/Database Administrator		
Sam Griffiths-Jones - Research and Development		
Rob Finn - Research and Development		
Kevin Howe - PhD student		
Corin Yeats - PhD student		
Lachlan Coin - PhD student		
Previous contributors:		
Ewan Birney - now working at EBI		
Lorenzo Cerrutti		
William Misud - Summer student from University of Malta		
Nina Mian		
Matthew Bashton		

Pfam: Protein Families Database
http://www.sanger.ac.uk/Software/Pfam/



Pfam: Protein Families Database
http://www.sanger.ac.uk/Software/Pfam/



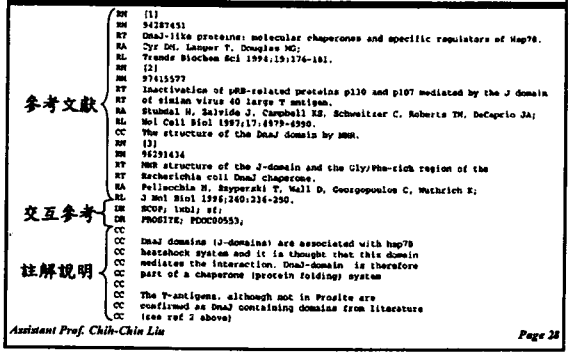
Pfam: Protein Families Database
http://www.sanger.ac.uk/Software/Pfam/

- Protein Sequence Alignment 採用Stockholm格式
  (http://www.cgb.ki.se/cgb/groups/sonnhammer/Stockholm.html)
- Pfam 原始資料:
  - **Pfam-A.seed**: 包含種子蛋白質序列比對結果
  - **Pfam-A.full**: 包含完整蛋白質序列比對結果
  - **Pfam**: 包含 Profile HMM 蛋白質 Motif 樣板
- PfamDB 以 MySQL 實作 (10 個 tables, 資料不全)

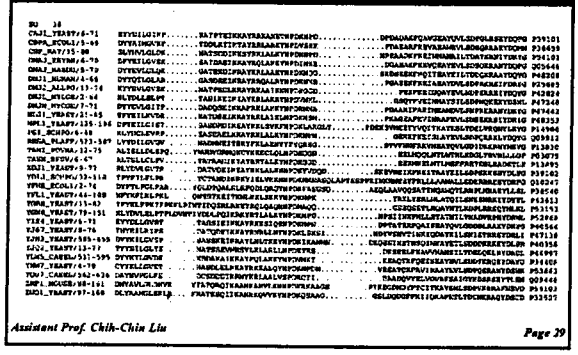
Pfam資料樣本
基本資料與Seed Alignment

Motif名稱	ID	DnaJ
Motif編號	AC	PF00226
Motif描述	DE	DnaJ domain
作者	AU	Finn RD, Birney E
資料來源	SE	Prosite
比對方法	AL	Clustalw
建構方法	BM	hmmbuild HMM SEED
	BM	hmmcalibrate --seed 0 HMM
建構條件	GA	hmmsearch -T 10 --domT 10
最低相似度	TC	10.50 10.50
最高相似度	NC	7.90 7.90

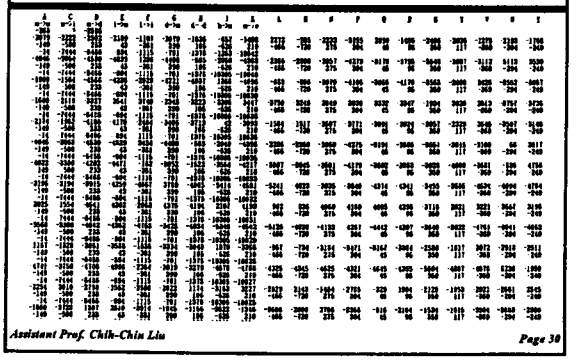
Pfam資料樣本
Annotation



Pfam資料樣本
Full Alignment



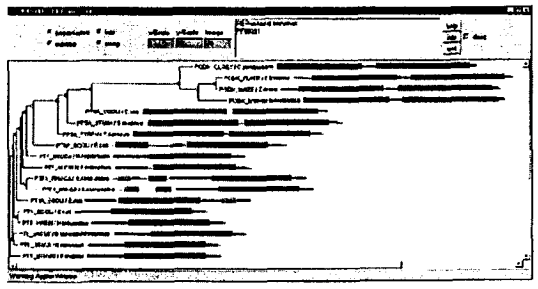
Pfam資料樣本
Profile HMM



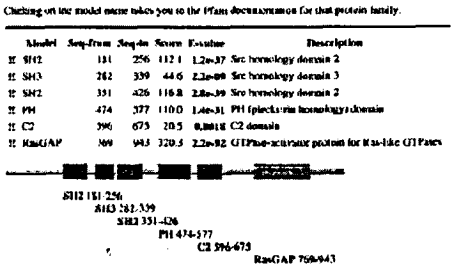
Pfam資料格式說明

- [Bateman02] Bateman, A., *et al.*, "The Pfam Protein Families Database," *Nucleic Acids Research*, Vol. 30, No. 1, pp. 276-280, 2002. (Pfam資料庫簡介)
- [Krogh94] Krogh, A., et al, "" *J. Mol. Biol.*, Vol. 235, pp. 1501-1531, 1994. (Profile HMM 方法)
- [Eddy96] Eddy, S.R., "" *Curr. Opin. Struct. Biol*, Vol. 6, pp. 361-365, 1996. (Profile HMM 方法)

Pfam



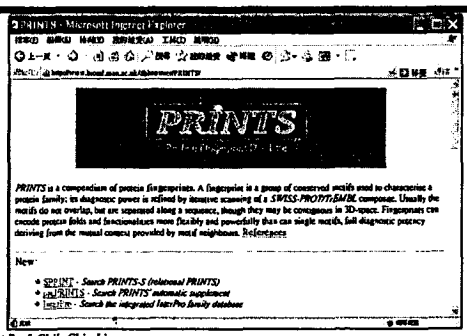
Pfam
用 Profile HMM 比對蛋白質序列



References
Pfam資料庫簡介

- [Bateman02] Bateman, A., *et al.*, "The Pfam Protein Families Database," *Nucleic Acids Research*, Vol. 30, No. 1, pp. 276-280, 2002. (Pfam資料庫簡介)
- [Krogh94] Krogh, A., et al, "" *J. Mol. Biol.*, Vol. 235, pp. 1501-1531, 1994. (Profile HMM 方法)
- [Eddy96] Eddy, S.R., "" *Curr. Opin. Struct. Biol*, Vol. 6, pp. 361-365, 1996. (Profile HMM 方法)
- [Krogh01] Krogh, A., *et al*, "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes," *J. Mol. Biol.*, Vol. 305, pp. 567-580, 2001. (Profile HMM 方法的應用)

PRINTS: Protein Fingerprint Database
http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/



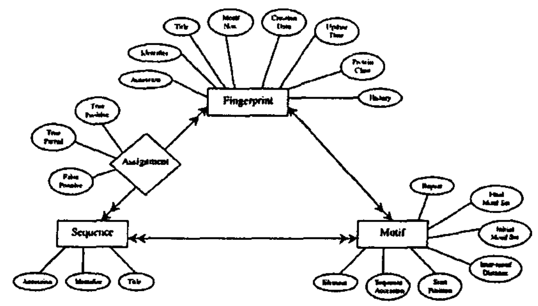
PRINTS: Protein Fingerprint Database
http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/

- 動機: 許多蛋白質功能相似, 但序列差異度很大
  - E.g.: globins, SH2 domains, SH3 domains
  - 傳統的 PSI-BLAST 序列相似性比對方法無法辨識
- 辨識機制: 與 Protein Profile 比對
- ***Fingerprints*** are ***groups of conserved sequence motifs*** that together provide diagnostic signatures for protein families.
- 與 Blocks, PROSITE(profile) 類似
- 特點: hierarchical classification

PRINTS: Protein Fingerprint Database
http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/

DATABASE VERSION	DATE	NO.OF DATABASE ENTRIES	TOTAL NO. OF MOTIFS	DATABASE VERSION	DATE	NO.OF DATABASE ENTRIES	TOTAL NO. OF MOTIFS
1.0	18.10.1991	29	116	18.0	31.03.1998	865	4892
1.1	16.10.1992	28	110	19.0	20.07.1998	925	5334
2.0	13.03.1993	59	279	20.0	25.10.1998	990	5701
2.1	16.04.1993	59	279	21.0	03.01.1999	1050	6142
3.0	22.07.1993	104	486	22.0	23.03.1999	1100	6510
4.0	29.10.1993	150	702	23.0	30.06.1999	1160	6938
5.0	20.04.1994	200	951	23.1	15.08.1999	1159	6935
6.0	22.07.1994	250	1211	24.0	01.10.1999	1210	7241
7.0	22.11.1994	300	1433	25.0	07.01.2000	1260	7539
8.0	22.03.1995	350	1686	26.0	29.03.2000	1310	7897
9.0	22.07.1995	400	1942	27.0	31.05.2000	1360	8244
9.1	14.11.1995	400	1942	28.0	25.09.2000	1410	8550
10.0	21.12.1995	450	2227	29.0	01.01.2001	1460	8880
11.0	08.04.1996	500	2559	30.0	30.03.2001	1500	9136
12.0	21.06.1996	550	2875	31.0	30.06.2001	1550	9531
12.1	22.08.1996	550	2874	32.0	23.09.2001	1600	9800
13.0	29.09.1996	600	3197	33.0	01.01.2002	1650	10085
13.a	03.11.1996	614	3280	34.0	01.04.2002	1700	10342
14.0	16.12.1996	650	3504	35.0	14.07.2002	1750	10626
15.0	15.04.1997	700	3838				
16.0	16.06.1997	750	4136				
17.0	13.09.1997	800	4460				

PRINTS: Protein Fingerprint Database
ERD



PRINTS: Protein Fingerprint Database
樣本資料

- 基本資料與交互參考

Identifier	APOLIPOA1
Creation Date	29-JUL-2002
Accession	PP00099
No. of Motifs	4
Title	Apolipoprotein a-i (APO-AI) signature
Database References	PFAM; PF01442 Apolipoprotein INTERPRO; IPR000074 PDB; 1ODP; 1ODQ; 1ODR; 1GW3; 1GW4; 1AV1 SCOP; 1ODP; 1ODQ; 1ODR; 1GW3; 1GW4; 1AV1 CATH; 1ODP; 1ODQ; 1ODR; 1GW3; 1GW4; 1AV1 MIM; 107680; 205400; 105200

PRINTS: Protein Fingerprint Database
樣本資料

- 參考文獻

Literature References

1. WEISGRABER, K.H., RALL, S.C., BERSOT, T.P., MAHLEY, R.W., FRANCESCHINI, G. AND SIRTORI, C.R. Apolipoprotein A-IMilano. Detection of normal A-I in affected subjects and evidence for a cysteine for arginine substitution in the variant A-I. J.BIOL.CHEM. 258 2508-2513 (1983)

2. NAKAI, T., WHAYNE, T.F. AND TANG, J. The amino- and carboxyl-terminal sequences of canine apolipoprotein A-I. FEBS LETT. 64 409-411 (1976).

3. NICHOLS, W.C., GREGG, R.E., BREWER, H.B., JR. AND BENSON, M.D. A mutation in apolipoprotein A-I in the Iowa type of familial amyloidotic polyneuropathy. GENOMICS 8 318-323 (1990).

4. PROLI, R.P., ORDOVAS, J.M., ROSENBERG, I., SCHAEFER, E.J. AND PEREIRA, M.E.A. Similarity of cruzin, an inhibitor of Trypanosoma cruzi neuraminidase, to high-density lipoprotein. SCIENCE 238 1417-1419 (1987).

5. BORHANI, D.W., ROGERS, D.P., ENGLER, J.A. AND BROUILLETTE, C.G. Crystal structure of truncated human apolipoprotein A-I suggests a lipid-bound conformation. PROC.NATL.ACAD.SCI.USA 94 12291-12296 (1997).

6. WANG, G., TRELFAVEN, W.D. AND CUSHLEY, R.J. Conformation of human serum apolipoprotein A-I(166-185) in the presence of sodium dodecyl sulfate or dodecylphosphocholine by 1H-NMR and CD. Evidence for specific peptide-SDS interactions. BIOCHIM.BIOPHYS.ACTA 1301 174-184 (1996).

PRINTS: Protein Fingerprint Database
樣本資料

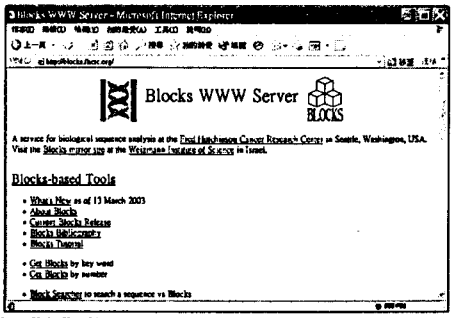
- 功能描述

[illegible]

References
PRINTS資料庫簡介

- [Attwood03] Attwood, L., *et al.*, "PRINTS and Its Automatic Supplement, prePRINTS," *Nucleic Acids Research*, Vol. 31, No. 1, pp. 400-402, 2003. (PRINTS資料庫簡介)
- [Parry-Smith92] Parry-Smith, D. J. and T. K. Attwood, "ADSP - A new package for computational sequence analysis," *Comput. Applic. Biosci.*, Vol. 8, No.5, pp. 451-459, 1992. (提出Fingerprint 的觀念)
- [Attwood94] Attwood, T. K. and J. B. C. Findlay, "Fingerprinting G-protein-coupled receptors," *Protein Engineering*, Vol. 7, No. 2, pp. 195-203, 1994. (提出Fingerprint 的觀念)
- [Scordis99] Scordis, P., D. R. Flower, and T. K. Attwood, "FingerPRINTScan: Intelligent Searching of the PRINTS Motif Database," Bioinformatics, Vol. 15, No. 10, pp. 799-806, 1999. (提出功能未知蛋白質序列與PRINTS比對的演算法)

Blocks: Protein Block Database
http://blocks.fhcrc.org/



Blocks: Protein Block Database
http://blocks.fhcrc.org/

- ***Blocks*** **are *ungapped multiple alignments*** corresponding to the most conserved regions of proteins.

Blocks: Protein Block Database
Sample Data

Block BL01173A

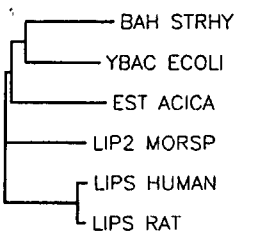
Block BL01173B

Block BL01173C

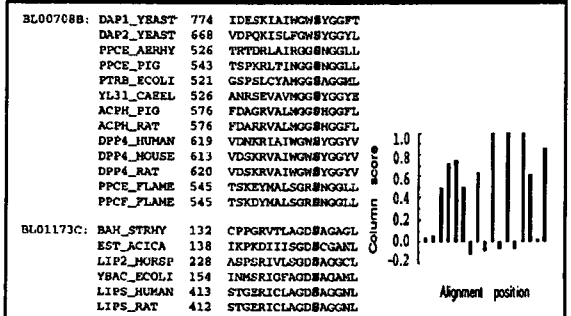
Blocks: Protein Block Database
Sample Data

Block BL01173A | Block BL01173B | Block BL01173C

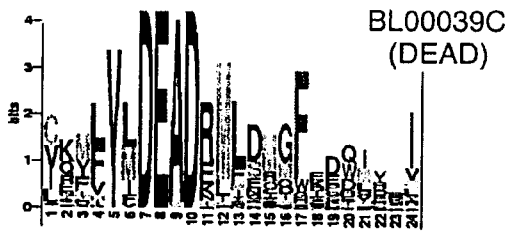
Blocks: Protein Block Database
Sample Data



Blocks: Protein Block Database
Block to Profile



Blocks: Protein Block Database
Block to Profile



Sequence logo of block BL00039C

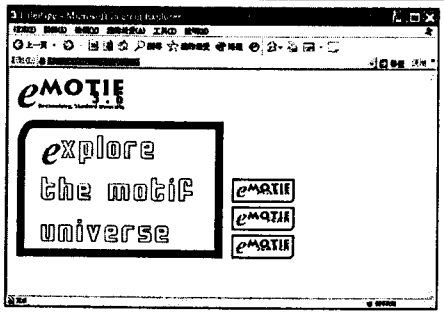
Blocks: Protein Block Database
Sample Data

```
ID   Kringle; BLOCK
AC   IPB000001B; distance from previous block=(13,361)
DE   Kringle
BL   RGS;  width=25; seqs=45; 99.5%=1167; strength=1320
FA12_CAVPO|Q04962  ( 216) CYEGRGVSYRGMARTTVSGAKCQRW  17

FA12_HUMAN|P00748  ( 217) CYDGRGLSYRGLARTTLSGAPCQPW  19

PLMN_ERIEU|Q29485  ( 103) CKVGNGKYYRGTVSKTKTGLTCQKW  26

PLMN_MOUSE|P20918  ( 103) CKTGIGNGYRGTMSRTKSGVACQKW  27

UROK_CHICK|P15120  (  79) CYSGNGEDYRGMAEDPGCLYWDHPS 100

TPA_RAT|P19637     ( 213) CYVGKGVTYRGTHSFTTSKASCLPW  31

ROR1_HUMAN|Q01973  ( 313) CYNSTGVDYRGTVSVTKSGRQCQPW  17
ROR1_MOUSE|Q9Z139  ( 313) CYNSTGVDYRGTVSVTKSGRQCQPW  17

HGFA_HUMAN|Q04756  ( 286) CFLGNGTGYRGVASTSASGLSCLAW  21
HGFA_MOUSE|Q9R098  ( 283) CFLGNGTEYRGVASTAASGLSCLAW  22
```


References
Blocks資料庫簡介

- [Henikoff00] Henikoff, J. G., *et al.*, "Increased Coverage of Protein Families with the Blocks Database Servers," *Nucleic Acids Research*, Vol. 28, No. 1, pp. 228-230, 2000. (Blocks資料庫簡介)
- [Henikoff99] Henikoff, S., *et al.*, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations," *Bioinformatics*, Vol. 15, No. 6, pp. 471-479, 1999. (Blocks資料庫簡介)

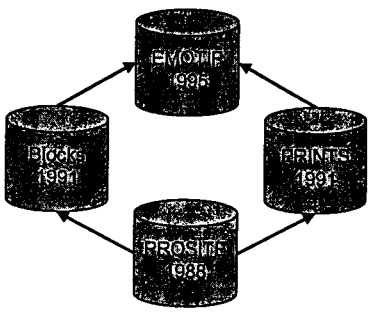
EMOTIF: Protein Motif Database
http://motif.stanford.edu/emotif/



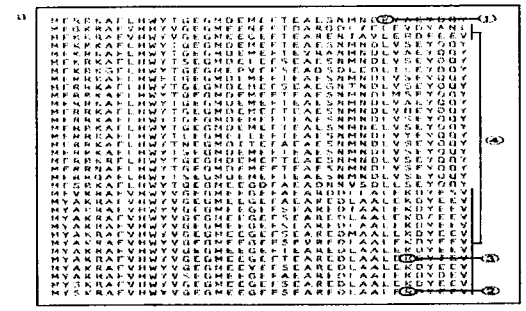
EMOTIF/IDENTITY
Protein Motif Database

- ***Motifs*** are useful for assigning functions to proteins even in the absence of any homology apart from the limited motif regions.
- ***IDENTIFY*** is a valuable tool for assignment of function to newly sequenced proteins, especially in those cases where there are no significant sequence similarities by
  - Alignment (PSI-BLAST)
  - Profile (PROSITE, Blocks, PRINTS)
  - Hidden Markov Methods (Pfam)

EMOTIF/IDENTITY
Protein Motif Database



EMOTIF/IDENTITY
Protein Motif Database



EMOTIF/IDENTITY
Protein Motif Database

b

```
MFGKKAFVHHFVGEGMDENEFAEARQNVAALVKEFQQL
 YKRRR L WYTS   EPGD SDYEEDLLDPLSDYASV
  R HG     L    IM  T    D ISE EY   EEY
  V O      N    L        A MNV IA   IN
  A             V        N TE  L
  S             T        S Y   V
```

多重比對結果

c

```
MF.K..FVH.F...EGMQ...QFPQ....Q.......QF...
 Y R   L Y       N  N AN     N        NY
 W     I W       E  E GE     E        EW
                 D  D SD     D        D
                    T
```

化學性質一般化

d

```
MF.K..FVH.F...EGMQ...QFPQ....Q..L..QF...
 Y R   L Y       N  N AN     N        NY
 W     I W       E  E GE     E        EW
                 D  D SD     D        D
                    T
```

第一條序列視作雜訊移除後結果

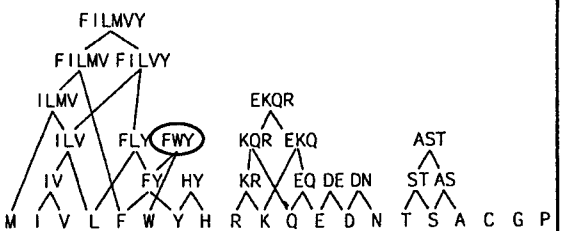
EMOTIF/IDENTITY
Protein Motif Database

胺機酸依化學性質分類表

AG	small
ST	small hydroxyl
KR	basic
FWY	aromatics
HKR	basic
ILV	small hydrophobic
ILMV	medium hydrophobic
EDNQ	acidic/amid
AGPST	small polar
.	all amino acids

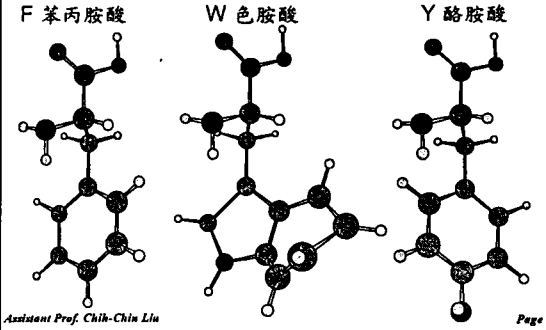
EMOTIF/IDENTITY
Protein Motif Database

胺機酸替換群組



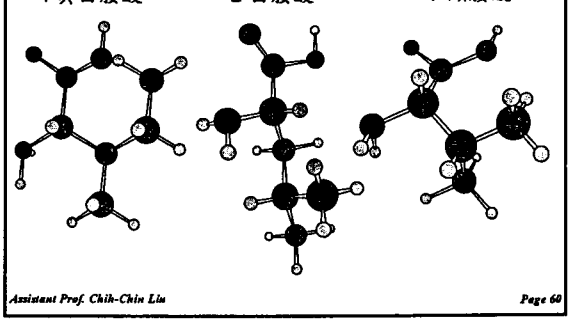
EMOTIF/IDENTITY
Protein Motif Database

F 苯丙胺酸　W 色胺酸　Y 酪胺酸

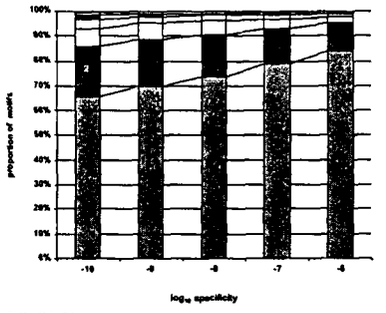


EMOTIF/IDENTITY
Protein Motif Database

I 異白胺酸　L 白胺酸　V 纈胺酸



EMOTIF/IDENTITY
實驗結果: 專一性 vs. 涵蓋率



EMOTIF/IDENTITY
實驗結果: 專一性 vs. 涵蓋率

Specificity	# ORFs assigned	# ORFs assigned with no annotations	# Motifs assigned	Expected # of false motif assignments
10^{-10}	61	41	179	0.02
10^{-9}	86	59	238	0.2
10^{-8}	103	69	301	1.7
10^{-7}	172	121	488	17

833個功能未知且無法以序列比對法辨識的酵母菌ORFs, 用EMOTIF可辨識172個

EMOTIF/IDENTITY
實驗結果

Genome	Specificity	Total motifs assigned & verified	Motifs verified manually	Assignments unverified	Expected false assignments	ORFs identified	total ORFs	% of total ORFs identified
S. cerevisiae	10^{-10}	4442	909	9	0	1345	6220	22%
	10^{-9}	4879	1027	31	5	1468		24%
	10^{-8}	4604	1114	124	42	1621		26%
H. influenzae	10^{-10}	1604	644	11	0	479	1697	28%
	10^{-9}	1668	708	33	0	503		30%
M. jannaschii	10^{-10}	349	115	3	0	157	1680	9%
	10^{-9}	403	126	21	0	182		11%
M. genitalium	10^{-10}	297	75	4	0	98	467	21%
	10^{-9}	331	87	7	0	108		23%
Syn. sp.	10^{-10}	1360	380	21	2	447	3169	14%
	10^{-9}	1508	461	34	20	513		16%
M. pneumoniae	10^{-10}	304	75	8	0	101	677	15%
	10^{-9}	360	89	8	0	117		17%
H. pylori	10^{-10}	476	100	18	0	200	1566	13%
	10^{-9}	579	121	19	0	233		15%

References
EMOTIF/IDENTITY資料庫簡介

- [Huang01] Huang, J. Y. and D. L. Brutlag, "The EMOTIF Database," *Nucleic Acids Research*, Vol. 29, No. 1, pp. 202-204, 2001. (EMOTIF資料庫簡介)
- [Wu95] Wu, T. D. and Brutlag, D. L., "Identification of protein motifs using conserved Amino Acid Properties and Partitioning Techniques," *Proc. Intelligent Systems for Molecular Biology*, pp. 402-410, 1995. (EMOTIF資料庫前身)
- [Nevill-Manning98] Nevill-Manning, C. G., Wu, T. D. and Brutlag, D. L., "Highly Specific Protein Sequence Motifs for Genome Analysis," *Proc. Natl. Acad. Sci.* USA, Vol. 95, No. 11, pp. 5865-5871, 1998. (EMOTIF產生演算法)
- [Nevill-Manning97] Nevill-Manning, C.G., Sethi, K.S., Wu, T.D. and Brutlag, D.L., "Enumerating and ranking discrete motifs," *Proc. Intelligent Systems for Molecular Biology*, 1997. (EMOTIF產生演算法)

Motif Database 綜合比較

Motif DB	Developer	Since	Motif Pattern	Syntax	Data	DB
PROSITE	Bucher 瑞士 生物資訊學院	1988 1994	profile	RE weighted-matrix	1,614 profiles	
Pfam	Durbin Eddy Sonnhammer 英國 桑格中心	1996	profile HMM	HMM	5,193 families	MySQL
PRINTS	Attwood 英國 曼徹斯特大學	1991	fingerprint	unweighted sequences	10,626 motifs	PostgreSQL
Blocks	Henikoff 美國 Fred Hutchinson 癌症研究中心	1991	block	weighted-matrix (PSSM)	8,656 Blocks	
EMOTIF	Brutlag 美國 史丹佛大學	1995	eMOTIF	RE	170,294 REs	

Chap 6 HMM Tools

劉 志 俊 (Chih-Chin Liu)

中華大學 資訊工程系

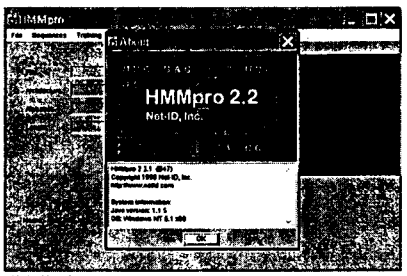
May 2003

Outline

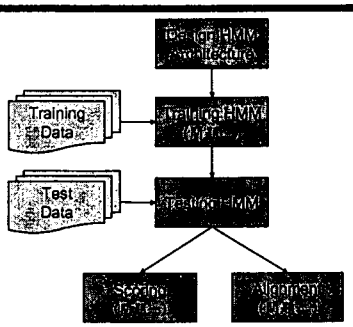
- Books and Tutorials About HMM
- Biological Applications of HMM
  - Profile HMM: [Eddy98][Haussler93]
  - Protein 2nd Structure Prediction: [Asai93][Karplus97]
  - Gene Prediction: [Krogh94][Kulp96]
  - Multiple Alignment: [Eddy95]
  - ncRNA: [Eddy02][Eddy01]

HMMpro

- Download HMMpro from http://www.netid.com/

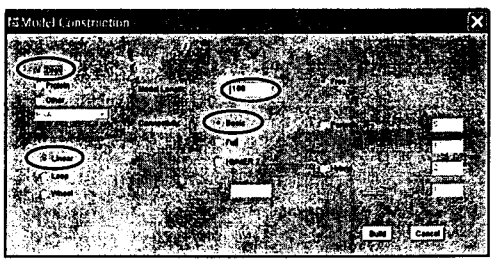


HMMpro



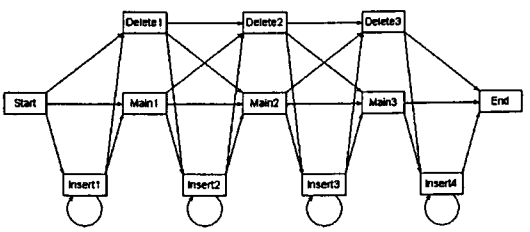
HMMpro

- Step 1: Design HMM Architecture



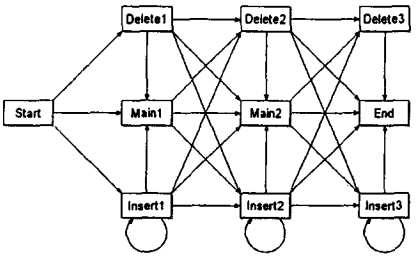
HMMpro HMM Architecture

- Linear HMM



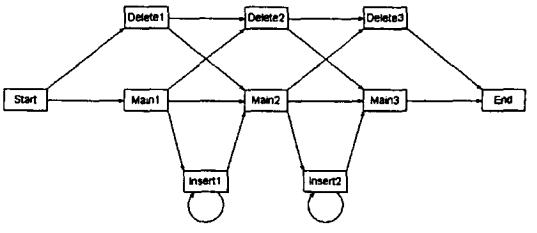
HMMpro
HMM Architecture

- Linear Full HMM



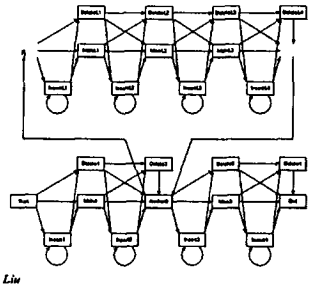
HMMpro
HMM Architecture

- HMMER 2



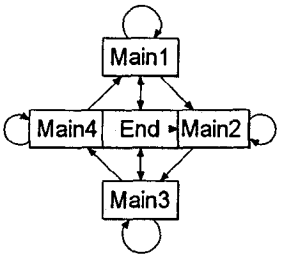
HMMpro
HMM Architecture

- Loop HMM



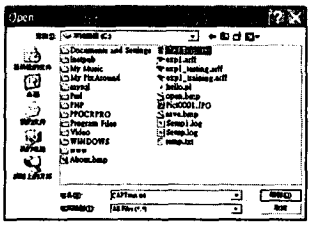
HMMpro
HMM Architecture

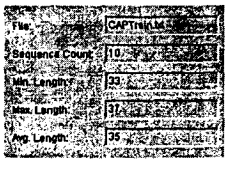
- Wheel HMM



HMMpro
Step 2: Training HMM

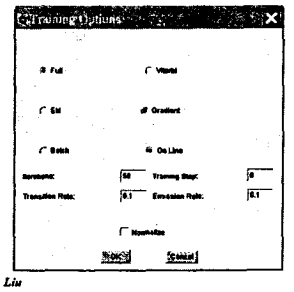
- Step 2a: Select Training Data





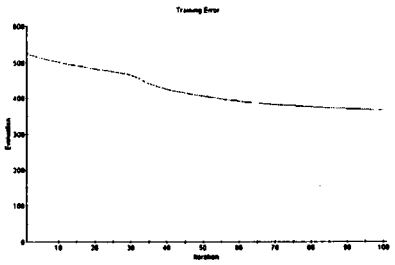
HMMpro
Step 2: Training HMM

- Step 2b: Training Options Setting



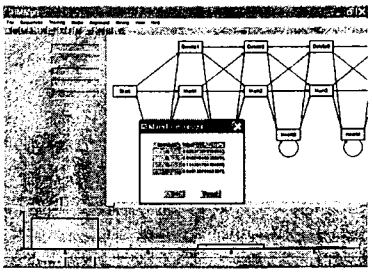
HMMpro
Step 2: Training HMM

- Step 2c: View Training Process



HMMpro
Step 2: Training HMM

- Step 2d: Check Training Result



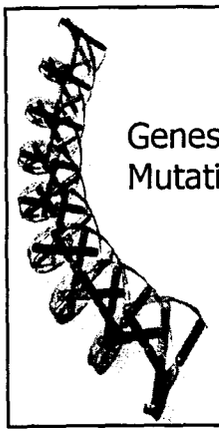
HMMpro
Step 3: Scoring

- Step 3a: Select Training Data

References
HMM 在生物資訊各類應用

- [Kulp96] Kulp, D., *et al.* "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," *Nucleic Acids Research*, Vol. 22, pp. 4768-4778, 1994. (使用HMM對大腸桿菌DNA作基因預測)
- [Eddy95] Eddy, S. R., "Multiple Alignment Using Hidden Markov Models," in *Proc. Intl. Conf. Intelligent Systems for Molecular Biology*, AAAI Press, pp. 114-120, 1995. (使用HMM來作序列多重比對)
- [Eddy02] S.R. Eddy, "Computational genomics of noncoding RNA genes," *Cell*, 109:137-140, 2002. (Eddy 鼓吹使用計算基因體學對ncRNA 進行研究, 以進一步了解基因體)
- [Rivas01] E. Rivas and S.R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," *BMC Bioinformatics*, 2:8, 2001. (使用三個 HMMs 預測非編碼RNA)

生子序列與演化樹分析教材內容



Genes , Genetic Codes , Mutation

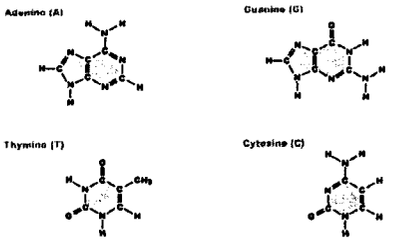
鄭銘杰

2003/09/17

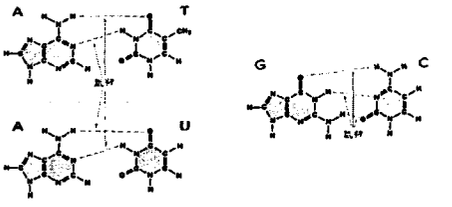
Nucleotide sequence

- The hereditary information of all living organisms, with the exception of some viruses, is carried by deoxyribonucleic acid (DNA) molecules.
- Purines : adenine(A), guanine(G)
- Pyrimidines : thymine(T), cytosine(C)
- Weak bond → A:T
- Strong bond → C:G

Nucleotide sequence



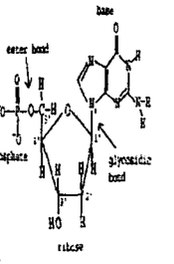
Nucleotide sequence

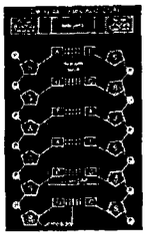


Nucleotide sequence

- Each nucleotide in a DNA sequence contains a pentose sugar, a phosphate group, and a purine or pyrimidine base.
- The backbone of the DNA molecule consists of sugar and phosphate moieties.
- The 5′ and the 3′ directions are also referred to as upstream and downstream

Nucleotide sequence





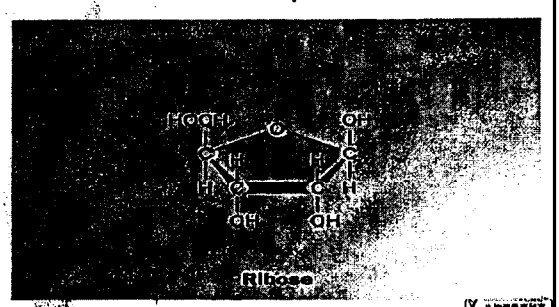
Nucleotide sequence

- The heavy strand is the one that contain more than 50% of the heavier nucleotides, the purines. The light strand is the one that contain more than 50% of the lighter nucleotides, the pyrimidines.

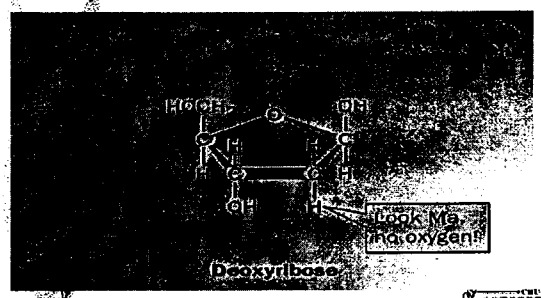
Nucleotide sequence

- RNA differ from DNA by having ribose instead of deoxyribose as its backbone sugar moiety, and by using the nucleotide uracil in place of thymine

Nucleotide sequence



Nucleotide sequence



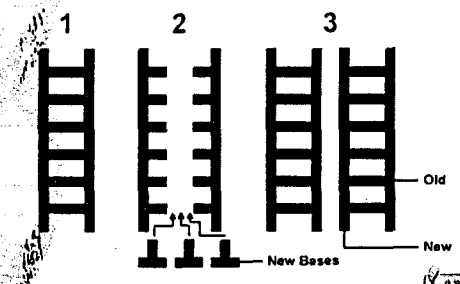
Genomes and DNA replication

- The entire complement of genetic material carried by an individual is called the genomes
- Ad hoc term that may include genic sequences whose function has not yet been determined.

Genomes and DNA replication

- All organisms replicate their DNA before every cell division
- Each of the two DNA strands servers as a template for the formation of a new strand
- DNA is replicated semiconservatively
- Replication starts at a structure called a replication bubble or origin of replication

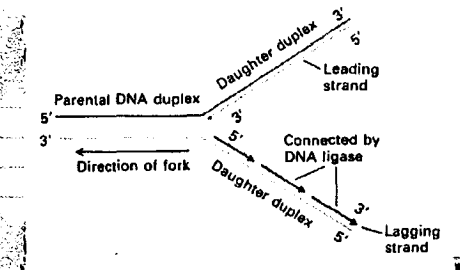
Genomes and DNA replication



Genomes and DNA replication

- Replication proceeds in both directions as two replication forks
- Replication DNA occurs only in the 5′ to 3′
- Leading strand , lagging strand and Okazaki fragment

Genomes and DNA replication



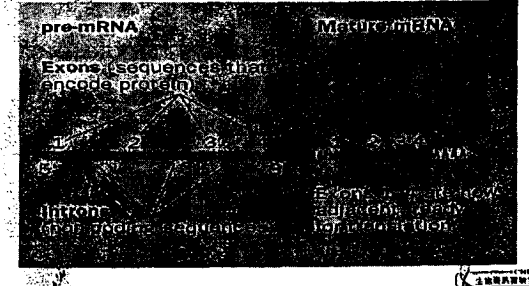
Genes and Gene Structure

- DNA → RNA → Protein
- A standard eukaryotic protein-coding genes consists of transcribed and untranscribed parts
- untranscribed parts are designated according to their location relative to the transcribed parts as 5′ and 3′ flanking regions

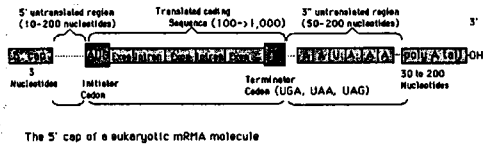
Genes and Gene Structure

- The promoter region consists of the following signals :
  - TATA BOX , GC BOX , CAAT BOX
- With few exceptions, all eukaryotic nuclear introns begin with GT and end with AG (the CT – AG rule)

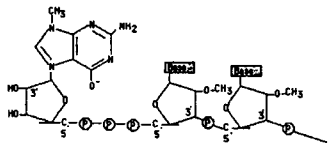
Genes and Gene Structure



Genes and Gene Structure



The 5' cap of a eukaryotic mRNA molecule



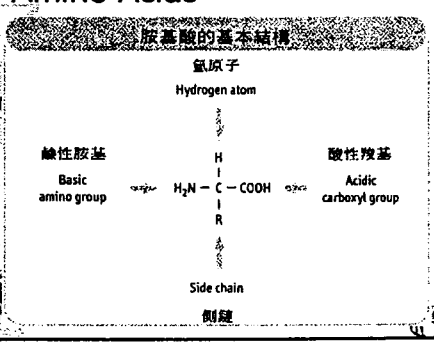
Genes and Gene Structure

- RNA-specifying genes are transcribed into RNAs but are not translated into protein.

Amino Acids

- Amino Acids are the elementary structure units of proteins
- Each amino acid has an – NH2 group and a HOOH group on either side of a central carbon called the α carbon. Also attached to the α carbon are a hydrogen atom and a side chain, also denoted as the – R group

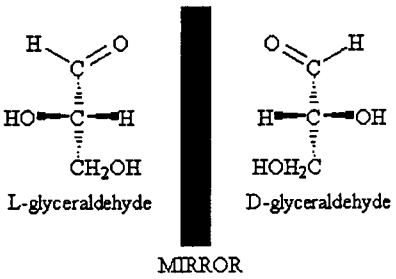
Amino Acids



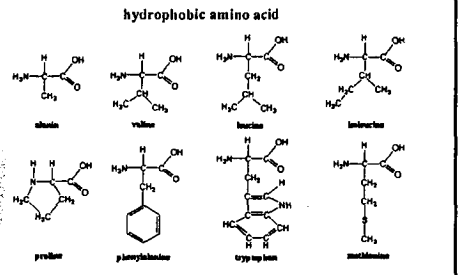
Amino Acids

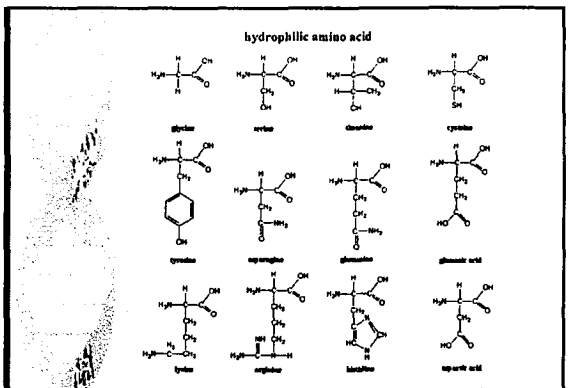
- Mirror-image : levorotatory (L) and dextrorotatory (D)
- Only L-amino acid are used in the process of translation of mRNAs into protein
- The classification of amino acids is made on the basis of their – R group

Amino Acids



Amino Acids



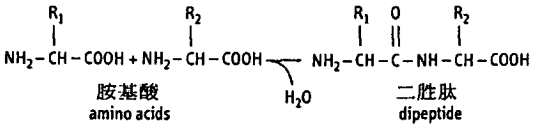


Proteins

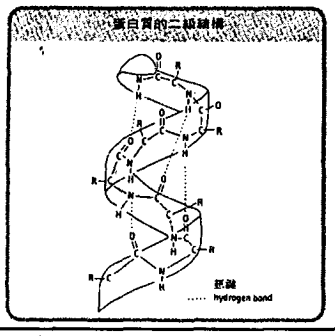
- A protein is a macromolecule that consists of one or more polypeptide chains.
- In a polypeptide chain, the α-carboxyl group of one amino acid is joined to the α-amino group of another amino acid by a peptide bond.

Proteins

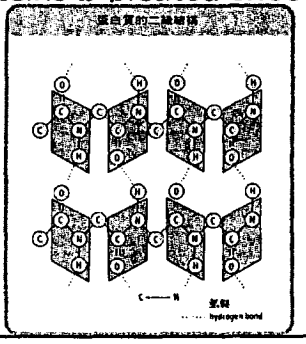
二肽肽的结构



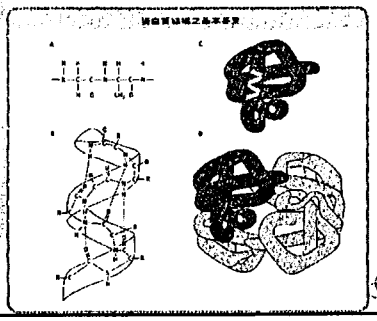
Proteins α-helix



Proteins ß-pleated sheet



Proteins



Translation and Genetic Codes

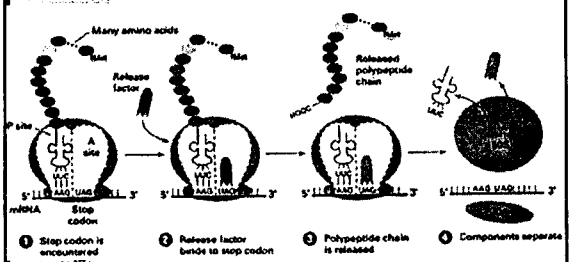
- 依照mRNA的密碼，藉tRNA攜帶amino acid ，在ribosome上合成polypeptide的過程。
- Translation過程分為:

起始(initiation)

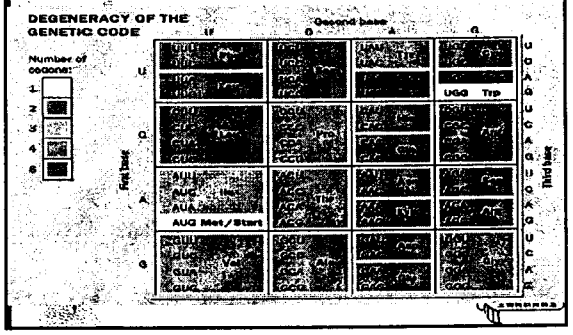
延長(elongation)

終止(termination)

Translation and Genetic Codes



Translation and Genetic Codes



Mutation

Error in either DNA replication or repair occur giving rise to new sequence

- A. 依遺傳物質的改變可分為

1. 染色體的突變(chromosomal mutation)
染色體的數目或構造改變所引起的變化。

2. 點突變(point mutation)
DNA分子內一個或數個鹼基發生改變，可分為以下兩類:

a. 鹼基取代突變(base substitution mutation)
DNA內一個鹼基被另一個鹼基所取代。

Mutation

b. 框架轉移突變(frameshift mutation)
可分為以下二種:

(1). 鹼基插入(base insertion): DNA分子內多插了一個或數個鹼基。

(2). 鹼基缺失(base deletion): DNA分子內缺了一個或數個鹼基。

AAACCCGGG
TTTGGGCCC — Original sequence

AAACTCGGG
TTTGAGCCC — substitution

AAACCCTGGG
TTTGGGACCC — insertion

AAACCGGG
TTTGGCCC — deletion

Frameshift mutation

Wild type

5'—C A A U C C C G G----------3'
Gln Ser Arg

Mutant

5'---C A A A U C C C G----------3'
Gln Ile Pro

Mutation

3. DNA base改變

有兩種形式:

a. 轉換突變(transition mutation)

The purine (A and G) in a base pair is replaced by the other purine and the pyrimidine (C and T) is replaced by the other pyrimidine. For example, AT→GC

b. 顛換突變(transversion mutation)

the purines change to pyrimidine and vice versa. For example, AT→CG

Mutation

B. 依氨基酸改變造成的影響可分為:

1. missense mutation

DNA分子內鹼基改變，結果合成的蛋白質中其中一個氨基酸發生改變，使得蛋白質帶有一個錯誤的氨基酸。

2. nonsense mutation

DNA分子內鹼基改變形成一個終止密碼(stop codon)，例如CAG→UAG，造成蛋白質合成半途終止，產生一條不完整的polypeptide chain。

Mutation

3. silent mutation

DNA分子內鹼基改變，但外表型無影響。通常triplet codon第三個鹼基(wobble位置)改變，容易造成silent mutation，因為通常不會影響密碼的訊息。例如UCA→UCG，結果氨基酸不變，因二者均製造出serine。

4. neutral mutation

DNA分子內鹼基改變，新的氨基酸與原有的氨基酸構造類似。例如AAA→AGA，使得lysine變成arginine，此二種氨基酸性質類似，通常這種突變不會影響蛋白質的功能。

Mutation

5. leaky mutation

DNA分子內鹼基改變，造成一個氨基酸的改變，結果產生的酵素活性降低或者使得微生物的生長降低。

Mutation

C. 依外表型或營養型的改變可分為:

1. morphological mutation

外表型態的改變，例如果蠅(fruit fly)由紅眼變白眼。白色松鼠或白烏鴉等都是製造色素的基因突變所造成。

2. Biochemical and nutritional mutation

產生某種生化產物的能力發生改變，造成外表型改變。

a. sugar utilization mutation

$lac^+ \rightarrow lac^-$

Mutation

b. amino acid production mutation

$arg^+ \rightarrow arg^-$

3. antibiotic resistant mutation

$amp^s \rightarrow amp^r$ (s: sensitivity, r: resistant)

4. lethal mutation

突變結果會造成生物體死亡。例如Tay-Sachs disease。

5. conditional mutation

突變是否表現由環境決定

MUTATION

指導教授：吳哲賢
M9202055 研一甲 朱廷翰

Mutation

- **Def :**
  The errors in either DNA replication or repair occur giving rise to new sequences.
- **Type :**
  1. substitution
  2. recombination
  3. deletion & insertion
  4. inversion

Mutation

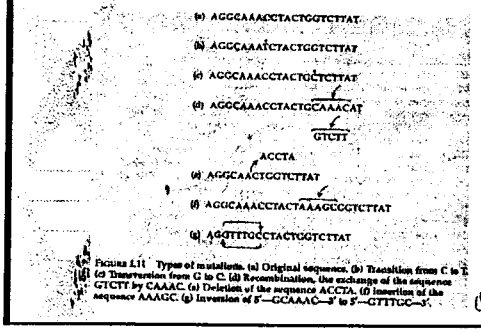
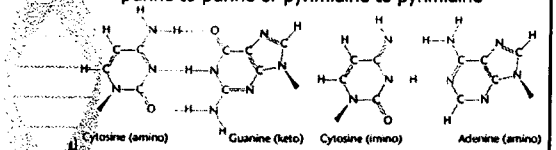


FIGURE 1.11 Types of mutations. (a) Original sequence. (b) Transition from C to T. (c) Transversion from G to C. (d) Recombination, the exchange of the sequence GTCTT by CAAAC. (e) Deletion of the sequence ACCTA. (f) Insertion of the sequence AAAGC. (g) Inversion of 5'—GCAAAC—3' to 5'—GTTTGC—3'.

