生物資訊



生物資訊概述

什麼是生物資訊？

生物資訊學（bioinformatics）係指利用應用數學、資訊學、統計學和計算機科學的方法研究生物學的問題。目前的生物資訊學基本上只是分子生物學與資訊技術（尤其是網際網路技術）的結合體。生物資訊學的研究材料和結果就是各種各樣的生物學數據，其研究工具是計算機，研究方法包括對生物學數據的搜索（收集和篩選）、處理（編輯、整理、管理和顯示）及利用（計算、模擬）。目前主要的研究方向有：序列比對、基因識別、基因重組、蛋白質結構預測、基因表達、蛋白質反應的預測，以及建立進化模型。

生物學技術往往生成大量的**嘈雜數據**。與**數據挖掘**類似，生物資訊學利用數學工具從大量數據中提取有用的生物學資訊。生物資訊學所要處理的典型問題包括：重新組裝在**霰**(ㄒㄧㄢˋ)**彈槍定序法**測序過程中被打散的DNA序列，從**蛋白質**的**胺基酸**序列預測蛋白質結構，利用mRNA**微陣列**或**質譜儀**的數據檢驗**基因調控**的假說。

某些人將**計算生物學**作為生物資訊學的同義詞處理；但是另外一些人認為計算生物學和生物資訊學應該被當作不同的條目處理，因為生物資訊學更側重於**生物學領域中計算方法的使用和發展**，而計算生物學強調**應用資訊學技術對生物學領域中的假說進行檢驗，並嘗試發展新的理論**。

生物資訊學的發展

生物資訊學是探討有關生物資訊的學問，主要是源自於電腦在分子生物科學之應用，於1980年代中期開始漸漸成形。事實上，生物資訊學之發展與分子生物科學及網路科技是密不可分同步進行的。自1970年代基因重組技術問世後，開始了現代分子生物科學的蓬勃發展。到1980年代陸續成功之生物科技產業，更為人類帶來了莫大的福祉與震憾。1990年代開始之人類基因體計劃，則將生命科學推向更廣的角度，並產生了大量的DNA序列資訊。有鑒於生物科技快速發展所累積之大量資訊需要解析，美國國家衛生院NIH於1982年建立了GenBank，初期由Los Alamos National Lab運作，1987年轉由IntelliGenetics經營。NIH亦於1984-1989委託IntelliGenetics成立Bionet，將生物資訊網路化。

然而直到1988年建立了NCBI（National Center for Biotechnology Information），才得以更穩健的推展生物資訊。在同一時期，網際網路亦平行發展起來。自1972年因軍事需求建立了ARPAnet (由國防現代化計畫推動)。1979年進一步發展成用於新聞傳播之USEnet。經過1982年之TCP/IP及1987年之NSFnet等原為科學用之網路系統，漸漸發展成為如今更為普及且快速之WWW網路系統。可以說，生物資訊學是乘著分子生物科學及網際網路這兩個強有力的翅膀而起飛的。若是捨去其中之一，生物資訊學都不會像今天這般熱門且後勁十足。

生物資訊的特性與內容

生物資訊學之主要工作可分為三大部份：

(1)生物資料庫及網站之建立

(2) 生物資訊的搜尋

(3) 生物資訊的分析

其中值得注意的是，除了生物資訊原始數據之取得以外，生物資料庫如何建立，搜尋與分析方法的設計原理等，均為資訊科學之專業領域，但其目的在於**快速解析生物學上複雜的種種問題**。

生物資訊的內容型態繁多，與其他資訊大不相同。生物資訊可略分為四類 (一)有關生物之結構、形態、顏色等巨觀及微觀之資訊。這些傳統生物學之資訊對於生物分類 演化與欽緣關係之比較研究很有助益。這類資訊的發展亦促成了數位化標本館的實現，將許多不易維持的實體標本，以數位圖形資訊加以取代。 (二)生物遺傳物質DNA及基因體 (生物所具有之遺傳物質之總和，稱為基因體 genome) DNA序列及其特性的資訊。在序列資訊中，以基因註解及尋找致病基因最受重視。此外，基因splicing之模式，基因調控機制之結構，亦都是分子生物學的核心問題。(三)生物大分子如蛋白質及碳水化合物結構與特性之資訊。蛋白質是扮演生命現象的要角，其結構與功能之解析一向是生物化學的重心。如何從序列資訊預測蛋白之二級結構，motif及domain，甚至3D結構，都期望能由生物資訊軟體加以解析。 (四)其他有關生物之生化、生理、遺傳、演化等之特性。其中生物的各種生化反應途徑(pathway)是生物化學的基礎。其比較分析則有助於基因解碼及比較基因體學之研究。基因之變異與各種疾病關係之研究，則為分子醫學及醫學檢驗的基礎與研究重點。事實上，現代生物資訊學所專注的生物資訊以DNA及蛋白質為主。其中，所有生物的遺傳資訊均以包含核酸分子ACGT之序列編碼，而蛋白質之資訊除了氨基酸之序列外，亦包含有3D之結構，已經不只是平面資訊，而是立體資訊了。

