生物資訊簡介

交通大學生物科技系及生物資訊所　黃鎮剛
e-mail: jkhwang@cc.nctu.edu.tw

一、簡介

1990 年 10 月一日，美國國家衛生總署與能源部重新開始人類基因體計畫（Human Genome Project, 簡稱 HGP 或 HUGO），這個計畫對科學界乃至整個社會產生重大影響。今年 2 月 15、16 日，HGP 與 Celera 公司分別在自然(Nature)與科學(Science)雜誌出版了人類基因體圖譜草稿(working draft)[1]。截至筆者寫本篇報導止，生物學家已解出上千種基因體（詳細數字見http://www.ebi.ac.uk/genomes/），包括了病毒與類病毒(viroids)、質體(plasmids)、細胞器(organelles)、細菌(bacteria)、古菌(archaea)及真核生物(eukaryota)。在人類基因圖譜中有許多重要的發現[2]：估計人類約有三萬兩千多個基因[3]，遠低於先前的估計值。雖然人類基因體是果蠅的 30 倍，酵母菌的 250 倍，但是人類基因數目僅為果蠅的兩到三倍。這由於人類的基因，像 alternate splicing sites 遠比果蠅、線蟲來的複雜，組合不同的編碼順序（exon）來表達不同的蛋白質[4]；另外像基因、CG 成分、CpG 島(island)、重組速度、single-nucleotide polymorphism (SNP)（目前在人類基因體發現了 140 萬 SNP）分佈密度隨著區域不同而有很大的改變。基因體研究，累積了大量序列資料。資料的運用，是緊接著的重要的工作。這些資訊對於未來新藥的研發、基因治療、生物機制的探討、癌症研究、蛋白質相互作用、蛋白質結構預測等都大有助益。在“後基因體時代“，序列的註解、功能的預測，為將來生物領域中最主要的研究的題目，這些都是生物資訊學研究的課題。在基因體解序的研究，生物資訊已經佔了極重要的地位。人類基因體計畫需要組合大量序列片段，比對序列，預測基因位置，預測序列與功能的相關性。Celera 在 1998 成立，在很短的時間，利用所謂的「全基因體霰彈槍方法」(whole-genome shutgun approach)[5]，立刻能夠與進行多年 HGP 競爭解碼人類基因體圖譜—而「全基因體霰彈槍方法」的可行性倚賴著計算生物演算法[6]與高速電腦硬體。繼深藍(Deep Blue)之後，IBM 在 1999 年 12 月宣布投資一億美金五年的計畫，發展速度超過 10^{15} 運算/秒 (petaop/s)、一百萬平行處理器的超級 "藍色基因" 電腦(Blue gene)及相關軟體，來解決蛋白質摺疊的相關問題[7]。”科學”雜誌稱生物資訊為二十一世紀生物學的”絕對必要條件”(*sine qua non*)[8]，確實點出生物資訊在現代生物的重要地位。

二、生物資訊

生物資訊是一個整合性的系統，包括了資料庫管理、資料擷取、資料庫存、分析引擎之發展及網路使用介面。現在網路上有許多關基因體的網站[9]，各式各樣資料庫更是散佈各處。大部分的生物相關的資料庫如基因體序列、蛋白質結構資料庫等皆為公有領域(public domain)，學術界可自由下載。這些資料時時更新，各有特定資料內容與格式，例如：EMBL 的 Nucleotide Sequence Database, SWISSPROT 的 Protein Sequence Database 與電腦註解的 TrEMBL； ENSEMBL[10]的自動註解的真核生物基因體資料庫；National Library Medicine (NLM)的 MEDLINE 為生命科學與醫學文獻的資料庫，涵蓋 4000 生物醫學期刊，超過 1000 萬個生物醫學文獻引索；InterPro (Integrated Resource of Protein Domains and Functional Sites)資料庫包含蛋白質家族、功能區(domain)、功能位置等資訊；Pfam 包含對蛋白質區域多重序列比對，利用 Hidden Markov Models (HMM)[11]對蛋白質家族的分類；PROSITE 為蛋白質序列形式(pattern)的資料庫；PRINTS 為蛋白質指紋圖譜資料庫；USPO—PRT 收集 10280（截至 2001 年 3 月 2 日）美國專利的蛋白質序列；PRODOM 是蛋白質區域資料庫 (Protein Domain Database)；DSSP (Definition of Secondary Structure of Proteins)， HSSP，FSSP 為結合蛋白質一級、二級、三級結構的資料庫；PDB 收集蛋白質與 DNA 三級構造等；SCOP[12], CATH[13]為蛋白質結構分類資料庫；SRS[14]是一個資料庫管理系統。限於篇幅，我們只列