Substituion

- **Type :**
  1. Transition
     purine to purine or pyrimidine to pyrimidine



  2. Transversion
     purine ⇔ pyrimidine

Substituion

- **Effect :**
  1. synonymous
     nucleotide change that does not alter the amino acid sequence.
  2. nonsynonymous
     1.missense
     mutation changes one amino acid into another amino acid
     2.nonsense
     changes amino acid codon into a stop codon.

Substituion

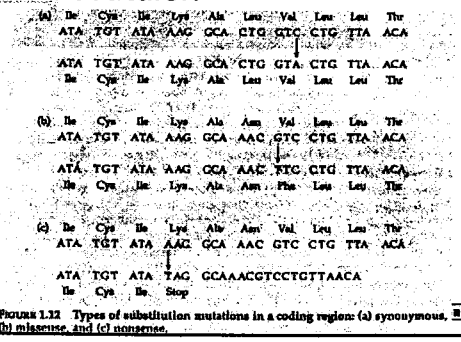
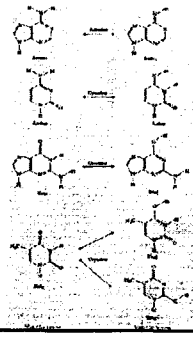


FIGURE 1.12 Types of substitution mutations in a coding region: (a) synonymous, (b) missense, and (c) nonsense.

Substituion

Substitution	Number	Percent
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4
Total in first codons	183	100
Synonymous	8	4
Nonsynonymous	175	96
Missense	166	91
Nonsense	9	5
Total in second codons	183	100
Synonymous	0	0
Nonsynonymous	183	100
Missense	176	96
Nonsense	7	4
Total in third codons	183	100
Synonymous	126	69
Nonsynonymous	57	31
Missense	50	27
Nonsense	7	4

Substituion

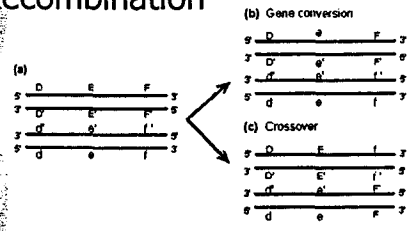


Recombination

- **Homologous recombination**
  It occurs between two homologous DNA molecules.

Two types of homologous recombination
1. crossing over
2. gene conversion

Recombination



- Gene conversion - the red DNA **donates** part of its genetic information (e-e' region) to the blue DNA.
- DNA crossing over - the two DNAs **exchange** part of their genetic information (f-f' and F-F').

Recombination

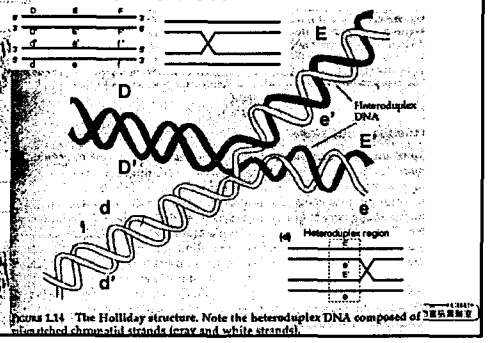
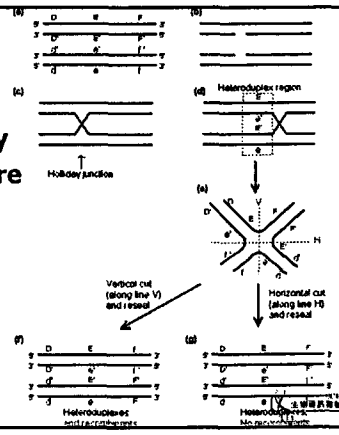
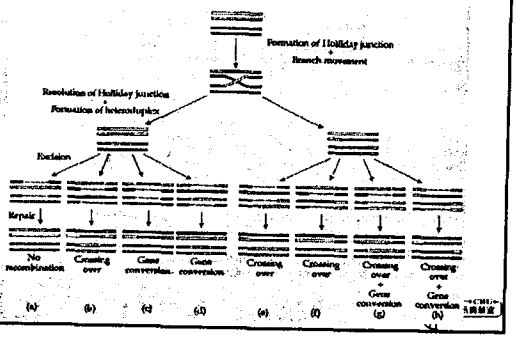


FIGURE 1.14 The Holliday structure. Note the heteroduplex DNA composed of

Holliday structure



Recombination



Recombination

- **Site-specific recombination**

  ex: the integration of phage genomes into bacterial chromosomes

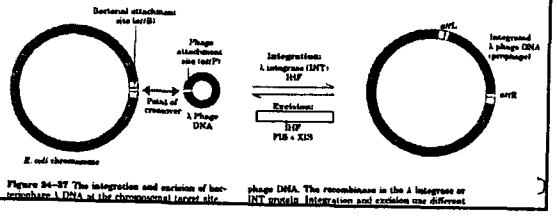


Figure 24–37 The integration and excision of bacteriophage λ DNA at the chromosomal target site ... phage DNA. The recombinases in the λ integrase or INT protein. Integration and excision use different

Deletion and insertion

- **Mechanisms**

  Deletion and insertion can occur by several mechanisms :

  1. unequal crossing over
  2. intrastrand deletion
  3. replication slippage
  4. transposition

Deletion and insertion

- **unequal crossing over**

  Crossing over between homologous chromosomes that are not precisely paired, resulting in non-reciprocal exchange of material and chromosomes of unequal length.

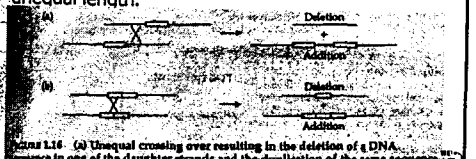


Figure 1.16 (a) Unequal crossing over resulting in the deletion of a DNA sequence in one of the daughter strands and the duplication of the same sequence in the other strand. (b) When a DNA segment is duplicated in tandem, the chromo

Deletion and insertion

- **Intrastrand deletion**

  a repeated sequence pairs with another in the same orientation on the same chromatid , and an intra-chromosomal crossing over event occurs

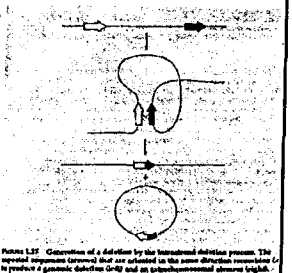


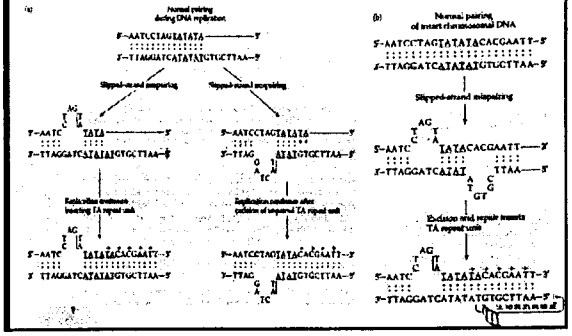
Figure 1.17 Generation of a deletion by the intrastrand deletion process. The repeated sequences (arrows) that are oriented in the same direction recombine (...) to produce a genomic deletion (left) and an extrachromosomal element (right).

Deletion and insertion

- **replication slippage**

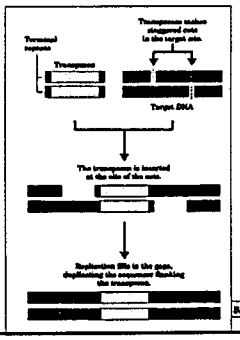
  It occurs at the repetitive sequences when the new strand mispairs with the template strand.

Deletion and insertion

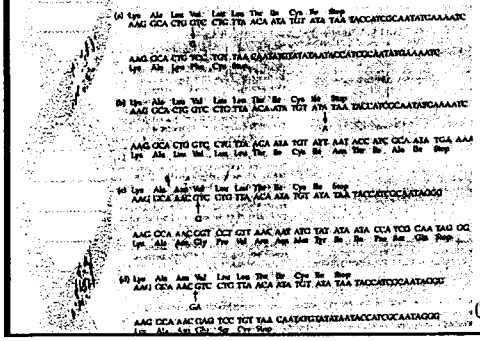


Deletion and insertion

- **Transposition**
  It is the process that a transponse is inserted into a target DNA.

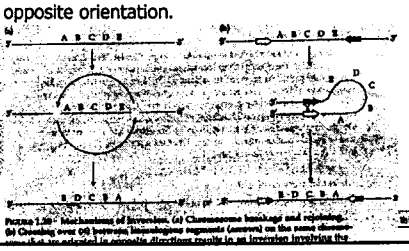


Deletion and insertion



Invertion

- **Inversion**
  Chromosome segment excises and reintegrates in opposite orientation.



Mutation Rate

- **Spontaneous**
  mammalian nuclear DNA :
  3 ~ 5 x 10^-6 per nucleotide per year
  mammalian mitochondrial DNA :
  3 ~ 5 X 10^-5 per nucleotide per year
  virus :
  10^-2 per nucleotide per year

Mutation Rate

- **Type**
  nuclear DNA
  Transition : 60 ~ 70 %
  Transvertion : 33%
  mitochondrial DNA
  Transition : transvertion = 15 ~ 20 : 1
- **Hotspots**
  Some region has higher frequency

Chapter 2

Dynamics of Genes in Populations

M09202050
顏勝茂

INTRODUCTION

- Evolution is the process of change in the genetic makeup of populations.
- A basic problem in evolutionary population genetics is to determine how the frequency of a mutant allele will change in time under the effect of various evolutionary forces.

CHANGES IN ALLELE FREQUENCIES

- The chromosomal or genomic location of a gene is called a **locus.**
- Alternative forms of the gene at a given locus are called **alleles.**
- The relative proportion of an allele is referred to as the **allele frequency** or **gene frequency.**(某種基因在一個族群裏所佔的多寡，我們稱爲基因頻率)

CHANGES IN ALLELE FREQUENCIES

- EX:a haploid population of size N individuals, tow alleles , A_1 and A_2.

n_1爲 A_1複製的數目,n_2爲 A_2複製的數目.

則allele frequencies爲

$A_1 = n_1 / N, A_2 = n_2 / N$

Note: $n_1 + n_2 = N$ and $n_1 / N + n_2 / N = 1$

CHANGES IN ALLELE FREQUENCIES

- The set of all alleles existing in a population at all loci is called the **gene pool.**(族群的所有基因稱爲基因庫)
- For a new mutation to become significant from an evolutionary point of view it must increase in frequency and ultimately become fixed in the population.

CHANGES IN ALLELE FREQUENCIES

- The major factors affecting the frequency of alleles in populations are **natural selection** and **random genetic drift.**

CHANGES IN ALLELE FREQUENCIES

Two mathematical model to studying genetic changes

1. Deterministic model
2. Stochastic models

CHANGES IN ALLELE FREQUENCIES

Deterministic model(預測模式)

- The deterministic model is simpler.
- It assumes that changes in the frequencies of alleles in population form generation to generation occur in a unique manner and can be unambiguously predicted from knowledge of initial conditions.

CHANGES IN ALLELE FREQUENCIES

Need two conditions:

1. The population is infinite in size.
2. The environment either remains constant with time or changes according to deterministic rules.

CHANGES IN ALLELE FREQUENCIES

Stochastic models(機率模式)

Stochastic models assume that changes in allele frequencies occur in a probabilistic manner.

NATURAL SELECTION

Natural selection(天擇) is defined as the differential reproduction of genetically distinct individuals or genotypes within a population.

NATURAL SELECTION

The **fitness of a genotype,** commonly denoted as w, is a measure of the individual's ability to survive and reproduce.

- In nature, the fitness of a genotype is not expected to remain constant for all generations and under all environmental circumstances.

NATURAL SELECTION

- **Deleterious**:reduce the fitness of genotype.This type of selection is called **negative** or **purifying selection**.
- **Neutral**:as fit as the best allele in population.
- **Advantageous**:increase the fitness of genotype. Is call **positive** or **advantageous selection**.

NATURAL SELECTION

- 2個 alleles 則有3個 genotypes 分別爲:
  A_1A_1, A_1A_2, A_2A_2
  Fitness分別爲 w_{11}, w_{12}, w_{22}
  A_1 allele frequency -> p
  A_2 allele frequency -> $q = 1 - p$

NATURAL SELECTION

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	w_{11}	w_{12}	w_{22}
Frequency	p^2	$2pq$	q^2

NATURAL SELECTION

- A_2 allele frequency in the next generation(q_{t+1}).

$$q_{t+1} = \frac{pqw_{12} + q^2w_{22}}{p^2w_{12} + 2pqw_{12} + q^2w_{22}} \qquad (2.1)$$

NATURAL SELECTION

- A2每世代之間frequency的差異

$$\Delta q = q_{t+1} - q_t$$

$$= \frac{pq[p(w_{12} - w_{11}) + q(w_{22} - w_{12})]}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \qquad (2.2)$$

$$\frac{pqw_{12} + q^2w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} - q$$

$$= \frac{pqw_{12} + q^2w_{22} - p^2qw_{11} - 2pq^2w_{12} - q^3w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}}$$

$$= \frac{pqw_{12} - 2pq^2w_{12} - p^2qw_{11} + q^2w_{22} - q^3w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}}$$

$$= \frac{pqw_{12}(p - q) - p^2qw_{11} + pq^2w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}}$$

$$= \frac{pq[p(w_{12} - w_{11}) + q(w_{22} - w_{12})]}{p^2w_{11} + 2pqw_{12} + q^2w_{22}}$$

NATURAL SELECTION

Five common modes
1. codominance or genic selection
2. complete recessiveness
3. complete dominance
4. overdominance
5. underdominance

NATURAL SELECTION

- Assume population is diploid.

Allele	A_1		
Genotype	A_1A_1		
Allele	A_1 & A_2		
Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	1	1+s	1+2s

Codominance

- Two homozygotes fitness different.
- Heterozygote is the mean of the fitness of the two homozygous genotypes.
- 異行結合子(Heterozygote)同時兼具親代的外表型.

Codominance

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	1	1+s	1+2s

- fitness代入 Equation 2.2

$$\Delta q = \frac{spq}{1+2spq+2sq^2} \qquad (2.3)$$

$$\Delta q = \frac{pq[p(w_{12}-w_{11})+q(w_{22}-w_{12})]}{p^2w_{11}+2pqw_{12}+q^2w_{22}}$$
$$= \frac{pq[p(1+s-1)+q(1+2s-(1+s))]}{p^2+2pq(1+s)+q^2(1+2s)}$$
$$= \frac{pq[ps+qs]}{p^2+2pq+2pqs+q^2+2q^2s}$$
$$= \frac{pqs(p+q)}{1+2pqs+2q^2s}$$
$$= \frac{spq}{1+2spq+2sq^2}$$

Codominance

- If s is small, the denominator in Equation 2.3 approximately 1, and the equation is reduced to $\Delta q = spq$.
- 近似於微分方程式：

$$\frac{dq}{dt} = spq = sq(1-q) \qquad (2.4)$$

Codominance

$$q_t = \frac{1}{1+\left(\frac{1-q_0}{q_0}\right)e^{-st}} \qquad (2.5)$$

- The frequency qt is expressed as a function of time t.

Codominance

$$t = \frac{1}{s}\ln\frac{q_t(1-q_0)}{q_0(1-q_t)} \qquad (2.6)$$

- The number of generation form q_0 to q_t.

$$q_t = \frac{1}{1+\left(\frac{1-q_0}{q_0}\right)e^{-st}}$$

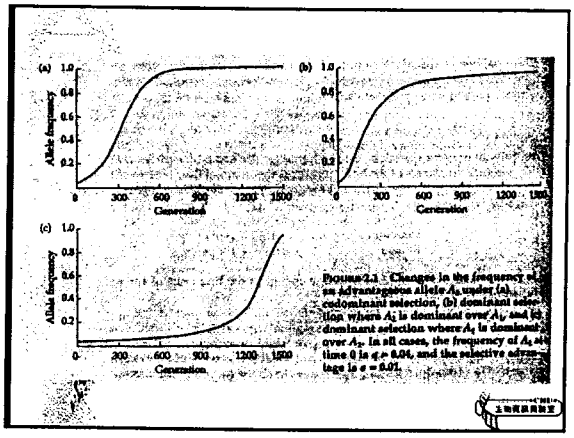
$$\left(\frac{1-q_0}{q_0}\right)e^{-st} = \frac{1}{q_t} - 1 = \frac{1-q_t}{q_t}$$

$$\left(\frac{1-q_0}{q_0}\right)\left(\frac{q_t}{1-q_t}\right) = e^{st}$$

$$\frac{q_t(1-q_0)}{q_0(1-q_t)} = e^{st}$$

$$\ln\frac{q_t(1-q_0)}{q_0(1-q_t)} = st$$

$$t = \frac{1}{s}\ln\frac{q_t(1-q_0)}{q_0(1-q_t)}$$



Dominance

- The two homozygotes have different fitness values,
- The fitness of the heterozygote is the same as the fitness of one of the two homozygous genotypes

Dominance

- A_2 is dominant over allele A_1.
- Genotype A_1A_1 A_1A_2 A_2A_2
- Fitness 1 1+s 1+s
- fitness代入 Equation 2.2

$$\Delta q = \frac{p^2qs}{1-s-p^2s} \qquad (2.7)$$

$$\Delta q = \frac{pq[p(w_{12}-w_{11})+q(w_{22}-w_{12})]}{p^2w_{11}+2pqw_{12}+q^2w_{22}}$$
$$= \frac{pq[p(1+s-1)+q(1+s-(1+s))]}{p^2+2pq(1+s)+q^2(1+s)}$$
$$= \frac{pq[ps]}{p^2+2pq+2pqs+q^2+q^2s}$$
$$= \frac{p^2qs}{1+2pqs+q^2s}$$
$$= \frac{p^2qs}{1+2sp(1-p)+(1-p)^2s}$$
$$= \frac{p^2qs}{1+2sp-2sp^2+s-2sp+sp^2}$$
$$= \frac{p^2qs}{1+s-p^2s}$$

Dominance

- A_1 dominant over A_2.

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	1	1	1+s

- fitness代入 Equation 2.2

$$\Delta q = \frac{pq^2s}{1+q^2s} \qquad (2.8)$$

$$\Delta q = \frac{pq[p(w_{12}-w_{11})+q(w_{22}-w_{12})]}{p^2w_{11}+2pqw_{12}+q^2w_{22}}$$
$$= \frac{pq[p(1-1)+q(1+s-1)]}{p^2+2pq+q^2(1+s)}$$
$$= \frac{pq[qs]}{p^2+2pq+q^2+q^2s}$$
$$= \frac{pq^2s}{1+q^2s}$$

Overdominance and underdominance

- Overdominance selection,the heterozygote has the highest fitness.

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	1	1+s	1+t

- Note:s>0 & s>t,t can be +,zero,-

Overdominance and underdominance

- fitness代入 Equation 2.2

$$\Delta q = \frac{pq(2sq-tq-s)}{1+2spq+tq^2} \qquad (2.9)$$

$$\Delta q = \frac{pq[p(w_{12}-w_{11})+q(w_{22}-w_{12})]}{p^2w_{11}+2pqw_{12}+q^2w_{22}}$$
$$= \frac{pq[p(1+s-1)+q(1+t-(1+s))]}{p^2+2pq(1+s)+q^2(1+t)}$$
$$= \frac{pq[ps+qt-gs]}{p^2+2pq+2pqs+q^2+q^2t}$$
$$= \frac{pq[s(p-q)+qt]}{1+2pqs+q^2t}$$
$$= \frac{pq[s(1-2q)+qt]}{1+2pqs+q^2t}$$
$$= \frac{pq(-2sq+tq+s)}{1+2spq+tq^2}$$

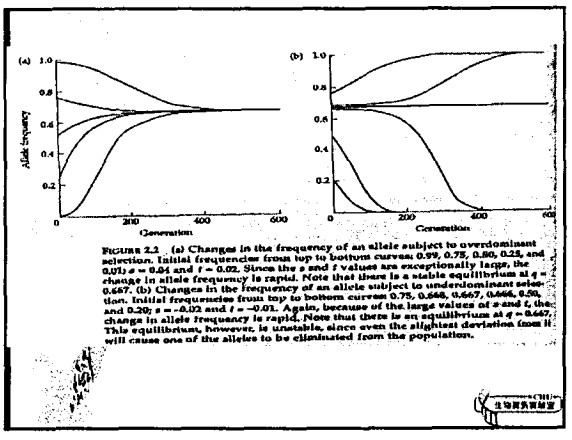


FIGURE 2.2 (a) Changes in the frequency of an allele subject to overdominant selection. Initial frequencies from top to bottom curves: 0.99, 0.75, 0.50, 0.25, and 0.01; $s = 0.04$ and $t = 0.02$. Since the s and t values are exceptionally large, the change in allele frequency is rapid. Note that there is a stable equilibrium at $q = 0.667$. (b) Changes in the frequency of an allele subject to underdominant selection. Initial frequencies from top to bottom curves: 0.75, 0.668, 0.667, 0.666, 0.50, and 0.20; $s = -0.02$ and $t = -0.01$. Again, because of the large values of s and t, the change in allele frequency is rapid. Note that there is an equilibrium at $q = 0.667$. This equilibrium, however, is unstable, since even the slightest deviation from it will cause one of the alleles to be eliminated from the population.

Overdominance and underdominance

- Stable equilibrium
- Overdominant selection called **balancing** or **stabilizing selection**.
- A_2 frequency allele at equilibrium, $\hat{q}$
- $\Delta q = 0$,由(2.9)得

$$\hat{q} = \frac{s}{2s - t} \qquad (2.10)$$

Overdominance and underdominance

- Underdominant selection,the heterozygote has the lowest fitness.
- Genotype A_1A_1 A_1A_2 A_2A_2
- Fitness 1 1+s 1+t
- s<0 & s<t
- Δq 和(2.9)同

Overdominance and underdominance

- Unstable equilibrium.
- Below the equilibrium frequency to extinction.

RANDOM GENETIC DRIFT

- Allele frequency changes can occur by chance.
- The process of change in allele frequency due to solely to chance effects is called **random genetic drift.**

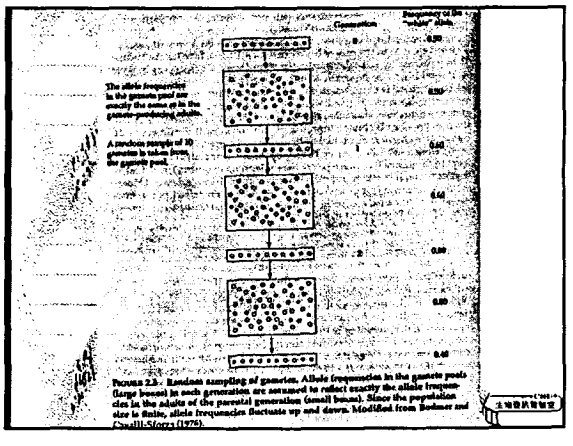


FIGURE 2.3 Random sampling of gametes. Allele frequencies in the gamete pools (large boxes) in each generation are assumed to reflect exactly the allele frequencies in the adults of the parental generation (small boxes). Since the population size is finite, allele frequencies fluctuate up and down. Modified from Bodmer and Cavalli-Sforza (1976).

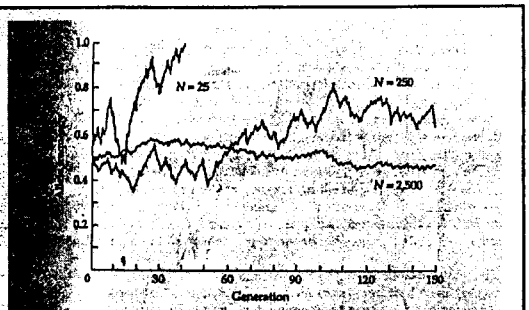
RANDOM GENETIC DRIFT

$$P_i = \frac{(2N)!}{i!(2N-i)!} p^i q^{2N-i} \qquad (2.11)$$

- $P_i > 0, 0 < q < 1$

RANDOM GENETIC DRIFT

- Diploid population
- N individuals
- The population contains 2N genes.
- 2 alleles, A_1 and A_2
- A_1 frequency p
- A_2 frequency $q = 1 - p$
- i：樣本中 A_1 allele type的數目



Changes in frequencies of alleles subject to random genetic drift in ...ions of different sizes. The smallest population (N_e = 25) reached fixation ...generations. The other two populations were still polymorphic after 150 ...ions, but will ultimately reach fixation if the experiment is continued long ...From Bodmer and Cavalli-Sforza (1976).

RANDOM GENETIC DRIFT

- The direction of change is random at any point in time.
- The random genetic drift is that fluctuations in allele frequencies are much more pronounced in small populations than in larger ones.

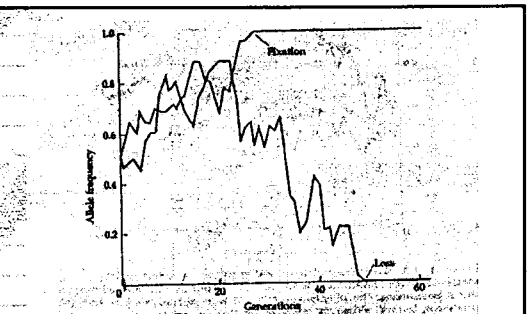


FIGURE 2.5 Two possible outcomes of random genetic drift in populations of size 25 and p_0 = 0.5. In each generation, 25 alleles were sampled with replacement from the previous generation. In the population represented by the gray line, the allele becomes fixed in generation 27; in the other population, the allele is lost in generation 49. Modified from Li (1997).

RANDOM GENETIC DRIFT

- The mean of frequency of allele
  $A_1 \rightarrow \overline{p}_0$
- The variance of the frequency of allele

$$A_1 \rightarrow V(p_t) \quad \overline{p}_t = p_0 \qquad (2.12)$$

$$V(p_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right] \qquad (2.13)$$

$$\approx p_0(1-p_0)\left[1-e^{-t/(2N)}\right]$$

RANDOM GENETIC DRIFT

- The mean frequency does not change with time.
- The variance increase with time.

RANDOM GENETIC DRIFT

The cumulative effect of random genetic drift.

5 diploid individuals. alleles frequency A_1 and A_2 are each 50%.

- Ask:what is the probability of obtaining the same allele frequencies in the next generation ?

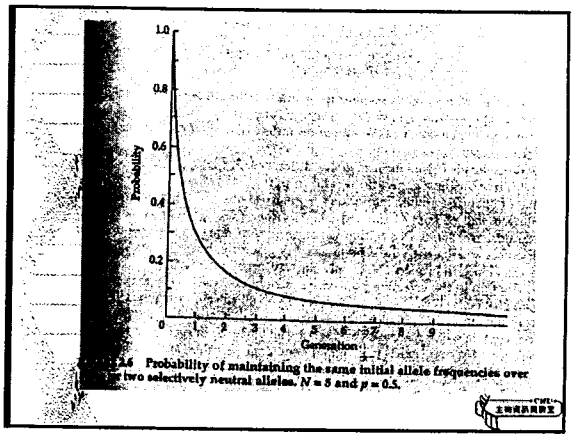
RANDOM GENETIC DRIFT

Using Equation 2.11, we obtain probability of 25 %. In other words, 75% will be different form the initial allele frequency.

$$\frac{(2N)!}{i!(2N-i)!}p^i q^{2N-i}$$

$$=\frac{10!}{5!5!}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5$$

$$=\frac{252}{1024}=0.24609375\approx 25\%$$



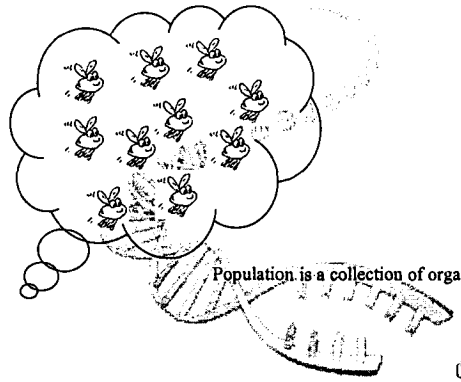
Probability of maintaining the same initial allele frequencies over two selectively neutral alleles. $N = 5$ and $p = 0.5$.

RANDOM GENETIC DRIFT

- The probability of either A_1 or A_2 being lost increase with time.
- The frequency of an allele reaches either 0 or 1, its frequency will not change in subsequent generations.
- The first case is refereed to as **loss** or **extinction**, and the second **fixation**

Chapter 2
Dynamics of Genes in Populations

CHU,CSIE
M09102048
賴韋丞



One locus with two alleles, A_1 and A_2

$f(A_1) = p$
$f(A_2) = q$

$p + q = 1$

$A_1A_1 = 6$
$A_1A_2 = 3$
$A_2A_2 = 1$

$f(A_1) = (6 \times 2 + 3) / 2 \times 10 = 0.75$
$f(A_2) = (1 \times 2 + 3) / 2 \times 10 = 0.25$

The Hardy-Weinberg equilibrium

One locus with two alleles, A_1 and A_2

A_1A_1

$f(A_1A_1) = p^2$
$f(A_1A_2) = 2pq$
$f(A_2A_2) = q^2$

Natural Selection — the differential reproduction of genetically distinct individuals or genotypes within a population.

Fitness (ω) — a measure of the individual's ability to survive and reproduce.

- The evolutionary success of an individual is determined not by it absolute fitness but by its relative fitness

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness

$$q_{t+1} = \frac{pq\,\omega_{12} + q^2\,\omega_{22}}{p^2\,\omega_{11} + 2pq\,\omega_{12} + q^2\,\omega_{22}}$$

$$q_{t+1} = \frac{pq\,\omega_{12} + q^2\,\omega_{22}}{p^2\,\omega_{11} + 2pq\,\omega_{12} + q^2\,\omega_{22}}$$

$$\Delta q = q_{t+1} - q_t = \frac{pq\,[p\,(\omega_{12} - \omega_{11}) + q\,(\omega_{22} - \omega_{12})]}{p^2\,\omega_{11} + 2pq\,\omega_{12} + q^2\,\omega_{22}}$$

Δq is frequency of allele A_2 per generation

Consider a recessive allele, A_1, double the fitness of its carrier
If $f(A_2) = 0.2$ in this generation, what is it in next generation?

Genotype A_1A_1 A_1A_2 A_2A_2

$(0.8)^2$ $2(0.8)(0.2)$ $(0.2)^2$

Fitness 1 0.5 0.5

$p^2\omega_{11}$ $2pq\,\omega_{12}$ $q^2\omega_{22}$

$$q_{t+1} = \frac{0.16 \times 0.5 + 0.04 \times 0.5}{0.64 + 0.32 \times 0.5 + 0.04 \times 0.5} = 0.12$$

Dominance(A_2 is dominant over the old A_1)

relative fitness value of 1

WW 白 -> 1
WB 黑 -> 1+s
BB 黑 -> 1+s

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness 1 1+s 1+s

p^2 $2pq(1+s)$ $q^2(1+s)$

$$\Delta q = \frac{p^2qs}{1 + qs - p^2s}$$

Dominace(A_1 is dominant over A_2, A_2 is recessive)

WW 白 -> 1
WB 白 -> 1
BB 黑 -> 1+s

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness 1+s 1 1+s

$p^2(1+s)$ $2pq$ $q^2(1+s)$

$$\Delta q = \frac{p^2qs}{1 + q^2s}$$

Codominance

WW 白 -> 1
WB 灰 -> 1+s
BB 黑 -> 1+2s

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness 1 1+s 1+2s

p^2 $2pq(1+s)$ $q^2(1+2s)$

$$\Delta q = \frac{spq}{1 + 2spq + 2sq^2}$$

Overdominace and underdominace

$s < t$, $s < 0$ Underdominance
$s > t > 0$ Overdominance

AA -> 1
Aa -> 1+s
aa -> 1+t

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness 1 1+s 1+t

p^2 $2pq(1+s)$ $q^2(1+t)$

$$\Delta q = \frac{pq(2sq - tq - s)}{1 + 2spq + 2tq^2}$$

For mathematical convenience,
we shall assign a ***relative fitness*** value of 1
$\omega_{11} = 1$, $\omega_{12} = 1 + s$, and $\omega_{22} = 1 + t$

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness 1 1+s 1+t

p^2 $2pq(1+s)$ $q^2(1+t)$

$$q_{t+1} = \frac{pq(1+s) + q^2(1+t)}{p^2 + 2pq(1+s) + q^2(1+t)}$$

$$q_{t+1} = \frac{pq(1+s) + q^2(1+t)}{p^2 + 2pq(1+s) + q^2(1+t)}$$

$$\Delta q = q_{t+1} - q_t = \frac{pq(1+s) + q^2(1+t)}{p^2 + 2pq(1+s) + q^2(1+t)} - q_t$$

$$\frac{-pq(2sq - tq - s)}{p^2 + 2pq(1+s) + q^2(1+t)}$$

At equilibrium, $\Delta q = 0$

$$\frac{-pq(2sq - tq - s)}{p^2 + 2pq(1+s) + q^2(1+t)} = 0$$

$$2sq - tq - s = 0$$

$$q = \frac{s}{2s - t}$$

Genotype A_1A_1 A_1A_2 A_2A_2

p^2 $2pq$ q^2

Fitness 1 1+s 1+t

$t = s$ Dominant

$t = 2s$ Codominance

$t > s$, $s < 0$ Underdominance

$s > t > 0$ Overdominance

Random Genetic Drift

Allele frequency changes can also occur by chance, in which case the changes are not directional but random. The process of changes in allele frequency due solely to chance effects is called **random genetic drift.**

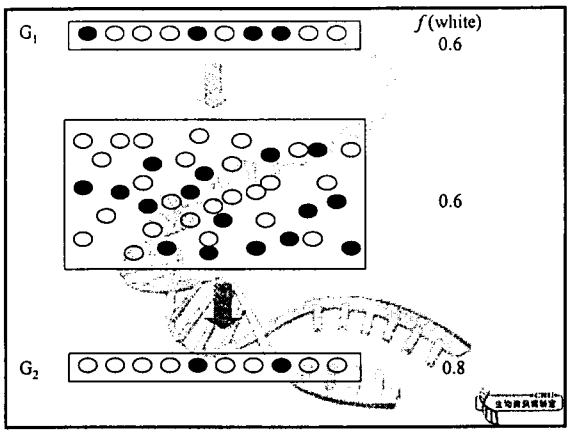
f(white)
0.5

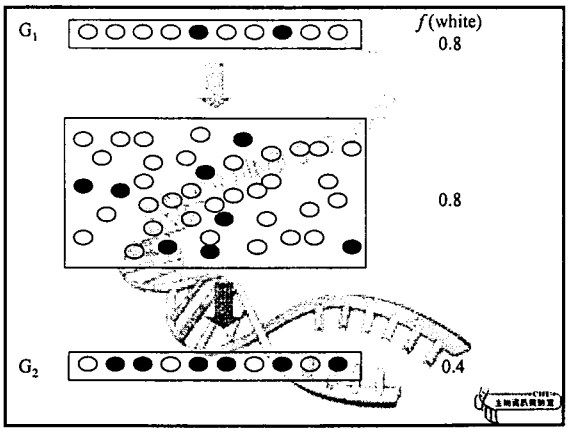
0.5

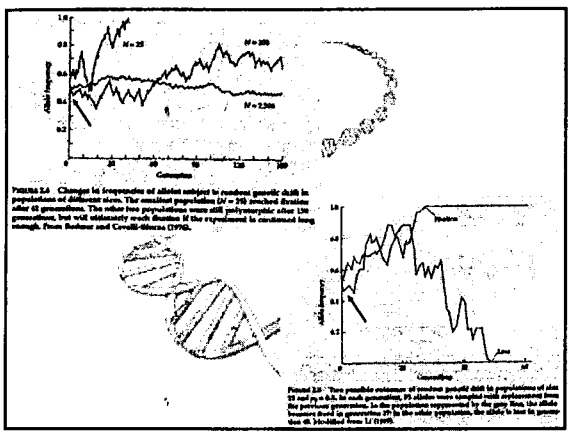
G_0 f(white) 0.5

0.5

G_1 0.6







A diploid population with N individuals
⇒ at any given locus, there are $2N$ genes
one locus with two alleles, A_1 and A_2

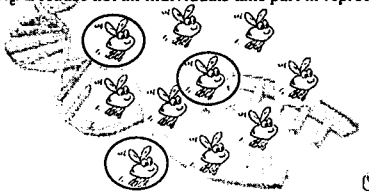
$f(A_1) = p$
$f(A_2) = 1 - p = q$

When $2N$ gametes are sampled from the infinite gamete pool, the probability, P_i, that the sample contains exactly i alleles of genotype A_1 is given by the binomial probability function

$$P_i = \frac{(2N)!}{i!\,(2N-i)!} p^i q^{2N-i}$$

Census population size, N, defined as the total number of individuals in a population.

From the point of view of population genetics and evolution, however, the relevant number of individuals to be considered consists of only those individuals that actively participate in reproduction. This part is called the **effective population size** and is denoted by N_e. Because not all individuals take part in reproduction.



In general, N_e is smaller, sometimes much smaller, than N.
In a population consist of N_m males and N_f females ($N = N_m + N_f$), N_e is given by

$$N_e = \frac{4 N_m N_f}{N_m + N_f}$$

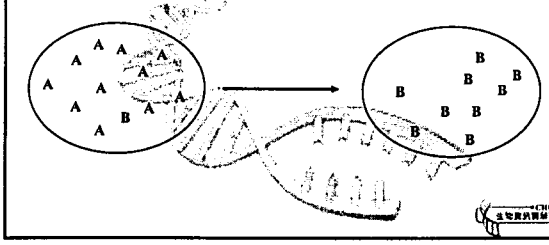
The long-term effective population size in a species for a period of n generations is given by

$$N_e = \frac{n}{1/N_1 + 1/N_2 + \cdots + 1/N_n}$$

Where N_i is the population size of the ith generation. In other words, N_e equals the harmonic mean of the N_i values, and consequently it is closer to the smallest value of N_i than to the largest one.

Gene Substitution

Gene substitution is defined as the process whereby a mutant allele completely replaces the predominant or wild type allele in a population. In this process, a mutant allele arises in a population as a single copy and become fixed after a certain number of generations.



Fixation probability

The probability that a particular allele will become fixed in a population depends on 1) its frequency, 2) its selective advantage or disadvantage, s, and 3) the effective population size, N_e.

In the following, we shall consider that the relative fitness of the three genotypes A_1A_1, A_1A_2, and A_2A_2 are 1, 1+ s, and 1+ 2s, respectively. Kimura (1962) showed that the probability of fixation of A_2 is

$$P = \frac{1 - e^{-4N_e sq}}{1 - e^{-4N_e s}}$$

$$P = \frac{1 - e^{-4N_e sq}}{1 - e^{-4N_e s}}$$

Since $e^{-x} \approx 1 - x$ for small values of x, P equals to q as s approaches 0. Thus for a neutral allele, the fixation probability equals its initial frequency in the population.

We note that a new mutant arising as a single copy in a diploid population of size N has an initial frequency of $1/(2N)$.
For a **neutral mutation**, i.e., s = 0, the fixation probability

$$\frac{1}{2N}$$

When $s \neq 0$,

$$P = \frac{1 - e^{-(2N_e s/N)}}{1 - e^{-4N_e s}}$$

If the population size is equal to the effective population size,

$$P = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}}$$

If the absolute value of s is small, we obtain

$$P = \frac{2s}{1 - e^{-4N_e s}}$$

For positive value of s and large value of N,

$$P \approx 2s$$

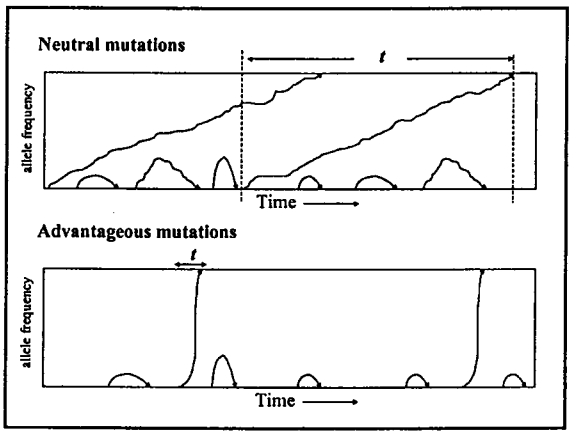
Fixation time

The time required for the fixation or loss of an allele depends on 1) the frequency of the allele, 2) its selective advantage, and 3) the size of the population.
In the following, we deal with the mean fixation time of those mutations that will eventually become fixed in the population. This variable is call conditional fixation time. In the case of a new mutation whose initial frequency in a diploid population is be definition $q = 1/(2N)$, the mean conditional fixation time, t, was calculated by Kimura and Ohta (1969). For a neutral mutation, it is approximated by

$$t = 4N \text{ generations}$$

and for a mutation with a selective advantage of s, it is approximated by

$$t = (2/s)\ln(2N) \text{ generations}$$



Rate of gene substitution

Rate of gene substitution: the number of mutations reaching fixation per unit time.

Neutral mutations: If neutral mutations occur at a rate of u per gene per generation, then the number of mutants arising at a locus in a diploid population of size N is $2Nu$ per generation. Since the probability of fixation for each of these mutations is $1/(2N)$, we obtain the rate of substitution of neutral alleles by multiplying the total number of mutations by the probability of their fixation:

$$K = 2Nu\frac{1}{2N} = u$$

Advantageous mutation: the rate of substitution can also be obtained by multiply the rate of mutation by the probability of fixation for advantageous alleles. When $s > 0$

$$K = 2Nu \times 2s = 4Nsu$$

Genetic polymorphism

Monomorphic: A locus in population is monomorphic if there exists only one allele at the locus.
Polymorphic: A locus in population is polymorphic if two or more alleles coexist. Polymorphic commonly defined as the frequency of the most common allele is less than 99%. Having genetic diversity.

Gene diversity

Proportion of polymorphic loci(P): divide the number of polymorphic loci by the total number of loci smpled.

Example: If 4 of the 20 loci are polymorphic then;
If you survey genetic variation at 20 loci and only 4 loci are polymorphic then,

$P = 4/20 = 0.20$

Measure of genetic variability is the **mean expected heterozygosity** or **gene diversity** or **single-locus expected heterozygosity**, is defined as

$$h = 1 - \sum_{i=1}^{m} x_i^2$$

where x_i is the frequency of allele i and m is the total number of alleles at the locus.

The average of the h values over all the loci studied, H, can be used as an estimate of the extent of genetic variability within the population. That is,

$$H = \frac{1}{n}\sum_{i=1}^{n} h_i$$

Where h_i is the gene diversity at locus i, and n is the number of loci.

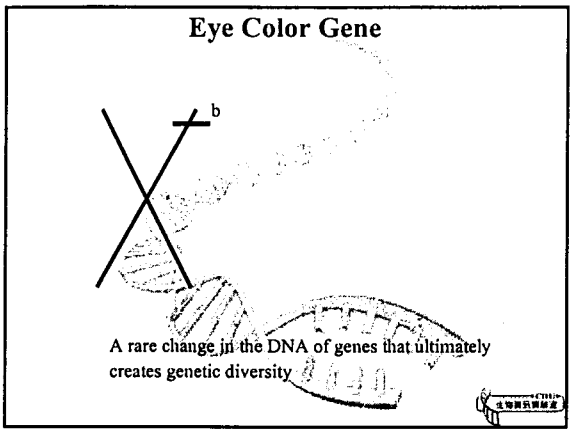
Nucleotide diversity

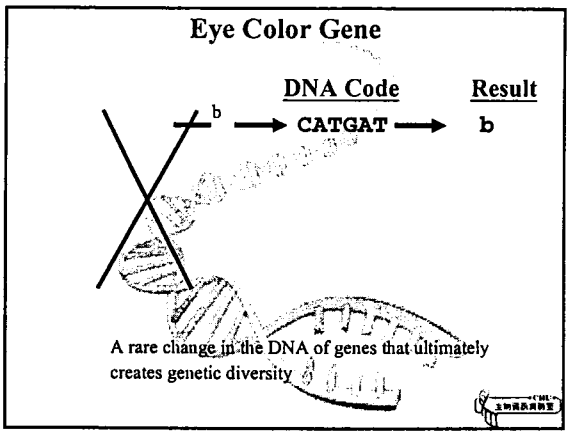
For DNA sequence data, a more appropriate measure of polymorphism in a population is the average number of nucleotide differences per site between any two randomly chosen sequences. This measure is called **nucleotide diversity** and its denoted by Π

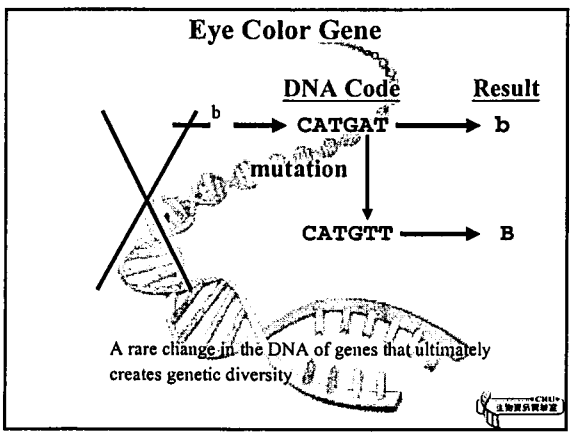
$$\Pi = \sum_{ij} x_i x_j \pi_{ij}$$

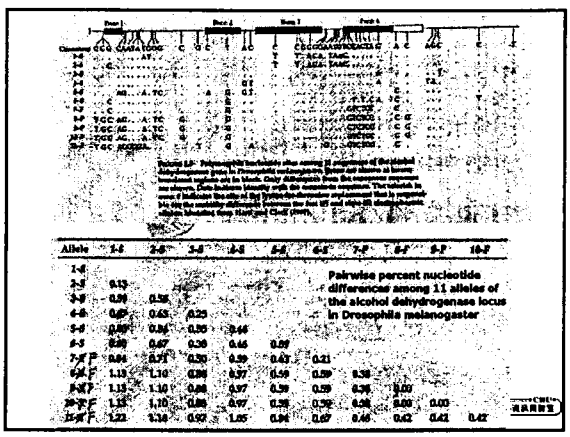
where $x_i x_j$ are the frequencies of the ith and jth type of DNA sequences, respectively, and π_{ij} is the proportion of the different alleles between the ith and jth types.

This is equivalent to gene diversity at the nucleotide level.


Eye Color Gene
b
A rare change in the DNA of genes that ultimately creates genetic diversity


Eye Color Gene
DNA Code
Result
b
CATGAT
b
A rare change in the DNA of genes that ultimately creates genetic diversity


Eye Color Gene
DNA Code
Result
b
CATGAT
b
mutation
CATGTT
B
A rare change in the DNA of genes that ultimately creates genetic diversity


Pairwise percent nucleotide differences among 11 alleles of the alcohol dehydrogenase locus in Drosophila melanogaster

Evolutionary Changes in Nucleotide Sequences

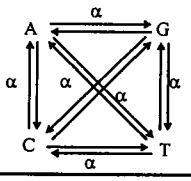
劉家輝　　廖國良
M09102051
2003.9.24

Jukes and Cantor's one-parameter model

This simple model assumes that substitutions occur with equal probability among the four nucleotide types.

Conversely, the probabilities of either T,C or G changing to A are also equal, and the rate of substitution is in each of the three possible directions of change is α.

Because the model involves a single parameter, α, it is called the one-parameter model.

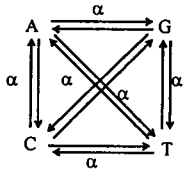


Since we start with A, the probability hat this site is occupied by A at time 0 is $P_{A(0)}=1$.

At time 1, the probability of still having A at this site is given by

$$P_{A(1)} = 1 - 3\alpha$$

In which 3α is the probability of A changing to T, C or G, and $1-3\alpha$ is the probability that A has remained unchanged.

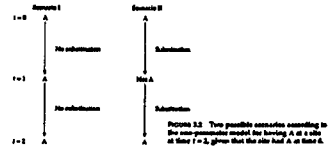


The probability of having A at time 2 is

$$P_{A(2)} = (1-3\alpha)\ P_{A(1)} + \alpha\,[1-P_{A(1)}]$$

To derive this equation, we consider two possible scenarios:
1) The nucleotide has remained unchanged from time 0 to time 2

2) The nucleotide has changed to T, C, or G at time 1, but has subsequently reverted to A at time 2.