生物資訊主要研究方向

可分為10種研究方向

(一)序列分析

1977年，噬菌體Φ-X174成為第一個被完整測定**基因組序列**的生物體。自此以後，越來越多生物體的DNA序列被人類測定。通過對這些序列的分析，人們希望獲知其中**對應蛋白質編碼的基因**和**基因調控序列**。不同物種間的基因比對既能夠解釋和預測他們蛋白質的功能的相似性，又能夠揭示不同物種間的聯繫。由於數據量巨大，依靠人工分析DNA序列早已變得不切實際，這使得人們不得不採用計算機分析數千種生物體的數十億個核苷酸組成的DNA序列。由於DNA序列中普遍存在**變異現象**，這些計算機程序需要識別大量相關但是不完全相同的序列。即便是在DNA測序的過程中，也存在著不確定的因素。在DNA測序的**霰彈槍定序法**（The Institute for Genomic Research依此技術測定了第一個細菌的基因組）中，完整的DNA鏈被打散為成千上萬條長約600到800個核苷酸的DNA片段。這些DNA片段的兩端相互重疊，只有依照正確的順序組合，才能還原為完整的DNA序列。對於較大的基因組，霰彈槍定序法能夠迅速的測定DNA片段的序列，但將它們組裝起來的工作則相當複雜。在人類基因組計劃中，該基因重組過程花費了幾個月的CPU時間（on a circa-2000 vintage DEC Alpha computer）。由於現今幾乎所有基因序列均由霰彈法測定，**基因重組**演算法是資訊生物學研究的重點課題。

基因組中並不是所有的核苷酸都構成基因，所以序列分析的另一個研究課題是對基因組中的**基因**和**基因調控序列**進行自動識別。在較高等的生物體中，DNA序列的大部分並沒有明顯的作用。但是，這些所謂的「**垃圾**DNA」卻可能具有未被識別的功能。

【基因組注釋】

在基因組學中，**對基因和其他生物特徵的標註**稱為**基因組注釋**。1995年，Owen White（在1995年完成的人類首次為獨立的生物體——流感嗜血桿菌——測序的工作中，他是組員之一）設計出了第一套基因組注釋軟體系統。該系統能夠自動識別基因、轉錄基因和其他生物學特徵，並能夠初步的分析它們的功能。大多數現今的注釋系統的與之類似，但用以分析DNA序列的軟體在不斷進化之中。

在尋找基因和致力於發現新蛋白的努力中，人們習慣於把新的序列同已知功能的蛋白序列作比對。由於這些比對通常都希望能夠推測新蛋白的功能不管它們是雙重比對還是多序列比對，都可以回答大量的其它的生物學問題。舉例來說，**面對一堆搜集的比對序列，人們會研究隱含於蛋白之中的系統發生的關係，以便於更好地理解蛋白的進化。人們並不只是著眼於某一個蛋白，而是研究一個家族中的相關蛋白，看看進化壓力和生物秩序如何結合起來創造出新的具有雖然不同但是功能相關的蛋白**。研究完多序列比對中的高度保守區域，我們可以對蛋白質的整個結構進行預測，並且猜測這些保守區域對於維持三維結構的重要性。

顯然，分析一群相關蛋白質時，很有必要了解比對的正確構成。發展用於多序列比對的程序是一個很有活力的研究領域，絕大多數方法都是基於**漸進比對**（progressive alignment）的概念。漸進比對的思想依賴於使用者用作比對的蛋白質序列之間確實存在的生物學上的或者更準確地說是系統發生學上的相互關聯。不同演算法從不同方面解決這一問題，但是當比對的序列大大地超過兩個時（雙重比對），對於計算的挑戰就會很令人生畏。在實際操作中，演算法會在計算速度和獲得最佳比對之間尋求平衡，常常會接受足夠相近的比對。不管最終使用的是什麼方法，使用者都必須審視結果的比對，因為再次基礎上作一些手工修改是十分必要的，尤其是對保守的區域。