出這一些的例子。但是可以看出，整合這麼多資料為生物資訊中重要課題。現在雖有很多物種的基因圖譜已經被解出。但是大部份基因圖譜只是所謂的“原始資料“（raw data)。接下來的工作是如何將這些“原始資料“變成有用的知識—所謂的資料掘礦[15] (data mining)。因此分析工具的發展是生物資訊極重要的研究。例如說，在人類基因體中，編碼區域（coding region）只佔所有 DNA 3%（而重複序列(repeat sequences)佔了 46%）其中可以確定的人類基因至少有兩萬五仟個[5]，但是它的上限為何？卻很難說，雖然現在大家認可的估計值約為三萬兩千基因。因為基因預測軟體如 GenScan[16], GrailEXP[17] GeneWise[18], Genie[19]等，除了利用類神經網路或 HMM 等方法辨認 splicing sites，編碼區域(coding region)的特徵，也結合例如 Expressed Sequence Tag（EST）資訊，幫助預測基因部份是根據所謂的(EST)，但是一些不活躍的基因可能不會在 EST 資料庫中，導致被軟體漏掉（對於這些基因，生物資訊學家套用了物理名詞，稱之為冷暗物質(dark matter)）。因此基因的預測，仍是一活躍研究的領域。

在生物資訊中，一般人最常用的是序列比對的工具，如 Altschul 等人發展的局部序列比對軟體 BLAST(Basic Local Alignment Search Tool)[20]—NCBI 提供了許多不同用途的 BLAST 版本（blastn：核酸比對、 Megablast：基因體序列比對、blastp：蛋白質比對等）；計算演化樹的 PHYLIP[21]（the *PHYL*ogeny *I*nference *P*ackage）Pearson 與 Lipman 發展的 FASTA[22]，是根據 Smith-Waterman 演算法[23]，雖然比較慢但敏感度高；Thompson 等人發展的多重序列比對軟體 CLUSTAL W[24]等。偵測蛋白質區域軟體工具如 SMART(Simple Modular Architecture Research Tool)[25], NCBI 的 DART(Domain Architecture Retrieval Tool)[26]。工研院生醫中心最近發展的 FLAG[27]則是針對基因體與基因體之間序列比對。

對於蛋白質二級結構及其他特性的預測，有用的軟體工具，如 PHD[28]軟體組，可預測二級結構、溶劑接觸面積(solvent accessible area)、跨膜(transmembrane)部位等；PREDATOR 軟體[29]（宣稱準確率可到達 75%），利用support vector machine 方法預測二級結構[30]。在蛋白質側鍊預測工具，有用統計方法 SCRWL[31]、類神經網路 NETROT[32]、利用演化法預測側鍊[33]。在蛋白質結構的研究方面：蛋白質結構比對的工具，有 VAST(Vector Alignment Search Tool) [34]、CE(Combinatorial Extension of the optimal path)[35]、Dali 伺服器[36]提供比較蛋白質立體結構、SWISS-MODEL 提供自動化由蛋白質序列構建同源蛋白質結構的服務，而其所發展的 Swiss-PdbViewer[37]軟體，讓使用者有更多的空間來調控所建構蛋白質的模型，如能量最適化、threading 能量的計算、蛋白質表面電位的計算等；MaxSprout 伺服器[38]根據 Cα 徑跡(trace)構建整的蛋白質座標。立體分子繪圖軟體如 Roger Sayle 多平台的 Rasmol，繪圖速度極快，NCBI 的 Cn3D 能夠與 NCBI 的 MMDB 網路連接，做結構比對。在產業界，生物資訊對於新藥的研發與生物晶片結果的分析極為重要。誰先從基因圖譜淘出新的基因寶藏，誰就是贏家。無怪乎，人稱今日有所謂的“生物資訊淘金熱“。