The recurrence equation applies to any *t*:

$$P_{A(t+1)} = (1-3\alpha)\ P_{A(t)} + \alpha\,[1-P_{A(t)}]$$

Rewrite this equation in terms of the amount of change in $P_{A(t)}$ per unit time as

$$\begin{aligned}\Delta P_{A(t)} &= P_{A(t+1)} - P_{A(t)} = \{(1-3\alpha)\ P_{A(t)} + \alpha\,[1-P_{A(t)}]\} - P_{A(t)} \\ &= -3\alpha\ P_{A(t)} + \alpha\,[1-P_{A(t)}] \\ &= -4\alpha\ P_{A(t)} + \alpha\end{aligned}$$

Approximate this process by a continuous-time model , by regarding

$$\frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha \qquad (3.5)$$

First-order linear differential equation, and the solution is given by $P_{A(0)}=1$

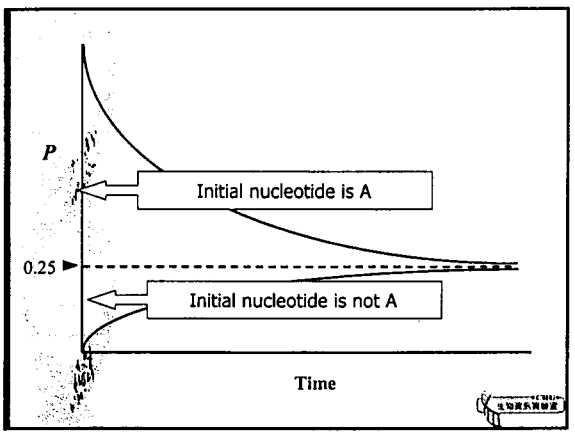
$$P_{A(t)} = \frac{1}{4} + \left(P_{A(0)} - \frac{1}{4}\right)e^{-4\alpha t} \qquad (3.6)$$

Started with A, the probability that the site has A at time 0 is 1, and $P_{A(0)}=1$

$$P_{A(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \qquad (3.7)$$

Not started with A, then

$$P_{A(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \qquad (3.8)$$



$$P_{AA(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \quad (3.9)$$

If the initial nucleotide is G instead of A, then

$$P_{CA(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \quad (3.10)$$

$$P_{ii(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \quad (3.11)$$

$$P_{ij(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \quad (3.12)$$

where $i \neq j$.

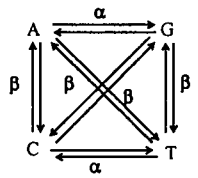
Kimura's two-parameter model

Transitions are generally more frequent then transversions.

In this model,

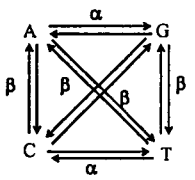
(1) The rate of transitional substitution at each nucleotide site is α per unit

(2) The rate of each transversional substitution is β per unit time.



Let us consider the probability that a site that has A at time 0 will have A time t. After one time unit, the probability of A changing to G is α, and the probability of A changing to either C or T is 2β. Thus the probability of A remaining unchanged after one time unit is

$$P_{AA(1)} = 1 - \alpha - 2\beta \quad (3.13)$$



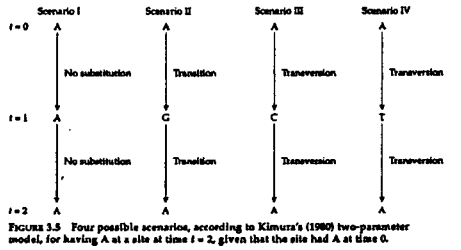


FIGURE 3.5 Four possible scenarios, according to Kimura's (1980) two-parameter model, for having A at a site at time $t = 2$, given that the site had A at time 0.

$$P_{AA(2)} = (1-\alpha-2\beta)P_{AA(1)} + \beta P_{TA(1)} + \beta P_{CA(1)} + \alpha P_{GA(1)} \quad (3.14)$$

By extention,

$$P_{AA(t+1)} = (1-\alpha-2\beta)P_{AA(t)} + \beta P_{TA(t)} + \beta P_{CA(t)} + \alpha P_{GA(t)} \quad (3.15)$$

Similarly, we can obtain

$$P_{T(t+1)} = \beta P_{A(t)} + (1-\alpha-2\beta) P_{T(t)} + \alpha P_{C(t)} + \beta P_{G(t)}$$

$$P_{C(t+1)} = \beta P_{A(t)} + \alpha P_{T(t)} + (1-\alpha-2\beta) P_{C(t)} + \beta P_{G(t)}$$

$$P_{G(t+1)} = \alpha P_{A(t)} + \beta P_{T(t)} + \beta P_{C(t)} + (1-\alpha-2\beta) P_{G(t)}$$

$$P_{AA(t)} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} \quad (3.17)$$

$$P_{AA(t)} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} \qquad (3.17)$$

$$P_{AA(t)} = P_{GG(t)} = P_{CC(t)} = P_{TT(t)}$$

$$X_{(t)} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} \qquad (3.18)$$

Let $Y_{(t)}$ = the probability that the initial nucleotide and the nucleotide at time t differ from each other by a ***transition***.

$$Y_{(t)} = P_{AG(t)} = P_{GA(t)} = P_{TC(t)} = P_{CT(t)}$$

$$Y_{(t)} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} \qquad (3.19)$$

The probability, $Z_{(t)}$, that the initial nucleotide and the nucleotide at time t differ by a specific type of ***transversion*** is given by

$$Z_{(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \qquad (3.20)$$

Number of nucleotide substitutions between two DNA sequences

If two sequences of length N differ from each other at n site, then the proportion of differences, n/N, is referred to as the degree of divergence or Hamming distance.

If the degree of divergence is substantial, then the observed number of differences is likely to be smaller than the actual number of substitutions due to multiple substitution or multiple hit at the same site.

ACTGAACGTAACGC		
↙	↘	
A	A	
C	C →A	**single substitution**
T	T	
G	G	
A →C →T →G	A	**sequential substitution**
A	A	
C →G	C →A	**Coincidental substitution**
G	G	
T →A	T →A	**Parallel substitution**
A	A	
A →C →T	A →T	**Convergent substitution**
C	C	
G	G	
C	C →T →C	**Back substitution**

Number of nucleotide substitutions between two noncoding sequences

Let us start with **one-parameter model.**
In this model, it is sufficient to consider only $I_{(t)}$, which is the probability that the nucleotide at a given site at the time t is the same in both sequences.
Suppose that the nucleotide at a given site was A at time 0. At time t, the probability that a descendant sequence will have A at this site is $P_{AA(t)}$, and consequently the probability that two descendant sequences have A at this site is $P^2_{AA(t)}$.
Similarly, the probabilities that both sequence have T, C, G at this site are $P^2_{AT(t)}$ $P^2_{AC(t)}$ $P^2_{AG(t)}$, respectively. Therefore,

$$I_{(t)} = P^2_{AA(t)} + P^2_{AT(t)} + P^2_{AC(t)} + P^2_{AG(t)} \qquad (3.21)$$

$$I_{(t)} = P_{AA(t)}^2 + P_{AT(t)}^2 + P_{AC(t)}^2 + P_{AG(t)}^2 \quad (3.21)$$

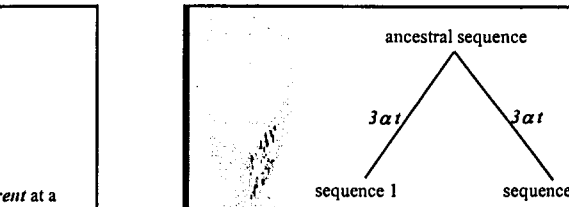
$$I_{(t)} = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t} \quad (3.22)$$

Note that the probability that the two sequences are *different* at a site at time t is $p = 1 - I_{(t)}$. Thus,

$$p = \frac{3}{4}(1 - e^{-8\alpha t}) \quad (3.23)$$

or

$$8\alpha t = -\ln\left(1 - \frac{4}{3}p\right) \quad (3.24)$$



The time of divergence between two sequences is usually given not known, and thus we can not estimate α. Instead, we compute K, which is the number of substitutions per site since the time of divergence between two sequences. In the case of the one parameter model, K= 2(3α t), where 3αt is the number of substitutions per site in a single lineage.

$$8\alpha t = -\ln\left(1 - \frac{4}{3}p\right) \quad (3.24)$$

$$K = 2(3\alpha\ t)$$

We can calculate *K* as

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \quad (3.25)$$

Where *p* is observed proportion of different nucleotides between two sequences.

In the case of **two-parameter model**, the differences between two sequences are classified into transitions and transversions.

Let *P* and *Q* be the proportion of transitional and transversonal differences between two sequences.

Then the number of nucleotide substitutions per site between two sequences, *K*, is estimated by

$$K = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right) \quad (3.27)$$

One-parameter

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \quad (3.25)$$

Two-parameter

$$K = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right) \quad (3.27)$$

Ex.1: 2 sequences with 200 nucleotides that differ by 20 transitions and 4 transversions

One-parameter	*Two-parameter*
$L = 200$	$L = 200$
$p = 24/200 = 0.12$	$P = 20/200 = 0.10$
	$Q = 4/200 = 0.02$
$K \approx 0.13$	$K \approx 0.13$

One-parameter	*Two-parameter*
$L = 200$	$L = 200$
$p = 24/200 = 0.12$	$P = 20/200 = 0.10$
	$Q = 4/200 = 0.02$
$K \approx 0.13$	$K \approx 0.13$

In this example, the two models give essentially the same estimate because the degree of divergence is small enough that the corrected degree of divergence (i.e., the number of nucleotide substitutions, K) is only slightly larger than the uncorrected value (i.e., the number of nucleotide differences, p).

One-parameter

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}P\right) \qquad (3.25)$$

Two-parameter

$$K = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right) \qquad (3.27)$$

Ex.2: 2 sequences with 200 nucleotides that differ by 50 transitions and 16 transversions

One-parameter	*Two-parameter*
$L = 200$	$L = 200$
$p = 66/200 = 0.33$	$P = 50/200 = 0.25$
	$Q = 16/200 = 0.08$
$K \approx 0.43$	$K \approx 0.48$

One-parameter	*Two-parameter*
$L = 200$	$L = 200$
$p = 66/200 = 0.33$	$P = 50/200 = 0.25$
	$Q = 16/200 = 0.08$
$K \approx 0.43$	$K \approx 0.48$

When the degree of divergence between two sequences is large, and especially in cases where there are prior reasons to believe that the rate of transition differs from the rate of transversion, the two parameter model tends to be more accurate than the one-parameter model.

Violation of assumptions

Several assumptions have been made that are not necessary met by the sequences under study.

1) The rate of substitution was assumed to be the same at all sites. This assumption might not hold, as the rate may vary greatly from site to site.

2) The substitution occur in an independent manner.

3) The substitution matrix was assumed not to change in time, so that the nucleotide frequencies are maintained at a constant equilibrium value throughout their evolution.

Substitution mutations

Transition — changes beween A and G, or between T and C
Transversion — changes between a purine and a pyrimidine

Synonymous (silent mutations)
Nucleotide changes do not effect amino acid sequence.

Nonsynonymous (replacement mutations)
A change in single nucleotide in a codon can result in an amino acid replacement.

DNA	CCG	CTG	CTC
mRNA	CCG	CUG	CUC
Amino acid	Proline	Leucine	Leucine

Number of substitutions between two protein-coding genes
(Li, 1985)

nondegernerate (L_0): all the possible changes at this site are nonsynonymous

twofold degenerate (L_2): one of the three possible changes is synonymous

fourfold degenerate(L_4): all possible changes at the site are synonymous

The nucleotide differences in each class are further classified into transitional (S_i) and transversional (V_i) differences, where $i = 0$, 2, and 4 denoted nondegerneracy, twofold degeneracy and fourfold degeneracy, respectively.
All the substitutions at nondegenerate sites are nonsynonymous.
All the substitutions at fourfold degenerate sites are synonymous.
At twofold degenerate site, transitional changes are synonymous, whereas transversitional changes are nonsynonymous.

CCC (Pro)
CAA (Gln)

Path I
S V
CCC ↔ CCA ↔ CAA
(Pro) (Pro) (Gln)

Path II
V V
CCC ↔ CAC ↔ CAA
(Pro) (His) (Gln)

The proportion of transitional differences at *i*-fold degenerate sites between two sequences is calculated as

$$P_i = \frac{S_i}{L_i} \qquad (3.31)$$

Similarly, the proportion of transversional differences at *i*-fold degenerate sites between two sequences is calculated as

$$Q_i = \frac{V_i}{L_i} \qquad (3.32)$$

Kimura's two-parameter method is used to estimate the number of transitional (A_i) and transversional (B_i) substitutions per *i*th type site.

$$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i) \qquad (3.33)$$

$$B_i = \frac{1}{2}\ln(b_i) \qquad (3.34)$$

Where $a_i = 1/(1-2P_i-Q_i)$,
$b_i = 1/(1-2Q_i)$

The total number of substitutions per ith type of degenerate site, K_i, is given by

$$K_i = A_i + B_i \qquad (3.37)$$

A_2 and B_2 denote the numbers of synonymous and nonsynonymous substitutions per twofold degenerate site, respectively.

$K_4 = A_4 + B_4$ denote the numbers of synonymous substitutions per fourfold degenerate site.

$K_0 = A_0 + B_0$ denote the numbers of nonsynonymous substitutions per nondegenerate site.

The number of synonymous substitutions per synonymous site (K_S)

The number of nonsynonymous substitutions per nonsynonymous site (K_A) can be obtained by

$$K_S = \frac{3(L_2A_2 + L_4K_4)}{L_2 + 3L_4} \qquad (3.39)$$

$$K_A = \frac{3(L_2B_2 + L_0K_0)}{2L_2 + 3L_0} \qquad (3.40)$$

Li (1993) and Pamilo and Bianchi (1993) proposed to calculated the number of synonymous substitution by taking $(L_2A_2 + L_4K_4)/(L_2 + L_4)$ as an estimate of the transition component of nucleotide substitution at twofold and fourfold degenerate site

$$K_S = \frac{L_2A_2 + L_4A_4}{L_2 + L_4} + B_4 \qquad (3.41)$$

$$K_A = A_0 + \frac{L_2B_0 + L_2B_2}{L_0 + L_2} \qquad (3.42)$$

Indirect estimations of the number of nucleotide substitution

Indirect estimate of K values are subject to much larger sampling errors than those based on direct comparisons of nucleotide sequence.

Number of Amino acid replacements between two proteins

From the comparison of two amino acid sequences, we can calculate the observed proportion of different amino acid between two sequences as

$$p = \frac{n}{L} \qquad (3.45)$$

where n is the number of amino acid differences between two sequences an L is the length of the aligned sequences.

A simple model that can be used to convert p into the number of amino acid replacements between two sequences is the Poisson process. The number of amino acid replacements per site, d, is estimated as

$$d = -\ln(1-p) \qquad (3.46)$$

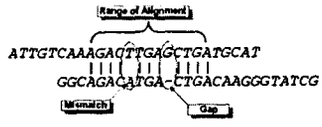
Comparison of two homologous sequences involves the identification if the location of deletions and insertions that might have occurred in either of the two lineages since their divergence from a common ancestor. This process is referred to as sequence alignment.

There are three types of aligned pairs:
A matched pair is on in which that same nucleotide appears in both sequence.

A mismatched pair is a pair in which different nucleotides are found in the two sequences.

A gap is a pair consisting a base from one sequence and a null base from the other. Null base are denoted by -.

A gap indicates that a deletion has occurred in one sequence or an insertion has occurred in the other.



Manual alignment by visual inspection

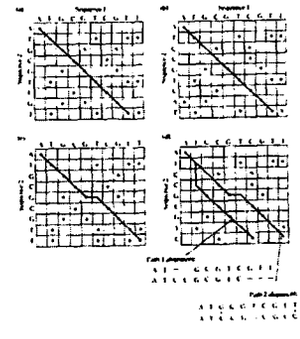
Advantages:
1) it uses the most powerful and trainable of all tools – the brain,
2) it allows the direct integration of additional data.

The main disadvantage of this method is that it is subjective and unscalable, i.e., its results cannot be compared to those derived from other methods.

The dot matrix

In a dot matrix, the two sequences to be aligned are written out as column and row headings of a two-dimensional matrix. A dot is put in the dot matrix plot at a position where the nucleotides in the two sequences are identical. The alignment is defined by a path through the matrix starting with the upper-left element and ending with the lower-right element.

There are four possible types of steps in this path:
1) a diagonal step through a dot indicates a match,
2) a diagonal step through an empty element of the matrix indicates a mismatch,
3) a horizontal step indicates a null nucleotide in the sequence on the top of the matrix,
4) a vertical step indicates a null nucleotide in the sequence on the left of the matrix.



Distance and similarity methods

The best possible alignment between two sequences, or the optimal alignment, is the one in which the numbers of mismatches and gaps are minimized according to certain criteria. Unfortunately, reducing the number of mismatches usually results in an increase in the number of gaps, and vice versa.

A: TCAGACGATTG L_A=11

B: TCGGAGCTG L_B=9

(I) TCAG -ACG-ATTG # of mismatches = 0
TC- GGA-GC-T- G # of gaps = 6

(II) TCAGACGATTG # of mismatches = 5
TCGGAGCTG— # of gaps = 1

(III) TCAG-ACGATTG # of mismatches = 2
TC- GGA-GCTG- # of gaps = 4

As a consequence, we must find a common denominator with which to compare gaps and mismatches. The common denominator is called the gap penalty or gap cost. The gap penalty is a factor by which gap values are multiplied to make the gaps equivalent in value to the mismatches.

Fro any given alignment, we can calculate a **distance** or **dissimilarity index** (D) between the two sequences in the alignment as

$$D = \Sigma\, m_i y_i + \Sigma\, w_k z_k$$

where y_i is the number of mismatches of type i,
m_i is the mismatches penalty for an i-type of mismatch,
z_k is the number of gaps of length k,
w_k is a positive number representing the penalty of gaps of length k.

Alternatively, the similarity between two sequences in an alignment may be measured by a **similarity index** (S). For any given alignment, the similarity between two sequences is

$$S = x - \Sigma\, w_k z_k$$

where x is the number of matches,
z_k is the number of gaps of length k,
w_k is a positive number representing the penalty of gaps of length k.

In the most frequent used gap penalty systems, it is assumed that the gap penalty has two components, a **gap-opening penalty** and a **gap-extension penalty**.

Using a linear gap penalty system in which the mismatch penalty is 1, the gap-open penalty is 2 and the gap-extension penalty is 6.

(I) TCAG-ACG-ATTG # of mismatches = 0
TC-GGA-GC-T-G # of gaps = 6

D = (0 x 1)+(6 x 2)+6 (1–1)=12

(II) TCAGACGATTG # of mismatches = 5
TCGGAGCTG- - # of gaps = 1

D = (5 x 1)+(1 x 2)+6 (2–1)=13

(III) TCAG -ACGATTG # of mismatches = 2
TC- GGA- GCTG- # of gaps = 4

D = (2 x 1)+(4 x 2)+6 (1–1)=10

Using a different penalty system in which the mismatch penalty is 1, the gap-open penalty is 3 and the gap-extension penalty is 0.

(I) TCAG-ACG-ATTG # of mismatches = 0
TC-GGA-GC-T-G # of gaps = 6

D = (0 x 1)+(6 x 3)=18

(II) TCAGACGATTG # of mismatches = 5
TCGGAGCTG- - # of gaps = 1

D = (5 x 1)+(1 x 3)=8

(III) TCAG -ACGATTG # of mismatches = 2
TC- GGA- GCTG- # of gaps = 4

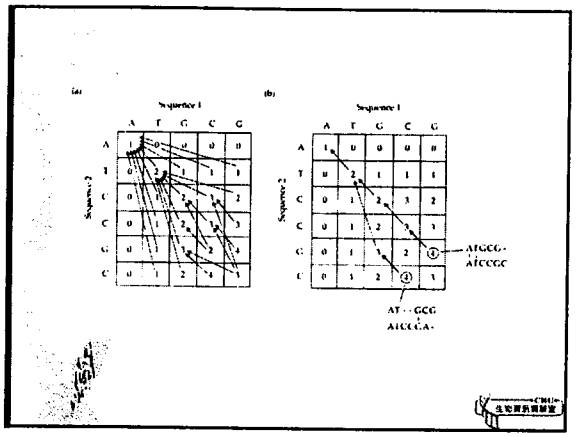
D = (2 x 1)+(4 x 3)=14

Alignment algorithms

The Needleman-Wunsch algorithm used dynamic programming, which is a general computational technique used in many fields of study.
Dynamic programming can be applied to alignment problems because similarity indices obey the following rule:

$$S_{1-x,\,1-y} = \max S_{1-x-1,\,1-y-1} + S_{x,y}$$

In which $S_{1-x,\,1-y}$ is the similarity index for the two sequences up to residue x in the first sequence and residue y in the second sequence, max $S_{1-x-1,\,1-y-1}$ is the similarity index for the best alignment up to residue x-1 in the first sequence and y-1 in the second sequence, $S_{x,y}$ is the similarity score for aligning residues x and y.



Multiple sequence alignment

Multiple sequence alignment can be viewed as an extension of pairwise sequences alignment, but the complexity of the computation grows exponentially with the number of sequences being considered and, therefore, it is not feasible to search exhaustively for optimal alignment.

Most of the programs use some sort of incremental or progressive algorithm, in which a new sequences is added to a group of already aligned sequences in order of decreasing similarity.

It is usually advisable to take a look at the final multiple alignment, as such alignments can be frequently improved by visual inspection.

Rates and Patterns of Nucleotide Substitution

指導教授：吳哲賢
報告者：M9202055
朱廷翰

Introduction

- It present data on the rates and patterns of nucleotide substitution and discuss three factors affecting them
  - Functional constraint
  - Positive selection
  - Mutation input
- It also dissect the substitution rate into its constituent parts in order to infer the pattern of substitution, in particular the pattern of spontaneous mutation

Rates of Nucleotide Substitution

- The rate of nucleotide substitution
  - r : the number of substitution per site per year
  - K : the number of substitutions
  - T : the time of divergence between the two sequence
- Function

4.1 only holds when dealing with distantly related species

$$r = \frac{K}{2T} \qquad (4.1)$$

Rates of Nucleotide Substitution

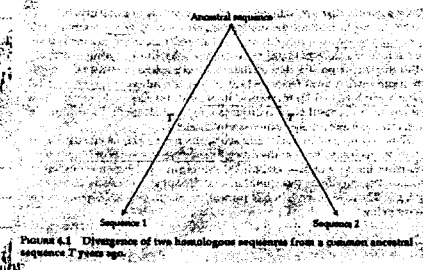
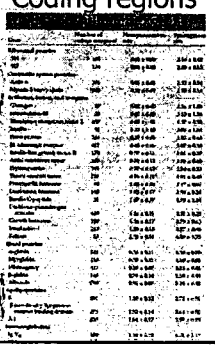


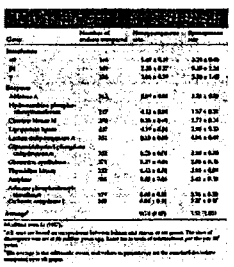
FIGURE 4.1 Divergence of two homologous sequences from a common ancestral sequence T years ago.

Coding regions

- In dealing with protein–coding sequence, it is important to deal discriminate between nucleotide changes that affect the primary structure of the encoded protein
- Nonsynonymous nucleotide substitution are reflected in the rates of protein evolution
  - Extremely conservative
  - Intermediate rate
  - High rate

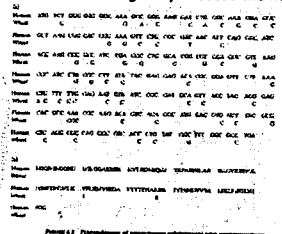
Coding regions





Coding regions

In the vast majority of genes, the synonymous substitution rate greatly exceeds the nonsynonymous rate.



Coding regions

- Nondegenerate site : low rate
- Twofold degenerate : intermediate rate
- Fourfold degenerate : highest rate

Type of substitution	Nondegenerate	Twofold degenerate	Fourfold degenerate
Transition	0.40	1.86	2.24
Transversion	0.38	0.38	1.47
Total	0.78	2.24	3.71

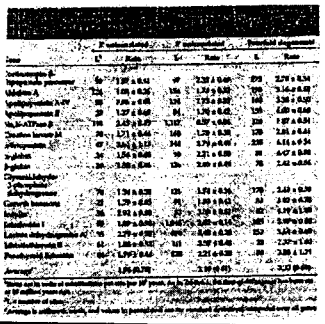
From Li (1997).

[a] The rates are averages over the genes in Table 4.1.

Noncoding regions

- The most published sequence are cDNA sequence derived from mRNAs. (only include 5' and 3' untranslated regions)
- The rates vary greatly among genes, but this variation may largely represent sampling effects due to the fact that both these regions are usually very short

Noncoding regions



Noncoding regions

- Pseudogenes are DNA sequence that were derived from functional genes but have been rendered nonfunctional by mutations that prevent their proper expression.

Region	K^a
5' flanking region	5.3 ± 1.2
5' untranslated region	4.0 ± 2.0
Fourfold degenerate sites	8.6 ± 2.5
Introns	8.1 ± 0.7
3' untranslated region	8.8 ± 2.2
3' untranslated region	8.0 ± 1.5
Pseudogenes	9.1 ± 0.9

[a] Means and standard errors.

Noncoding regions

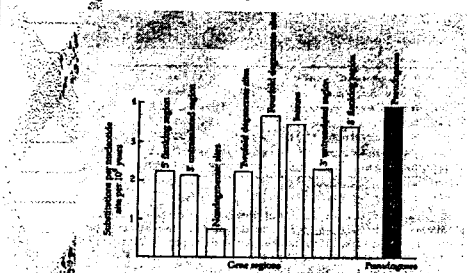


FIGURE 4.3 Average rates of substitution in different parts of genes (white) and in pseudogenes (gray). From Li (1997).

Similarity profile

A **similarity profile** is a graph of similarity along an alignment of two homologous sequence.

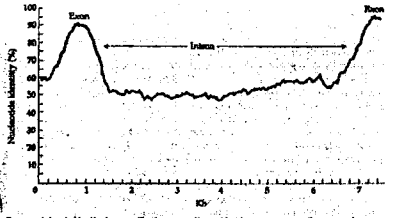
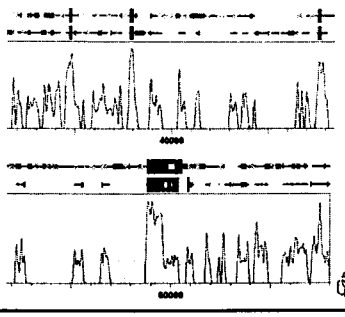


FIGURE 4.4 A similarity profile for two aligned DNA sequences. Conserved regions show up as peaks, which in protein-coding genes may allow us to identify the junctions between exons and introns.

Similarity profile



Cause of variation in substitution rates

- The rate of mutation
- The probability of fixation of a mutation, it depends on whether the mutation is advantageous, neutral, or deleterious.

Functional constraints

- **Functional or selective constraint** defines the range of alternative nucleotides that is acceptable at a site without affecting negatively the function or structure of the gene product.
- The stronger the functional constraints on a macromolecule are, the slower the rate of substitution will be.

Functional constraints

Kimura(1977, 1983) illustrated a simple model.

- A certain fraction f_0 of all mutations in a certain molecule are selectively neutral or nearly neutral and the rest are deleterious.
- V_T is the total mutation rate per unit time.
- The rate of neutral mutation V_0 is

$$V_0 = V_T f_0 \qquad (4.2)$$

- According to the neutral theory of molecular evolution, the rate of substitution $K = V_0$ is

$$K = V_T f_0 \qquad (4.3)$$

Functional constraints

- **Function density** is the proportion of amino acids that are subject to stringent functional constaints.
  - n_s is the number of sites committed to specific functions.
  - N is the total number of sites.
  - The functional density of a gene is F

$$F = n_s / N$$

The higher the functional density, the lower the rate of substitution is expected to be.

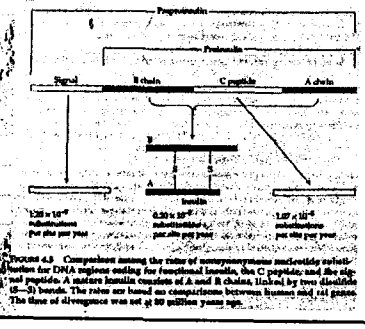
Functional constraints

- Advantage selection : the rate of synonymous substitution is higher.
- Purifying selection : the rate of nonsynonymous substitution is higher.
- Neutral selection : the rate should be the same, or at least very similar.

Variation among different gene regions

- Within a protein, the different structural or functional domains are likely to be subject to different **functional constraints** and to evolve at different rate.
- If structural or functional domains have less constraint, it should have higher substitution rate.
  ex : the signal peptide and the c peptide

Variation among different gene regions



Variation among genes

Two possible reason

- The rate of mutation
- The intensity of selection

- Ex : the apolipoproteins and histone H3
  the lax structural requirement may explain the fairly high substitution rate

Variation among genes

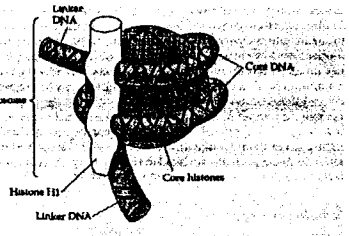


FIGURE 4.6 Schematic diagram of a nucleosome. The DNA double helix is wound around the core histones (two each of histones H2A, H2B, H3, and H4). Histone H1 binds to the outside of this core particle and to the linker DNA.

Acceleration of nucleotide substitution rates following partial loss of function

- When selection constraints are only partially, rather than entirely, removed. Such a phenomenon is called relaxation of selection.
- The substitution rates following partial loss of function is higher than fully functional.
  Ex : the *αA-crystallin* gene in the blind mole rat and other rodents, such as mouse, rat, gerbil, which possess fully functional eyes.

Estimating the intensity of purifying selection in protein-coding genes

- Ophir et al. (1999) proposed a simple method by which the intensity of purifying selection on a functional protein-coding gene can be quantified.
- Their method requires three homologous sequences
  - A pseudogene (ψA)
  - A functional homologous gene from the same species (A)
  - A functional homolog from a different species (B)

Estimating the intensity of purifying selection in protein-coding genes

- The number of nucleotide substitutions along the branches leading to ψA and A, i.e., $K_{\psi A}$ and K_A
- Following Kimura's (1977) model in equation 4.3, the numbers of nucleotide substitutions along the two branches leading to ψA and A are given by
  - $K_{\psi A} = \nu_{\psi A} f_{\psi A}$ (4.4)
  - $K_A = \nu_A f_A$ (4.5)
  - ν is the total mutation rate per unit time
  - f is the fraction of mutation that are selectively neutral

Estimating the intensity of purifying selection in protein-coding genes

- If we assume that the mutation rate is the same in the gene and the pseudogene, i.e., $\nu_A = \nu_{\psi A}$, and if we further assume that mutations occurring in a pseudogene do not affect the fitness of the organism, i.e., $f_{\psi A}=1$, we obtain

$$f_A = \frac{K_A}{K_{\psi A}} \qquad (4.6)$$

- By definition, the fraction of deleterious mutations that are subject to purifying selection (or the intensity of selection) is $1-f_A$

Estimating the intensity of purifying selection in protein-coding genes

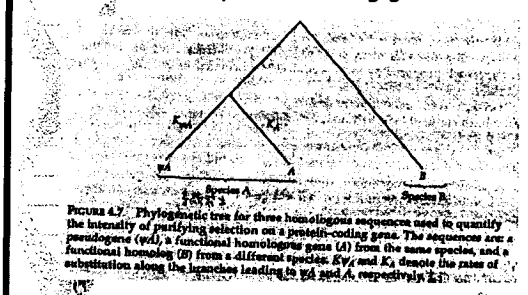


FIGURE 4.7 Phylogenetic tree for three homologous sequences used to quantify the intensity of purifying selection on a protein-coding gene. The sequences are a pseudogene (ψA), a functional homologous gene (A) from the same species, and a functional homolog (B) from a different species. $K_{\psi A}$ and K_A denote the rates of substitution along the branches leading to ψA and A, respectively.

Mutation input: Male-driven evolution

- u_m : the mutation rates in males
- u_f : the mutation rates in females
- α : the ratio of male to female mutation rates
- A : the mutation rate per generation for an autosomal sequence

$$\alpha = \frac{u_m}{u_f} \qquad (4.7)$$

$$A = \frac{u_m + u_f}{2} \qquad (4.8)$$

Mutation input: Male-driven evolution

- X : the mutation rate per generation for a sequence located on the X chromosome
- Y : the mutation rate per generation for a sequence located on the X chromosome
- An X-linked sequence is carried 2/3 of the time by females and 1/3 of the time by males.

$$X = \frac{u_m + 2u_f}{3} \qquad (4.9)$$

- A Y-linked sequence is only carried by males.

$$Y = u_m \qquad (4.10)$$

Mutation input: Male-driven evolution

- The ratio of Y to A is

$$Y/A = \frac{2\alpha}{1+\alpha} \quad (4.11)$$

- The ratio of X to A is

$$X/A = \frac{2(2+\alpha)}{3(1+\alpha)} \quad (4.12)$$

- The ratio of Y to A is

$$Y/X = \frac{3\alpha}{2+\alpha} \quad (4.13)$$

Mutation input: Male-driven evolution

- In human, u_f was estimated to be about 33
  u_m was estimated to be about 200
- In mice, u_f was estimated to be about 28
  u_m was estimated to be about 57
- Their evolution is male-driven.

Positive selection

Detecting positive selection

- K_A : the number of substitution per nonsynonymous site
- K_S : the number of substitution per synonymous site
- V : the variance

$$t = \frac{K_A - K_S}{\sqrt{V(K_A)+V(K_S)}} \quad (4.14)$$

- Under the null hypothesis of neutral evolution, i.e. no positive selection $K_A = K_S$

Positive selection

Detecting positive selection

- M_S , M_A the numbers of synonymous and nonsynonymous differences between the two protein-coding sequence
- N_S , N_A the average numbers of synonymous and nonsynonymous site

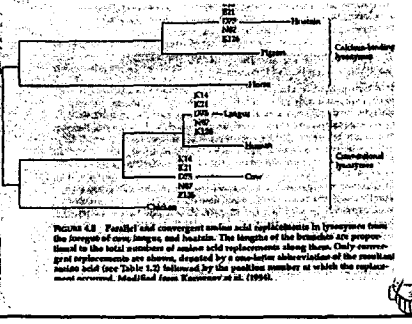
	Nonsynonymous	synonymous	Total
Changes	M_A	M_S	M_A+M_S
No changes	N_A-M_A	N_S-M_S	$L-(M_A+M_S)$
Total	N_A	N_S	L

- Under the null hypothesis of neutral evolution, i.e. no positive selection

Positive selection

- **Parallelism** at the molecular level is defined as the independent occurrence of two or more nucleotide substitutions of the same type at homologous sites in different evolutionary lineages.
- Molecular **convergence** is the occurrence of two or more nucleotide substitutions at homologous sites in different evolutionary lineages resulting in the same outcome.

Positive selection



Positive selection

Prevalence of positive selection

- According to a survey by endo et al. (1996), positive selection affecting entire protein-coding sequences is suspected in only very few cases. In their study of 3595 groups of homologous sequence, they found only 17 gene groups (about 0.45%).
- The highest ratio of nonsynonymous to synonymous substitution ($K_A / K_S = 5.15$) for a full-length protein was found in the 10-kilo-dalton protein in the acrosomal vesicle at the anterior of the sperm cell of several abalone species. It is thought that sex-related genes are subject to positive selection for short period of time during speciation as a means of erecting reproductive barriers that restrict gene flow between the speciating populations

Patterns of substitution and replacement

- Definition: the relative frequency with which a certain nucleotide changes into another during evolution.
- The pattern is usually shown in the form of a 4x4 matrix, in which each of the 12 elements of the matrix.

Patterns of substitution and replacement

- P_{ij} : the proportion of base changes from the i th type to the j th type of nucleotide (i , j =A, T, C or G, and $i \neq j$)
- n_{ij} : the number of substitutions from i to j
- n_i : the number of the i nucleotides in the ancestral sequence.

$$P_{ij} = \frac{n_{ij}}{n_i} \qquad (4.15)$$

- : the relative substitution frequency from nucleotide i to nucleotide j

$$f_{ij} = \frac{P_{ij}}{\sum_i \sum_{j \neq i} P_{ij}} \qquad (4.16)$$

Pattern of spontaneous mutation

One way to study the pattern of point mutation is to examine the pattern of substitution in regions of DNA that are subject to no selective constraint. **Pseudogenes** are particularly useful in this respect.

- In figure 4.9 we can assume that the nucleotide in the pseudogene sequence has changed from G to A if sequence 3 ha G, but that the nucleotide in sequence 2 has changed from A to G if sequence 3 has A. If sequence 3 has T or C, then we cannot decide the direction of change.

Pattern of spontaneous mutation

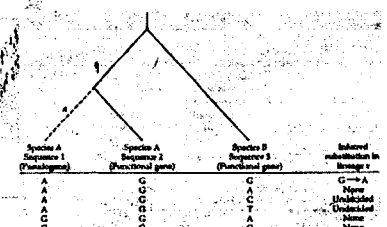


FIGURE 4.9 A tree for inferring the pattern of nucleotide substitution in a pseudogene sequence. The dashed line *x* implies "nonfunctional." In cases where the nucleotides occupying homologous sites in sequences 2 and 3 are identical, but different from the nucleotide in sequence 1, the type of substitution in lineage *x* can be unambiguously inferred.

Pattern of spontaneous mutation

From	To A	T	C	G	Row totals
A	—	3.4 ± 0.7 (3.6 ± 0.70)	4.5 ± 0.8 (4.8 ± 0.9)	12.5 ± 1.1 (13.5 ± 1.1)	20.3 (21.6)
T	3.3 ± 0.6 (3.5 ± 0.6)	—	13.8 ± 1.9 (14.7 ± 2.0)	3.3 ± 0.6 (3.5 ± 0.6)	20.4 (21.7)
C	4.2 ± 0.3 (4.2 ± 0.5)	20.7 ± 1.5 (16.4 ± 1.3)	—	4.6 ± 0.6 (4.4 ± 0.6)	29.5 (25.1)
G	20.4 ± 1.4 (21.9 ± 1.5)	4.4 ± 0.6 (4.6 ± 0.6)	4.9 ± 0.7 (5.2 ± 0.8)	—	29.7 (31.6)
Column totals	27.9 (29.5)	28.5 (24.6)	23.2 (23.2)	20.5 (21.3)	

Courtesy of Dr. Ron Ophir.

"Table entries are the inferred percentages (f_{ij}) of nucleotide changes from *i* to *j* based on 105 processed pseudogene sequences from humans. Values in parentheses were obtained by excluding all CG dinucleotides from the comparison.

Strand inequalities: Pattern of substitution in human mitochondrial

From	To A	T	C	G	Row totals
A	—	0.4	1.1	14.1	15.6
T	0.3	—	33.8	0.3	34.4
C	1.1	25.8	—	0.5	27.4
G	20.0	1.1	1.6	—	22.7
Column total	21.4	27.3	36.5	14.9	

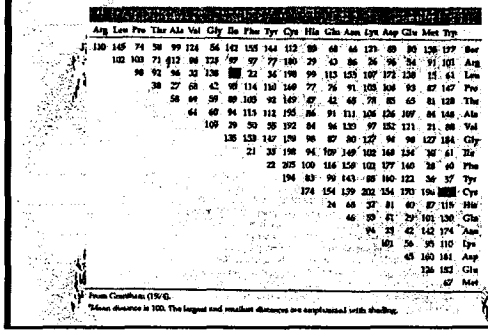
From Tamura and Nei (1993).

[illegible] are the inferred percentages (f_{ij}) of nucleotide changes from i to j based on 95 sequences.

Patterns of amino acid replacement

- These so-called **physicochemical distance** are based on such properties of the amino acids as polarity, molecular volume, and chemical composition.
- Grantham's (1974) physicochemical distances are shown in Tabel 4.7

Patterns of amino acid replacement



From Grantham (1974).

[illegible] divance is 100. The largest and smallest distances are emphasized with shading.

Patterns of amino acid replacement

- A replacement of an amino acid by a similar one (e.q., lecucine to isoleucine or leucine to methionine; Figure 4.10a) is called a conservative replacement.
- A replacement of an amino acid by a dissimilar one (e.q., glycine to tryptophan or cyysteine to tryptophan; Figure 4.10b) is called a radical replacement.

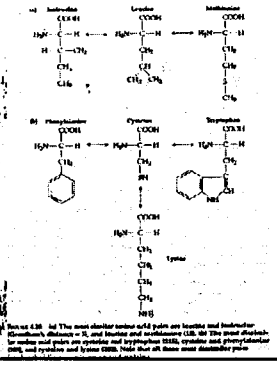
Patterns of amino acid replacement

- Some amino acid, such as leucine, isoleucine, glutamine, and methionine are typical amino acids, since they have a number of similar alternative amino acids with which they can be replaced through a single nonsynonymous substitution.
- Other amino acids, such as cysteine tryptophan, tyrosine, and glycine, are idiosyncratic amino acids; they have few similar alternative amino acids with which they can be replaced.

Patterns of amino acid replacement

- The conservative replacements occur more frequently than radical replacements in protein evolution.
- Argyle (1980) devised a circular graphical representation of amino acid exchangeability. A modified version by Pieber and Toha (1983) is shown in Figure 4.12. Depending on the protein, 60-90% of observed amino acid replacement involve the nearest or second-nearest neighbors on the ring.

Patterns of amino acid replacement



Patterns of amino acid replacement

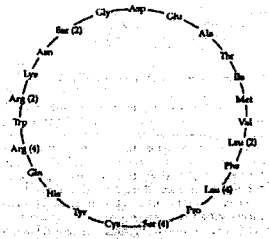


FIGURE 4.12 A circular graphical representation of amino acid exchangeability according to Argyle's (1980) method. Numbers in parentheses denote number of codons in a family in cases where an amino acid is encoded by two codon families. About 50–90% of the observed amino acid replacements involve the nearest or second-nearest neighbors on the ring. Data from Fitch and Ibbd (1968).

Protein properties conserved in evolution

- The two most highly conserved properties are bulkiness (volume) and refractive index (a measure of protein density).
- Hydrophobicity and polarity also seem to be moderately well conserved, whereas optical rotation seems to be an irrelevant property in the evolution of proteins.

Rate and Patterns of Nucleotide Substitution

Speaker:鄭銘杰

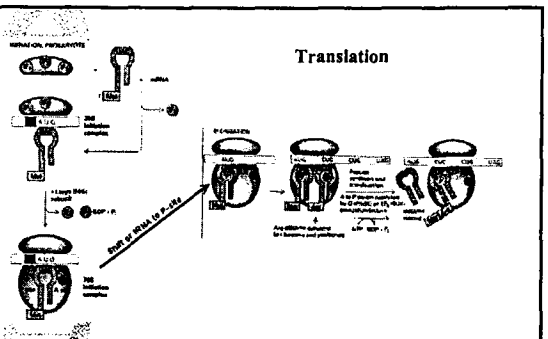
2003/10/29

Nonrandom Usage of Synonymous Codons

- Synonymous mutations do not cause any change in amino acid sequence
- If all synonymous mutations are indeed selectively neutral, and if the pattern of mutation is symmetrical , then the synonymous codons for an amino acid should be used with equal frequencies.

Nonrandom Usage of Synonymous Codons

- It became evident that the usage of synonymous codons is distinctly nonrandom in both prokaryotic and eukaryotic



Source: www.esb.utexas.edu/dr325/ Supplements/Ch10.htm

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

Genetic Code

Source: www.cimr.cam.ac.uk/ links/codon.htm

Measures of codon-usage bias

生物體在合成protein的時候，是由mRNA translate出protein，此時須藉由tRNA利用其上的anticodon辨識mRNA上的codon，並攜帶與該codon對應的amino acid至ribosome，一連串正確amino acid的連結即可形成protein的1級結構。

Measures of codon-usage bias

- 在這裡，mRNA上的codon(64種)與amino acid(20種)並不是全屬於一對一的關係，可能由多個(2、3、4、6)codon去對應到一個amino acid。就機率上來說，如果是4個codon對應到一個amino acid，那麼對於表現這個amino acid而言，這四個codon出現的機率應個爲25%，但我們或發現實際上的情況並非如此。

Measures of codon-usage bias

- 舉例來說： UUA、UUG、CUU、CUC、CUA、CUG這6個codon都可以對應到leucine，而在*E.coli*的outer membrane protein II(*ompA*)中，23個leucine residues中有21個是encoded by the codon CUG。此種nonrandom codon usage就稱爲codon usage bias。

Measures of codon-usage bias

- Relative synonymous codon usage (RSCU) a simple measure of nonrandom usage of synonymous codons in a gene.

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i} \qquad (4.17)$$

- n is the number of synonymous codons ($1 \le n \le 6$)
- X_i is the number of occurrences of codon i

Measures of codon-usage bias

- CAI codon adaptation index
- The relative adaptiveness of a codon w_i
- L is the number of codon

$$w_i = \frac{RSCU_i}{RSCU_{max}} \qquad (4.18)$$

$$CAI = \left(\prod_{k=1}^{L} w_i\right)^{\frac{1}{L}} \qquad (4.19)$$

Measures of codon-usage bias

- $RSCU_{max}$ is the RSCU value for the most frequently used codon for an amino acid
- The CAI value for a gene is calculated as the geometric mean of w_i values for all the codons used in the gene
- CAI value are frequently used to identify genomic regions that have been horizontally transferred among species. (Chapter 8)

Measures of codon-usage bias

- Another measure of codon bias is the effective number of codons (ENC)

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \qquad (4.20)$$

- F_i (i=2,3,4,or 6) is the average probability that two randomly chosen codons for an amino acid with i codons will be identical.

Measures of codon-usage bias

- ENC values range from 20 (the number of amino acid) to 61 (the number of sense codons)

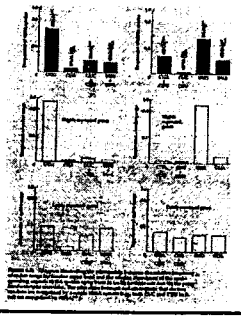
Universal and species-specific patterns of coden usage

- According to genome hypothesis, the genes in any given genome use the same strategy with respect to choices among synonymous coden; in other word, the bias in coden usage is species-specific.

Codon usage in unicellular organisms

- E. coli ribosomal protein-coding genes preferentially use synonymous codons that are recognized by the most abundant tRNA species.
- Suggest that the preference resulted from natural selection because using a codon that is translated by an abundant tRNA species will increase translational efficiency and accuracy

Codon usage in unicellular organisms



Codon usage in unicellular organisms

- In E. coli CUG (the codon recognized by this rRNA) is much more frequently used than the other five codons.

Codon usage in unicellular organisms

- U in the first position of anticodons can pair with both A and G.
- G can pair with both C and U.
- C in the first anticodon position can only pair with G at the third position of codons, and A can only pair with U.

Codon usage in unicellular organisms

Amino acid	Codon	E. coli High	E. coli Low	S. cerevisiae High	S. cerevisiae Low
Leu	UUA	0.06	1.24	0.49	1.49
	UUG	0.07	0.87	5.34	1.48
	CUU	0.13	0.72	0.02	0.73
	CUC	0.17	0.65	0.00	0.54
	CUA	0.04	0.31	0.15	0.95
	CUG	5.54	2.20	0.02	0.84
Val	GUU	2.41	1.09	2.07	1.13
	GUC	0.08	0.99	1.91	0.76
	GUA	1.12	0.63	0.00	1.13
	GUG	0.40	1.29	0.02	0.98
Ile	AUU	0.48	1.38	1.26	1.29
	AUC	2.51	1.12	1.74	0.66
	AUA	0.01	0.50	0.00	1.05
Phe	UUU	0.34	1.33	0.19	1.38
	UUC	1.66	0.67	1.81	0.62

From Sharp et al. (1988)

For each group of synonymous codons, the sum of the relative frequencies equals the number of codons in the group. For example, there are six codons for leucine, and so the sum of the relative frequencies for these six codons is 6. Under equal codon usage, the relative frequencies for each codon in a group should be 1, and so the degree of deviation from 1 indicates the degree of bias in codon usage. "High" and "Low" denote genes with high and low expression levels, respectively.

Codon usage in unicellular organisms

- Factors determining the choice of optimal codons in unicellular organism

1. tRNA availability.
2. Preference for A over G when thiolated uridine or 5-carboxymethyl are at the anticodon wobble position.
3. Preference for T and C over A when inosine is at the anticodon wobble position.
4. Preference for C in the third position of codons AAN, ATN, TAN, and TIN

After Ikemura and Ozeki (1983).

Codon usage in multicellular organisms

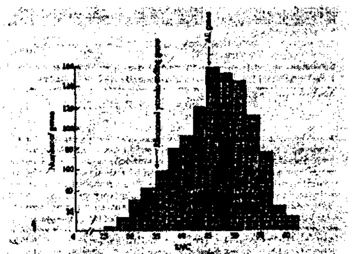


FIGURE 4.14 Distribution of effective number of codons (ENC) for 1,117 Drosophila melanogaster genes. Mean ENC values for all these genes, as well as for a subset of 38 ribosomal protein-coding genes are shown. Modified from Moriyama and Powell (1997).