(二)計算進化生物學

進化生物學研究物種的起源和演化。引入資訊學到進化生物學中，使得研究者能夠：

★通過度量DNA序列的改變研究眾多生物體間的進化關係（超越了以前基於身體和生理特徵觀察的研究方法）

★通過整個基因組的比對，研究更為複雜的進化論課題，如**基因複製**，**基因橫向遷移**等

★為種群進化建立複雜的計算模型，以預測種群**隨時間的演化**

★保存大量物種的**遺傳資訊**

未來的研究工作包括重建業已相當複雜的**進化樹**。

計算進化生物學常常與採用**遺傳演算法**的計算機科學相混淆。後者受到生物進化原理的啟發，發展出一套軟體用於改進**配方**、**演算法**、**集成電路設計**等。

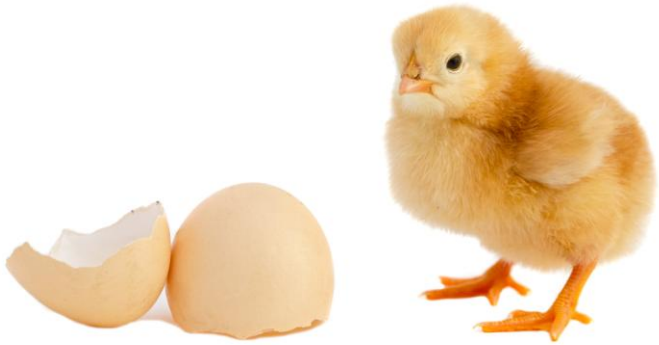
(三)生物多樣性的度量

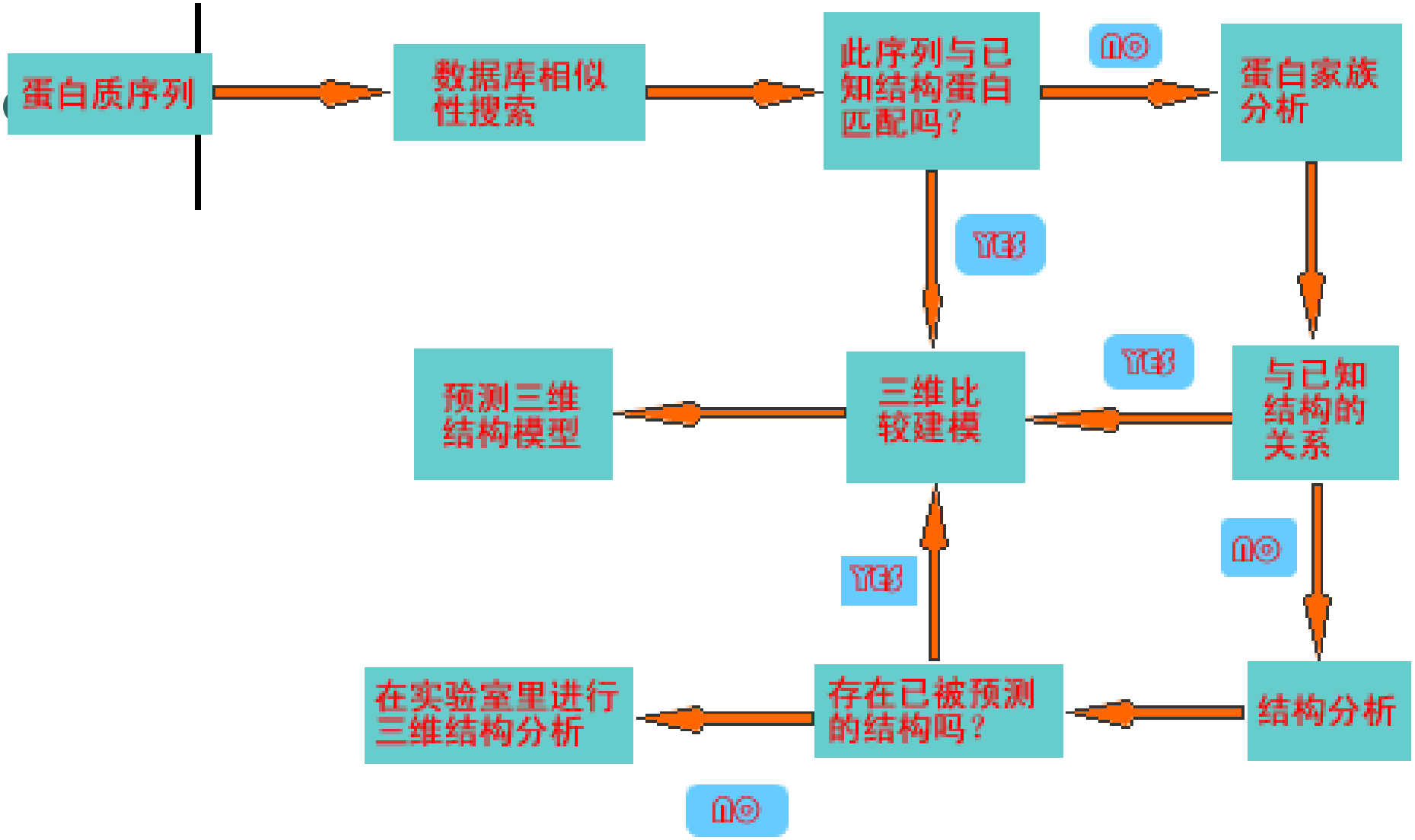
- 對一個特定的**生態系統**，小到一層生物膜、一滴海水、一鏟泥土，大到整個地球，其中全體物種的基因組成分可被定義為這個生態系統的**生物多樣性**。搜集各物種的**名稱**、**描述**、**分布**、**遺傳資訊**、**地位**、**種群大小**、**棲息地**，和**各生物體間的相互作用**等資訊，可以建立一個資料庫。有專門的軟體用於搜尋、分析和可視化這些資訊，更重要的是，它們還能夠幫助人們相互交流這些資訊。計算機能夠模擬相應的模型，以**計算種群動態演變**，**遺傳健康狀態**等。