三、結論

現在生物相關資料大量累積，但是由於現階段軟體工具及硬體的不足，一般生物學家電腦資訊及數理統計的訓練不夠，不少生物學家面對這些大量資料的累積，卻沒有能力利用這些資訊。因此生物資訊及計算生物的發展與教學更加迫切。生物資訊是個跨生物，電腦資訊與數學統計的領域。但是現有生物系所的課程無法涵蓋。大學已經體認到生物資訊的重要，如交通大學成立生物資訊研究所，有些學校（如陽明，清華，交通）也成立生物資訊學程。為促進生物資訊的教育與研究，台灣的生物資訊學會也已成立。由於將來未來生物科技必朝向跨領域的方向，結合理化，資電，工程，甚至人文社會，以生命科學為中心的研究。因此現代生物學的教育，必須反映出現代生命科學發展的趨勢，生科學生須要接受更多跨領域的訓練，非生科學生亦須認識生物學（分子層次）的內容。現代生物科技突飛猛進，而生物資訊為未來生物學的”絕對必要條件”(*sine qua non*)，因此政府對於台灣生物資訊教育與研究的的支持實在是刻不容緩。

參考文獻

[1] HGP 與 Celera 原本計畫一起發表在自然雜誌，但是後來談不妥，Celera 自行投稿

科學雜誌發表。見 http://www.sciencemag.org/cgi/content/full/291/5507/1195

[2] International Human Genome Sequencing Consortium, *Nature*, **409**, 860 (2001).

[3] 撇開預測的基因數不算，對於確定是基因的數目，Celera 估計為 26383，HGP 為 24500。

[4] 人類的一個基因，平均可表現三個蛋白質。

[5] J. L. Weber and E. W. Myers, *Genome Research*, **7**, 401 (1997); 參見 http://www.genome.org/cgi/content/full/7/5/401

[6] Celera 提供了一很好的非專業性介紹「散彈槍方法」的文章，詳見 http://www.celera.com/genomics/news/articles/03_00/assemble_genome_3_24.cfm

[7] http://www.research.ibm.com/bluegene/

[8] S. J. Spengler, *Science*, **287**, 1221(2000).

[9] http://www.ensembl.org/genome/central, http://www.ncbi.nlm.nih.gov/genome/central, http://genome.ucsc.edu/

[10] ENSEMBL 是由 EMBL-EBI 與 Sanger Center 共同發展的軟體系統。網址為 http://www.ensembl.org

[11] 對 HMM 的介紹，可見 R. Durbin, S. Eddy A. Krogh, G. Mitchison, Biological sequence analysis, *Cambridge University Press*, 1998.

[12] http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.1.html

[13]http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html

[14] http://srs.ebi.ac.uk/

[15] E. Birney, A. Bateman, M. E. Clamp and T. J. Hubbard, *Nature*, **409**, 827 (2001)

[16] Burge, C. and Karlin, S. J. Mol. Biol. **268,** 78 (1997)

[17]http://grail.lsd.ornl.gov/grailexp/references.html

[18] E. Birney and R. Durbin, *Genome Res*., **10**, 547 (2000).

[19] M. G. Reese, D. Kulp, H. Tammana and D. Haussler, *Genome Res.*, **10**, 529 (2000)

[20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol*., **215**, 403(1990).

[21]http://evolution.genetics.washington.edu/phylip.html

[22] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. USA*, **85**, 2444 (1988).

[23] T. F. Smith and M. S. Waterman, *J. Mol. Biol*., **147**, 195 (1981).

[24] J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, **22**, 4673 (1994); http://www.ebi.ac.uk/clustalw/

[25] J. Schultz, F. Milpetz, P. Bork and C. P. Ponting, *Proc. Natl. Acad. Sci. USA*, **95**, 5857(1998); J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting and P. Bork, *Nucleic Acids Res*., **27**, 229 (2000)

[26]http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps

[27] http://flag.itri.org.tw/

[28] http://www.embl-heidelberg.de/predictprotein/predictprotein.html

[29] http://www.embl-heidelberg.de/argos/predator/predator_info.html

[30] C. J. Lin (private communication), S. S. Sua and Z. Sun, *J. Mol. Biol*., **308**, 397 (2001).

[31]http://www.fccc.edu/research/labs/dunbrack/scwrl/

[32] J.-K. Hwang and W. F. Liao, *Protein Engineering*, **8**, 363 (1995).

[33] J. M. Yang (private communication).

[34]http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

[35] http://cl.sdsc.edu/ce.html

[36] http://www.ebi.ac.uk/dali/

[37] http://www.expasy.ch/spdbv/

[38] http://www.ebi.ac.uk/dali/maxsprout/