Codon usage in multicellular organisms

- The vast majority of genes show codon-usage biases, as indicated by a mean ENC value of 46.
- In the ribosomal protein-coding genes, the codon-usage bias is much stronger (mean ENC=35)

Codon usage in multicellular organisms

- Multicellular organisms, different cells produce different protein, and therefore a simple relationship between codon usage and tRNA abundance is not expected.

Molecular Clocks

- Comparative studies of hemoglobin and cytochrome c protein sequence form different species, first noticed that the rates of amino acid replacement were approximately the same among various mammalian lineages.

Molecular Clocks

- Therefore proposed that for any given protein , the rate of molecular evolution is approximately constant over time in all lineages or, in other words, that there exists a molecular clock.

Molecular Clocks

- They can be used to determine dates of species divergence and to reconstruct phylogenetic relationships among organisms.
- This estimated rate could then be used to date the divergence time is lacking.

Molecular Clocks

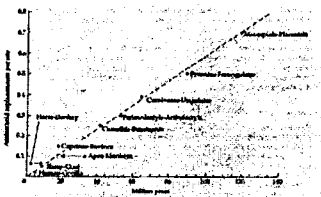
- Let us assume that the rate of nonsynonymous substitution for the a chain of hemoglobin is 0.56*10^-9 substitutions per site year, and that a-globins from rat and human differ by 0.093 substitutions per site.

1

Molecular Clocks

- Under the molecular clock hypothesis, the divergence time between the human and rat lineages is estimated to be approximately 0.093/(2*0.56*10^-9) = 80 million years ago.

Molecular Clocks



Relative Rate Tests

- The controversy over the molecular clock hypothesis often involves disagreements on dates of species divergence.
- To avoid this problem, several tests that do not require knowledge of divergence times have been developed.

Margoliash, Sarich, and Wilson's test

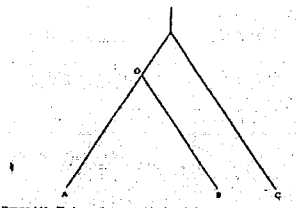


FIGURE 4.16 Phylogenetic tree used in the relative rate test. O denotes the common ancestor of species A and B. C is the outgroup.

Margoliash, Sarich, and Wilson's test

$$K_{AB} = K_{OA} + K_{OB} \quad (4.21)$$

Similarly,

$$K_{AC} = K_{OA} + K_{OC} \quad (4.22)$$

and

$$K_{BC} = K_{OB} + K_{OC} \quad (4.23)$$

$$K_{OA} = \frac{K_{AC} + K_{AB} - K_{BC}}{2} \quad (4.24)$$

$$K_{OB} = \frac{K_{AB} + K_{BC} - K_{AC}}{2} \quad (4.25)$$

$$K_{OC} = \frac{K_{AC} + K_{BC} - K_{AB}}{2} \quad (4.26)$$

Margoliash, Sarich, and Wilson's test

- The time that has passed since species A and B last shared a common ancestor is by definition equal for both lineages.
- According to the molecular clock hypothesis, K_{OA} and K_{OB} should be equal.

Margoliash, Sarich, and Wilson's test

- K_{OA}-K_{OB}=K_{AC}-K_{BC}, we compare substitution rates in A and B directly from K_{AC} and K_{BC}.
- We use K_{AC}-K_{BC} as an estimator of K_{OA}-K_{OB} represents the difference in branch length between the two lineages leading from O to species A and B.

Margoliash, Sarich, and Wilson's test

- We denote this difference by d.
- A positive d value means the molecule has evolved faster in lineage A than in lineage B.
- The variance of d is given below

Margoliash, Sarich, and Wilson's test

$$V(d) = V(K_{AC}) + V(K_{BC}) - 2V(K_{OC}) \quad (4.27)$$

$$p = \frac{3}{4}\left(1 - e^{-4/3K_{OC}}\right) \quad (4.28)$$

Margoliash, Sarich, and Wilson's test

- The one-parameter model, V(K_{AC}) and V(K_{BC}) can be obtained from Equation 3.27 (Chapter 3). V(K_{OC}) can be obtained by putting into Equation 3.27.

Margoliash, Sarich, and Wilson's test

- We note that Equations 4.21~4.26 also hold for the two alternative cases : (1) A and C diverged from each other after the divergence of B, and (2) B and C diverged from each other after the divergence of A

Margoliash, Sarich, and Wilson's test

- When the order of divergence among three species is unknown, two of the species must be more closely related to each other than either is to the third.

Tajima's 1D method

- Assume (1) that the substitution model is known (2) that the substitution rates among different sites vary according to some prespecified distribution
- We start with three aligned nucleotide sequence 1,2,3
- n_{ijk} be the number of site where sequence 1,2,3 have nucleotides i,j,k

Tajima's 1D method

- Sequence 3 is the outgroup

$$E(n_{ijk}) = E(n_{jik}) \qquad (4.29)$$

- This equality holds regardless of the substitution model or the pattern of variation in substitution rates among sites.

Tajima's 1D method

$$m_1 = \sum n_{ij}$$
$$= n_{AGG} + n_{ACC} + n_{ATT} + n_{GAA} + n_{GCC} + n_{GTT} + n_{CAA} + n_{CGG} + n_{CTT} + n_{TAA} + n_{TGG} + n_{TCC} \qquad (4.30)$$

$$m_2 = \sum n_{ji}$$
$$= n_{AGA} + n_{ACA} + n_{ATA} + n_{GAG} + n_{GCG} + n_{GTG} + n_{CAC} + n_{CGC} + n_{CTC} + n_{TAT} + n_{TGT} + n_{TCT} \qquad (4.31)$$

Tajima's 1D method

- When sequence 3 is the outgroup, the expectation of m1 is equal is that of m2 under the molecular clock

$$E(m_1) = E(m_2) \quad (4.32)$$

Tajima's 1D method

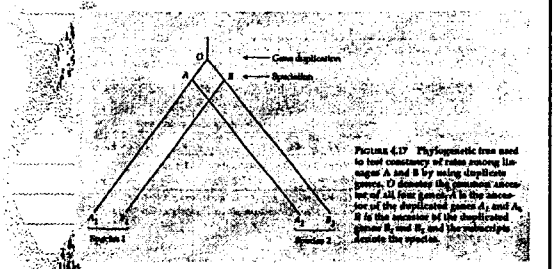
- The equality can be tested by using X^2 with one degree of freedom, namely

$$\chi^2 = \frac{(m_1 - m_2)^2}{m_1 + m_2} \quad (4.33)$$

Tests involving comparisons of duplicate genes

- The molecular clock hypothesis can also be tested by comparing homologous genes that originated through a gene duplication event.
- If we denote the number of substitutions between sequence i and sequence j by K_{ij}

Tests involving comparisons of duplicate genes



Tests involving comparisons of duplicate genes

$$K_{A_1B_1} = K_{AA_2} + K_{OA} + K_{OB} + K_{BB_2} \quad (4.34)$$

$$K_{A_2B_2} = K_{AA_2} + K_{OA} + K_{OB} + K_{BB_2} \quad (4.35)$$

$$K_{A_1B_1} - K_{A_2B_2} = K_{AA_1} + K_{BB_1} - K_{AA_2} - K_{BB_2} \quad (4.36)$$

Tests involving comparisons of duplicate genes

- If A_1 evolves at the same rate as A_2 and B_1 evolves at the same rate as B_2, then $K_{AA1} = K_{AA2}$ and $K_{BB1} = K_{BB2}$.
- $K_{A1B1} - K_{A2B2} = 0$

Nearly equal rates in mice and rats

- Table 4.10 shows a comparison of the rates of synonymous and non-synonymous substitution in mice and rats using the relative rate test.
- Species 1 : mouse
- Species 2 : rats
- Species 3 : hamster

Nearly equal rates in mice and rats

Type of substitution	Number of sites compared	Number of substitutions[a]			
		K_{12}[b]	K_{13}	K_{23}	$K_{13} - K_{12}$
Synonymous	4,855	19.9 ± 0.7	31.1 ± 0.9	32.4 ± 1.0	−1.3 ± 7.9
Nonsynonymous	17,440	1.9 ± 0.1	2.9 ± 0.1	2.7 ± 0.1	0.3 ± 1.3

Modified from O'hUigin and Li (1992).

[a]Mean ± standard error. K_{ij} = number of substitutions per 100 sites between species i and j.

Lower rates in humans than in African apes and monkeys

- Table 4.11 in the comparisons between Homo (species 1) on the one hand and P. paniscus, P. troglodytes, or Gorilla (species 2) on the other, we see that the value of m1 are significantly smaller than the m2 values.

Lower rates in humans than in African apes and monkeys

Species 1	Species 2	Species 3	m_1	m_2	χ^2
Pan troglodytes	*Pan paniscus*	*Homo sapiens*	21	14	1.40
Homo sapiens	*Pan paniscus*	*Gorilla gorilla*	37	55	3.50*
Homo sapiens	*Pan troglodytes*	*Gorilla gorilla*	34	61	7.67**
Homo sapiens	*Gorilla gorilla*	*Pongo pygmaeus*	56	38	3.15*
Pan troglodytes	*Gorilla gorilla*	*Pongo pygmaeus*	63	56	0.41
Pan paniscus	*Gorilla gorilla*	*Pongo pygmaeus*	56	58	0.04
Homo sapiens	*Pongo pygmaeus*	*Hylobates lar*	91	105	1.60
Pan troglodytes	*Pongo pygmaeus*	*Hylobates lar*	120	104	1.14
Pan paniscus	*Pongo pygmaeus*	*Hylobates lar*	109	104	0.12
Gorilla gorilla	*Pongo pygmaeus*	*Hylobates lar*	104	112	0.17

*Significant at the 5% level; **significant at the 1% level.

[a]Species 1 and 2 are the hominoids, species 3 is the outgroup.

Lower rates in humans than in African apes and monkeys

- In all cases, K13 – K12 is significantly larger than 0 and we may conclude that the rate of substitution in the noncoding regions, which presumably reflects the rate of mutation, is higher in the African monkeys than in humans.
- Species 1 : African monkey
- Species 2 : human
- Species 3 : New World monkey

Lower rates in humans than in African apes and monkeys

Type of sequence	Sequence length	K_{12}[a]	K_{23}	K_{13}	$K_{13} - K_{12}$	Ratio[b]
Pseudogene	8,781	10.7	11.8	10.7	1.1 ± 0.7**	1.4
Intron	8,478	7.1	14.7	13.9	0.8 ± 0.7**	1.5
Flanking and untranslated regions	930	9.9	14.9	13.7	3.1 ± 1.1**	2.3

Data from Bailey et al. (1991), Porter et al. (1995), and Ellsworth et al. (1993).

**Significant at the 1% level.

[a]K_{ij} = number of substitutions per 100 sites between species i and j.

[b]The ratio of the rate in the African monkey region to that in the human lineage.

Higher rates in rodents than in primates

- K_S : number of nucleotide substitutions per synonymous site
- K_A : number of nucleotide substitutions per non-synonymous site
- 血紅素蛋白(globin)

Higher rates in rodents than in primates

Gene pair	K_S	K_A
β-like globin genes[a]		
Human adult–Human fetal	0.73	0.18
Mouse adult–Mouse fetal	0.90	0.21
Human adult–Human embryonic	0.62	0.16
Mouse adult–Mouse embryonic	0.97	0.18
Human fetal–Human embryonic	0.56	0.10
Mouse fetal–Mouse embryonic	0.96	0.15
Aldolase *A* and *B* genes		
Human *A*–Human *B*	1.55	0.21
Rat *A*–Rat *B*	1.92	0.21

Data from Li *et al.* (1987a).

[a]The adult globin genes are β in human and β_{maj} in mouse; the fetal genes are $^G\gamma$ in human and β_H in mouse; and the embryonic genes are ε in human and $\beta 1$ in mouse.

"Primitive" versus advanced: A question of rates

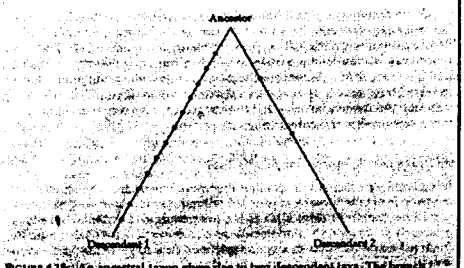


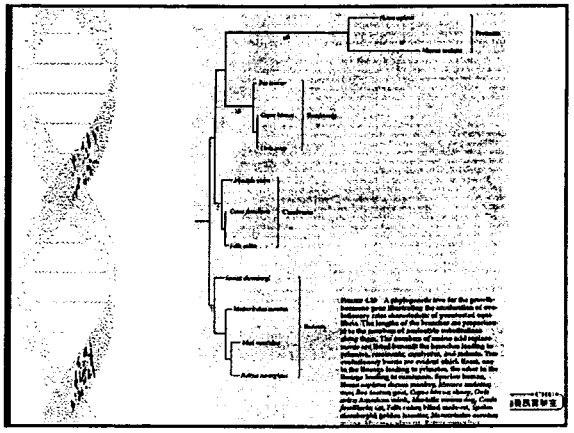
FIGURE 4.18 An ancestral taxon gives rise to two descendant taxa. The branch leading to descendant 1 has accumulated more substitutions (dots) than the branch leading to descendant 2.

"Primitive" versus advanced: A question of rates

- The lineage leading to descendant 1 evolevd faster (i.e., accumulated more substitutions) than the lineage leading to descendant 2.
- We may conclude that descendant 2 is more primitive than descendant 1.
- Ergo, Homo sapiens may be the most primitive mammal.

Phyletic gradualism versus punctuated equilibria at the molecular level

- In figure 4.19, we see that growth hormone genes evolve quite slowly throughout most mammalian evolution.
- Two independent bursts of rapid evolution.



Phyletic gradualism versus punctuated equilibria at the molecular level

- The lengths of the branches are proportional to the numbers of nucleotide substitutions along them.
- Two evolutionary bursts are evident, one in the lineage leading to primates, the other in the lineage leading to ruminants.

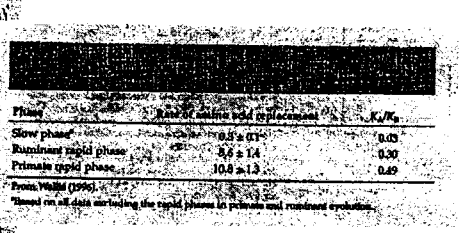
Phyletic gradualism versus punctuated equilibria at the molecular level

- Three possible explanations could account for the increased evolutionary rates in ruminants and primates
  - (1)an increase in the mutation rate
  - (2)positive selection for altered biological properties
  - (3)relaxation of purifying selection

Phyletic gradualism versus punctuated equilibria at the molecular level

- Table 4.14 we see that during the rapid phases of evolution, there is a significant increase in the K_A/K_S values, indicative of either positive selection or relaxation of selection.

Phyletic gradualism versus punctuated equilibria at the molecular level



Phase	Rate of amino acid replacement	K_A/K_S
Slow phase	0.5 ± 0.1*	0.03
Ruminant rapid phase	8.6 ± 1.4	0.30
Primate rapid phase	10.0 ± 1.3	0.49

From Wolfe (1994).
*Based on all data including the rapid phases in primate and ruminant evolution.

Rates of substitution in organelle DNA – Mammalian mitochondrial genes

- The synonymous rate of substitution in mammalian mitochondrial protein-coding genes has been estimated to be 5.7*10^-8 substitutions per synonymous site per year.
- This is about 10 times the value for synonymous substitution in nuclear protein-coding genes.

Rates of substitution in organelle DNA - Plant nuclear, mitochondrial, and chloroplast DNAs

- Since the plant and animal kingdoms diverged about 1 billion years ago, the pattern of evolution in plants might have become very different from that in animals.
- Plants differ from animals in the organization of their organelle by mitochondrial and chloroplast.

Plant nuclear, mitochondrial, and chloroplast DNAs

Species	Size (bp)	Open reading frames*	rRNAs	tRNAs	Introns
Mitochondria					
Chondrus crispus	25,836	36	3	23	1
Prototheca wickerhamii	55,328	36	3	26	5
Marchantia polymorpha	186,609	74	3	29	32
Arabidopsis thaliana	366,923	117	3	21	22
Chloroplasts					
Epifagus virginiana	70,028	54	8	28	12
Odontella sinensis	119,704	140	6	29	0
Pinus thunbergii	119,707	156	4	36	15
Marchantia polymorpha	121,024	89	8	34	21
Oryza sativa	134,525	108	8	44	14
Zea mays	140,387	111	8	39	25
Euglena gracilis	143,172	65	8	43	146
Nicotiana tabacum	158,939	102	7	36	25

*Putative protein-coding genes larger than 100 codons. Some of the differences in protein-coding gene number among chloroplast genomes may be attributed to the presence of variable numbers of duplicated genes.

Plant nuclear, mitochondrial, and chloroplast DNAs

- Table 4.16 shows a comparison of the substitution rates in the three genomes of vascular plants.
- Ks :number of synonymous substitutions per synonymous site
- KA : number of non-synonymous substitutions per non-synonymous site
- Ls :number of synonymous site
- LA :number of non-synonymous site

Plant nuclear, mitochondrial, and chloroplast DNAs

Genome	K_S	L_S	K_A	L_A
Comparison between monocot and dicot species				
Chloroplast	0.58 ± 0.02	4,177	0.05 ± 0.00	14,421
Mitochondrial	0.21 ± 0.01	1,219	0.04 ± 0.00	4,380
Comparison between maize and wheat or barley				
Nuclear	0.71 ± 0.04	1,475	0.06 ± 0.00	5,096
Chloroplast	0.17 ± 0.01	2,068	0.01 ± 0.00	7,001
Mitochondrial	0.03 ± 0.01	413	0.01 ± 0.00	1,326

From Wolfe et al. (1987, 1989b).

*K_S, number of synonymous substitutions per synonymous site; K_A, number of nonsynonymous substitutions per nonsynonymous site; L_S and L_A, are the number of synonymous and nonsynonymous sites, respectively.

Rates of Substitution In DNA Viruses - Estimation models

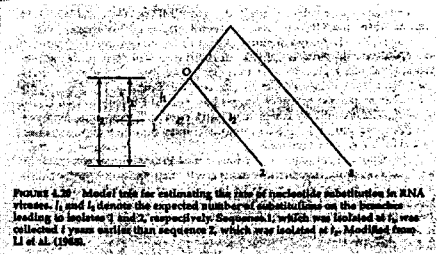


FIGURE 4.20 Model tree for estimating the rate of nucleotide substitution in RNA viruses. l_1 and l_2 denote the expected number of substitutions on the branches leading to isolates 1 and 2, respectively. Sequence 1, which was isolated at t_1, was collected t years earlier than sequence 2, which was isolated at t_2. Modified from Li et al. (1985).

Rates of Substitution In DNA Viruses - Estimation models

- Sequence 1 was isolated t years earlier than sequence 2.
- R is the rate of substitution per nucleotide site per year
- L_1 and L_2 are the expected numbers of substitutions per site from O to the time of isolation of sequence 1 and 2.

Rates of Substitution In DNA Viruses - Estimation models

$$l_2 - l_1 = rt_2 - rt_1 = rt \qquad (4.37)$$

$$l_2 - l_1 = d_{23} - d_{13} \qquad (4.38)$$

$$r = \frac{d_{23} - d_{13}}{t} \qquad (4.39)$$

Rates of Substitution In DNA Viruses - Estimation models

- d_{ij} denotes the number of substitutions per site between sequence i and j

Human immunodeficiency viruses

Coding region	Function	K_s (range)		K_a (range)	
gag	Group-specific antigen	9.7	(6.5–13.1)	1.7	(1.1–2.3)
pol	Polymerase	11.0	(7.4–14.8)	1.6	(1.0–2.1)
vif	Infectivity	9.1	(6.1–12.3)	4.7	(3.1–6.3)
tat (exon 2)	Regulatory	7.0	(4.7–9.5)	8.3	(5.6–11.2)
rev (exon 2)	Regulatory	7.4	(5.0–10.0)	6.6	(4.5–8.9)
gp120	Outer membrane protein	8.1	(5.5–10.9)	3.3	(2.2–4.4)
env	Hypervariable region	17.2	(11.6–23.2)	14.0	(9.4–18.8)
gp41	Transmembrane protein	9.6	(6.6–13.2)	5.1	(3.5–6.9)
env	Envelope	9.2	(6.3–12.4)	5.1	(3.5–6.9)
p27	Capsid protein	7.9	(5.2–10.7)	5.9	(4.0–8.0)
Average[a]		9.64	(2.92)	5.58	(3.60)

Modified from Li et al. (1988).

[a]The average is the arithmetic mean, and values in parentheses are the standard deviations computed over all genes.

Chapter 5

Molecular Phylogenetics

M09202050
顏勝茂

INTRODUCTION

- Molecular phylogenetics is the study of evolutionary relationships among organisms by using molecular data such as DNA and protein sequences, insertions of transposable elements, or other molecular markers.

INTRODUCTION

- This chapter will:(1)introduce the vocabulary of phylogenetics, (2)explain how to reconstruct a phylogenetic tree from molecular data, and (3) discuss some theoretical problems associated with molecular phylogenetic reconstruction.

TERMINOLOGY OF PHYLOGENETIC TREES

- The evolutionary relationships among a group of organisms are illustrated by means of a **phylogenetic tree** (or **dendrogram**).
- A phylogenetic tree is a graph composed of **nodes** and **branches**, in which only one branch connects any two adjacent nodes.

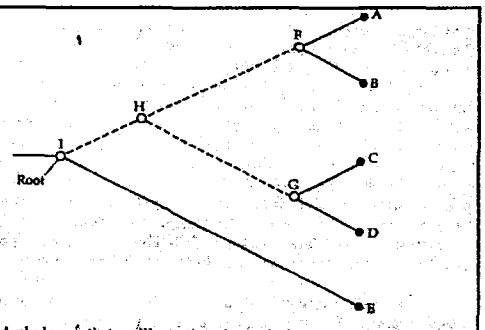


FIGURE 5.1 A phylogenetic tree illustrating the evolutionary relationships among five OTUs (A–E). Solid and white circles denote terminal and internal nodes, respectively. Solid and dashed lines denote terminal and internal branches, respectively. The internal nodes (F–H) represent the HTUs. I is the root.

TERMINOLOGY OF PHYLOGENETIC TREES

- We distinguish between **terminal** and **internal nodes**, and between **external branches** (branches that end in a tip) and **internal branches** (branches that do not end in a tip).

TERMINOLOGY OF PHYLOGENETIC TREES

- Terminal nodes represent the extant taxonomic units under comparison, which are referred to as **operational taxonomic units** (**OTUs**). Internal nodes represent inferred ancestral units, and since we have no empirical data pertaining to these taxa, they are sometimes referred to as **hypothetical taxonomic units** (**HTUs**).

TERMINOLOGY OF PHYLOGENETIC TREES

- A node is **bifurcating** if it has only two immediate descendant lineages, but **multifurcating** if it has more than two immediate descendant lineages.
- In evolutionary studies we assume that the process of speciation is usually a binary one.

Rooted and unrooted trees

- In a **rooted tree** there exists a particular node, called the root, from which a unique path leads to any other node.
- Au **unrooted tree** is a tree that only specifies the degree of kinship among the taxonomic units but does not define the evolutionary path.

Rooted and unrooted trees

- In an unrooted tree with four external nodes, the internal branch is frequently referred to as the **central branch**.

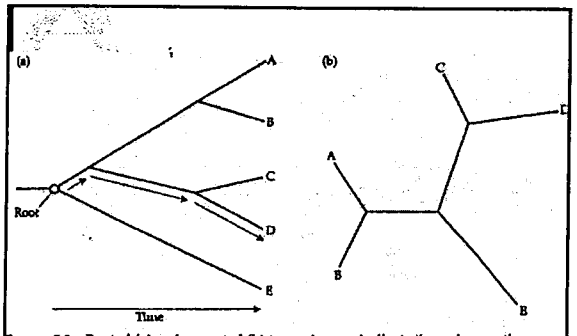


FIGURE 5.2 Rooted (a) and unrooted (b) trees. Arrows indicate the unique path leading from the root to OTU D.

Scaled and unscaled trees

- **Unscaled**:their lengths are not proportional to the number of changes, which are indicated on the branches.
- **Scaled**:each branch length is proportional to the number of changes(e.q.,nucleotide substitutions) that have occurred along that branch.

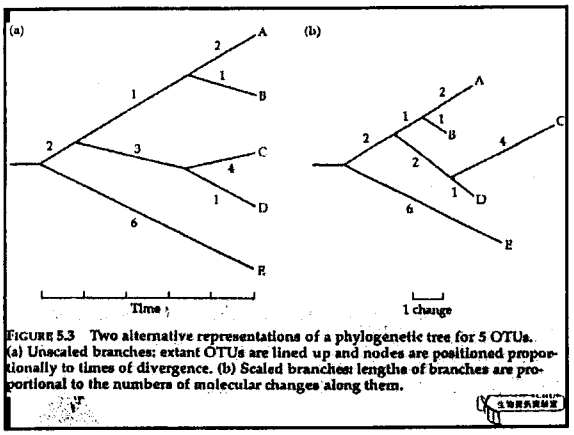


FIGURE 5.3 Two alternative representations of a phylogenetic tree for 5 OTUs. (a) Unscaled branches; extant OTUs are lined up and nodes are positioned proportionally to times of divergence. (b) Scaled branches: lengths of branches are proportional to the numbers of molecular changes along them.

The Newick format

- In computer programs, trees are represented in a linear form by a series of nested parentheses, enclosing names and separated by commas.

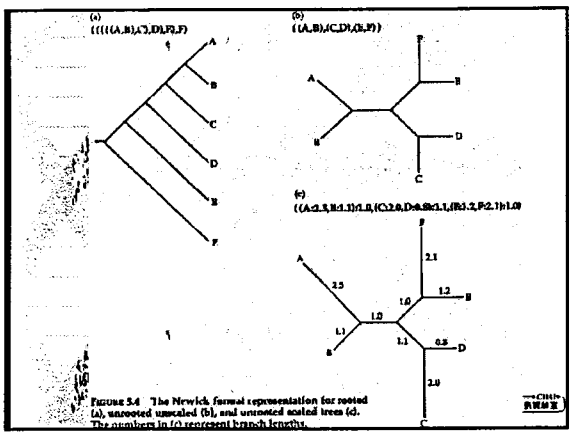


FIGURE 5.4 The Newick format representation for rooted (a), unrooted unscaled (b), and unrooted scaled trees (c). The numbers in (c) represent branch lengths.

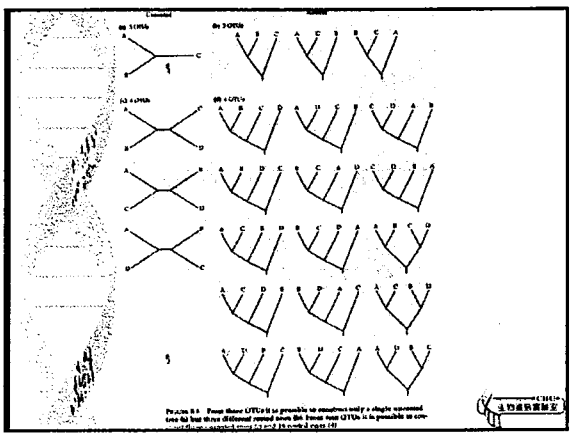
Number of possible phylogenetic trees

- The number of bifurcating rooted tree (N_R) for n OTUs is given by

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \qquad n>=2 \qquad (5.1)$$

- The number of bifurcating unrooted trees (N_U) for n >= 3 is

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad (5.2)$$



Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
14	7,905,853,580,625	316,234,143,225
15	213,458,046,676,875	7,905,853,580,625
16	6,190,283,353,629,375	213,458,046,676,875
17	191,898,783,962,510,625	6,190,283,353,629,375
18	6,332,659,870,762,850,625	191,898,783,962,510,625
19	221,643,095,476,699,771,875	6,332,659,870,762,850,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

Data from Felsenstein (1978b).

True and inferred trees

- **True tree**:only one of all the possible trees that can be built with a given number of OTUs represents the true evolutionary history.
- **Inferred tree**:a tree that is obtained by using a certain set of data and a certain method of tree reconstruction.

Gene trees and species trees

- The routers of inheritance represent the passage of genes from parents to offspring, and the branching pattern depicts a **gene tree.**
- When we trace back the history of many genes from different species, we infer the routers of inheritance for the species, and in this case we obtain a phylogenetic tree for the species, or a **species tree.**

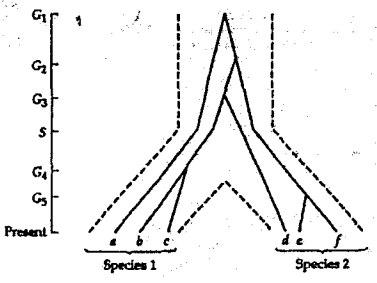


FIGURE 5.6 Diagram showing that in a genetically polymorphic population, gene splitting events (G_1–G_5) may occur before or after the speciation event (S). The evolutionary history of gene splitting resulting in the six alleles denoted *a*–*f* is shown in solid lines; speciation (i.e., population splitting), is shown by broken lines. Modified from Nei (1987).

Gene trees and species trees

- The gene tree can differ from the species tree in two respects.
  - First,the divergence of two genes sampled from two different species may have pre-dated the divergence of the species from each other.
- The second problem with gene trees is that the branching pattern of a gene tree may be different from that of the species trees.

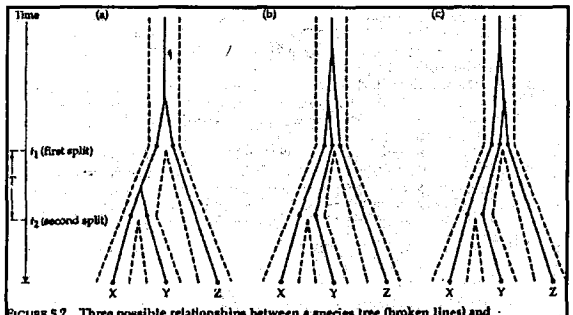


FIGURE 5.7 Three possible relationships between a species tree (broken lines) and a gene tree (solid lines). In (a) and (b), the topologies of the species trees are identical to those of the gene trees. Note that in (a) the time of divergence between the genes is roughly equal to the time of divergence between the populations. In (b), on the other hand, the time of divergence between genes X and Y greatly pre-dates the time of divergence between the respective populations. The topology of the gene tree in (c) is different from that of the species tree. Modified from Nei (1987).

Taxa and clades

- A **taxon** is a species or a group of species (e.g., a genus, family, order, or class).
- **Natural clades**:a group of all the taxa that have been derived from a common ancestor, plus the common ancestor itself.
- **Archosauria**:two taxa share a common ancestor not shared by any extant organism.

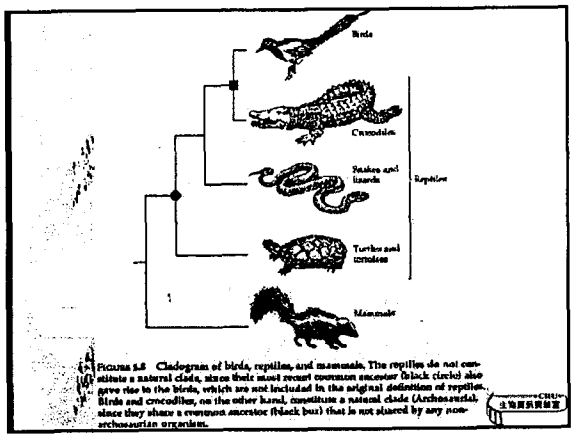


FIGURE 5.5 Cladogram of birds, reptiles, and mammals. The reptiles do not constitute a natural clade, since their most recent common ancestor (black circle) also gave rise to the birds, which are not included in the original definition of reptiles. Birds and crocodiles, on the other hand, constitute a natural clade (Archosauria), since they share a synapsus ancestor (black bar) that is not shared by any non-archosaurian organism.

TYPES OF DATA

- Molecular data fall into one of two categories:characters and distances.
- A characters provides information about an individual OTU.
- A distance represents a quantitative statement concerning the dissimilarity between two OTUs.

Character data

- Characters are either **quantitative** or **qualitative**.
- The character states of a quantitative character(e.g.,height) are usually **continuous** and are measured on an interval scale.
- The character states of a qualitative character(e.g.amino acid positions in a protein) are **discrete**.

Assumptions about caracter evolution

- (1).The number of discrete steps required for one character state to change into another.
- (2).The probability with which such a change may occur.
- A character is designated as **unordered** if a change from one character state to another occurs in one step.

Assumptions about caracter evolution

- **Ordered** if the number of steps from one state to another equals the absolute value of the difference between their state number.
- **Partially ordered** characters are those characters in which the number of steps varies for the different pairwise combination of character states, but for which no definite relationship exists between the number of steps and the character-state number.

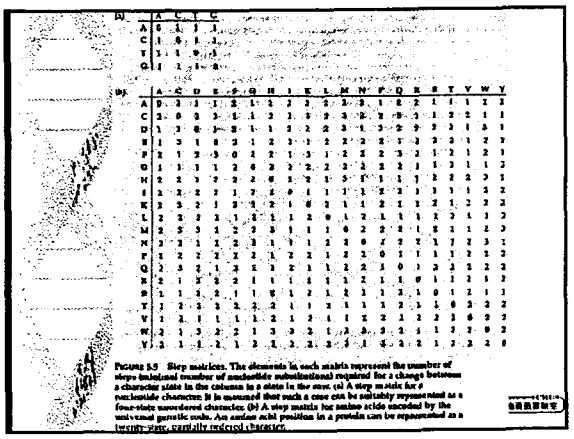


FIGURE 5.5 Step matrices. The elements in each matrix represent the number of steps (minimal number of nucleotide substitutions) required for a change between a character state in the column to a state in the row. (a) A step matrix for a nucleotide character. It is assumed that such a base can be suitably represented as a four-state unordered character. (b) A step matrix for amino acids encoded by the universal genetic code. An amino acid position in a protein can be represented as a twenty-state partially ordered character.

Distance data

- Three possible reasons for converting characters into distances.
  - 1.A long list of character states, such as a DNA sequence, is in itself meaningless in an evolutionary context.
  - 2.As pointed out in Chapter 3, one must take into account multiple substitutions at a site.
  - 3.Numerous methods exits for inferring phylogenetic trees from distance data.

METHODS OF TREE RECONSTRUCTION

- A phylogenetic reconstruction, therefore, consists of two steps:
  - Definition of an **optimality criterion**, or **objective function**.
  - Design of specific algorithms to compute the value of the objective function and to identify the tree (or set of trees) that have the best values according to this criterion.

DISTANCE MATRIX METHODS

- In the distance matrix methods, evolutionary distances (usually the number of nucleotide substitutions or amino acid replacements between two taxonomic units) are computed for all pairs of taxa, and a phylogenetic tree is constructed by using an algorithm based on some functional relationships among the distance values.

Unweighted pair-group method with arithmetic means(UPGMA)

- Identify from among all the OTUs (or **simple OTUs**) the two that are most similar to each other and treat these as a new single OTU. Such and OTU is referred to as a **composite OUT**.

OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

- Let us assume that d_{AB} is the smallest. Then, OTUs A and B are the first to be clustered, and the branching point, l_{AB}, is positioned at a distance of $d_{AB}/2$ substitutions.

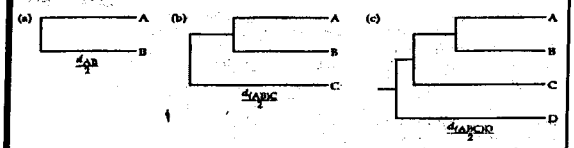


FIGURE 5.10 Diagram illustrating the stepwise construction of a phylogenetic tree for four OTUs by using UPGMA (see text).

Unweighted pair-group method with arithmetic means(UPGMA)

- The first clustering, A and B are considered as a single composite OUT (AB), and a new distance matrix is computed.

OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

- The branching point between two simple OTUs, I and j, is positioned at half the distance between them.

$$l_{ij} = \frac{d_{ij}}{2} \quad (5.3)$$

- The branching point between a simple OTU,I,and a composite OTU,(jm).

$$l_{(ij)(jm)} = \frac{(d_{ij} + d_{im})/2}{2} \quad (5.4)$$

- The position of the branching point between a composite OTU, (ij), and a composite OTU, (mn) is

$$l_{(ij)(mn)} = \frac{(d_{im} + d_{in} + d_{jm} + d_{jn})/4}{2} \quad (5.5)$$

- In the case of a tripartite composite OTU, (ijk), and a bipartite composite OTU, (mn)

$$l_{(ijk)(mn)} = \frac{(d_{im} + d_{in} + d_{jm} + d_{jn} + d_{km} + d_{kn})/6}{2} \quad (5.6)$$

Transformed distance method

- If the assumption of rate constancy among lineages does not hold, UPGMA may give an erroneous topology.

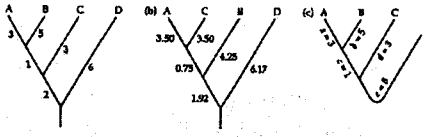


FIGURE 5.11 (a) The true phylogenetic tree. (b) The erroneous phylogenetic tree reconstructed by using UPGMA, which does not take into account the possibility of unequal substitution rates along the different branches. (c) The tree inferred by the transformed distance method. The root must be on the branch connecting OTU D and the node of the common ancestor of OTUs A, B, and C, but its exact location cannot be determined by the transformed distance method.

OTU	A	B	C
B	8		
C	7	9	
D	12	14	11

- A correction called the **transformed distance method.**
- An **outgroup** is an OTU or a group of several OTUs for which we have external knowledge, such as taxonomic or paleontological information, that clearly shows them to have diverged from the common ancestor prior to all the other OTUs under consideration (the **ingroup** taxa).

$$d'_{ij} = \frac{d_{ij} - d_{iO} - d_{jO}}{2} + \bar{d}_O \quad (5.7)$$

where d'_{ij} is the transformed distance between OTUs i and j, and $\bar{d}_O$ is a correction term. It is calculated as

$$\bar{d}_O = \frac{\sum_{k=1}^{n} d_{kO}}{n} \quad (5.8)$$

- n is the number of ingroup OTUs.

OTU	A	B
B	10/3	
C	13/3	13/3

- The transformed distance method dones not provide branch lengths.

Sattath and Tversky's neighbors-relation method

- In an unrooted bifurcating tree, two OTUs are said to be **neighbors** if they are connected through a single internal node.

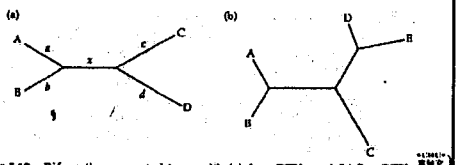


FIGURE 5.12 Bifurcating unrooted trees with (a) four OTUs and (b) five OTUs.

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2x = d_{AB} + d_{CD} + 2x \quad (5.9)$$

$$d_{AB} + d_{CD} < d_{AC} + d_{BD} \quad (5.10)$$

and

$$d_{AB} + d_{CD} < d_{AD} + d_{BC} \quad (5.11)$$

- These two conditions are collectively known as the **four-point condition.**
- This approach to phylogenetic reconstruction is called the **neighborliness approach.**
- Sattath and Tversky's method does not provide branch lengths.

Saitou and Nei's neighbor-joining method

- The **neighbor-joining** method is also a neighborliness method. It provides an approximate algorithm for finding the shortest **(minimum evolution)** tree.
- The sum of all the branch lengths is

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (d_{1k} + d_{2k}) + \frac{1}{2} d_{12} + \frac{1}{N-2} \sum_{3 \le i < j \le N}^{N} d_{ij} \qquad (5.12)$$

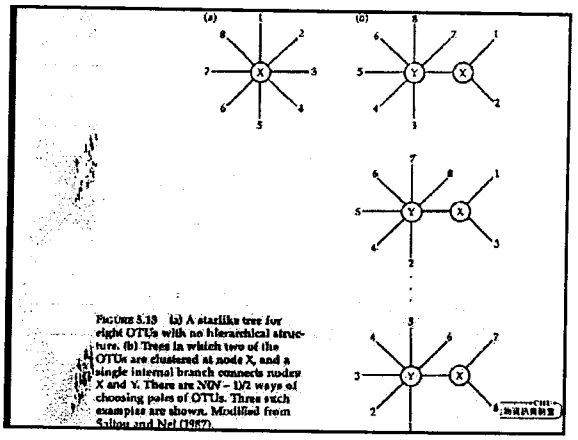


FIGURE 5.13 (a) A starlike tree for eight OTUs with no hierarchical structure. (b) Trees in which two of the OTUs are clustered at node X, and a single internal branch connects nodes X and Y. There are $NJV - 1)/2$ ways of choosing pairs of OTUs. Three such examples are shown. Modified from Saitou and Nei (1987).

MAXIMUM PARSIMONY METHODS

- The principle of **maximum parsimony** involves the identification of a topology that requires the smallest number of evolutionary changes (e.g., nucleotide substitutions) to explain the observed differences among the OTUs under study.
- Use discrete character states, and the shortest pathway leading to these character states is chosen as the best tree. Such a tree is called a **maximum parsimony tree.**

MAXIMUM PARSIMONY METHODS

- A site is defined as **invariant** if all the OTUs under study possess the same character state at this site.
- **Variable** sites may be **informative** or **uninformative.**
- A nucleotide site is phylogenetically informative only if it favors a subset of trees over the other possible trees.

MAXIMUM PARSIMONY METHODS

- A site is informative only when there are at least two different kinds of nucleotides at the site, each of which is represented in at least two of the sequences under study.

Sequence	Site 1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*

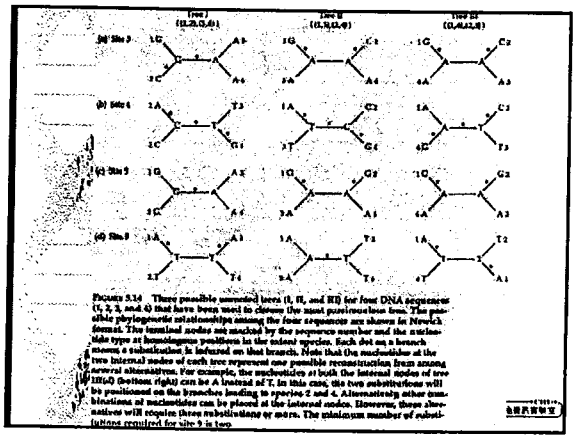


FIGURE 5.14 Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number and the nucleotide types at homologous positions in the related species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotides at both the internal nodes of tree III-(4) (bottom right) can be A instead of T. In this case, the two substitutions will be positioned on the branches leading to species 2 and 4. Alternatively, other combinations of nucleotides can be placed at the internal nodes. However, these alternatives will require three substitutions or more. The minimum number of substitutions required for site 9 is two.

MAXIMUM PARSIMONY METHODS

- infer a maximum parsimony tree
  - 1. identify all the informative site.
  - 2. for each possible tree we calculate the minimum number of substitutions at each informative site.
  - 3. We sum the number of changes over all the informative sites for each possible tree and choose the tree associated with the smallest number of changes.

MAXIMUM PARSIMONY METHODS

- The inference of the number of substitutions for a given tree can be made by using Fitch's method.
  - node相同 → 取node交集
  - node不相同 → 取node聯集

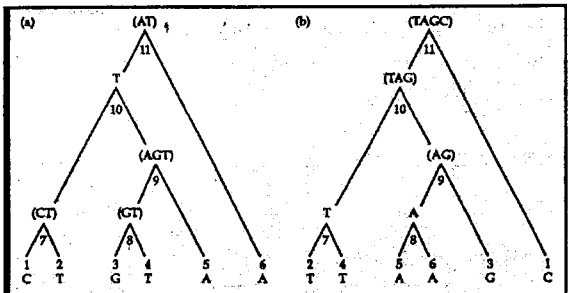


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal node is dependent on the tree. Modified from Fitch (1971).

MAXIMUM PARSIMONY METHODS

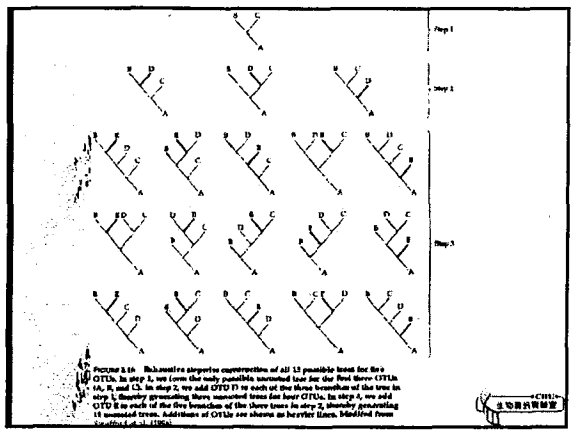
- The number of unions equals the minimum number of substitutions required to account for the descendant nucleotides from a common ancestor.
- The number of substitutions at an uninformative site is equal to the number of different nucleotides present at that site minus one.
- The total number of substitutions at both informative and uninformative sites in a particular tree is called **tree length.**

Weighted and unweighted parsimony

- **Unweighted parsimony** : All the different nucleotide substitutions were given equal weight.
- **Weighted parsimony** : assign different weights to the various character state changes.
  - We may wish to give a greater weight to transversions, since they occur less frequently than transitions.

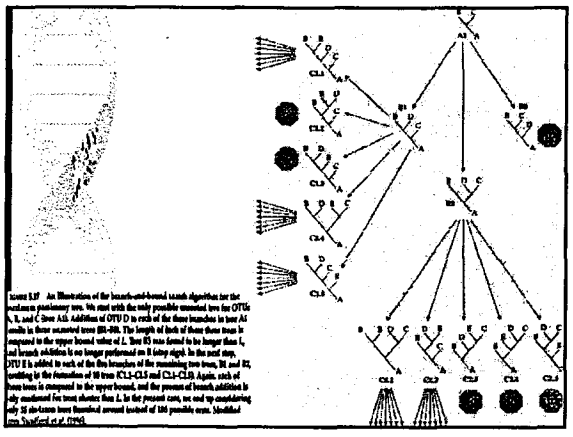
Searching for the maximum parsimony tree

- Determine all the trees length, and choose from among them the shortest one (or ones). This type of search for the maximum parsimony tree(s) is called an **exhaustive search.**



Searching for the maximum parsimony tree

- **Branch-and-bound method** :
  - consider an arbitrary tree or, better, a tree obtained from a fast method (e.g., the neighbor-joining method), and compute the minimum number of substitutions, L, for the tree.
  - L is then considered as the **upper bound** to which the length of any other tree is compared.
  - The rationale of the upper bound is that the maximum parsimony tree must be either equal in length to L or shorter.



Searching for the maximum parsimony tree

- **Heuristic searches** :
  - An initial tree is constructed by using a certain procedure, and we seek to find a shorter tree by examining trees that have a similar topology to the initial one.
  - If a shorter tree is found among the set of similar trees, a new round of exploration is initiated stating from this new tree.
  - This iterative quest terminated when at a certain round we fail to find a shorter tree within the set of similar ones.

Searching for the maximum parsimony tree

- **Branch swapping**(or rearrangement) : that can be used to generate toplogically similar trees from an initial one.

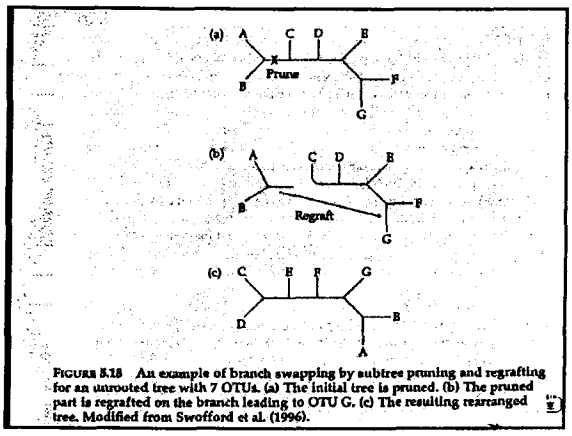


FIGURE 5.18 An example of branch swapping by subtree pruning and regrafting for an unrooted tree with 7 OTUs. (a) The initial tree is pruned. (b) The pruned part is regrafted on the branch leading to OTU G. (c) The resulting rearranged tree. Modified from Swofford et al. (1996).

MAXIMUM LIKELIHOOD METHODS

- The **likelihood**, L, of a phylogenetic tree is the probability of observing the data (e.g., the nucleotide sequences) under a given tree and a specified model of character state changes(e.g., the substitution pattern).
- The aim of maximum likelihood methods is to find the tree (from among all the possible tress) with the highest L value.

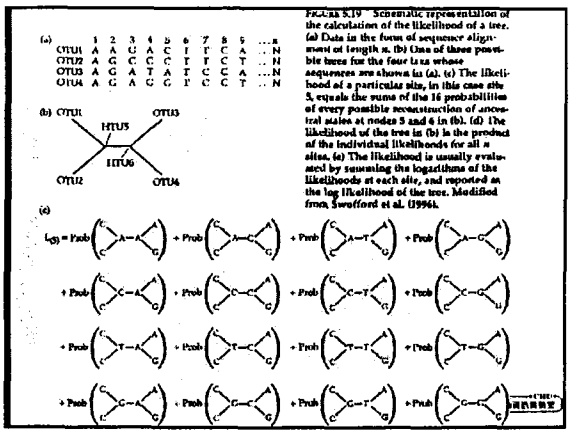


FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length n. (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood at a particular site, in this case site 5, equals the sums of the 16 probabilities of every possible reconstruction of ancestral states at nodes 5 and 6 in (b). (d) The likelihood of the tree in (b) is the product of the individual likelihoods for all n sites. (e) The likelihood is usually evaluated by summing the logarithms of the likelihoods at each site, and reported as the log likelihood of the tree. Modified from Swofford et al. (1996).

MAXIMUM LIKELIHOOD METHODS

- **Log likelihood** (lnL) : the likelihood is usually evaluated by the logarithmic transformation, which transforms multiplication into summation.
- Compute the likelihood values for the other possible tress, and the tree with the highest likelihood value is chosen as the **maximum likelihood tree.**

(d) $L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \ldots \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$

(e) $\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + \ldots + L_{(n)} = \sum_{i=1}^{n} \ln L_{(i)}$

ROOTING UNROOTED TREES

- To root an unrooted tree, we usually need an outgroup (an OTU for which external information, such as paleontological evidence, clearly indicates that it has branched off earlier than the taxa under study).
- The root is then placed between the outgroup and the node connecting it to the other OTUs, which are the ingroup.

ROOTING UNROOTED TREES

- In the absence of an outgroup, we may position the root by assuming that the rate of evolution has been approximately uniform over all the branches.
- Under this assumption we put the root at the midpoint of the longest pathway between two OTUs.

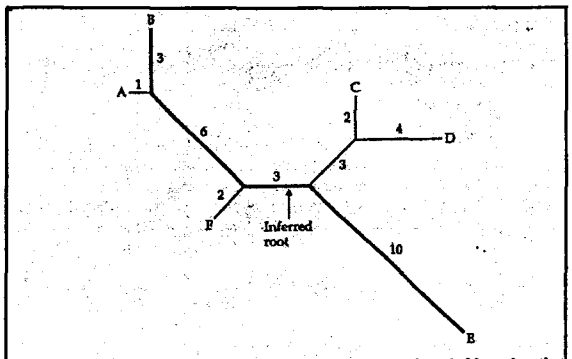


FIGURE 5.20 A hypothetical unrooted phylogenetic tree with scaled branches that has been rooted at the midpoint of the longest pathway (thick line) from among all possible pathways between two OTUs (i.e., B and E). The numbers of substitutions are marked on the branches.

ROOTING UNROOTED TREES

- An unrooted tree cannot be said to represent the evolutionary history of divergence among a group of taxa.

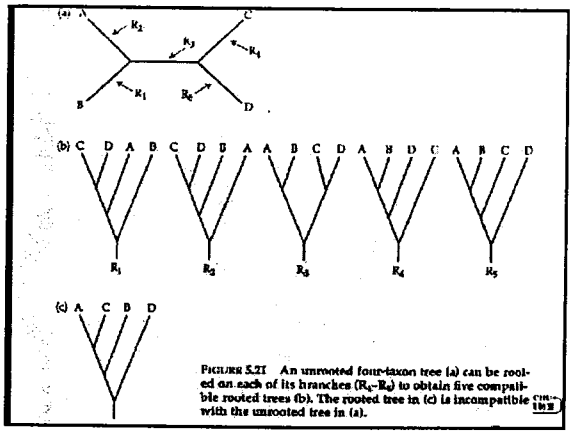


FIGURE 5.21 An unrooted four-taxon tree (a) can be rooted on each of its branches (R_1–R_5) to obtain five compatible rooted trees (b). The rooted tree in (c) is incompatible with the unrooted tree in (a).

ESTIMATING BRANCH LENGTHS

- Fitch and Margoliash's method
  - Assuming that the tree topology has already been inferred by a distance matrix procedure

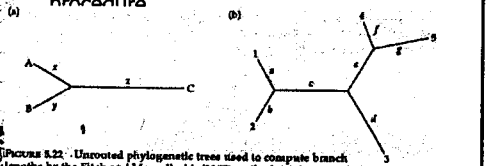


FIGURE 5.22 Unrooted phylogenetic trees used to compute branch lengths by the Fitch and Margoliash's (1967) method. (a) A tree with three OTUs. (b) A tree with five OTUs.

ESTIMATING BRANCH LENGTHS

$$d_{AB} = x + y \quad (5.13a)$$

$$d_{AC} = x + z \quad (5.13b)$$

$$d_{BC} = y + z \quad (5.13c)$$

$$x = \frac{d_{AB} + d_{AC} - d_{BC}}{2} \quad (5.14a)$$

$$y = \frac{d_{AB} + d_{BC} - d_{AC}}{2} \quad (5.14b)$$

$$z = \frac{d_{AC} + d_{BC} - d_{AB}}{2} \quad (5.14c)$$

ESTIMATING BRANCH LENGTHS

- Note that sometimes an estimated branch length can be negative. Since the true length can never be negative, it is better to replace such an estimate by 0.

ESTIMATING SPECIES DIVERGENCE TIMES

- Let us assume that the rate of evolution for a DNA sequence is known from a previous study to be r substitutions per site per year.
  - The divergence time, T.
  - The number of substitutions per site, K.
  - Chapter 4(Equation 4.1) $r = K/2T$

$$T = \frac{K}{2r} \quad (5.15)$$

ESTIMATING SPECIES DIVERGENCE TIMES

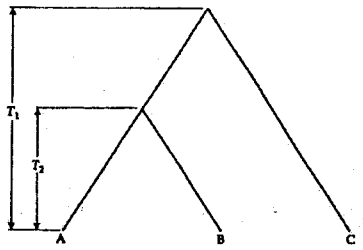


FIGURE 5.23 Model tree for estimating times of divergence. T_1 = divergence time between species C and the ancestor of species A and B. T_2 = divergence time between species A and B.

ESTIMATING SPECIES DIVERGENCE TIMES

Let K_{ij} be the number of nucleotide substitutions per site between species i and j .

$$r = \frac{K_{AC} + K_{BC}}{2(2T_1)} \tag{5.16}$$

The unknown divergence time between species A and B (T_2) is estimated by

$$T_2 = \frac{K_{AB}}{2r} = \frac{K_{AB}T_1}{K_{AC} + K_{BC}} \tag{5.17}$$

Conversely, in the case that T_2 is known but T_1 is not, T_1 is given by

$$T_1 = \frac{(K_{AC} + K_{BC})T_2}{2K_{AB}} \tag{5.18}$$

Molecular Phylogenetics

CHU.CSIE
M09102048
賴韋丞

Topological Comparisons

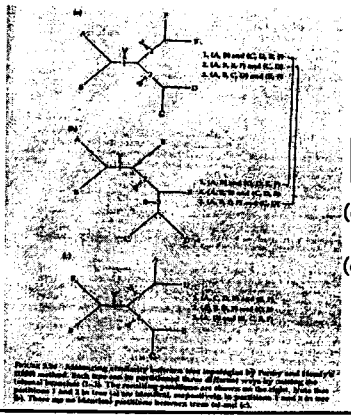
- It is sometimes Necessary to measure the similarity or dissimilarity among several tree topologies.
- Several methods of tree reconstruction(maximum parsimony) may produce many trees rather than a unique phylogeny.
- When two trees derived from different data sets or different methodologies are identical, they are said to be congruent.
- Congruence can sometimes be partial, i.e., limited to some parts of the trees, other parts being incongruent.

Penny and Hendy's topological distance

- A commonly used measure of dissimilarity between two tree topologies.
- The measure is based on tree partitioning, and is equal to twice the number of different ways of partitioning the OTUs between two trees.

$$d_T = 2c$$

- Where d_T is the topological distance and c is the number of partitions resulting in different divisions of the OTUs in the two tree.

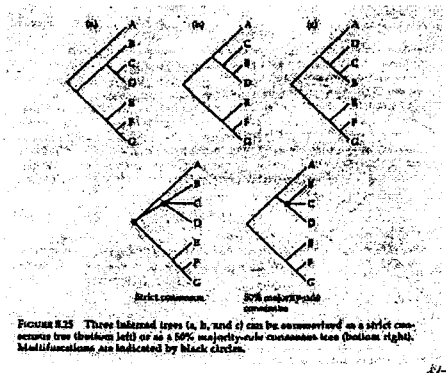


$$d_T = 2c$$

(a) and (b):
$d_T = 2x1 = 2$
(a) and (c):
$d_T = 2x3 = 6$

Consensus trees

- Consensus trees are trees that have been derived form a set of trees.
- The purpose of a consensus tree is to summarize several trees as a single tree.
- In consensus trees the points of agreement among the fundamental trees are shown as bifurcations, whereas the points of disagreement are collapsed into polytomies.
- The most commonly used are the strict consensus and majority-rule consensus trees.



Assessing Tree Reliability

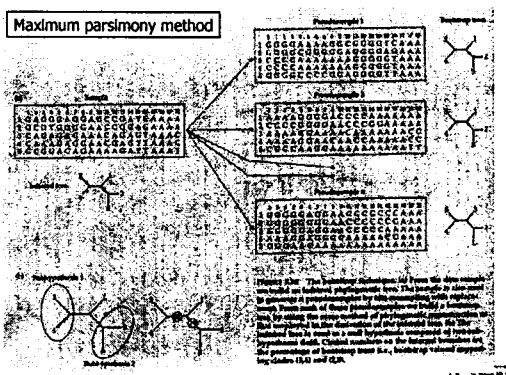
- Phylogenetic reconstruction is a problem of statistical inference.
- Assess the reliability of the inferred phylogeny and its component parts.
- After inferring a phylogenetic tree, two questions may be asked:
  - (1) How reliable is the tree? Which parts of the tree are reliable?
    Assess the reliability can be accomplished by several analytical or resampling
  - (2)Is this tree significantly better than another tree?

The bootstrap

- The bootstrap is a computational technique for estimating a statistic for which the underlying distribution is unknown or difficult to derive analytically.
- The bootstrap technique has been frequently uses as a means to estimate the confidence level of phylogenetic hypotheses.
- The bootstrap belongs to a class of methods called resampling techniques because it estimates the sampling distribution by repeatedly resampling data from the original sample data set.

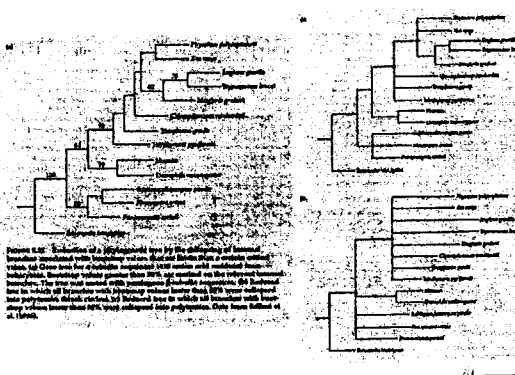
The bootstrap

- Each pseudosample is used to construct a tree by the same method used for the inferred tree.
- Subhypothesis (1) is given a score of 1 if OTUs 3 and 4 are sister taxa in a bootstrap tree, but a score of 0 otherwise. The score for subhypothesis (2) is similarly decided.
- Bootstrap values are expressed as percentages, and are indicated on the internal branches defining the clades.



The bootstrap

- A common practice in the literature is to "reduce" the inferred tree by collapsing branches that are associated with bootstrap values that are lower than a certain critical value.
- By using topological comparisons (see page 206) between simulated "true" trees and inferred ones, it has been shown that collapsed trees are more similar to the true tree than the original inferred tree.



Tests for two competing trees

- Several tests have been devised for testing whether one phylogeny is significantly better than another.
- Such tests exist for each of the three types of tree reconstruction methods (distance matrix, maximum parsimony, and maximum likelihood).
- In the following we present a simple test for testing maximum parsimony trees against alternative phylogenies.

Tests for two competing trees

- Kishino and Hasegawa Devised a parametric test for comparing two trees under the assumption that all nucleotide sites are independent and equivalent.
- The test uses the difference in the number of nucleotide substitutions at informative sites between the two trees, D, as a test statistic; where $D=\Sigma D_i$, and D_i is the difference in the minimum number of nucleotide substitutions between the two trees at the *i*th informative site.

Tests for two competing trees

- The sample variance of D is

$$V(D)=\frac{n}{n-1}\sum_{i=1}^{n}\left(D_i-\frac{1}{n}\sum_{k=1}^{n}D_k\right)^2$$

- Where n is the number of informative sites.
- The null hypothesis that D=0 can be tested with the paired t-test with n-1 degrees of freedom, where

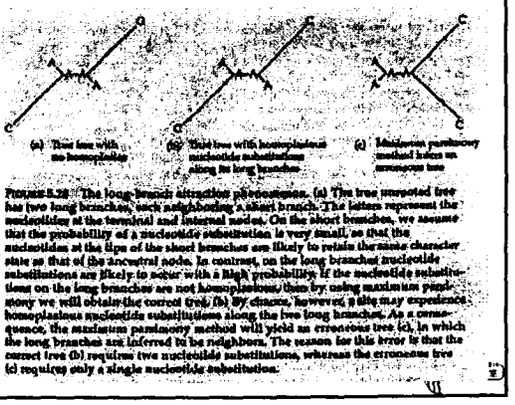
$$t=\frac{D/n}{\sqrt{V(D)\sqrt{n}}}$$

Problems Associated with Phylogenetic Reconstruction

- No method of phylogenetic reconstruction can be claimed to be better than others under all conditions.
- Each of the methods of phylogenetic reconstruction has advantages and disadvantages, and each method can succeed or fail depending on the nature of the evolutionary process, which is by and large unknown.
- In the following we will review the strengths and weaknesses of different methods and outline several strategies for minimizing error in phylogenetic analysis.

Strengths and weaknesses of different methods

- Maximum parsimony methods make no explicit assumptions except that a tree that requires fewer substitutions is better than one that requires more.
- When the degree of divergence between sequences is small so that homoplasies are rare, the parsimony criterion usually works well.
- However, when the degree of divergence is large so that homoplasies are common, maximum parsimony methods may yield faulty phylogenetic inferrences.



(a) Tree has with no homoplasies (b) That tree with homoplasious nucleotide substitutions along its long branches (c) Maximum parsimony method leads an erroneous tree

FIGURE 5.22 The long-branch attraction phenomenon. (a) The true unrooted tree has two long branches, each neighboring a short branch. The letters represent the nucleotides at the terminal and internal nodes. On the short branches, we assume that the probability of a nucleotide substitution is very small, so that the nucleotides at the tips of the short branches are likely to retain the same character state as that of the ancestral node. In contrast, on the long branches nucleotide substitutions are likely to occur with a high probability. If the nucleotide substitutions on the long branches are not homoplasious (that is, by using maximum parsimony we will obtain the correct tree (b) by chance, however, a tree may experience homoplasious nucleotide substitutions along the two long branches. As a consequence, the maximum parsimony method will yield an erroneous tree (c), in which the long branches are inferred to be neighbors. The reason for this error is that the correct tree (b) requires two nucleotide substitutions, whereas the erroneous tree (c) requires only a single nucleotide substitution.

Minimizing error in phylogenetic analysis

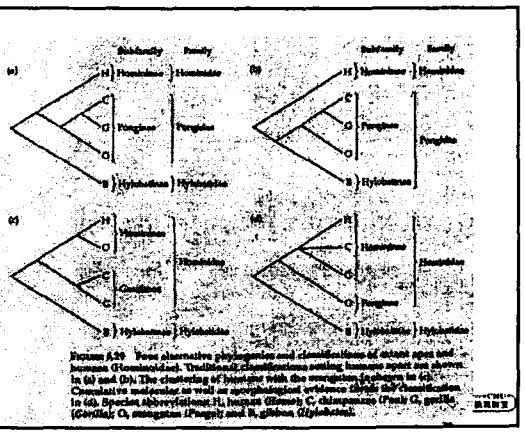
- The best way to minimize random errors is to use large amounts of data.
- A tree based on large amounts of molecular data is almost invariably more reliable than one based on a more limited amount of data.

Molecular Phylogenetic Examples

- In this section we present several examples where molecular studies have (1) resolved a longstanding issue, (2) led to a drastic revision of the traditional view, or (3) pointed to a new direction of research.

Phylogeny of humans and apes

- The issue of the closest living evolutionary relative of humans has always intrigued biologists.
- Darwin claimed that the African apes, the chimpanzee (*Pan*) and gorilla (*Gorilla*), are our closest relatives, and hence he suggested that the evolutionary origins of humans were to be found in Africa.
- Darwin's view fell into disfavor for various reasons.



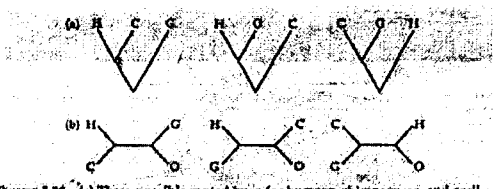
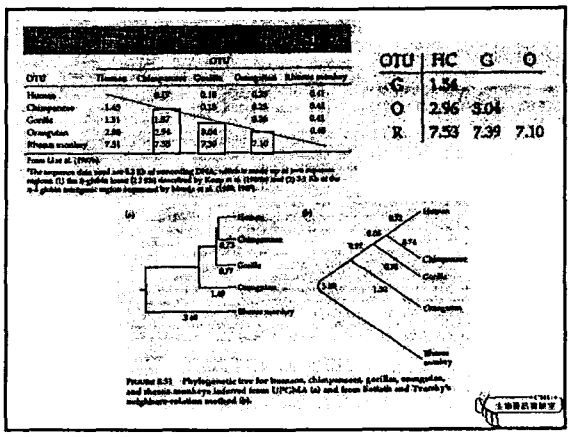


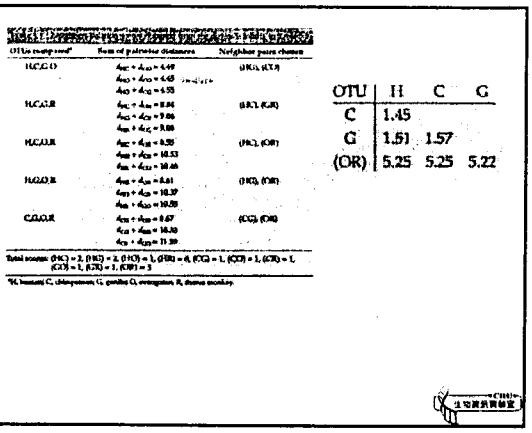
FIGURE 8.36 (a) Three possible rooted trees for humans, chimpanzees, and gorillas. (b) Comparable unrooted tree with the orangutan as an outgroup. Species abbreviations: H, human (*Homo sapiens*); C, chimpanzee (*Pan troglodytes*); G, gorilla (*Gorilla gorilla*); and O, orangutan (*Pongo pygmaeus*).

OTU	HC	G	O
G	1.54		
O	2.96	3.04	
R	7.53	7.39	7.10



OTU	HC	G	O
G	1.54		
O	2.96	3.04	
R	7.53	7.39	7.10

OTU	H	C	G
C	1.45		
G	1.61	1.57	
(OR)	5.25	5.25	5.22



Site*	Sequence: Human	Chimpanzee	Gorilla	Orangutan	Hypothesis supported*
Data from Miyamoto et al. (1987)					
34	A	G	A	G	III
566	C	C	A	A	I
1287	-	-	T	T	I
1326	C	G	A	A	I
3057–3060	----	----	TAAT	TAAT	I
3372	T	T	-	-	I
4673	C	C	T	T	I
5133	A	C	C	A	II
5154	A	G	G	A	II
5460	G	G	T	T	I
6268	C	T	C	T	III
8806	C	T	T	C	II
8971	G	G	T	T	I
Data from Maeda et al. (1988)					
127–132	------	------	AATATA	AATATA	I
1477	G	G	A	A	I
3131	A	A	G	G	I
3234	A	C	A	G	III
3241	G	C	G	C	III
3639	G	G	A	A-	I

Modified from Williams and Goodman (1989).

*Site numbers correspond to those given in the original sources. The total length of the sequences used is 10.5 kb, about twice that used in Table 8.2.

†Hypotheses: I, humans and chimpanzees are sister taxa; II, chimpanzees and gorillas are sister taxa; and III, humans and gorillas are sister taxa.

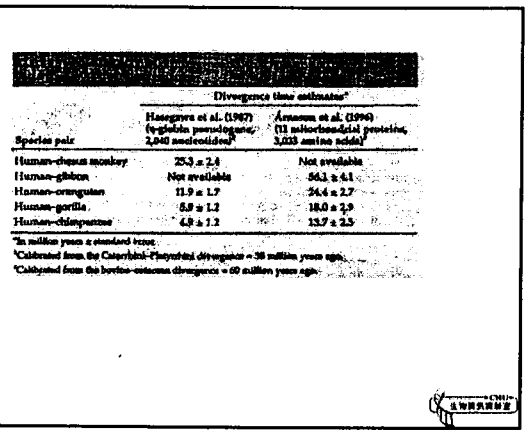
‡Each asterisk denotes a one-nucleotide indel at the site.

Species pair	Divergence time estimates*: Hasegawa et al. (1987) fg-globin pseudogene; 2,040 nucleotides†	Arnason et al. (1994) [1] mitochondrial protein; 3,033 amino acids†
Human–rhesus monkey	25.3 ± 2.6	Not available
Human–gibbon	Not available	56.3 ± 4.1
Human–orangutan	11.9 ± 1.7	24.4 ± 2.7
Human–gorilla	5.8 ± 1.2	18.0 ± 2.9
Human–chimpanzee	4.9 ± 1.2	13.7 ± 2.5

*In million years ± standard error.

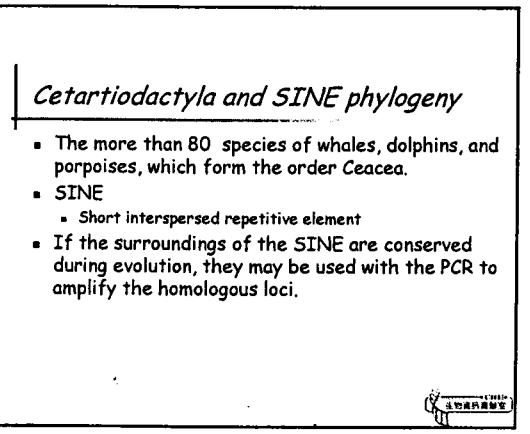
†Calibrated from the Catarrhini–Platyrrhini divergence = 38 million years ago.

‡Calibrated from the bovine–cetacean divergence = 60 million years ago.



Phylogeny of humans and apes

- Estimated the time of divergence between humans and chimpanzees to be almost three times as large.
- There are several possible reasons for the large difference between the two estimates.
  - The assumption of a constant rate (a molecular clock) my not hold.
  - The reference dates for calibration may not be accurate.
  - Each estimate is subject to stochastic errors.
- This example shows that divergence date estimates should be taken with extreme caution.



Cetartiodactyla and SINE phylogeny

- The more than 80 species of whales, dolphins, and porpoises, which form the order Ceacea.
- SINE
  - Short interspersed repetitive element
- If the surroundings of the SINE are conserved during evolution, they may be used with the PCR to amplify the homologous loci.

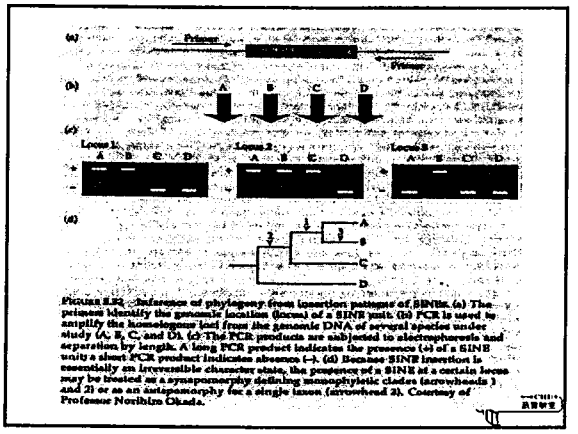
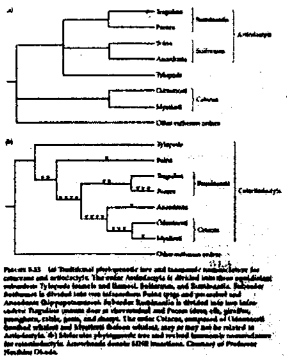


FIGURE 8.32 Inference of phylogeny from insertion patterns of SINEs. (a) The primers identify the genomic location (locus) of a SINE unit. (b) PCR is used to amplify the homologous loci from the genomic DNA of several species under study (A, B, C, and D). (c) The PCR products are subjected to electrophoresis and separation by length. A long PCR product indicates the presence (+) of a SINE unit; a short PCR product indicates absence (–). (d) Because SINE insertion is essentially an irreversible character state, the presence of a SINE at a certain locus may be treated as a synapomorphy defining monophyletic clades (arrowheads 1 and 2) or as an autapomorphy for a single taxon (arrowhead 3). Courtesy of Professor Norihiro Okada.



- SINE indicate that hippopotamuses (Ancodonta) are the closest extant relatives of whales.
- Incidentally, pigs and peccaries (Suina) were found to be unrelated to hippopotamuses.
- The molucular evidence therefore invalidates the monophyly of the suborder Suiformes.

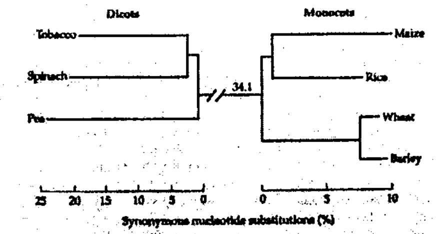


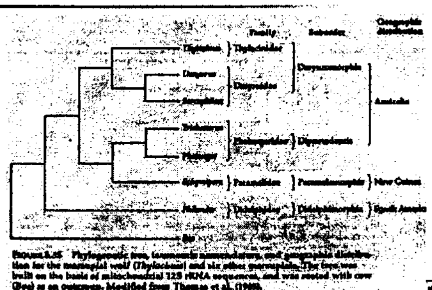
FIGURE 5.34 Phylogenetic tree for three dicot and four monocot species. The tree was inferred by the neighbor-joining method using the synonymous distances for three chloroplast genes: *rbcL*, *atpB*, and *atpE*. The length (0.7%) of the internal branch leading to the maize-rice pair is less than the standard error (~1.7%), and thus the maize, rice, and wheat/barley lineages are probably close to a trichotomy. Note that for purposes of clarity, the two scales are different from each other. From Wolfe et al. (1989b).

Molecular Phylogenetic Archeology

Sample	Maximum age (years)
Dry remains	
Museum skins	140
Naturally preserved skins	13,000
Human mummies	7,500
Bones and teeth	30,000
Natural animal mummies	13,000
Feathers	130
Hair	1,300
Herbarium plant specimens	120
Charred seeds and cobs	4,500
Mummified seeds	45,000
Frozen remains	
Muscle tissue	53,000
Wet remains	
Pickled museum specimens	100
Human remains preserved in peat	8,000

Data from Pääbo et al. (1989), Hanni et al. (1990), Lawlor et al. (1991), Brown and Brown (1994), Lin et al. (1995), and Krings et al. (1997).

Phylogeny of the marsupial wolf



Is the quagga extinct?

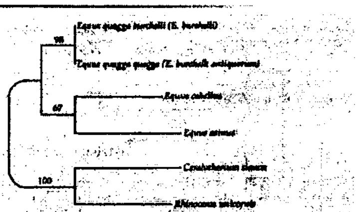
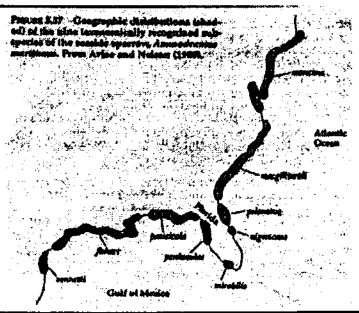
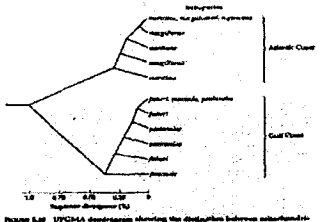


FIGURE 5.36: Scaled neighbor-joining phylogenetic tree for four *Equus* species (quagga, *E. quagga quagga*; plains zebra, *E. q. burchelli*; horse, *E. caballus*; and donkey, *E. asinus*). The tree has been rooted with two non-equid perissodactyl species (Asian one-horned rhinoceros, *Rhinoceros unicornis*; and white rhinoceros, *Ceratotherium simum*). The tree is based on a 229-bp concatenated mitochondrial sequence. The numbers on the internal branches represent bootstrap values based on 1,000 pseudosamples. Minor taxonomic synonyms are shown in parentheses.

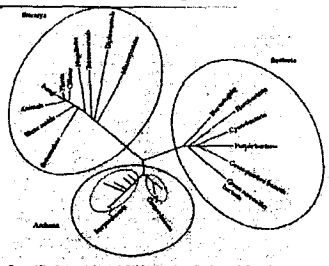
The dusky seaside sparrow: A lesson in conservation biology



The Universal Phylogeny



The first divergence events



The first divergence events

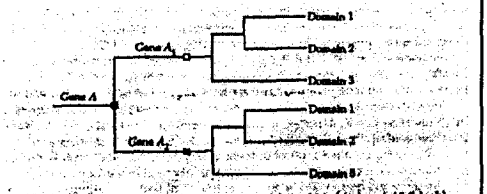


FIGURE 5.40 Duplication of gene *A* (gray square) into *A1* (white) and *A2* (black) prior to the divergence of three domains, will result in two identical topologies for the two subtrees. Modified from Li (1997).

The first divergence events

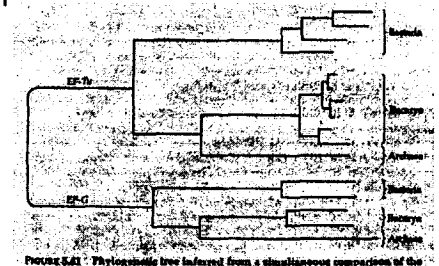


FIGURE 5.41 Phylogenetic tree inferred from a simultaneous comparison of the duplicated elongation factor genes, *EF-Tu* and *EF-G*, from Archaea, Bacteria, and Eucarya. Modified from Iwabe et al. (1989).

The first divergence events

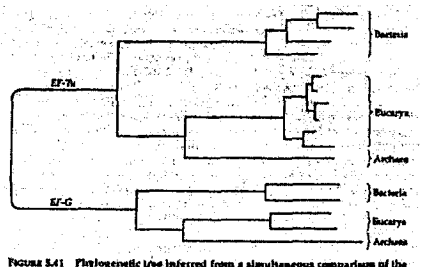


FIGURE 5.41 Phylogenetic tree inferred from a simultaneous comparison of the duplicated elongation factor genes, *EF-Tu* and *EF-G*, from Archaea, Bacteria, and Eucarya. Modified from Iwabe et al. (1989).

The first divergence events

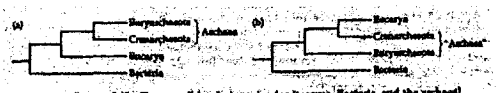


FIGURE 5.42 Two possible phylogenies for Eucarya, Bacteria, and the archaeal kingdoms Crenarchaeota and Euryarchaeota. (a) The Archaea is monophyletic. (b) The Eucarya arose from within the Archaea, which is therefore paraphyletic (indicated by the use of quotation marks). This tree is sometimes referred to as the Eocyte tree.

The first divergence events

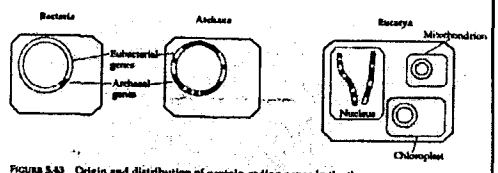
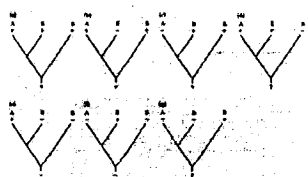


FIGURE 5.43 Origin and distribution of protein-coding genes in the three domains. Bacteria contain few archaeal genes (black segments), e.g., ATPase A in *Thermus* and *Enterococcus*. Archaea contain a large number of eubacterial genes (white segments), in particular genes involved in biosynthesis. The nuclear genome of Eucarya contains many archaeal genes, as well as several eubacterial genes derived from either the chimerical archaeal ancestor or from the organelles through horizontal gene transfer. The mitochondrial and chloroplast genomes are of exclusively eubacterial origin. Modified from Olendzenski et al. (1998).

The cenancestor



Endosymbiotic origin of mitochondria and chloroplasts

1. Histoneless DNA
2. 120,000–150,000 base pairs in size
3. Circular genome
4. Sensitivity of transcription to rifampicin
5. Inhibition of ribosomes by streptomycin, chloramphenicol, spectromycin, and paromomycin
6. Insensitivity of translation to cycloheximide
7. Translation starts with formylmethionine
8. Polyadenylation of mRNA absent or very short
9. Prokaryotic promoter structure

Endosymbiotic origin of mitochondria and chloroplasts

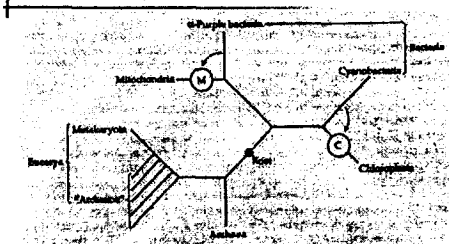
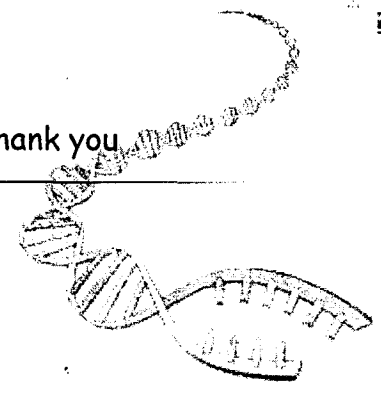


FIGURE 5.45 Schematic unrooted tree illustrating the phylogenetic affinities of chloroplasts and mitochondria. The mitochondria are derived from the α-purple bacteria via an endosymbiotic event (M). The chloroplasts are derived from the cyanobacteria via a second endosymbiotic event (C).

Thank you



蛋白質的結構與功能之分析教材內容

BIOLOGY

CONCEPTS & CONNECTIONS

Neil A. Campbell • Jane B. Reece • Lawrence G. Mitchell • Martha R. Taylor

Topic 1

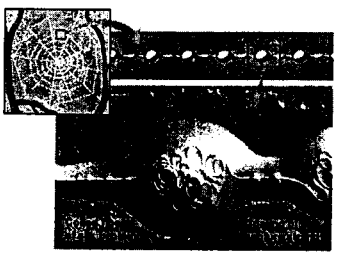
Fundamentals of protein structure

From PowerPoint® Lectures for *Biology: Concepts & Connections*

Spider Silk: Stronger than Steel

- Life's diversity results from the variety of molecules in cells
- A spider's web-building skill depends on its DNA molecules
- DNA also determines the structure of silk proteins
  - These make a spiderweb strong and resilient

- The capture strand contains a single coiled silk fiber coated with a sticky fluid
- The coiled fiber unwinds to capture prey and then recoils rapidly



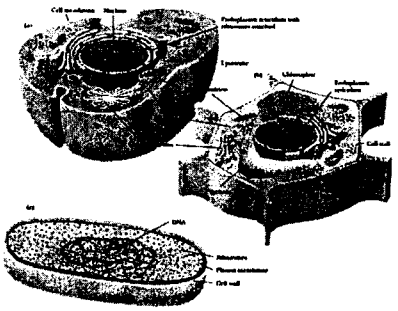
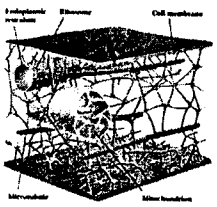
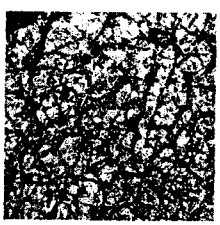


FIGURE 1.16
A comparison of (a) a typical animal cell, (b) a typical plant cell, and (c) a prokaryotic cell.





(a) (b)

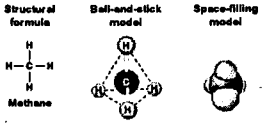
FIGURE 1.16
The microtrabecular lattice (a) This network of filaments, also called the cytoskeleton, provides the cytosol. Some filaments, called microtubules, are known to consist of the protein tubulin. Organelles such as mitochondria are attached to the filaments. (b) An electron micrograph of the microtrabecular lattice (magnified 87,150×).

INTRODUCTION TO ORGANIC COMPOUNDS AND THEIR POLYMERS

- Life's structural and functional diversity results from a great variety of molecules
- A relatively small number of structural patterns underlies life's molecular diversity

3.1 Life's molecular diversity is based on the properties of carbon

- A carbon atom forms four covalent bonds
  - It can join with other carbon atoms to make chains or rings



The 4 single bonds of carbon point to the corners of a tetrahedron.

Figure 3.1, top part

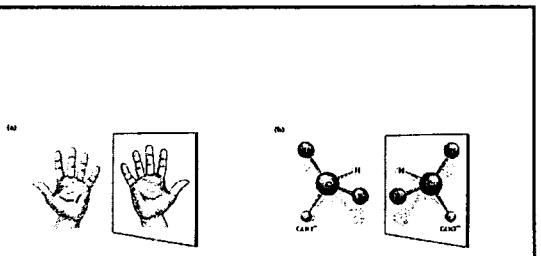


FIGURE 3.2
(a) Left and right hands are mirror images of each other. (b) A pair of mirror-image amino acids. Amino acids found in proteins have the configuration shown on the left side of the mirror.

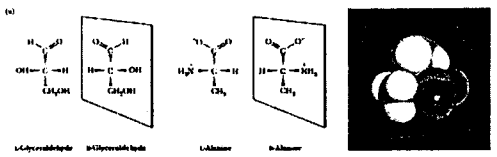
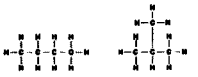


FIGURE 3.3
Stereochemistry of alanine and glyceraldehyde. The amino acids found in proteins have the same chirality as L-glyceraldehyde, which is opposite that of D-glyceraldehyde. (b) Space-filling model of L-alanine.

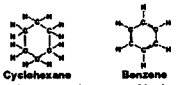
- Carbon skeletons vary in many ways

Carbon skeletons vary in length.



Skeletons may be unbranched or branched.

Skeletons may have double bonds, which can vary in location.



Skeletons may be arranged in rings.

Figure 3.1, bottom part

3.2 Functional groups help determine the properties of organic compounds

- Functional groups are the groups of atoms that participate in chemical reactions
  - Hydroxyl groups are characteristic of alcohols
  - The carboxyl group acts as an acid

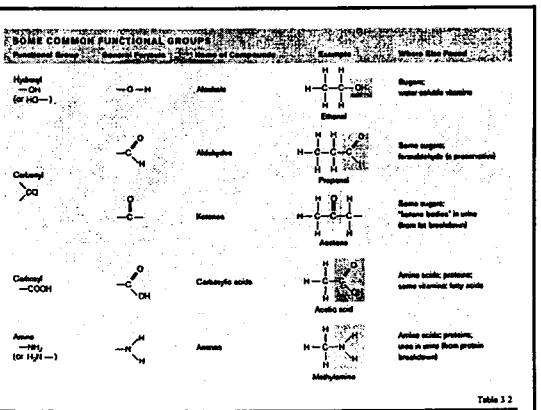


Table 3.2

3.3 Cells make a huge number of large molecules from a small set of small molecules

- Most of the large molecules in living things are macromolecules called polymers
  - Polymers are long chains of smaller molecular units called monomers
  - A huge number of different polymers can be made from a small number of monomers

- Cells link monomers to form polymers by dehydration synthesis

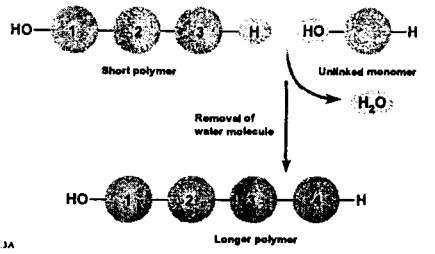


Figure 3.3A

- Polymers are broken down to monomers by the reverse process, hydrolysis

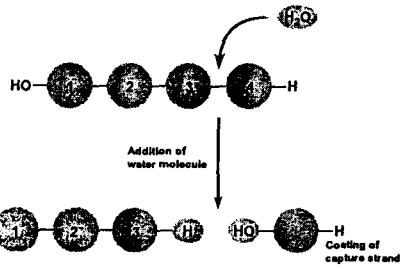


Figure 3.3B

PROTEINS

3.11 Proteins are essential to the structures and activities of life

- Proteins are involved in
  - cellular structure
  - movement
  - defense
  - transport
  - communication

Figure 3.11

- Mammalian hair is composed of structural proteins
- Enzymes regulate chemical reactions

3.12 Proteins are made from just 20 kinds of amino acids

- Proteins are the most structurally and functionally diverse of life's molecules
  - Their diversity is based on different arrangements of amino acids

- Each amino acid contains:
  - an amino group
  - a carboxyl group
  - an R group, which distinguishes each of the 20 different amino acids

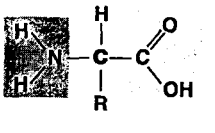


Figure 3.12A

Amino group

Carboxyl (acid) group

- Each amino acid has specific properties

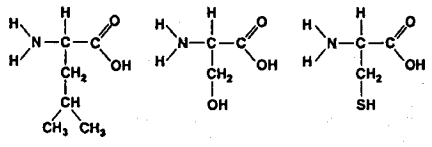
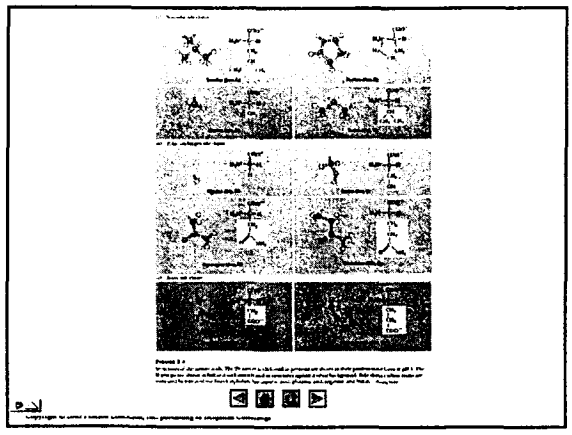


Figure 3.12B



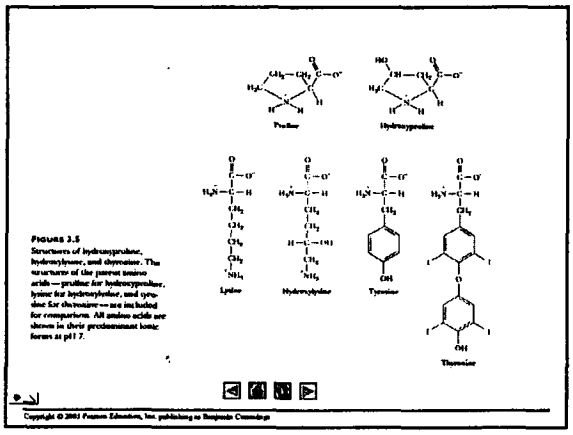


FIGURE 3.5
Structures of hydroxyproline, hydroxylysine, and thyroxine. The structures of the parent amino acids—proline for hydroxyproline, lysine for hydroxylysine, and tyrosine for thyroxine—are included for comparison. All amino acids are shown in their predominant ionic forms at pH 7.

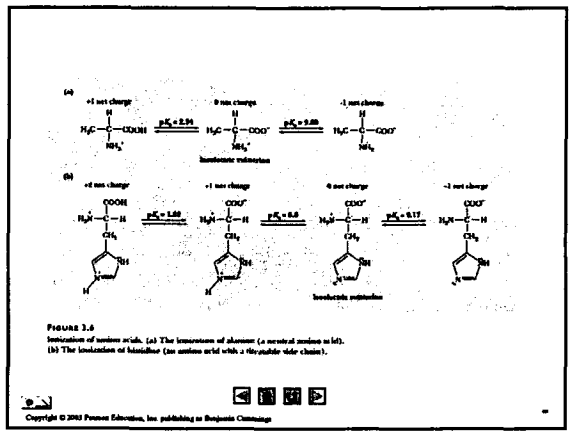


FIGURE 3.6
Ionization of amino acids. (a) The ionization of alanine (a neutral amino acid). (b) The ionization of histidine (an amino acid with a titratable side chain).

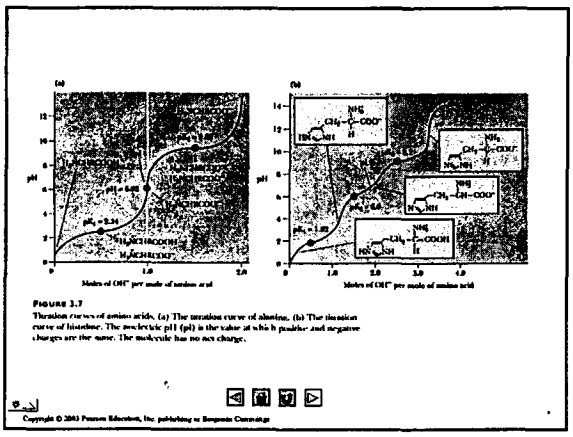


FIGURE 3.7
Titration curves of amino acids. (a) The titration curve of alanine. (b) The titration curve of histidine. The isoelectric pH (pI) is the value at which positive and negative charges are the same. The molecule has no net charge.

3.13 Amino acids can be linked by peptide bonds

- Cells link amino acids together by dehydration synthesis
- The bonds between amino acid monomers are called peptide bonds

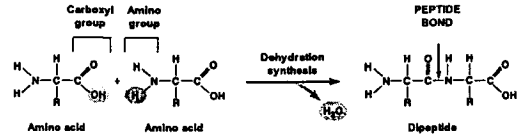
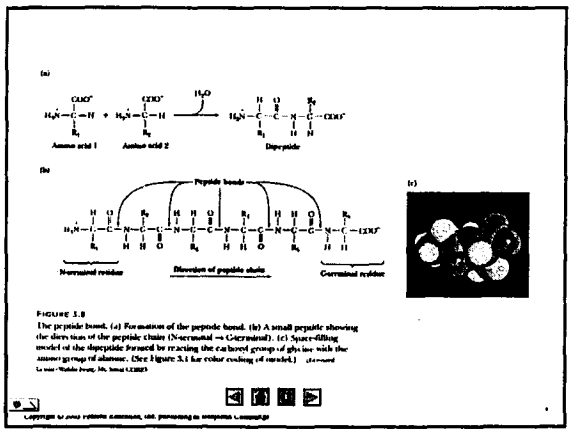
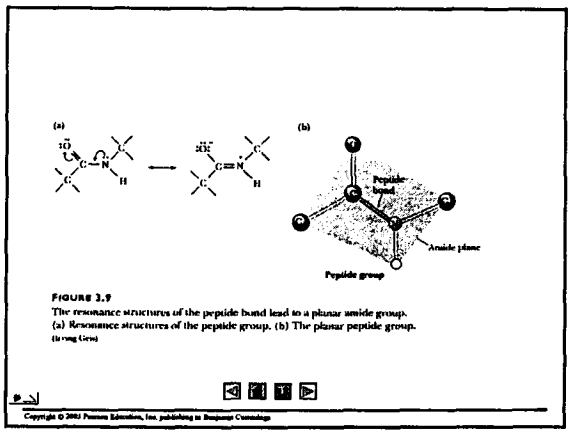
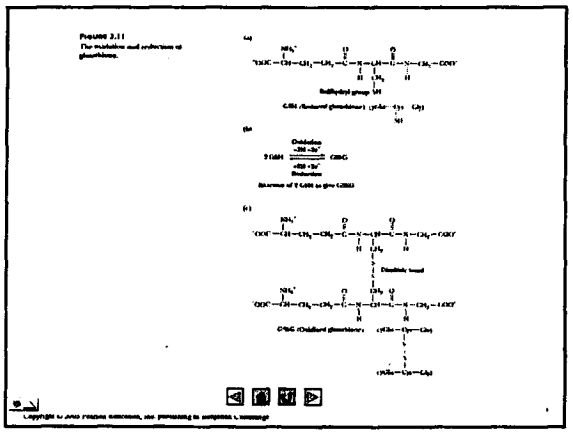
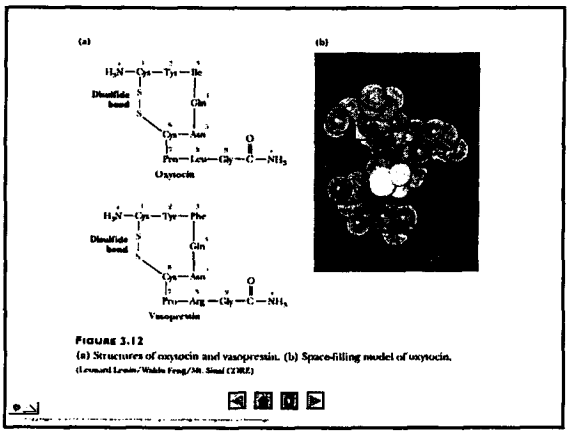


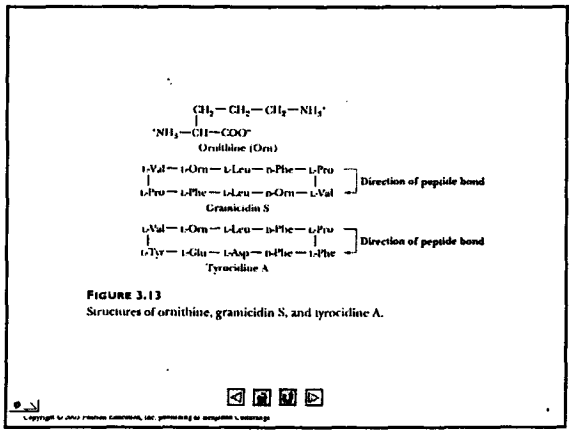
Figure 3.13











3.14 Overview: A protein's specific shape determines its function

- A protein, such as lysozyme, consists of polypeptide chains folded into a unique shape
  - The shape determines the protein's function
  - A protein loses its specific function when its polypeptides unravel

3.15 A protein's primary structure is its amino acid sequence

3.16 Secondary structure is polypeptide coiling or folding produced by hydrogen bonding

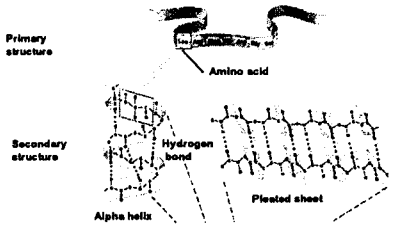


Figure 3.15, 16

3.17 Tertiary structure is the overall shape of a polypeptide

3.18 Quaternary structure is the relationship among multiple polypeptides of a protein

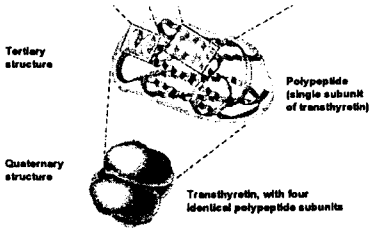


Figure 3.17, 18

Topic 2 Protein tertiary and quaternary structure

"Biochemistry" 3rd Edition, 2001.
Mary K. Campbell 原著，林順富、陳師瑩、顏瑞鴻、蕭慧美編譯，偉明圖書出版.

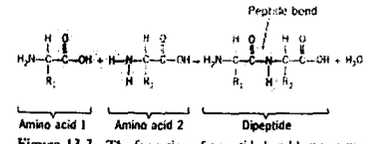
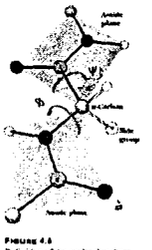


Figure 13.2 The formation of a peptide bond between two amino acids by the removal of water. Each peptide bond connects the amino group of one amino acid and the carboxyl group of the adjacent amino acid.



Secondary structure of proteins

- α- helix
- β-sheet
- Turns
- Combination of three
- Intramolecular interactions

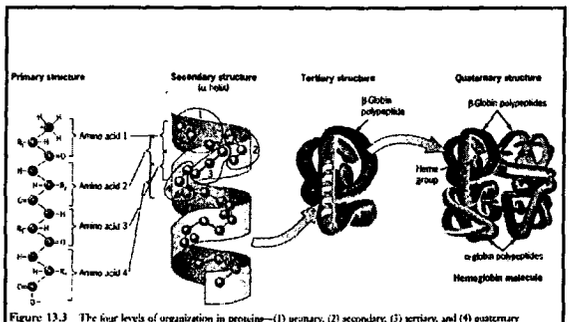
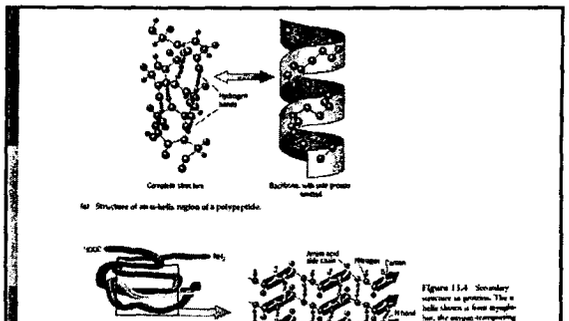


Figure 13.3 The four levels of organization in proteins—(1) primary, (2) secondary, (3) tertiary, and (4) quaternary structures—are illustrated using human hemoglobin as an example.



3.15 A protein's primary structure is its amino acid sequence

3.16 Secondary structure is polypeptide coiling or folding produced by hydrogen bonding

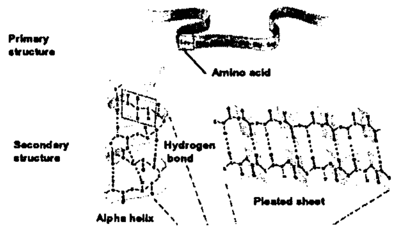
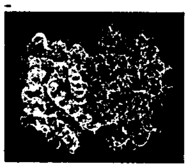


Figure 3.15, 16



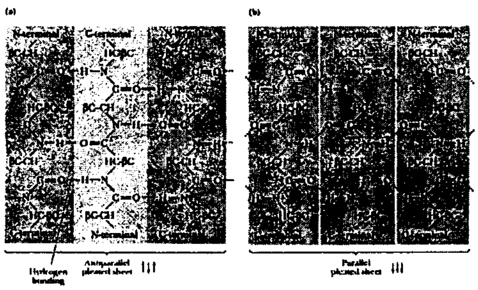


FIGURE 4.8
β-pleated sheet structures. (a) In an antiparallel pleated sheet, the peptide chains run in opposite directions from the N-terminal to the C-terminal ends. (b) In a parallel pleated sheet, the peptide chains run in the same direction.

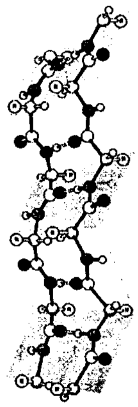


FIGURE 4.9
The three-dimensional form of the antiparallel β-pleated sheet arrangement. The chains do not fold back on each other but are in a fully extended conformation.

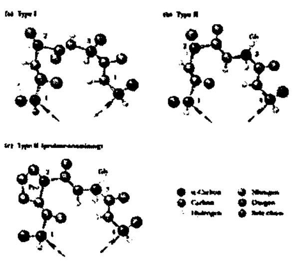
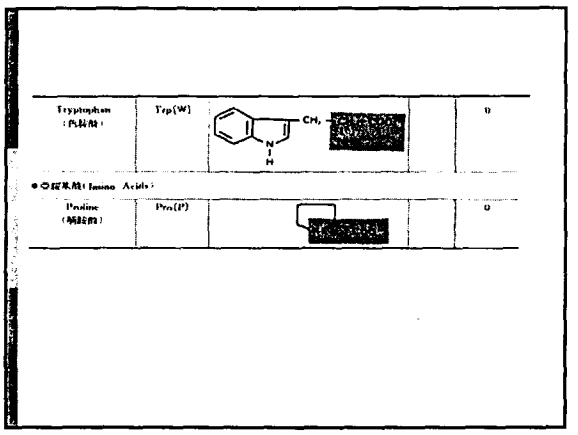


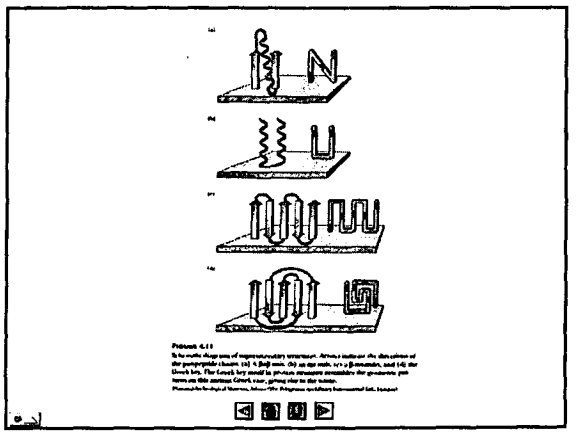
FIGURE 4.10
Structures of reverse turns. Arrows indicate the directions of the polypeptide chains. (a) A type I reverse turn. In residue 3 the side chain (green) lies outside the loop, and any amino acid can occupy this position. (b) A type II reverse turn. The side chain of residue 3 has been rotated 180 degrees from the position in the type I turn and is now on the inside of the loop. Only the hydrogen side chain of glycine can fit into the space available, so glycine must be the third residue in a type II reverse turn. (c) The five-membered ring of proline has the correct geometry for a reverse turn; this residue normally occurs as the second residue of a reverse turn. The turn shown here is type II, with glycine as the third residue.

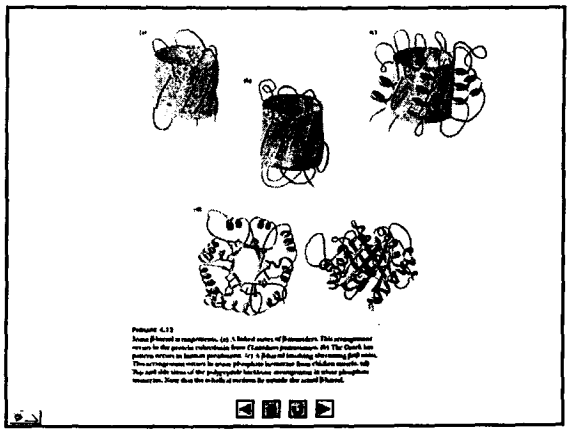
俗 名	符 號	結 構 式	極 性	非 極 性
● 支鏈烴基結構 (Aliphatic side chain)				
Glycine (甘胺酸)	Gly(G)	H-		0
Alanine (丙胺酸)	Ala(A)	CH_3-		0
Valine (纈胺酸)	Val(V)	$(H_3C)_2CH$-		0
Leucine (白胺酸)	Leu(L)	$(H_3C)_2CH-CH_2$-		0
Isoleucine (異白胺酸)	Ile(I)	$CH_3-CH_2-CH(CH_3)$-		0

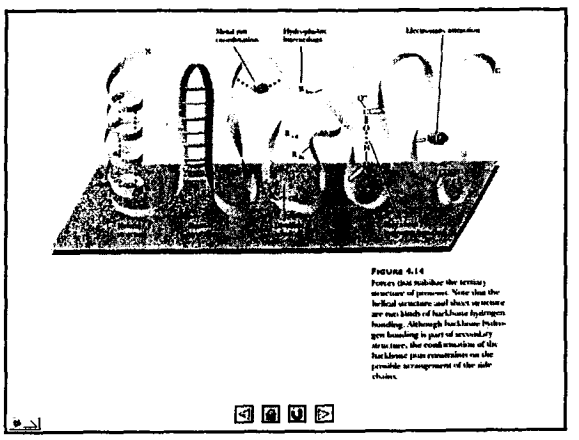


Proline is unique

- Proline has the amino nitrogen cyclized with the side chain terminal carbon which leads to many of the unique properties of proline.
- restricts the conformational space available to the peptide
- reduced barrier to cis-trans isomers of the peptide bond
- loss of H bonding capability of the immino nitrogen.
- Proline is a component of collagen, a major structural component of cells and animals.
- SH3(Src homology region 3) and WW domains of signal transduction pathway proteins recognize proline containing sequences and are used to mediate protein-protein interactions of signal transduction components.
- Proline specific peptidases have been found that are sensitive to the conformation of proline.
- Cis-trans isomerization of proline is the rate limiting step in folding of some proteins and many cyclophilins or racemases catalyze the cis-trans racemizaton of proline containing proteins.







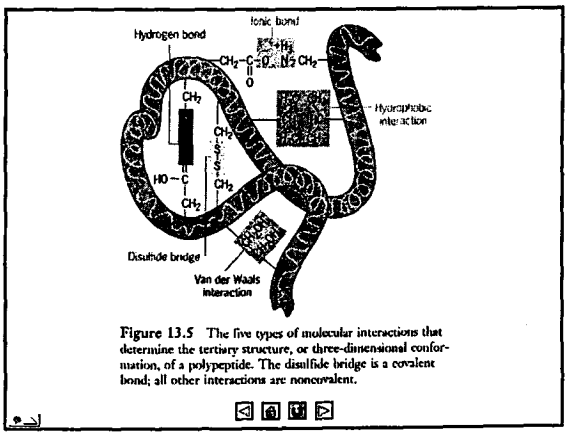
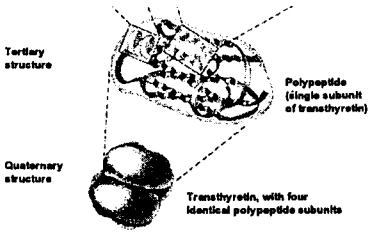


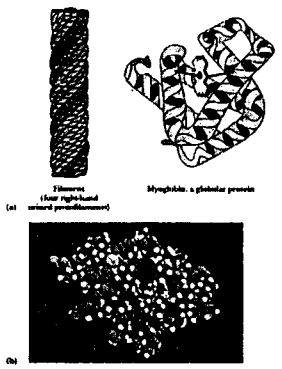
Figure 13.5 The five types of molecular interactions that determine the tertiary structure, or three-dimensional conformation, of a polypeptide. The disulfide bridge is a covalent bond; all other interactions are noncovalent.

Tertiary structure to quaternary structure

3.17 Tertiary structure is the overall shape of a polypeptide

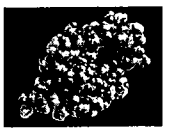
3.18 Quaternary structure is the relationship among multiple polypeptides of a protein





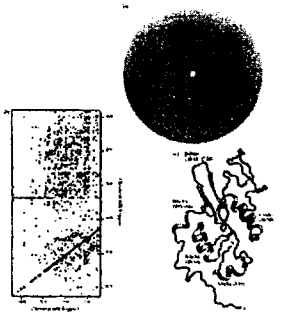
3.14 Overview: A protein's specific shape determines its function

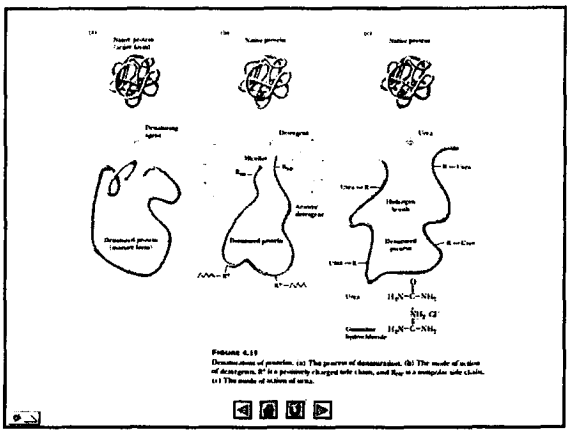
- ~~A protein, such as lysozyme, consists of~~ polypeptide chains folded into a unique shape
  - The shape determines the protein's function
  - A protein loses its specific function when its polypeptides unravel

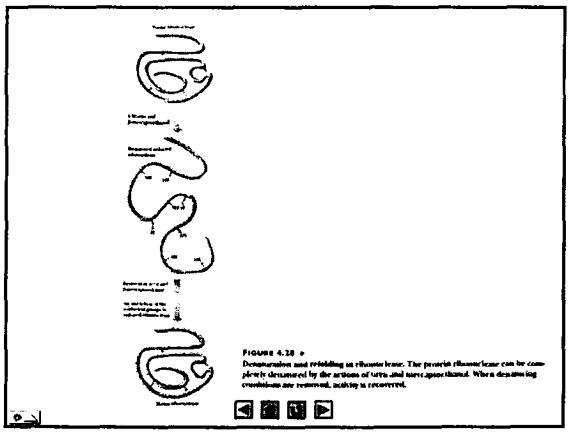



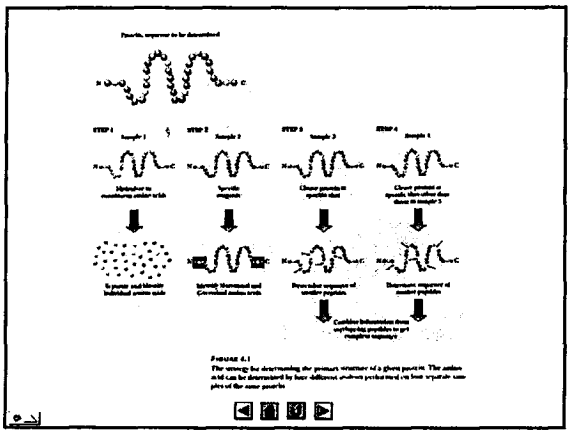
Protein structure analysis

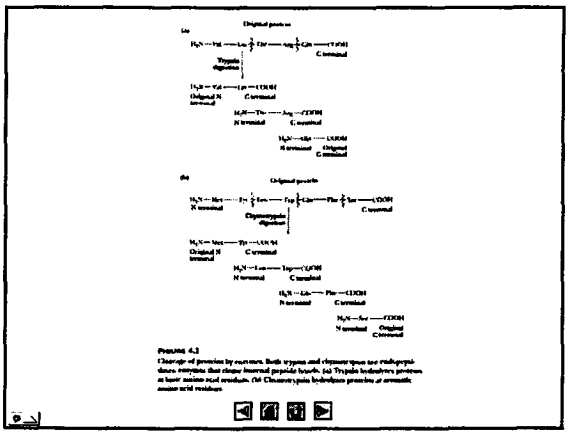
- Overall folding structure
- Amino acids sequencing

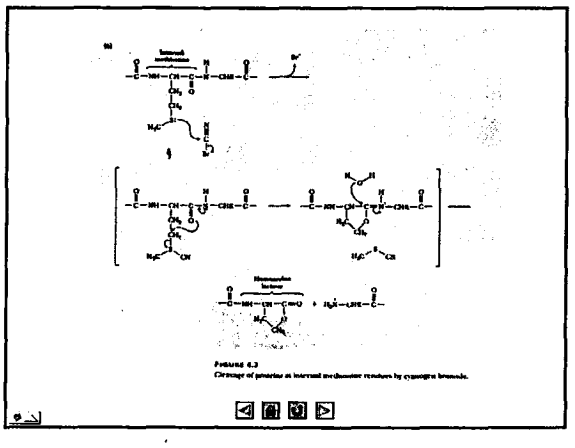


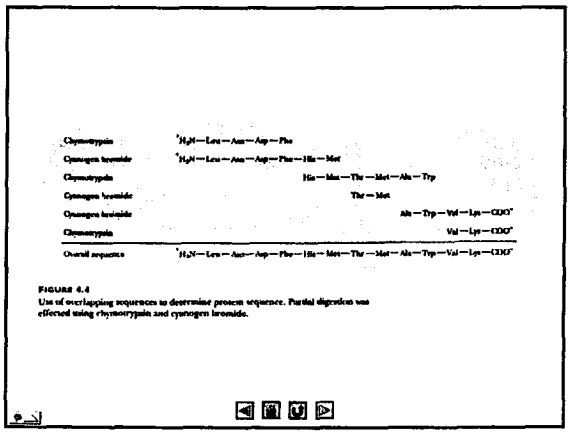






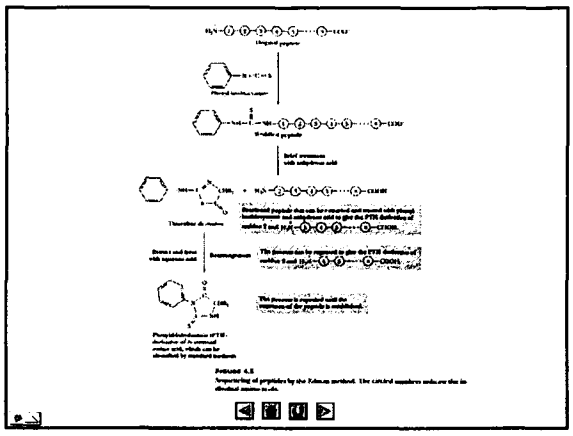






Chymotrypsin	$^{1}H_2N$—Leu—Asn—Asp—Phe
Cyanogen bromide	$^{1}H_2N$—Leu—Asn—Asp—Phe—His—Met
Chymotrypsin	His—Met—Thr—Met—Ala—Trp
Cyanogen bromide	Thr—Met
Cyanogen bromide	Ala—Trp—Val—Lys—COO^{-}
Chymotrypsin	Val—Lys—COO^{-}
Overall sequence	$^{1}H_2N$—Leu—Asn—Asp—Phe—His—Met—Thr—Met—Ala—Trp—Val—Lys—COO^{-}

FIGURE 4.4
Use of overlapping sequences to determine protein sequence. Partial digestion was effected using chymotrypsin and cyanogen bromide.



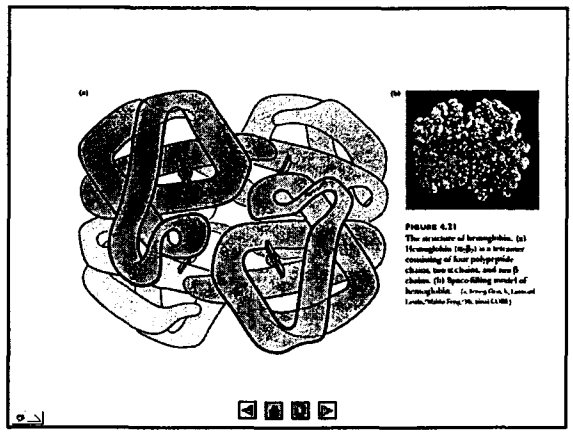
Examples of globular and fibrous proteins

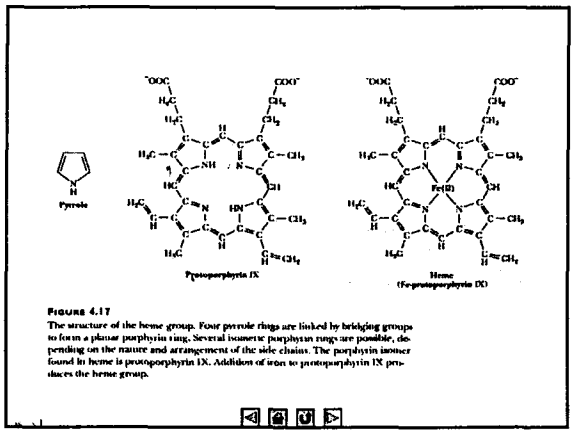
the globular protein

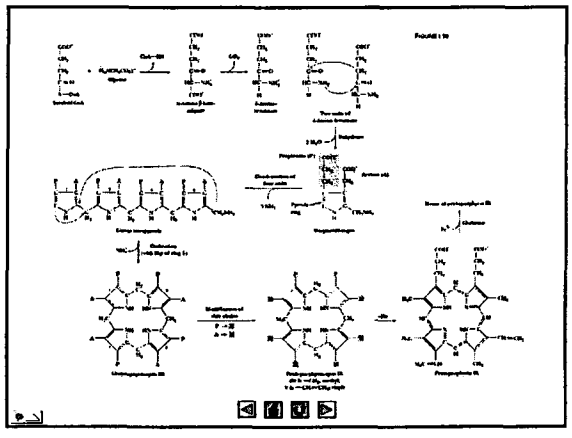
globular and fibrous proteins

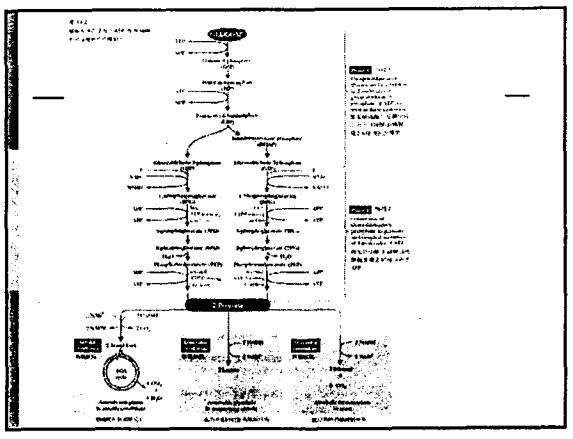
Examples of globular and fibrous proteins

- The globular protein- hemoglobulin, myoglobulin
- The fibrous protein-myofibril









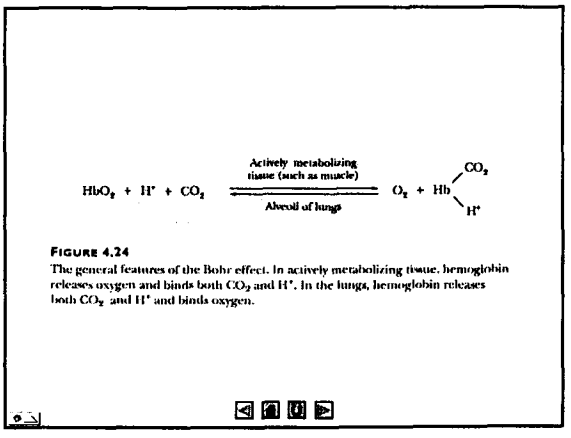


FIGURE 4.24
The general features of the Bohr effect. In actively metabolizing tissue, hemoglobin releases oxygen and binds both CO_2 and H^+. In the lungs, hemoglobin releases both CO_2 and H^+ and binds oxygen.

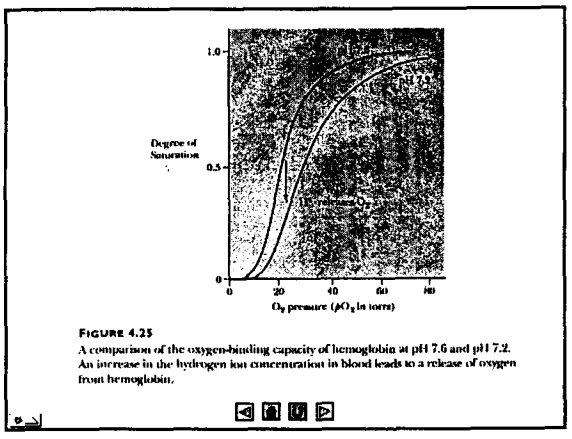
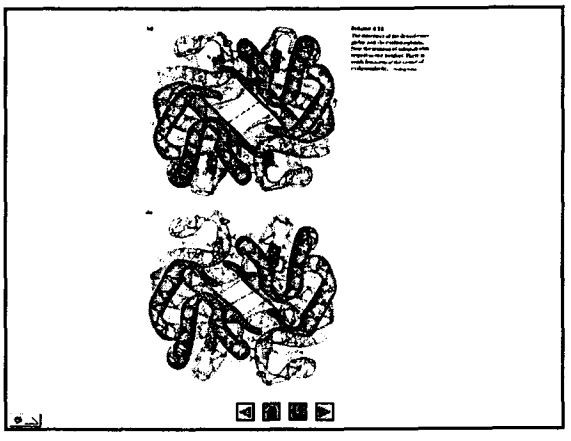


FIGURE 4.25
A comparison of the oxygen-binding capacity of hemoglobin at pH 7.6 and pH 7.2. An increase in the hydrogen ion concentration in blood leads to a release of oxygen from hemoglobin.



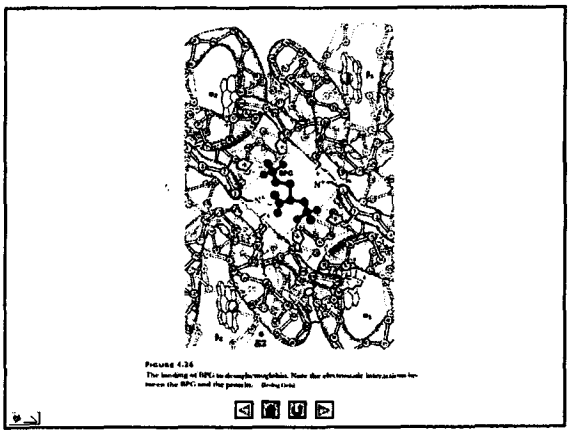


FIGURE 4.26
The binding of BPG to deoxyhemoglobin. Note the electrostatic interactions between the BPG and the protein.

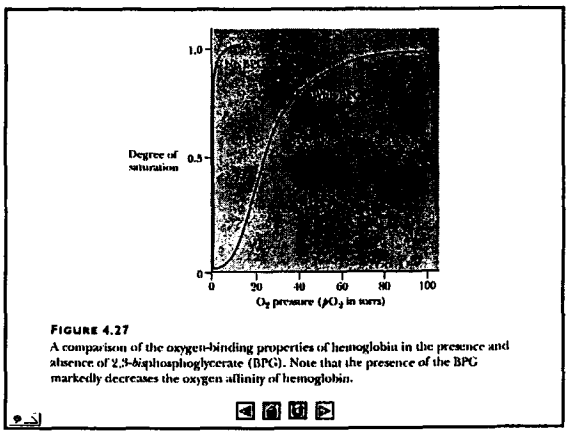


FIGURE 4.27
A comparison of the oxygen-binding properties of hemoglobin in the presence and absence of 2,3-*bis*phosphoglycerate (BPG). Note that the presence of the BPG markedly decreases the oxygen affinity of hemoglobin.

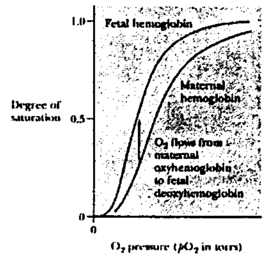


FIGURE 4.28
A comparison of the oxygen-binding capacity of fetal and maternal hemoglobins. Fetal hemoglobin binds less strongly to BPG and, consequently, has a greater affinity for oxygen, than does maternal hemoglobin.

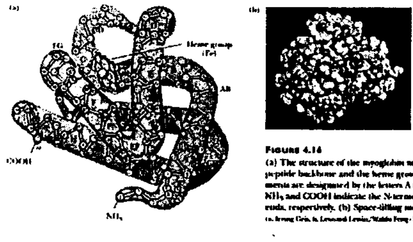


FIGURE 4.16
(a) The structure of the myoglobin molecule, showing the peptide backbone and the heme group. The helical segments are designated by the letters A through H. The terms NH_3 and COOH indicate the N-terminal and C-terminal ends, respectively. (b) Space-filling model of myoglobin.

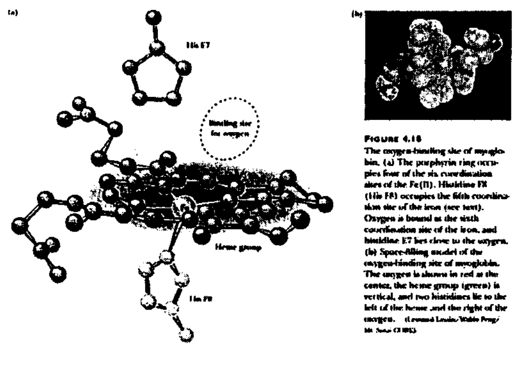


FIGURE 4.18
The oxygen-binding site of myoglobin. (a) The porphyrin ring occupies four of the six coordination sites of the Fe(II). Histidine F8 (His F8) occupies the fifth coordination site of the iron (see text). Oxygen is bound at the sixth coordination site of the iron, and histidine E7 lies close to the oxygen. (b) Space-filling model of the oxygen-binding site of myoglobin. The oxygen is shown in red at the center, the heme group (green) is vertical, and two histidines lie to the left of the heme and the right of the oxygen.

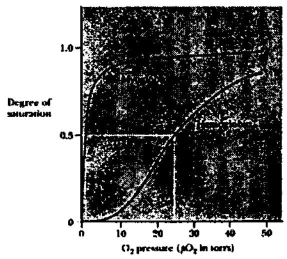


FIGURE 4.22
A comparison of the oxygen-binding behavior of myoglobin and hemoglobin. The oxygen-binding curve of myoglobin is hyperbolic, whereas that of hemoglobin is sigmoidal. Myoglobin is 50% saturated with oxygen at 1 torr partial pressure; hemoglobin does not reach 50% saturation until the partial pressure of oxygen reaches 26 torr.

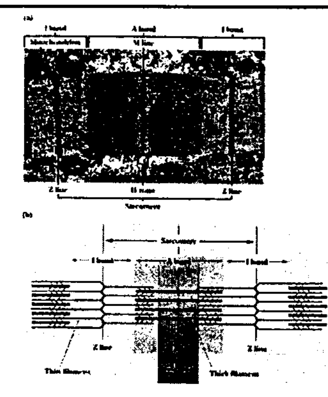


FIGURE 4.29
The arrangement of filaments in a myofibril. (a) Electron micrograph of a sarcomere. The overlapping thick and thin filaments give rise to the observed light and dark bands. (b) Schematic diagram of a sarcomere.

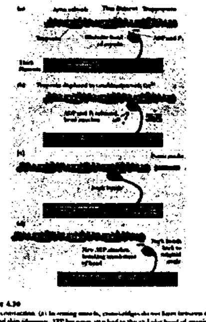


FIGURE 4.30

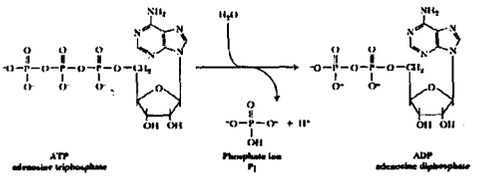
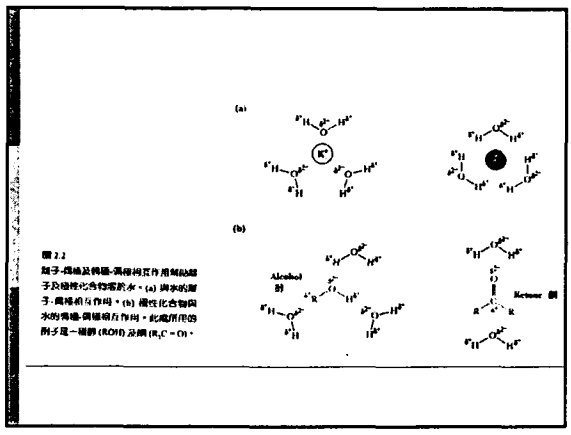


FIGURE 4.31
The hydrolysis of ATP to ADP and phosphate ion (P_i) releases energy which can be used in muscle contraction.

Topic 3: Protein stability and flexibility

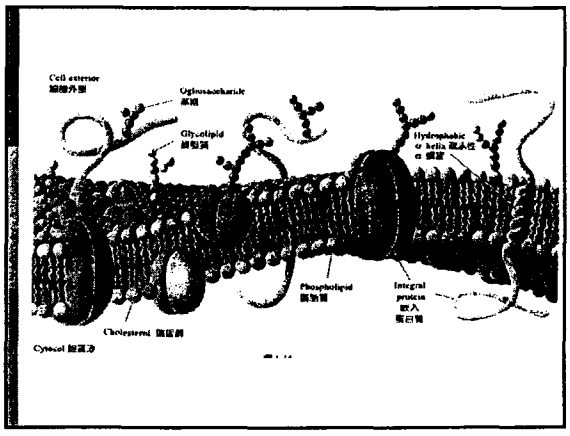
"Structure bioinformatics ", 2003. by Philip E. Bourne and Helge Weissig, by Wiley-Liss publishers, Inc. ，藝軒書局. (Chapter 2, 12, and 13)

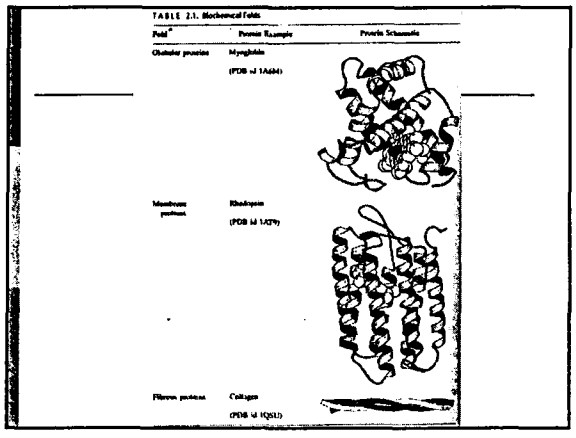


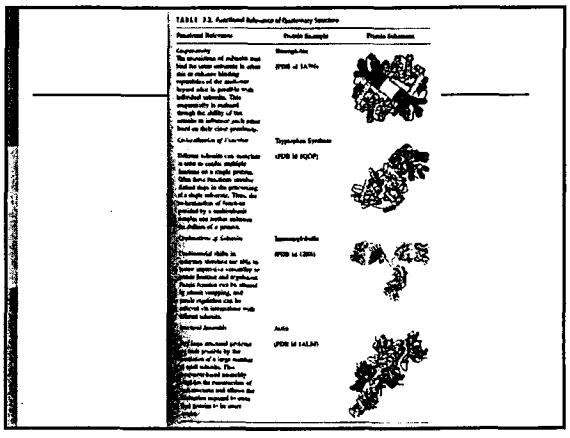
蛋白質中存在的一些氫鍵

SOME HYDROGEN BONDS FOUND IN PROTEINS

Amide–carbonyl 醯胺-羰	>N—H·····O=C<
Amide–hydroxyl 醯胺-羥	>N—H·····O—H
Amide–imidazole nitrogen 醯胺-咪唑氮	>N—H·····N NH
Hydroxyl–hydroxyl 羥-羥	—O—H·····O—H
Hydroxyl–carbonyl 羥-羰	—O—H·····O=C<

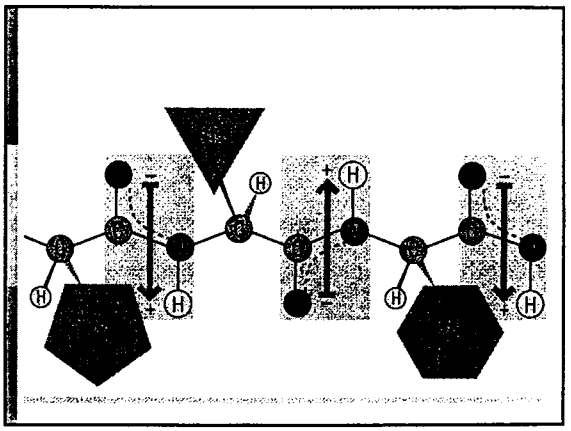


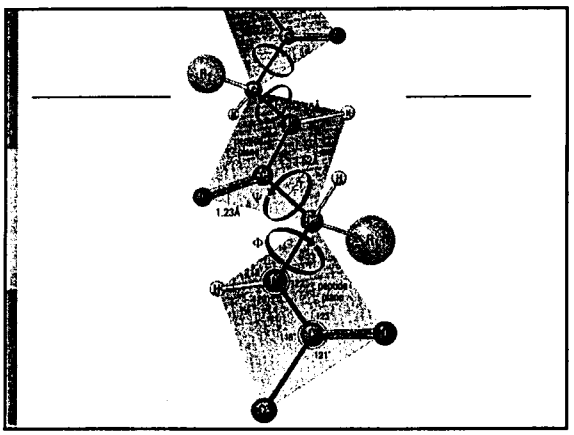


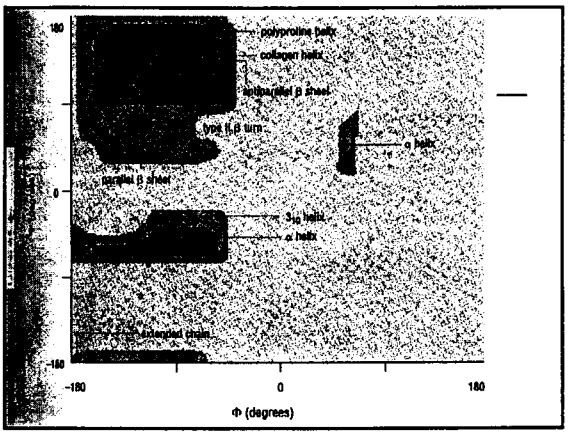


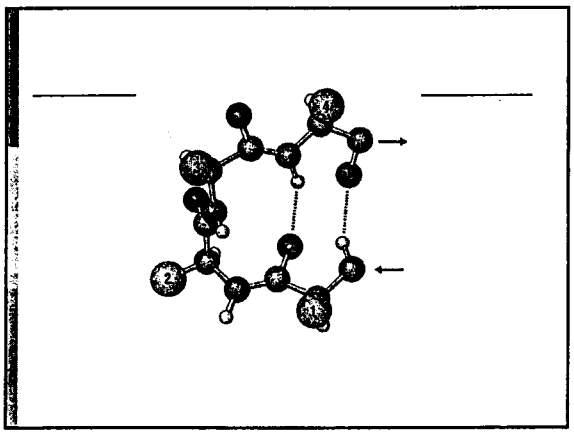
Topic 4: Protein function recognition

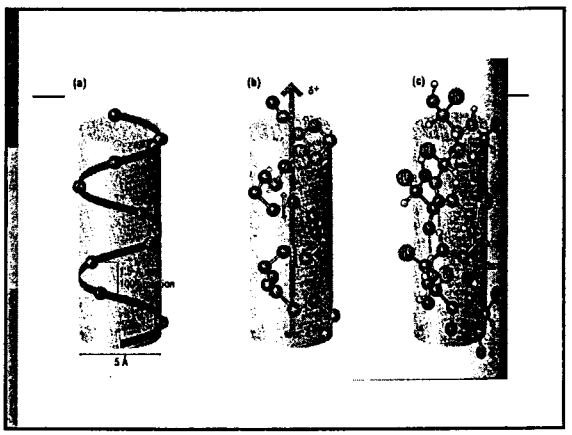
"Protein structure and function", 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局









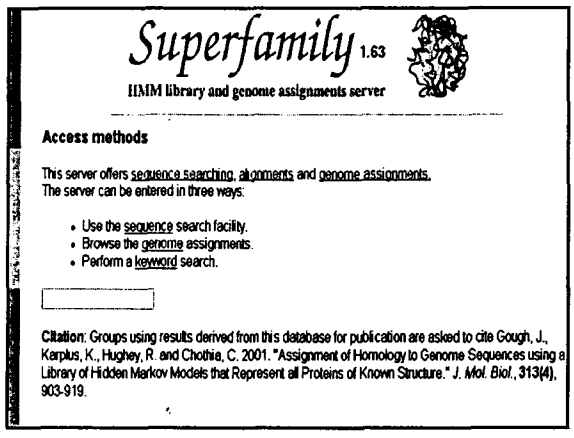


Protein can be classifies into four groups bases on their predominant secondary structural element:

- All alpha proteins [
- All beta proteins
- Alpha and beta proteins (a/b)
  Mainly parallel beta sheets (beta-alpha-beta units)
- Alpha and beta proteins (a+b)
  Mainly antiparallel beta sheets (segregated alpha and beta regions)
- Multi-domain proteins (alpha and beta)
  Folds consisting of two or more domains belonging to different classes
- Membrane and cell surface proteins and peptides
  Does not include proteins in the immune system
- Small proteins
  Usually dominated by metal ligand, heme, and/or disulfide bridges
- Coiled coil proteins
  Not a true class
- Low resolution protein structures
  Not a true class
- Peptides
  Peptides and fragments. Not a true class
- Designed proteins

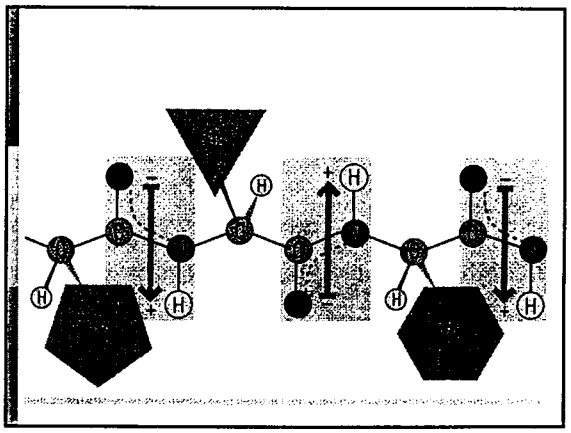
SCOP (structural classification of protein database from 1995) hierarchy

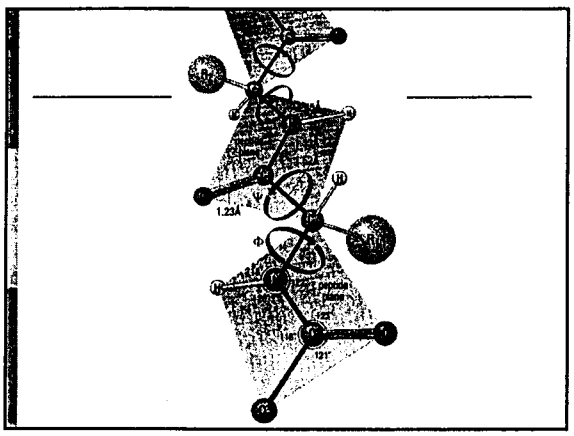
- The method used to construct the protein classification in SCOP is the visual inspection and comparison of structures first compared using automatic procedure.
  - The SCOP database is organized on a number of hierarchical levels, with the principle ones being family, superfamily, fold, and class.
  - Within this hierarchy, the unit of categorization is the protein domain since domains are typically the units of protein evolution, structure, and function.
  - Small- and medium-sized proteins usually have a single domain.
  - The domains in larger proteins are classified individually.
  - Thus, different regions of a single protein may appear in multiple places in the SCOP hierarchy under different folds or, in the case of repeated domains, several times under the same fold.

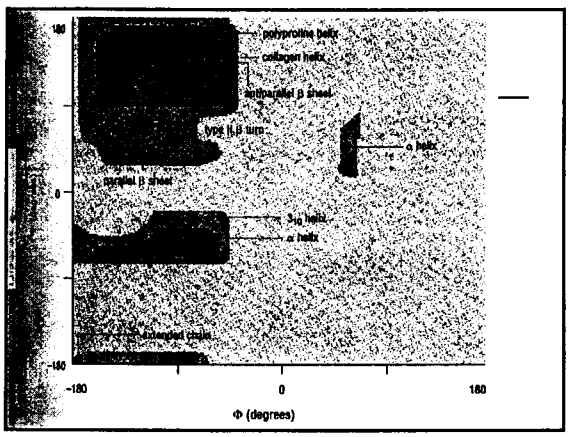


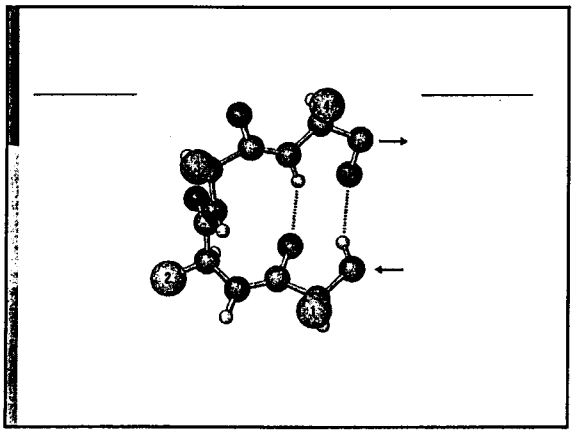
Topic 4: Protein function recognition

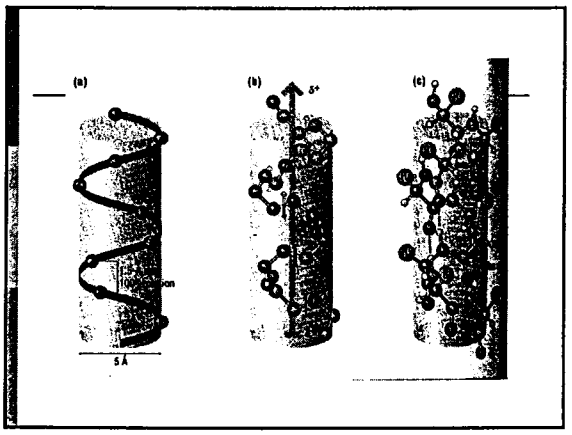
"Protein structure and function", 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局

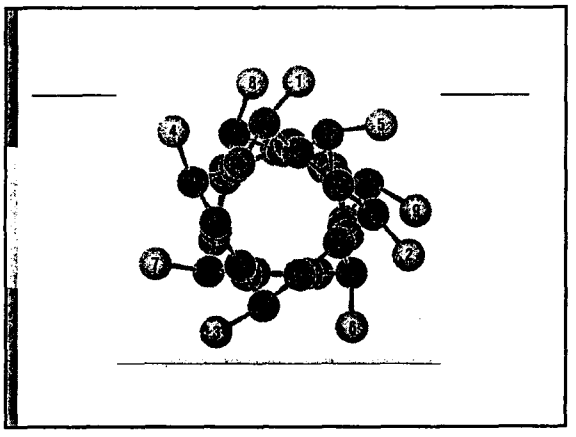






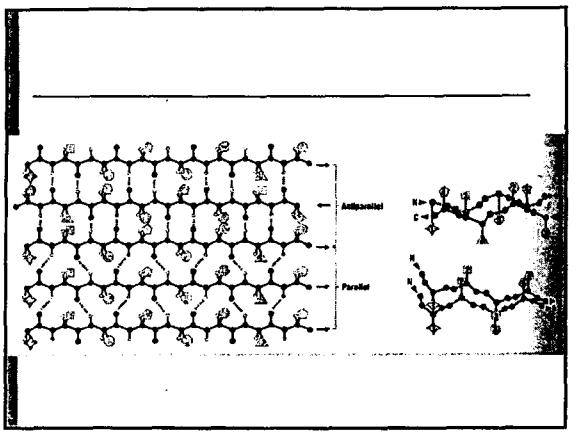


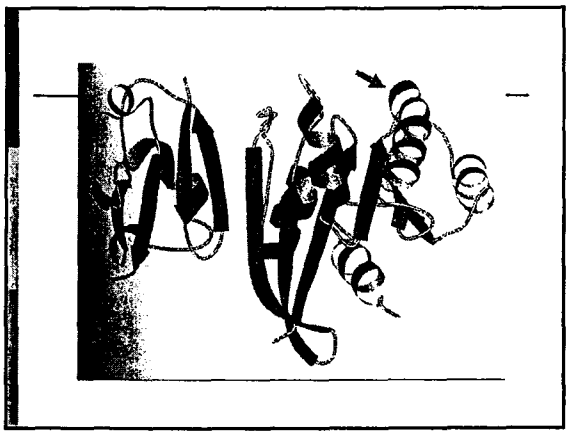


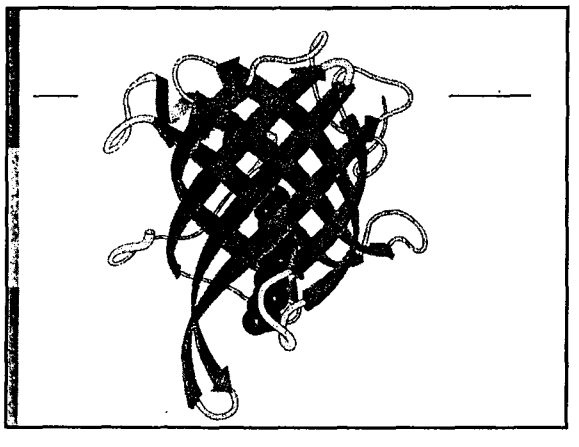


Typical Conformational Parameters of Helical Elements

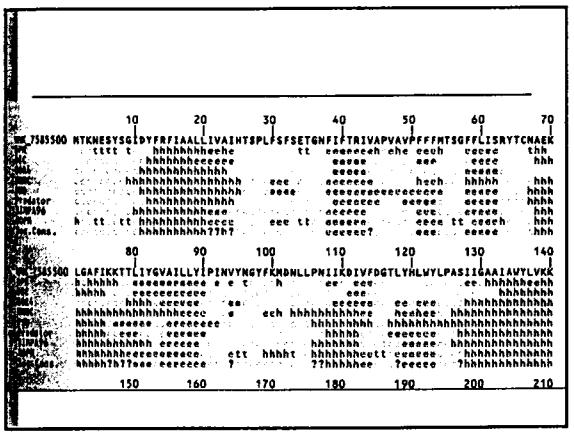
	Phi	Psi	Omega	Residues per turn	Translation per residue
Alpha helix	-57	-47	180	3.6	1.5
3-10 helix	-49	-26	180	3.0	2.0
Pi helix	-57	-70	180	4.4	1.15
Polyproline I	-83	+158	0	3.33	1.9
Polyproline II	-78	+149	180	3.0	3.12
Collagen III	-80	+150	180	3.0	3.1

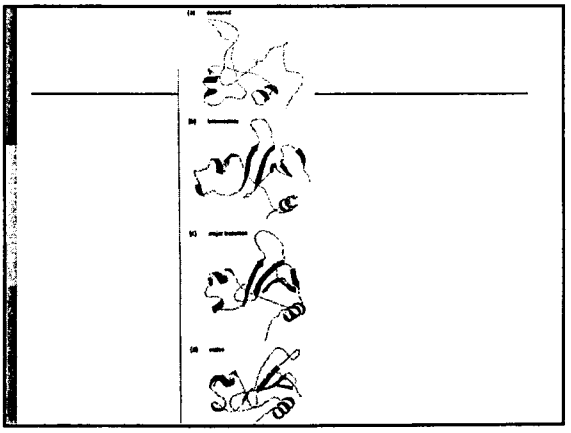


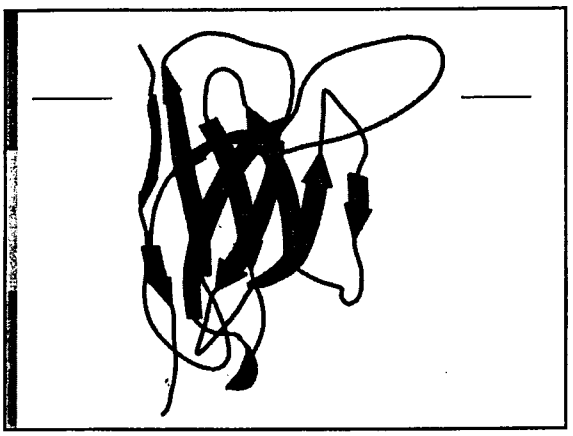




Glu	1.53	0.52	1.01
Ala	1.41	0.72	0.82
Leu	1.34	1.22	0.57
Met	1.30	1.14	0.52
Gln	1.27	0.96	0.84
Lys	1.23	0.69	1.07
Arg	1.21	0.84	0.90
His	1.09	0.80	0.81
Val	0.90	1.87	0.41
Ile	1.09	1.67	0.47
Tyr	0.74	1.48	0.76
Cys	0.66	1.40	0.54
Trp	1.02	1.35	0.65
Phe	1.16	1.33	0.59
Thr	0.76	1.17	0.96
Gly	0.43	0.58	1.77
Asn	0.76	0.48	1.34
Pro	0.34	0.31	1.32
Ser	0.57	0.96	1.22
Asp	0.99	0.39	1.24



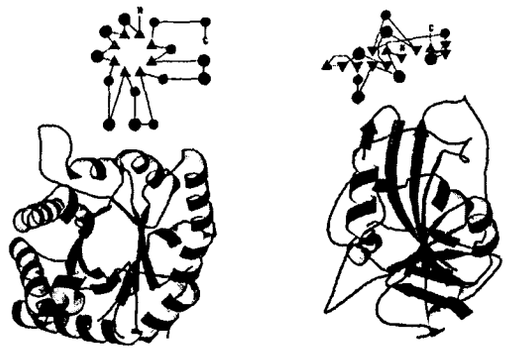




Topic 5 Catalysis function

"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

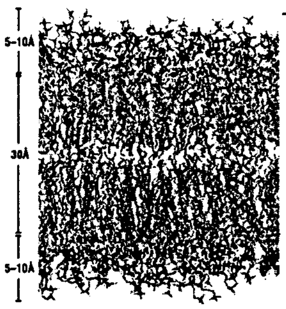
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

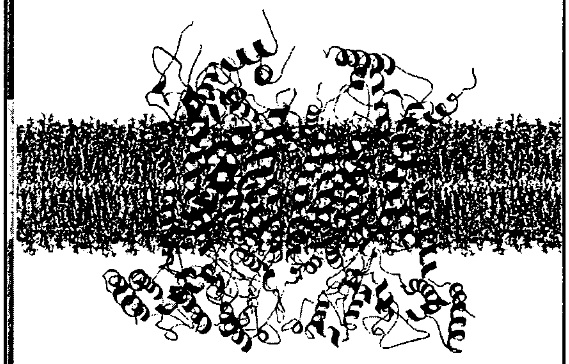
(a)

(b)

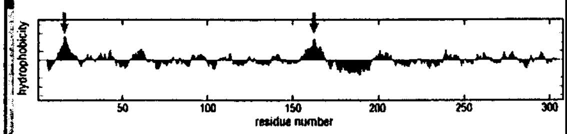
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



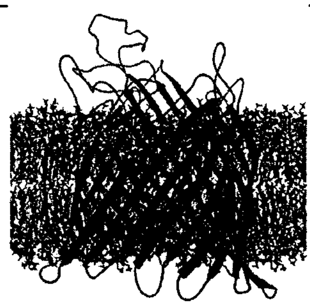
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



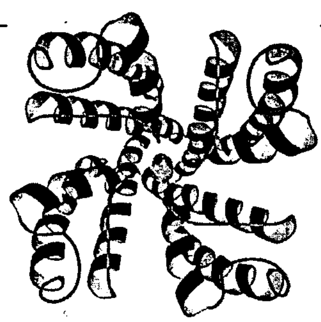
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



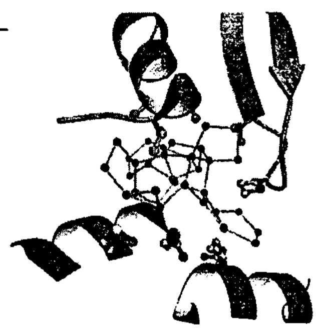
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



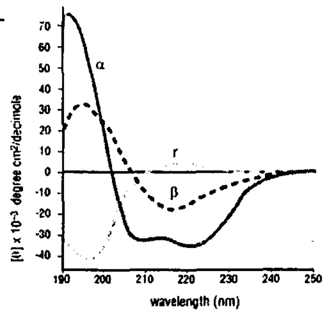
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



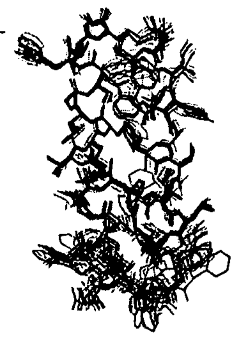
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



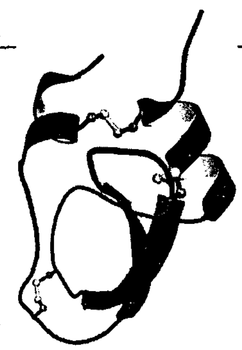
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



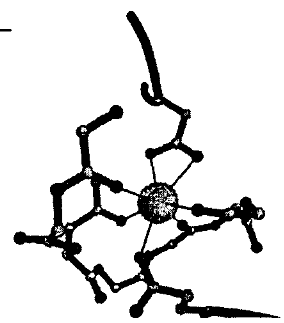
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



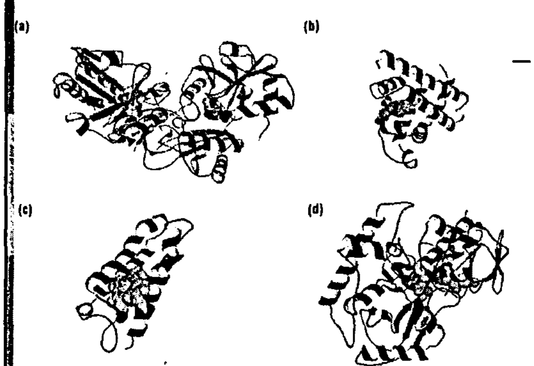
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

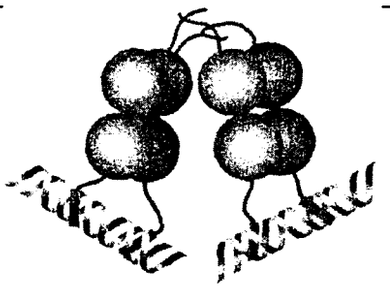


From **Protein Structure and Function**
by Gregory A Petsko and Dagmar Ringe

Most Common Post-translational Modifications

Reversible	Irreversible
disulfide bridge	cofactor binding
cofactor binding	proteolysis
glycosylation	ubiquitination
phosphorylation	peptide tagging
acylation	lysine hydroxylation
ADP-ribosylation	methylation
carbamylation	
N-acetylation	

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

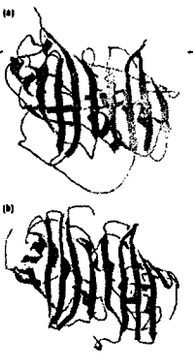


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

(a)

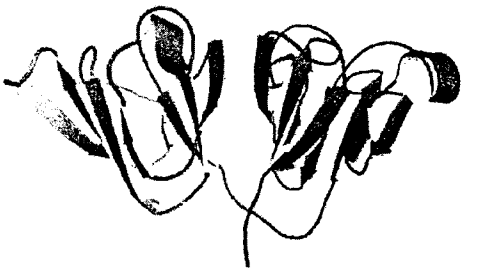
(b)



Topic 6: Ligand regulation

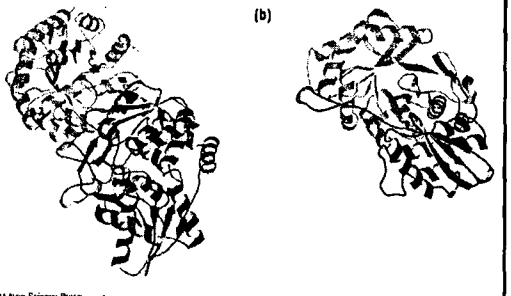
"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

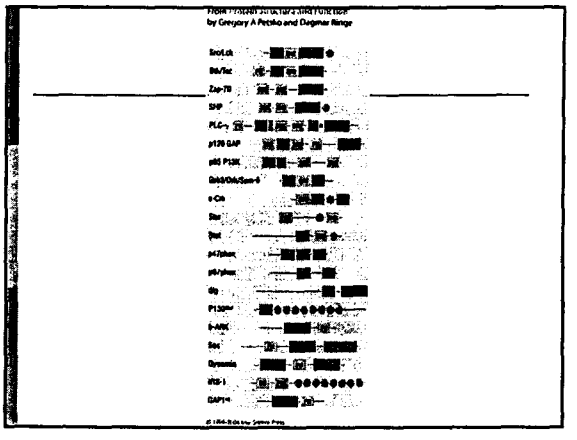
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

(a) (b)





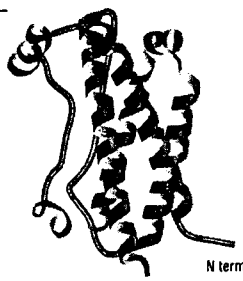
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

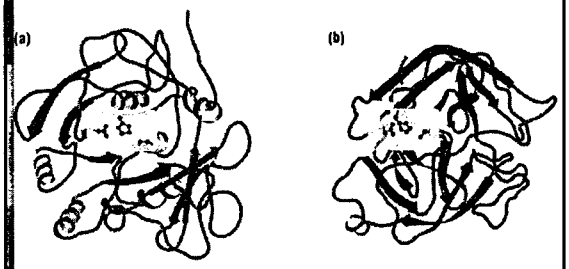
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

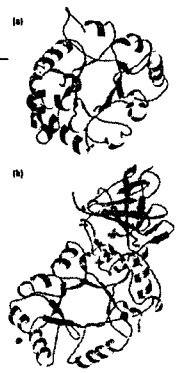
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



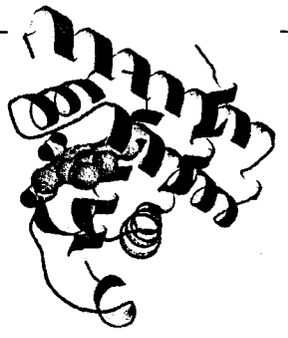
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



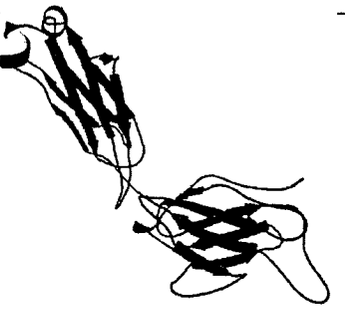


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

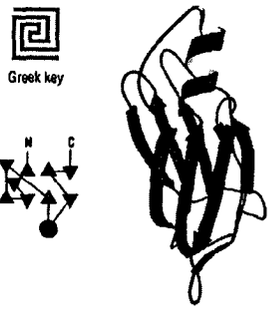


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



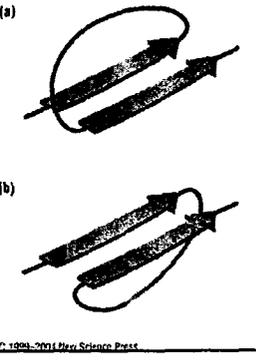
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

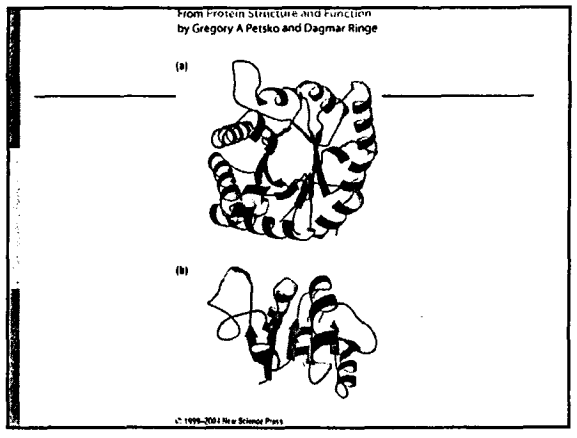
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

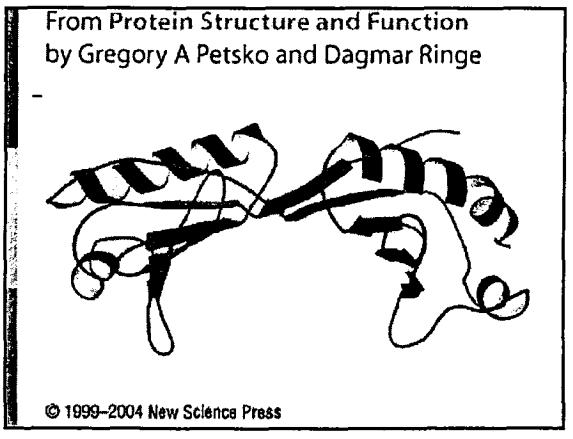


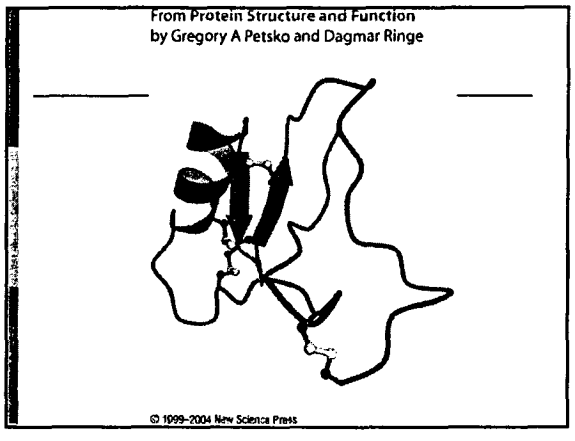
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

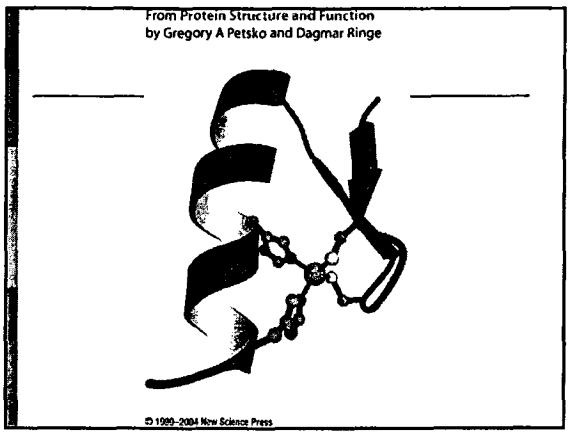
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

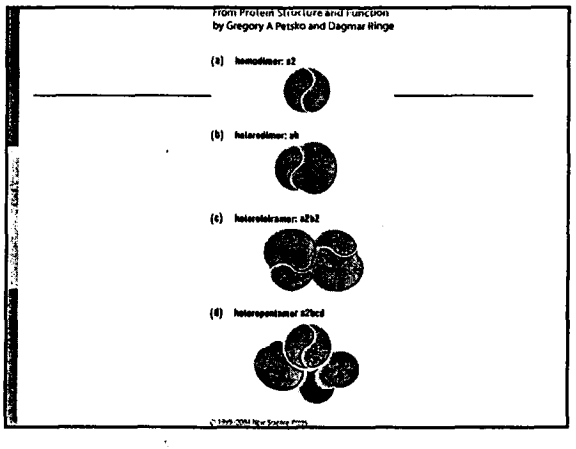


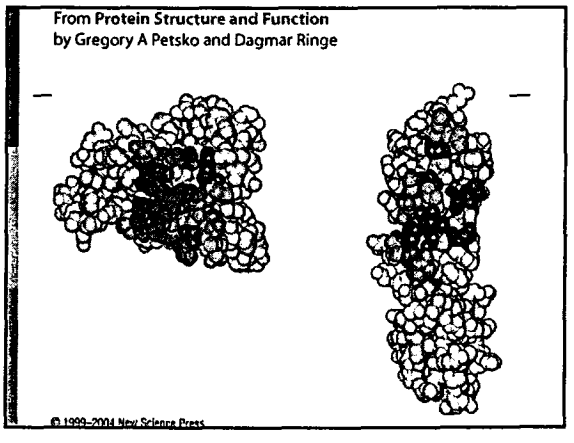

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
(a)
(b)


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
© 1999–2004 New Science Press

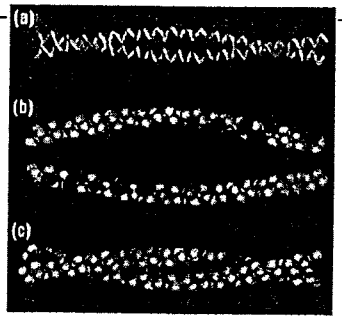

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
© 1999–2004 New Science Press


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

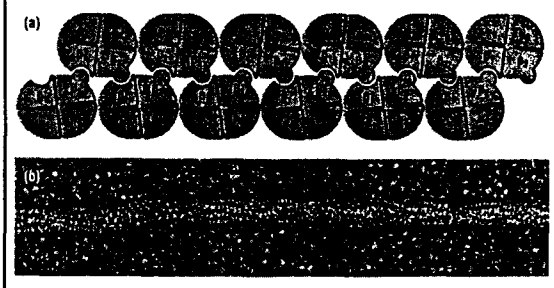
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

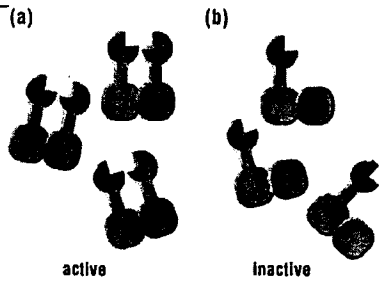
Topic 7: GTPase regulation

"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

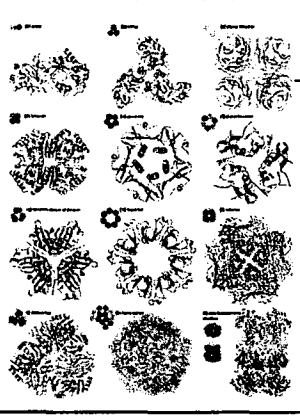
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



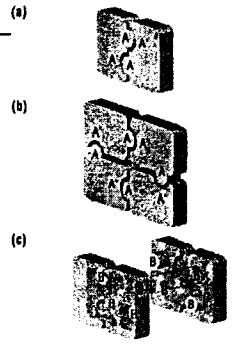
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

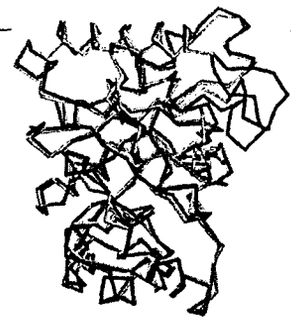


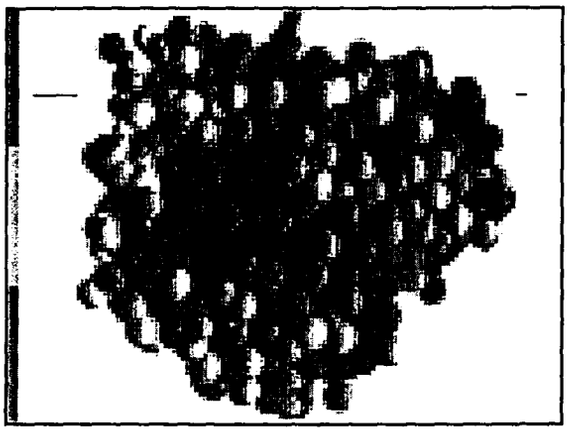
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

Types of Motion Found in Proteins (all values approximate)

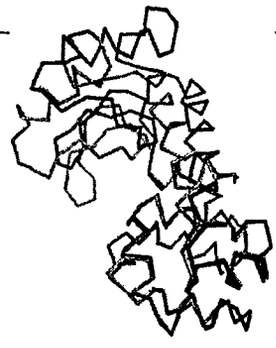
Motion	Spatial displacement (Å)	Characteristic time (s)	Energy source
Fluctuations (e.g., atomic vibrations)	0.01 to 1	10^{-15} to 10^{-11}	k_BT
Collective motions (A) fast, infrequent (e.g., Tyr, Phe ring flips) (B) slow (e.g., domain movement; hinge-bending)	0.01 to > 5	10^{-12} to 10^{-3}	k_BT
Triggered conformational changes	0.5 to > 10	10^{-9} to 10^{3}	Binding interactions

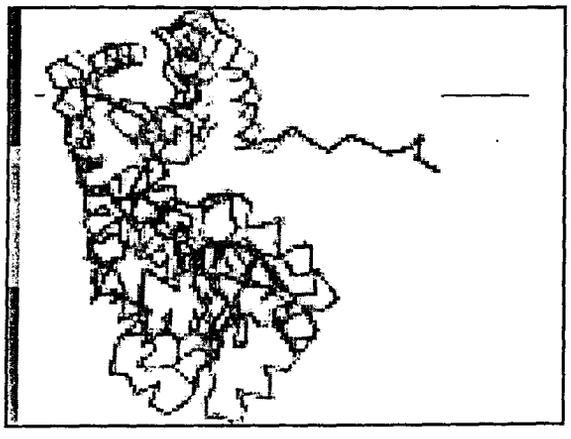
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



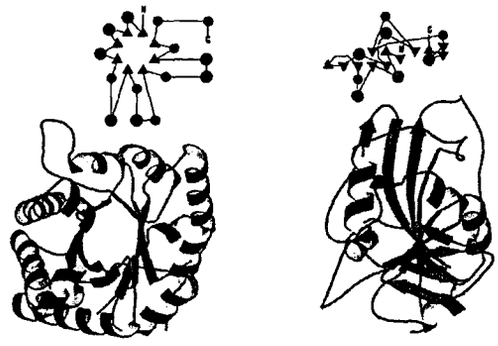


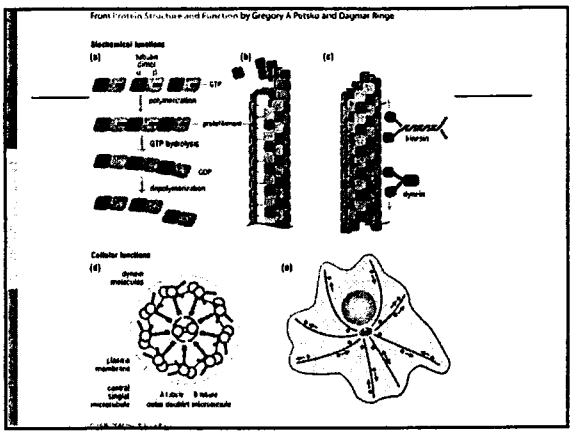
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



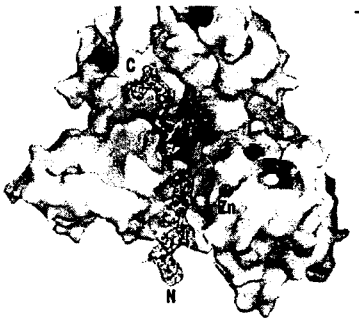


From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe





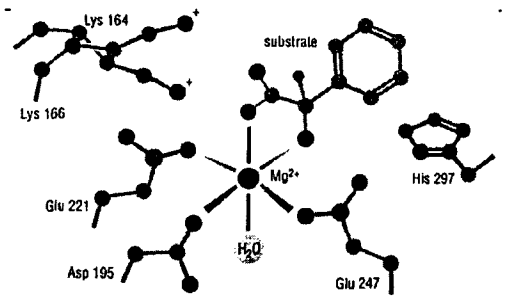
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



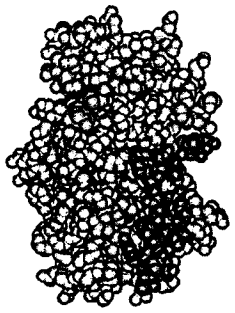
Topic 8: Degradation, phosphorylation, proteolysis, splicing and other PTM regulation

"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

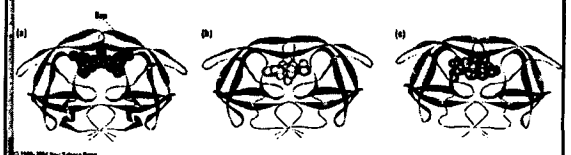
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



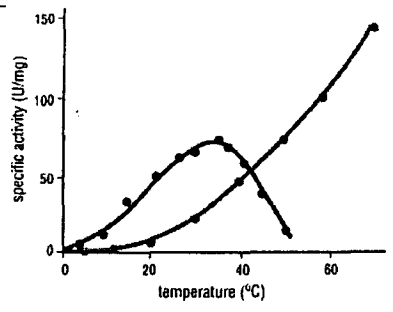
From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

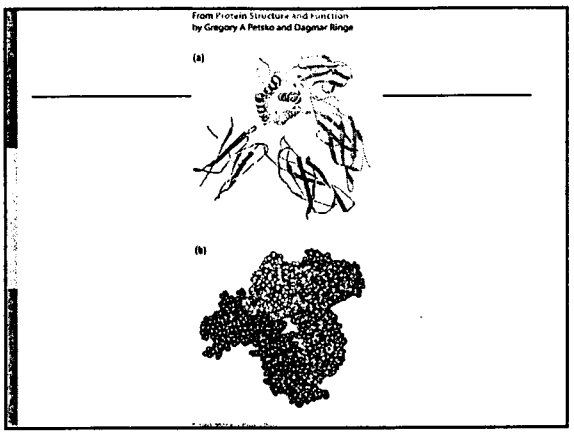


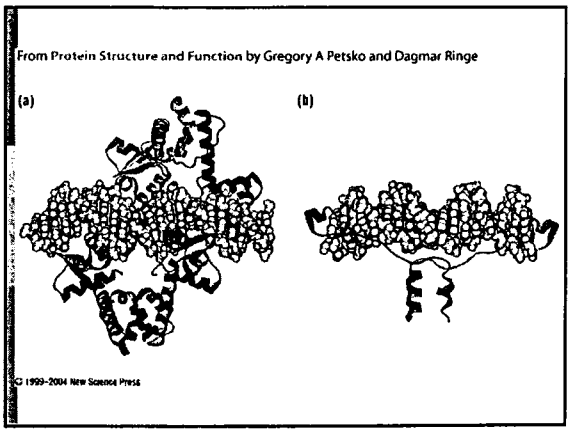
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

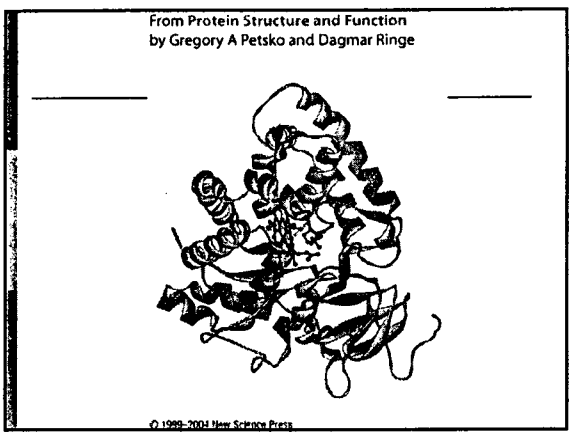


From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

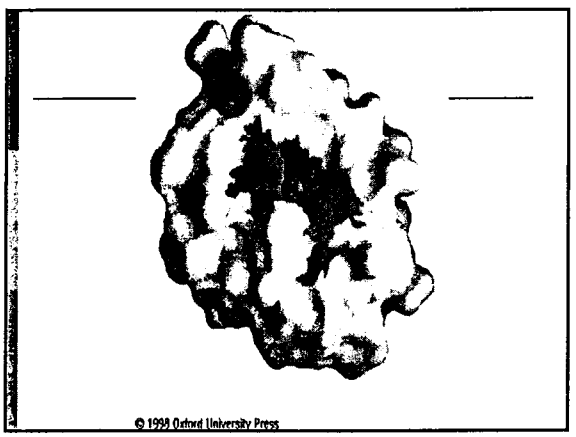


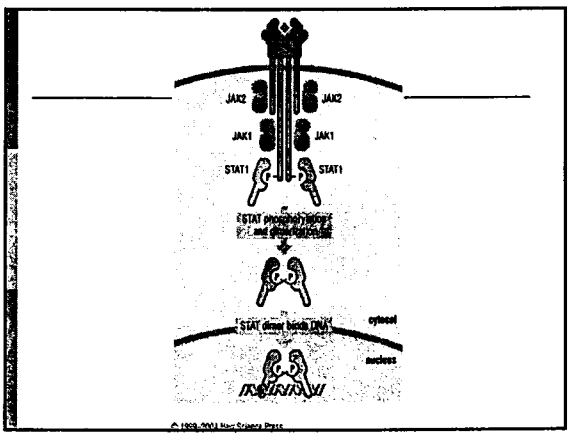

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
(a)
(b)

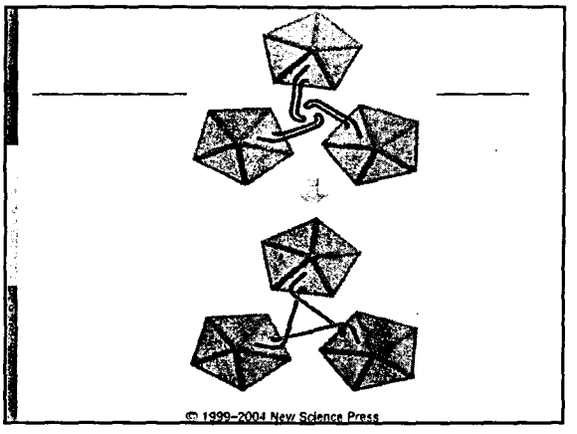

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe
(a)
(b)
© 1999-2004 New Science Press

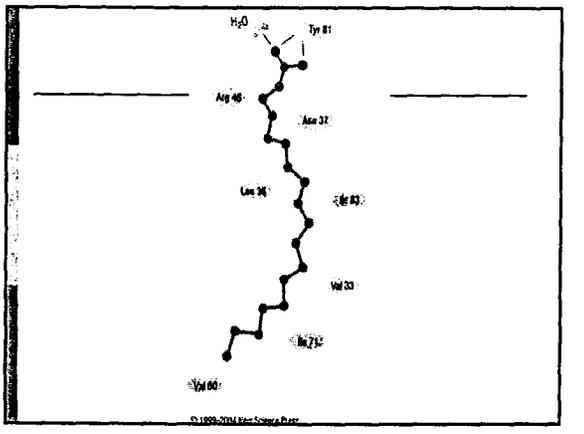

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
© 1999-2004 New Science Press

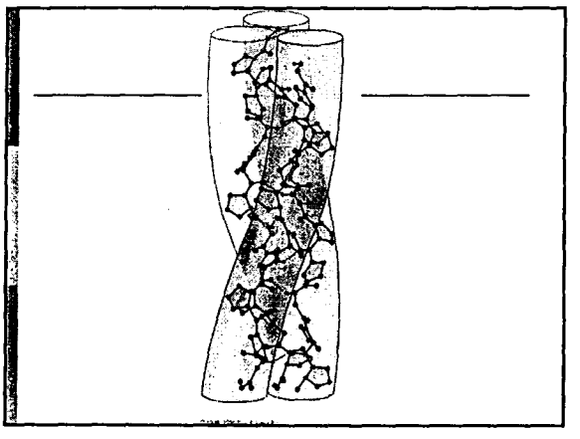

© 1999-2004 New Science Press

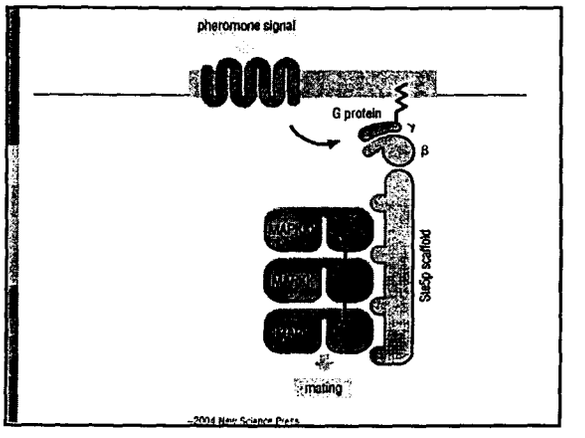

© 1998 Oxford University Press

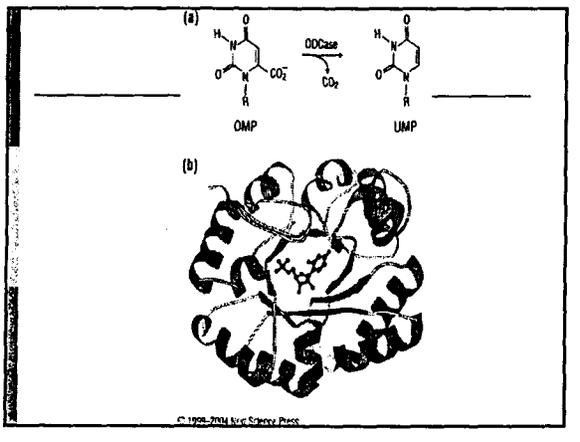

JAK2
JAK2
JAK1
JAK1
STAT1
STAT1
cytosol
nucleus

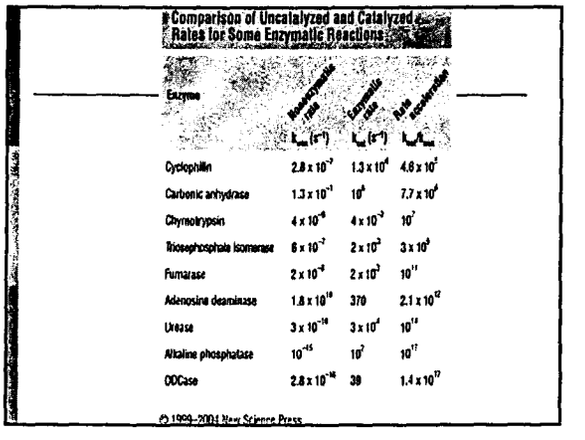






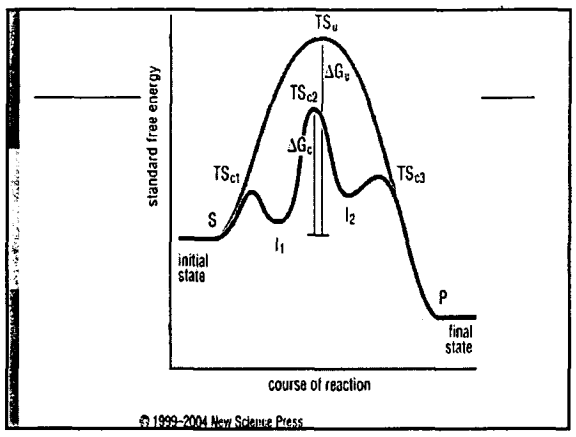






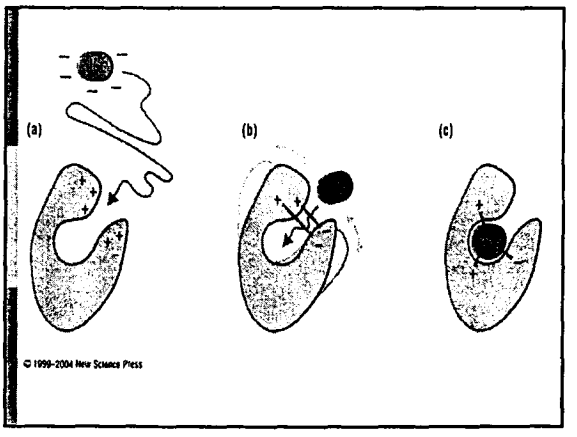
Comparison of Uncatalyzed and Catalyzed Rates for Some Enzymatic Reactions

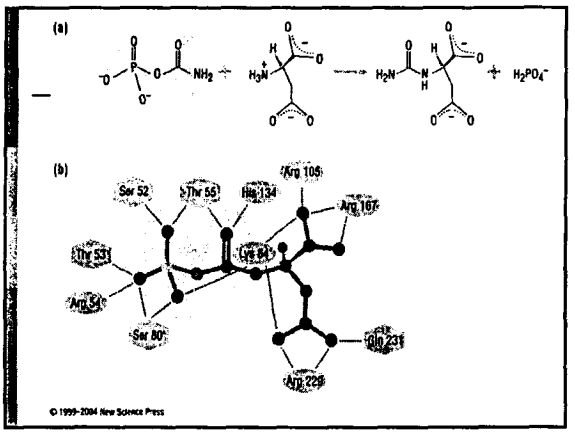
Enzyme	Nonenzymatic rate (k_{non} (s^{-1}))	Enzymatic rate (k_{cat} (s^{-1}))	Rate acceleration (k_{cat}/k_{non})
Cyclophilin	2.8×10^{-7}	1.3×10^{-2}	4.6×10^{4}
Carbonic anhydrase	1.3×10^{-1}	10^{6}	7.7×10^{6}
Chymotrypsin	4×10^{-8}	4×10^{-2}	10^{7}
Triosephosphate isomerase	6×10^{-7}	2×10^{3}	3×10^{9}
Fumarase	2×10^{-9}	2×10^{2}	10^{11}
Adenosine deaminase	1.8×10^{-10}	370	2.1×10^{12}
Urease	3×10^{-14}	3×10^{4}	10^{18}
Alkaline phosphatase	10^{-15}	10^{2}	10^{17}
ODCase	2.8×10^{-16}	39	1.4×10^{17}

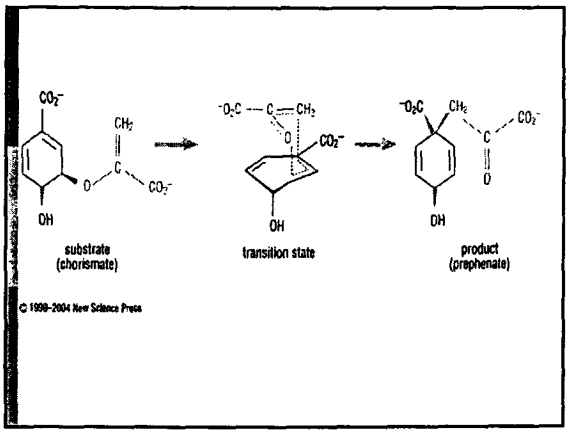

standard free energy
TS_u
ΔG_u
TS_{c2}
ΔG_c
TS_{c1}
TS_{c3}
S
I_1
I_2
initial state
P
final state
course of reaction
© 1999-2004 New Science Press

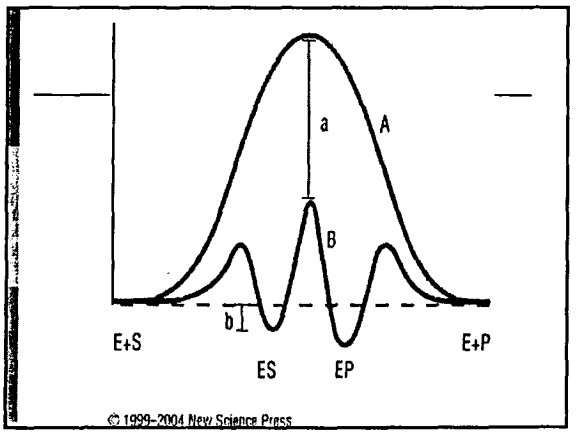
Topic 9: Olfactory proteins

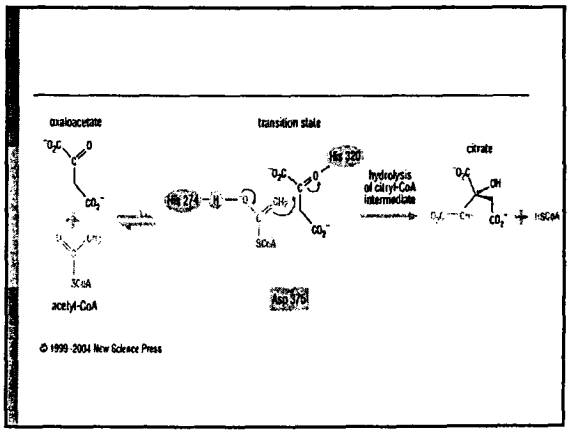
"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

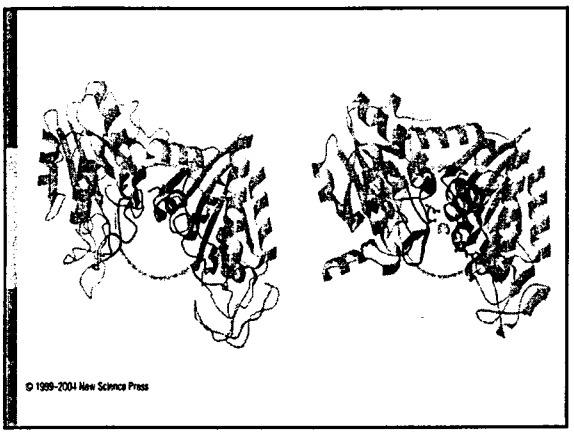


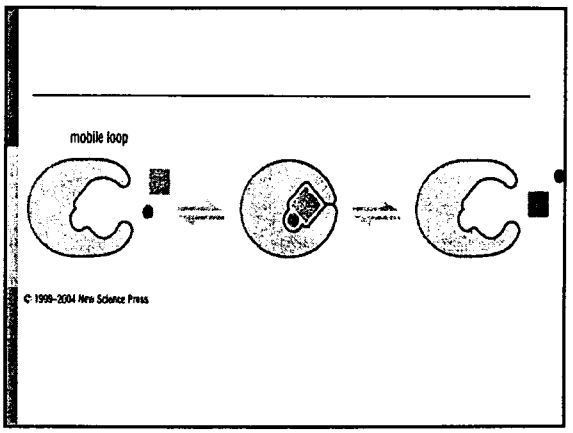


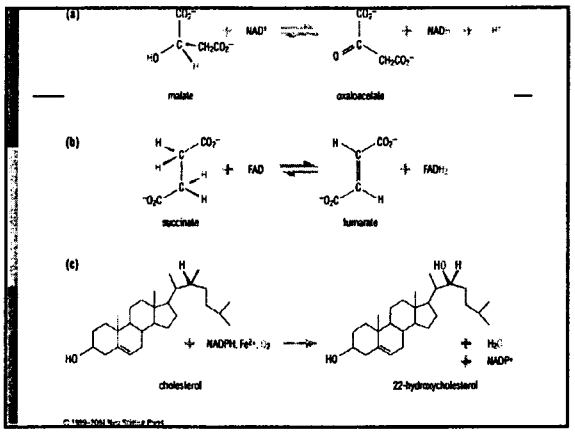


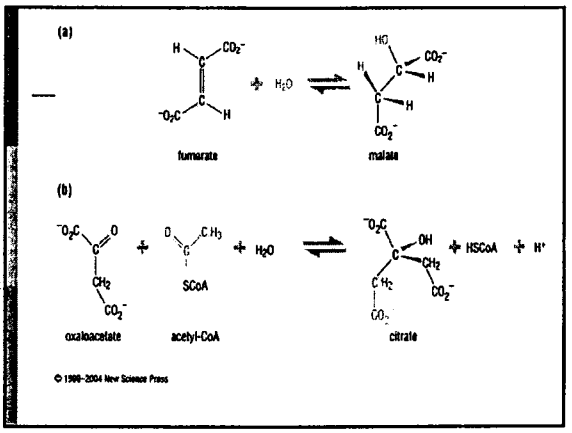


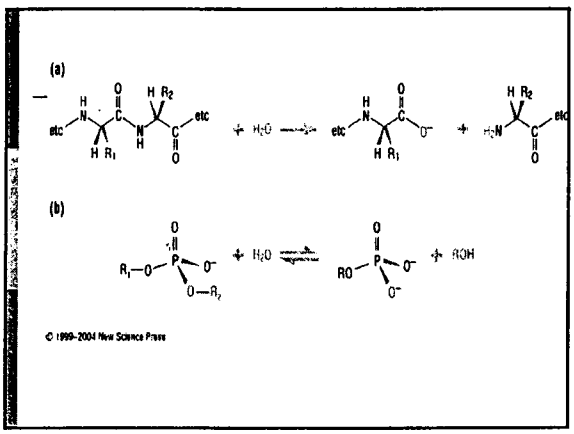



© 1999–2004 New Science Press


mobile loop
© 1999–2004 New Science Press


(a)
malate
oxaloacetate
(b)
succinate
fumarate
(c)
cholesterol
22-hydroxycholesterol


(a)
fumarate
malate
(b)
oxaloacetate
acetyl-CoA
citrate
© 1999–2004 New Science Press


(a)
(b)
© 1999–2004 New Science Press

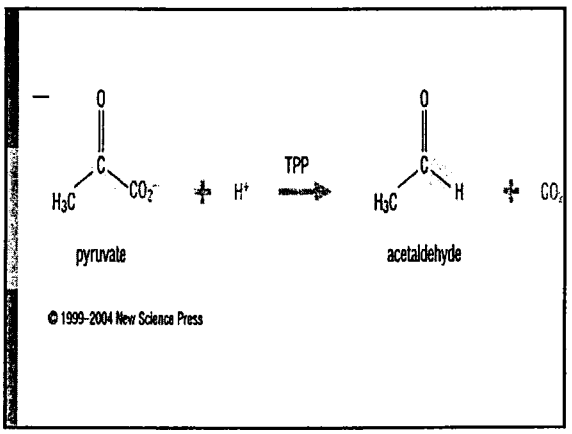
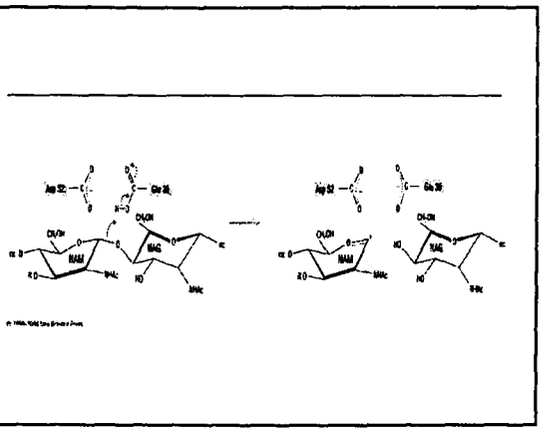
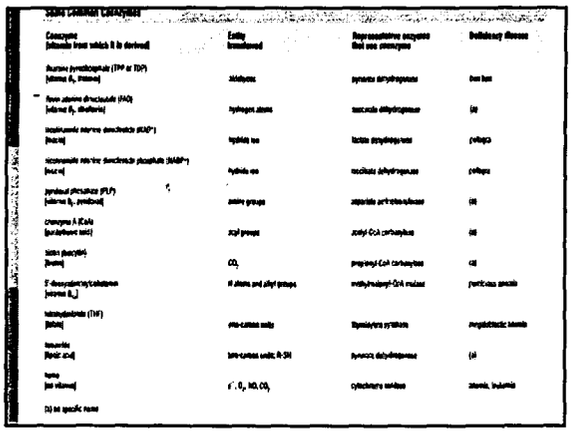

TPP
pyruvate
acetaldehyde
© 1999–2004 New Science Press

Table of Typical pKa Values

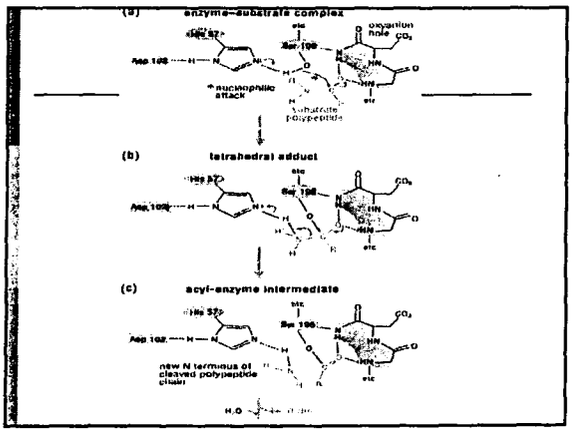
Acid (proton donor)	Conjugate base (proton acceptor)	pK_a
HCOOH formic acid	$HCOO^-$ formate ion	3.75
CH_3COOH acetic acid	CH_3COO^- acetate ion	4.76
OH $CH_3CH-COOH$ lactic acid	OH $CH_3CH-COO^-$ lactate ion	3.86
H_3PO_4 phosphoric acid	$H_2PO_4^-$ dihydrogen phosphate ion	2.14
$H_2PO_4^-$ dihydrogen phosphate ion	HPO_4^{2-} monohydrogen phosphate ion	6.86
HPO_4^{2-} monohydrogen phosphate ion	PO_4^{3-} phosphate ion	12.4
H_2CO_3 carbonic acid	HCO_3^- bicarbonate ion	6.37
HCO_3^- bicarbonate ion	CO_3^{2-} carbonate ion	10.25
C_6H_5OH phenol	$C_6H_5O^-$ phenolate ion	9.89
NH_4^+ ammonium ion	NH_3 ammonia	9.25
H_2O	OH^-	15.7

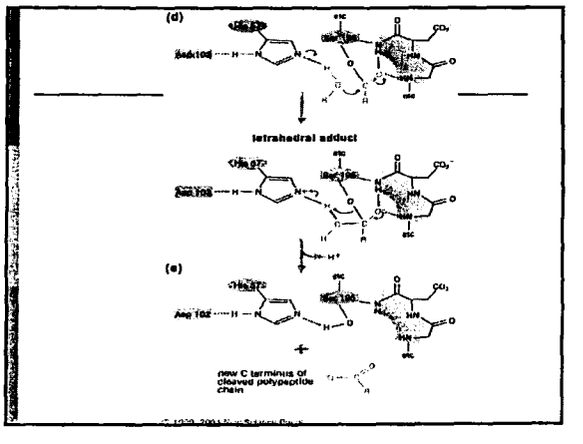


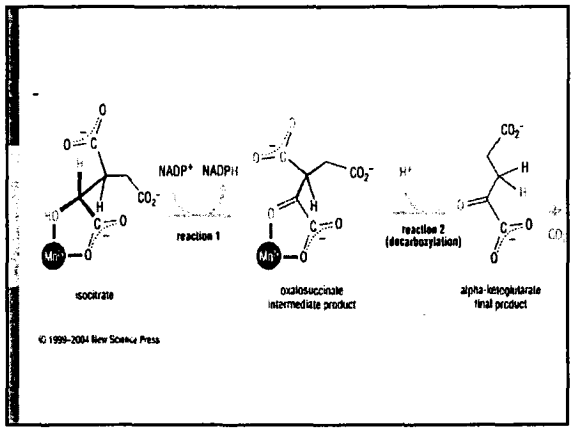


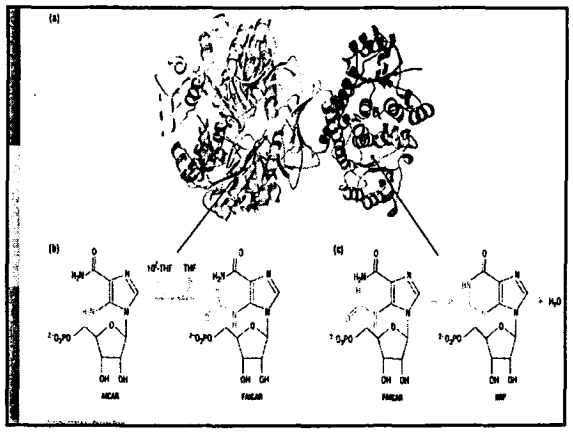
Metal Ions and Some Enzymes Requiring Them

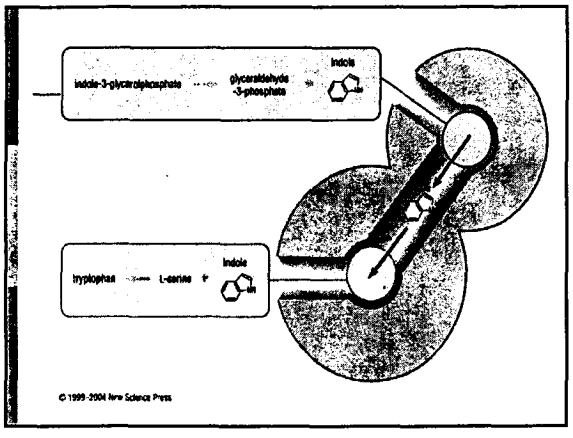
Metal ion	Enzyme
Fe^{2+} or Fe^{3+}	cytochrome oxidase catalase peroxidase
Cu^{2+}	cytochrome oxidase
Zn^{2+}	DNA polymerase carbonic anhydrase alcohol dehydrogenase
Mg^{2+}	hexokinase glucose-6-phosphatase pyruvate kinase
Mn^{2+}	arginase
K^+	pyruvate kinase
Ni^{2+}	urease
Mo	nitrate reductase
Se	glutathione peroxidase

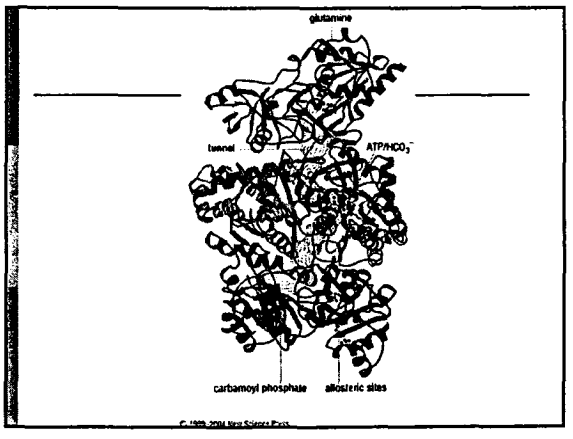





NADP+ NADPH
reaction 1
reaction 2
(decarboxylation)
isocitrate
oxalosuccinate
intermediate product
alpha-ketoglutarate
final product

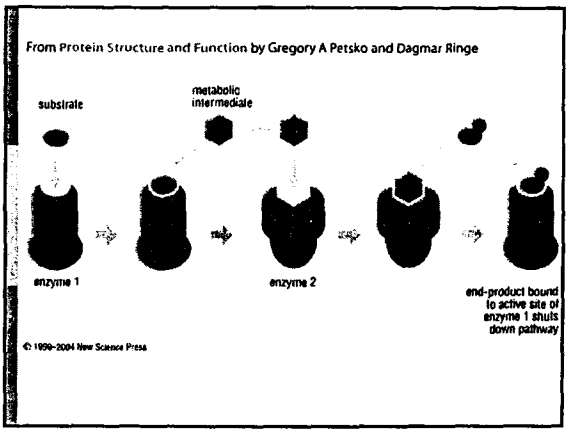

(a)
(b)
(c)

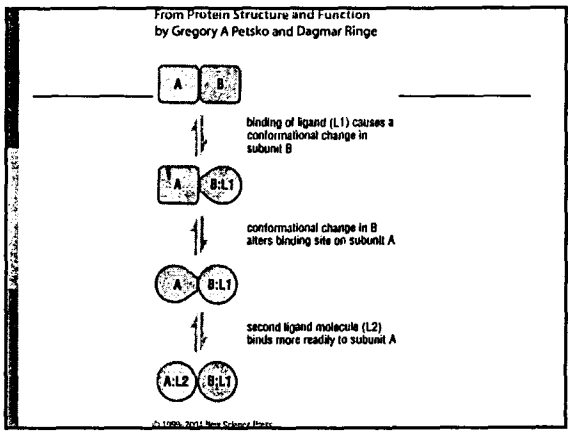

indole-3-glycerolphosphate
glyceraldehyde
-3-phosphate
indole
tryptophan
L-serine

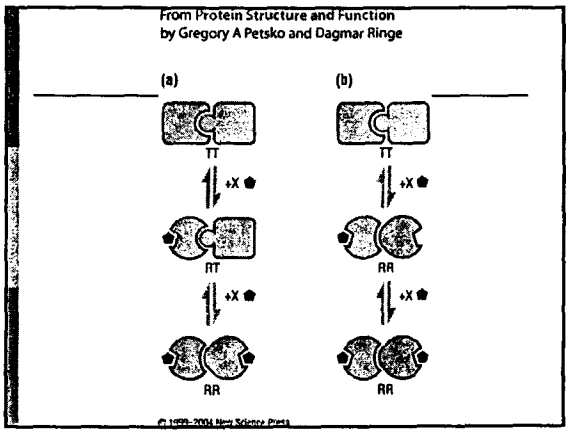

glutamine
tunnel
ATP/HCO3-
carbamoyl phosphate
allosteric sites

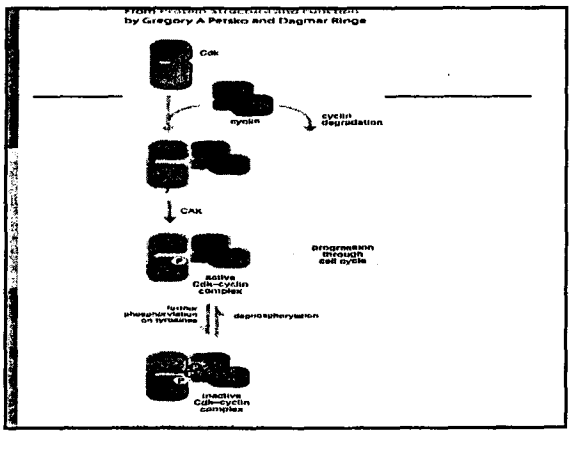
Topic 10: Experimental tools for probing protein function

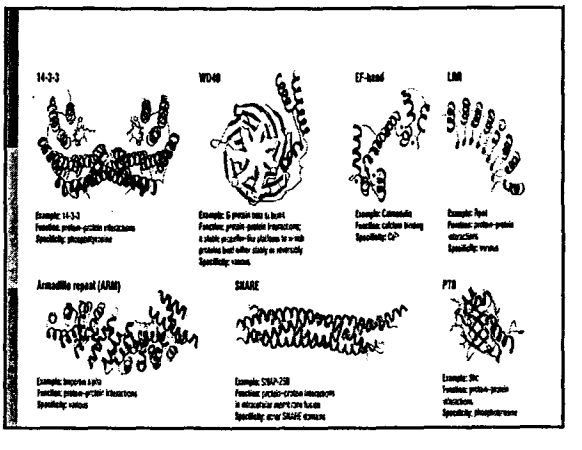
"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.

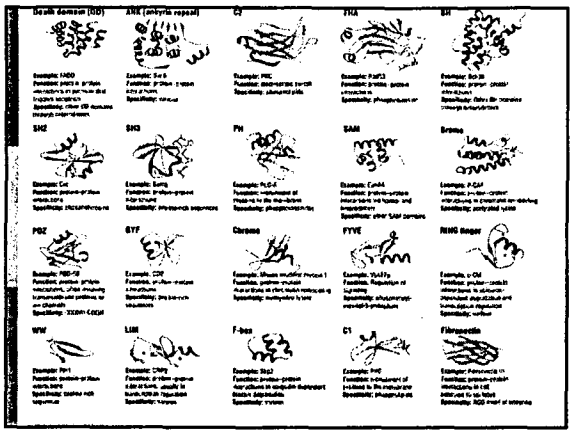


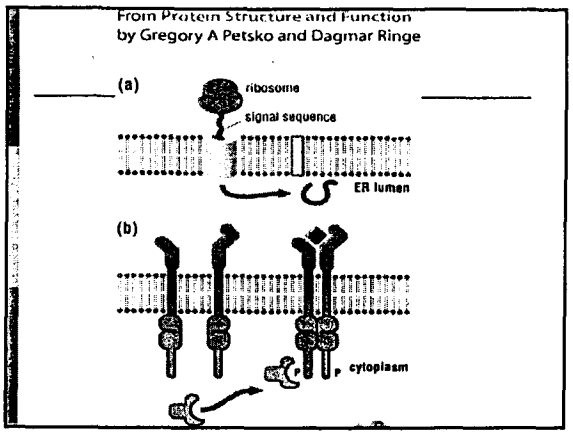


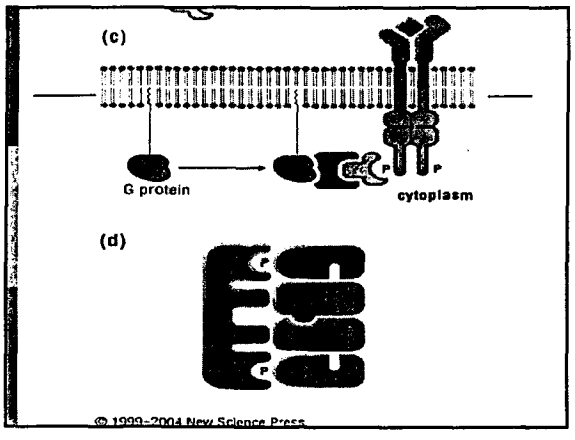


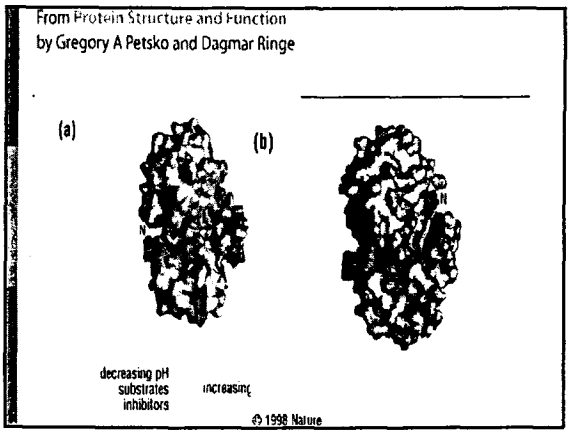


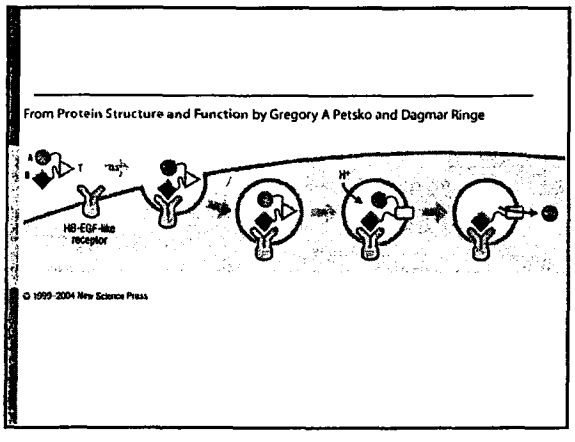


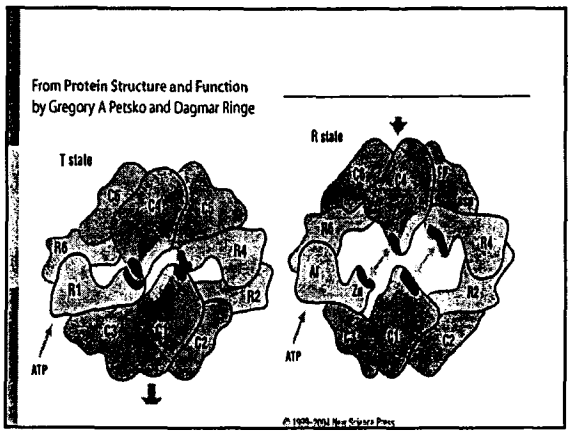



From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
(a)
ribosome
signal sequence
ER lumen
(b)
cytoplasm


(c)
G protein
cytoplasm
(d)
© 1999-2004 New Science Press

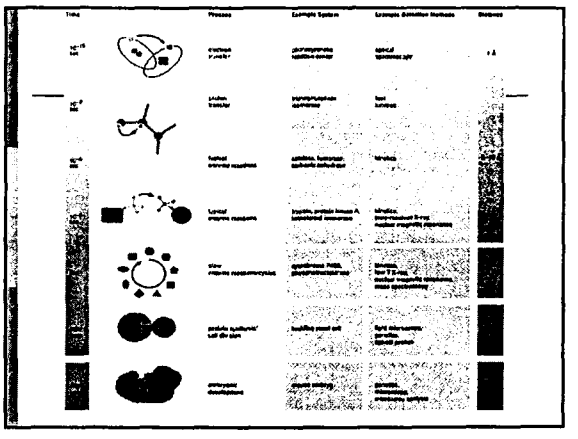

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
(a)
(b)
decreasing pH
substrates
inhibitors
increasing
© 1998 Nature


From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe
HB-EGF-like
receptor
© 1999-2004 New Science Press

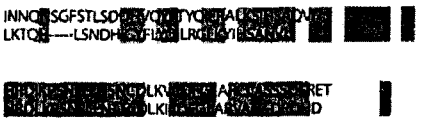

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
T state
R state
ATP
ATP

Topic 11: Sequence alignment, homology modeling, profile-based threading and rosetta

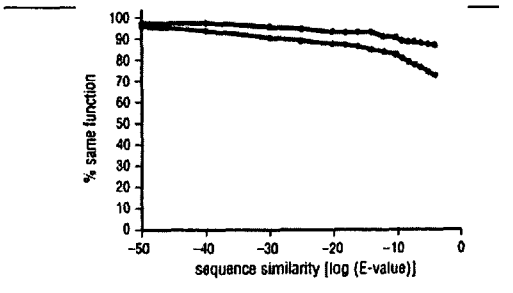
"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.



From **Protein Structure and Function** by Gregory A Petsko and Dagmar Ringe



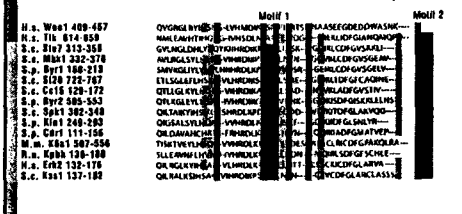
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



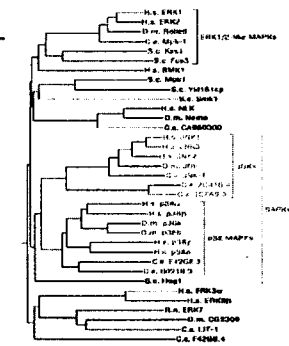
Alignment is the first step in determining whether two sequences are similar to each other

- A key step in comparing two sequences is to match them up in an alignment.
- As a quantitative measure of similarity, a pairwise alignment is given a score, which reflects the degree of matching.
  - Simplest case: only identical matched residues are counted known as percent identity.
  - The Hidden Markov Model is a statistical model that considers all possible combinations of matches, mismatches and gaps to generate the "best" alignment of two or more sequences.

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe



From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe



Multiple alignments and phylogenetic trees

- Multiple alignments of homologous proteins or gene sequences from different species are used to derive a so called evolutionary distance between each pair of species on the degree of difference.
- These distances can be used to build phylogenetic trees which are greatly influenced by the algorithm and specific assumptions used.

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

Some Examples of Small Functional Protein Domains

Domain	Function
SH2	binds phosphotyrosine
SH3	binds proline-rich sequences
Pleckstrin homology (PH)	binds to G proteins and membranes
WD40	protein–protein interaction
DH	guanine nucleotide exchange
EF-hand	binds calcium
Homeobox	binds DNA
TRBD	binds tRNA
Helix-turn-helix	binds DNA
PUA	RNA modification

Structural data can help sequence comparison find related proteins

- Prediction of secondary and tertiary structure from sequence alone are methods such as Chou-Fasman and profile-based threading (fitting unknown to any protein folds gathered).
- Knowledge of the variability (residue interplays, local structure influence, and superfamily features) of a sequence that can form closely similar structure can improve the prediction methods based on statistical analysis of sequences.
- All methods for prediction protein structure from sequence seem to have a maximum accuracy of about 70%.

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

Some Examples of Short Sequence Motifs and Their Functions

Contiguous motif	Consensus sequence	Function
Walker (P loop)	[A/G]XXXXGK[S/T]	binds ATP or GTP
Zn finger	$CX_{2-4}CX_{12}HX_{3-5}H$	binds Zn in a DNA-binding domain
Osteonectin	CX[D/N]XXCXXG[K/R/H]XCX_{6-7}PXCXCX_{5-6}CP	binds calcium and collagen
DEAD box helicase	XXDEAD[R/K/E/R]X	ATP-dependent RNA unwinding
MARCKS	GOENGNV[K/R]	substrate for protein kinase C
Calsequestrin	[E/Q][D/E]GL[D/N]FPXYDGXDRV	binds calcium

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe

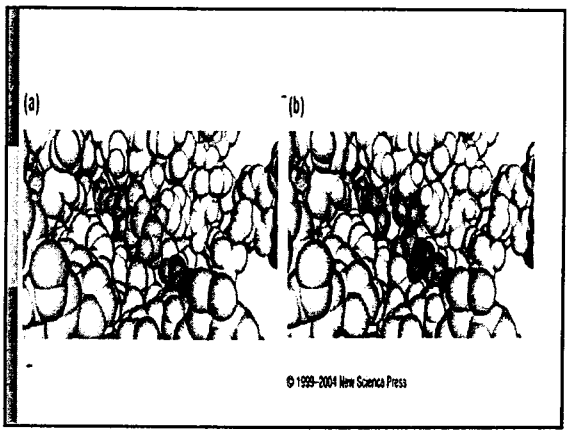
Constructing a Family Profile

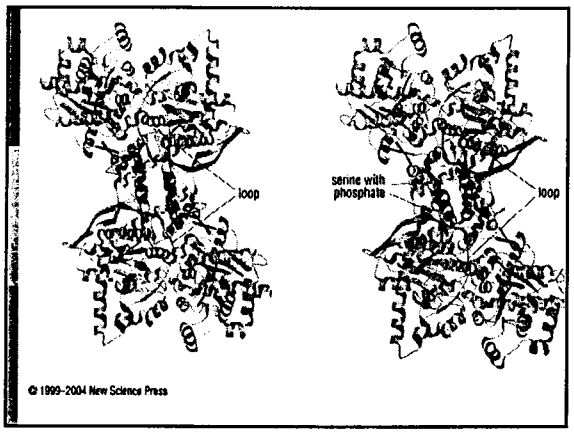
Position	1	2	3	4	5
	C	C	G	T	L
	C	G	H	S	V
	G	C	G	S	L
	C	G	G	T	L
	C	C	G	S	S

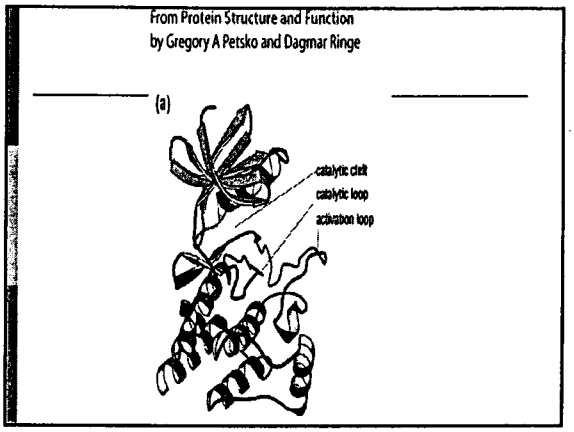
Position	1	2	3	4	5
Prob(C)	0.8	0.6	-	-	-
Prob(G)	0.2	0.4	0.8	-	-
Prob(H)	-	-	0.2	-	-
Prob(S)	-	-	-	0.6	0.2
Prob(T)	-	-	-	0.4	-
Prob(L)	-	-	-	-	0.6
Prob(V)	-	-	-	-	0.2

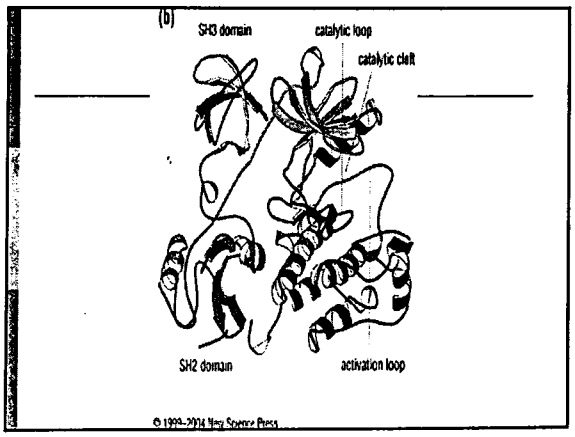
Topic 12: Identification of binding sites and catalytic residues

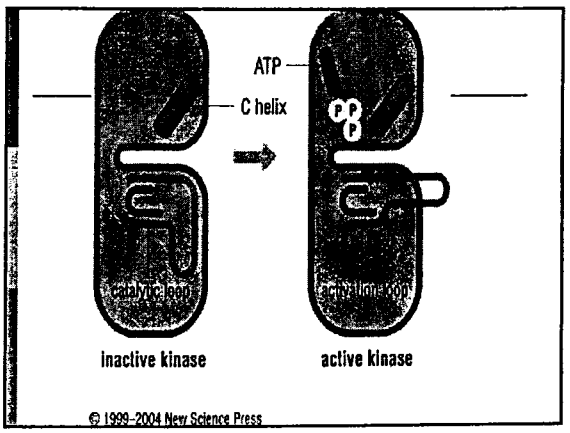
"Protein structure and function, 2004. by Gregory Petsko and Dagmar Ringe,藝軒書局.











Src kinases both activate and inhibit themselves

- Src-family kinases are activated early in many signaling pathways and once activated sustain their own activated states by autophosphorylation, providing a large amplification of the signal (prolonged activated src causes cancers).

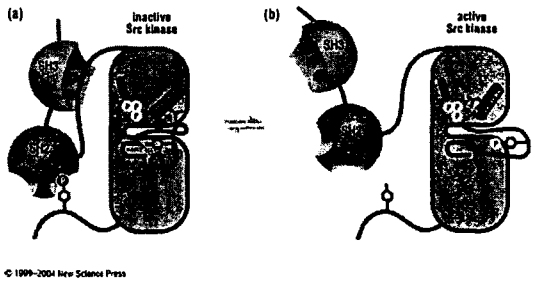
src kinase is inactivated

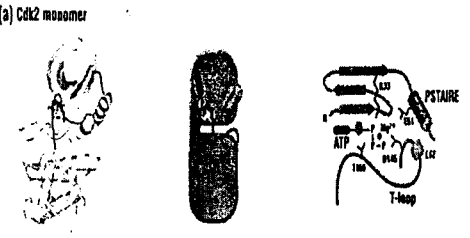
- In the absence of signal,
  - The SH2 domain binds to a phosphorylated tyrosine in the C-terminal, and
  - the linker region between SH2 and catalytic domain forms a polyproline helix binding to SH3 domain.

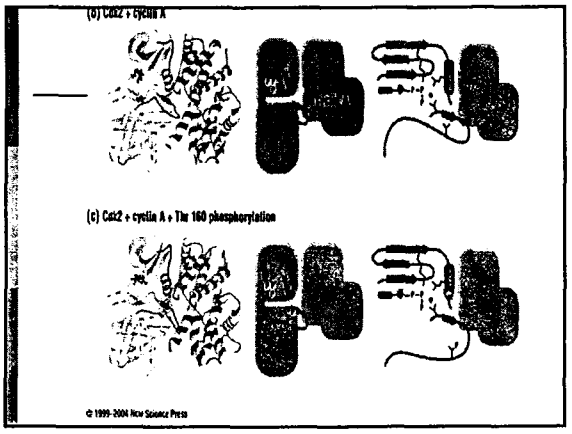
src kinase is activated

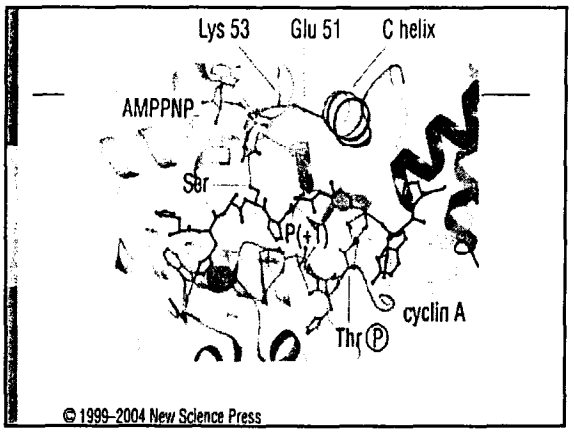
- In the presence of signal,.
  - The C-terminal phosphorylated tyrosine is dephosphorylated to release the SH2 or the SH2 is bound by a new phosphorylated tyrosine from a receptor protein.
  - The activation loop is phosphorylated to be active form.
  - The causes a conformational change to release SH3 from polyproline helix, and SH3 binds to the target proteins.

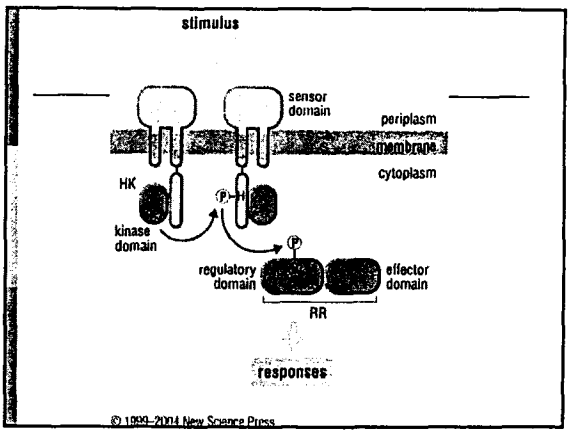
From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe

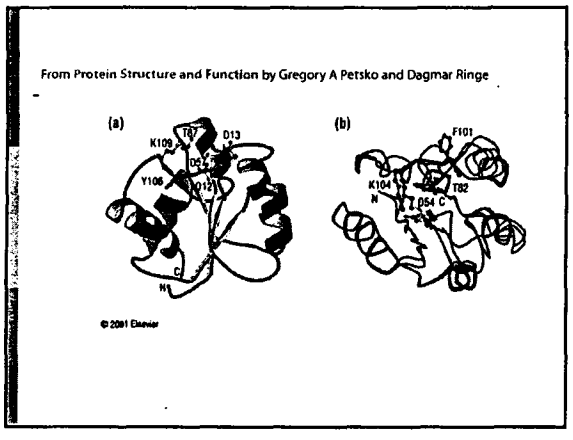


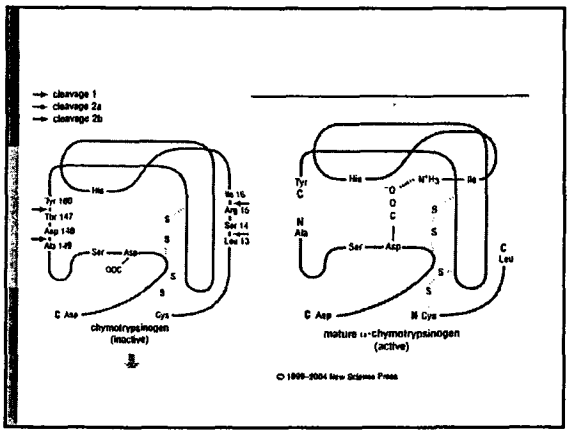


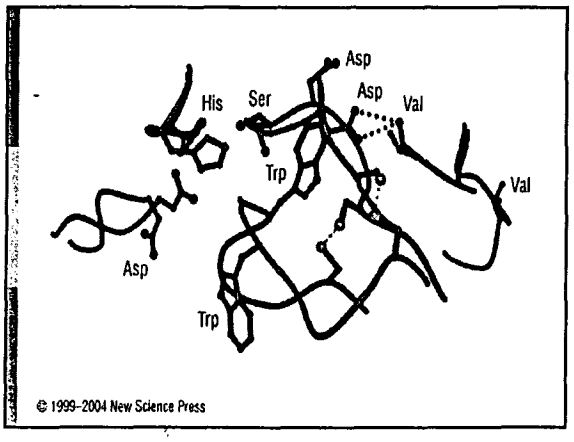


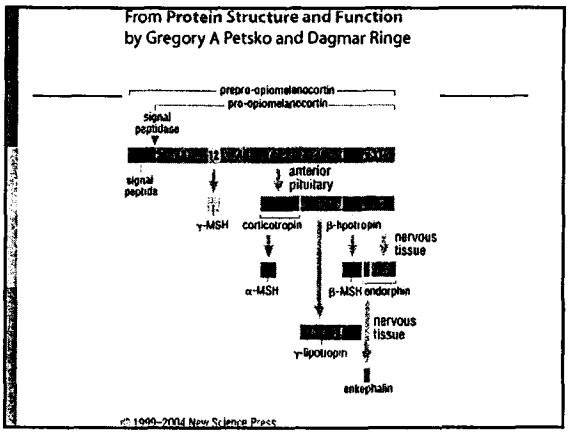

Lys 53
Glu 51
C helix
AMPPNP
Ser
P(+1)
cyclin A
Thr Ⓟ
© 1999–2004 New Science Press

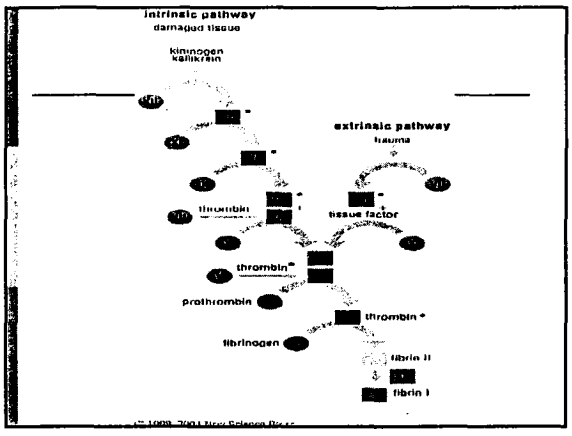

stimulus
sensor domain
periplasm
membrane
cytoplasm
HK
kinase domain
regulatory domain
effector domain
RR
responses
© 1999–2004 New Science Press

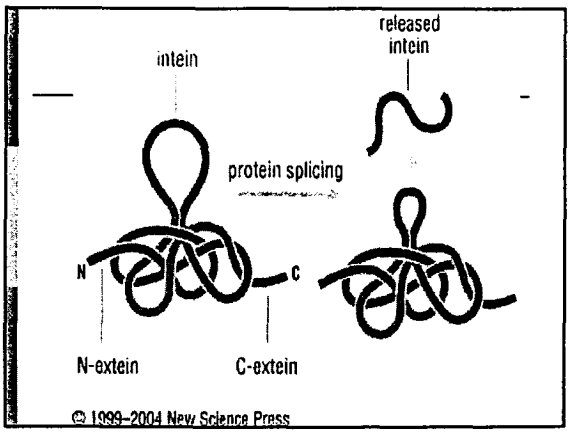

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe
(a)
(b)


cleavage 1
cleavage 2a
cleavage 2b
chymotrypsinogen (inactive)
mature α-chymotrypsinogen (active)

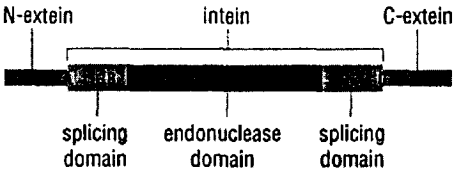

Asp
His
Ser
Asp
Val
Trp
Val
Asp
Trp
© 1999–2004 New Science Press


From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe
prepro-opiomelanocortin
pro-opiomelanocortin
signal peptidase
signal peptide
anterior pituitary
γ-MSH
corticotropin
β-lipotropin
nervous tissue
α-MSH
β-MSH endorphin
nervous tissue
γ-lipotropin
enkephalin

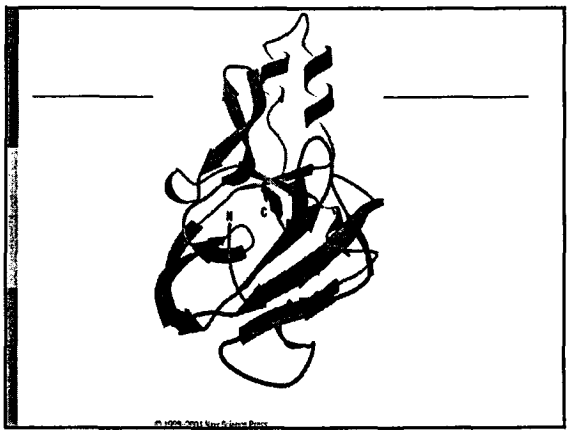


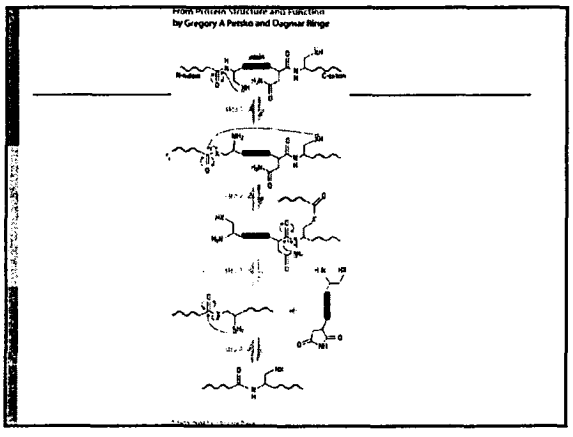


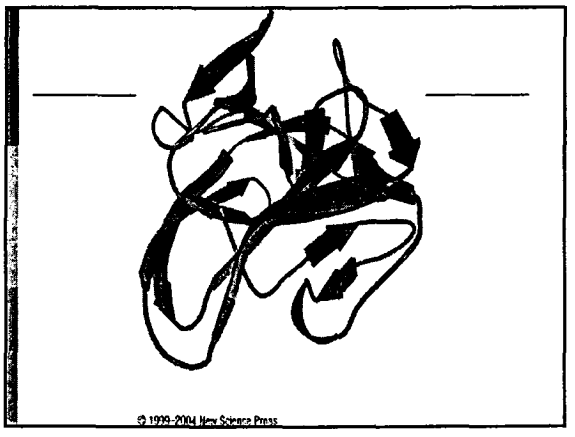
From **Protein Structure and Function**
by Gregory A Petsko and Dagmar Ringe

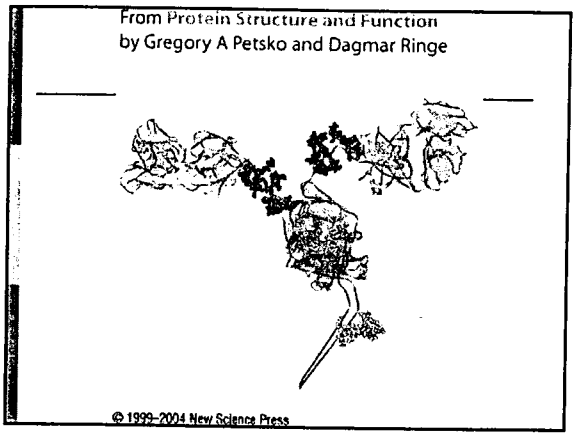


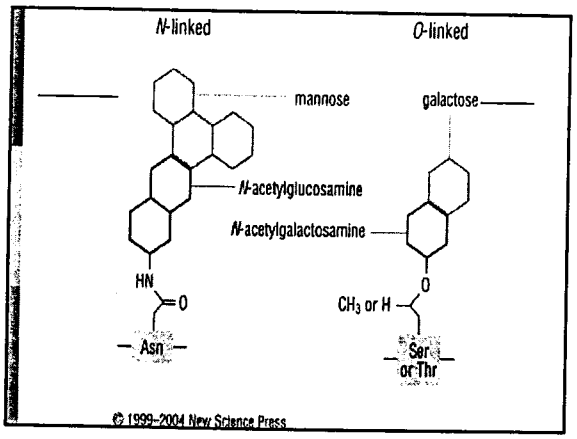
© 1999–2004 New Science Press

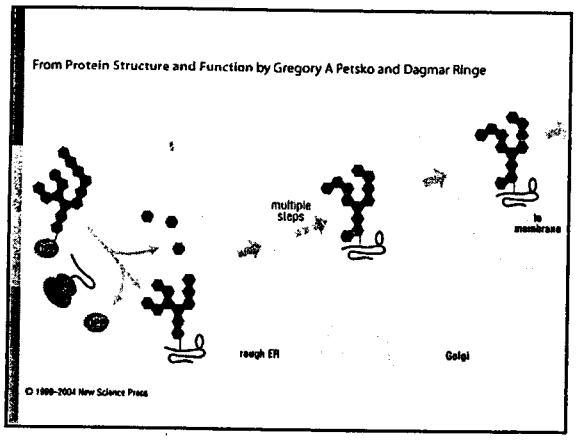


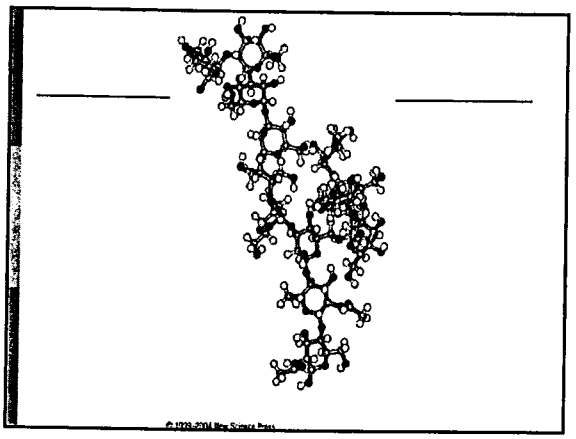


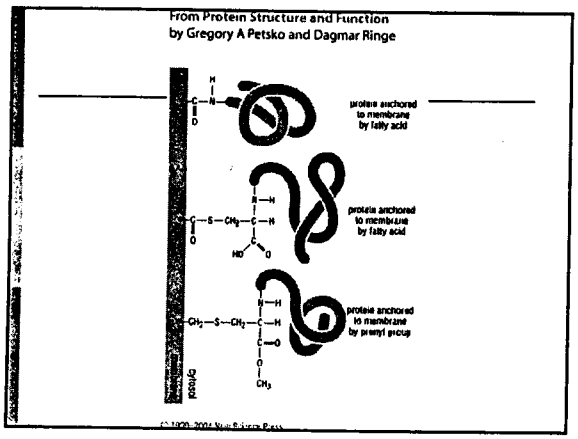


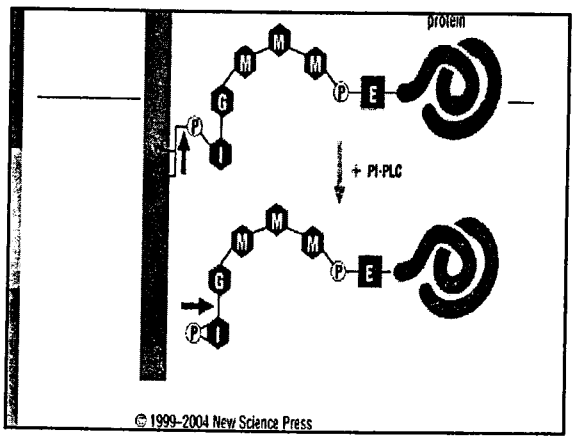

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
© 1999–2004 New Science Press

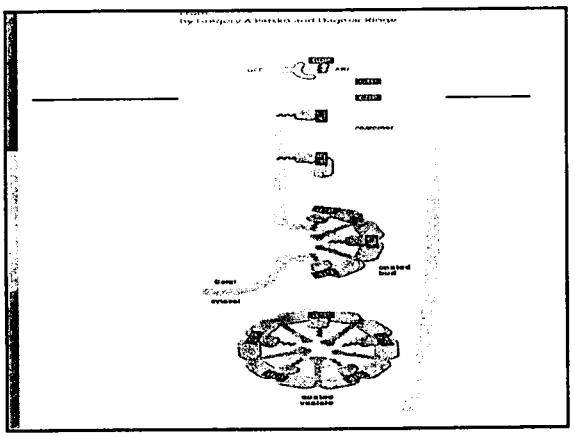

N-linked
O-linked
mannose
galactose
N-acetylglucosamine
N-acetylgalactosamine
HN
O
Asn
CH3 or H
Ser
or Thr
© 1999–2004 New Science Press

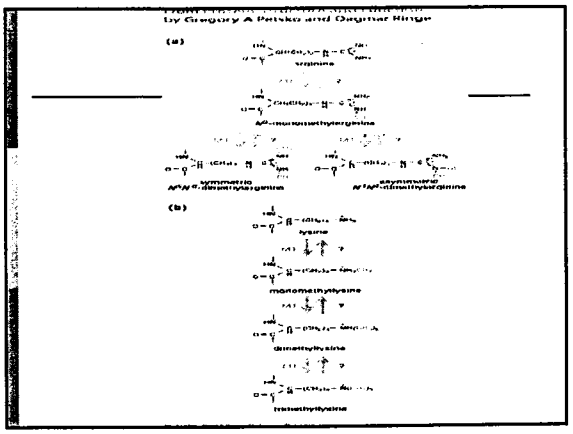

From Protein Structure and Function by Gregory A Petsko and Dagmar Ringe
multiple steps
to membrane
rough ER
Golgi

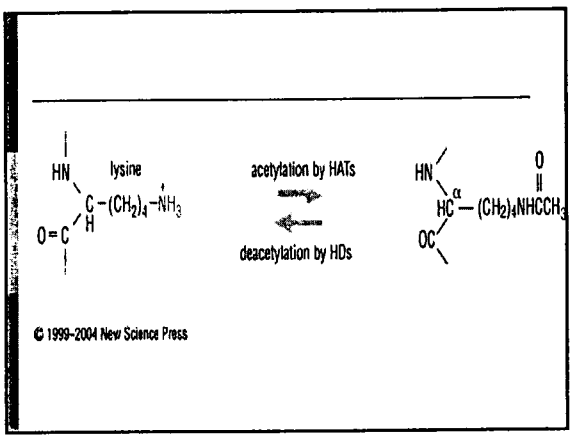


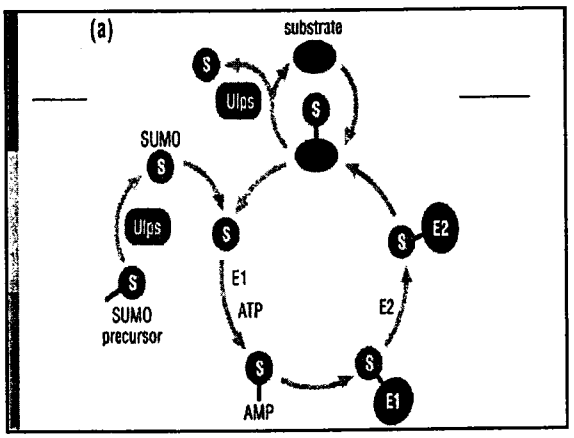

From Protein Structure and Function
by Gregory A Petsko and Dagmar Ringe
protein anchored to membrane by fatty acid
protein anchored to membrane by fatty acid
protein anchored to membrane by prenyl group
cytosol

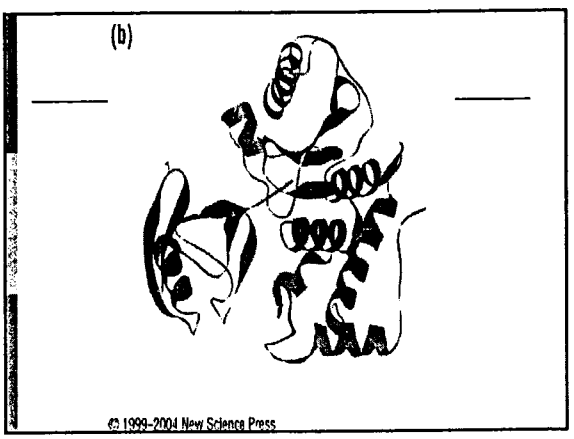

protein
+ PI-PLC
© 1999–2004 New Science Press











From **Protein Structure and Function**
by Gregory A Petsko and Dagmar Ringe