該領域的一個重要前景是為**瀕危物種**建立基因銀行，即將各物種的基因組資訊保存下來。這樣即便在將來這些物種滅絕了，人類也可能利用它們的基因組資訊重新創造出它們。

(四)蛋白質結構預測

蛋白質結構預測是生物資訊學的重要應用。蛋白質的胺基酸序列（也稱為一級結構）可以容易的由它的基因編碼序列獲得。在絕大多數情況下，在蛋白質的原生環境中，其結構由它的胺基酸序列唯一的決定。蛋白質的結構對於理解蛋白質的功能十分重要。這些結構資訊通常被稱為二級、三級、四級結構。目前尚沒有普遍可行方案實現這些結構的準確預測；大多數方案為啟發式的。




蛋白质序列
数据库相似性搜索
此序列与已知结构蛋白匹配吗？
否
蛋白家族分析
是
预测三维结构模型
三维比较建模
是
与已知结构的关系
否
是
否
在实验室里进行三维结构分析
存在已被预测的结构吗？
结构分析
否

▲蛋白質結構預測流程圖

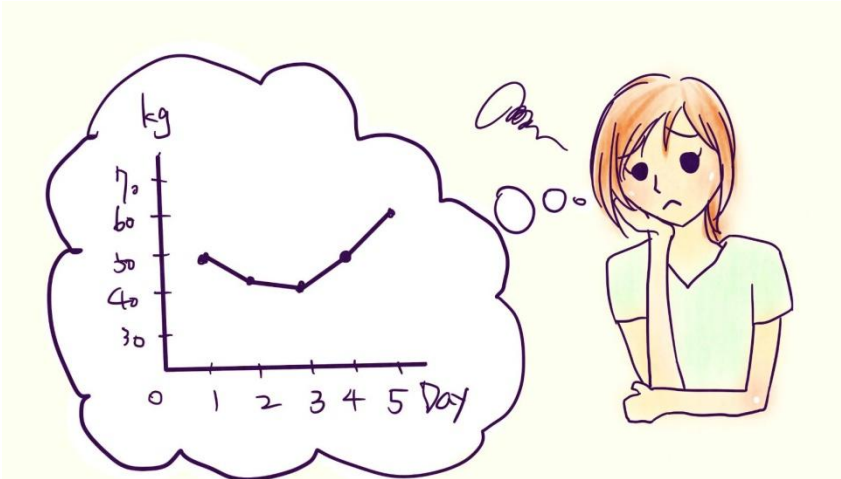
同源性是生物資訊學中的一個重要概念。在基因組的研究中，同源性被用以分析基因的功能：**若兩基因同源，則它們的功能可能相近**；在**蛋白質結構**的研究中，同源性被用於尋找在形成蛋白質結構和蛋白質反應中起關鍵作用的蛋白質片斷。在一種被稱為**同源建模**的技術中，這些資訊可與已知結構的蛋白質相比較，從而預測未知結構的蛋白質。目前為止，這是唯一可靠的預測蛋白質結構的方法。

人類**血色素**和**豆類血色素**間的相似性是以上方法的一個實例。兩種血色素有相同的功能，均能夠在各自的生物體內運輸氧氣。儘管它們的胺基酸序列大不相同，它們的蛋白質結構幾乎一樣。



(五)蛋白質表達分析

通過**蛋白質微陣列技術**或**高通量質譜分析**對生物標本進行測量所獲得的數據中，包含有大量生物標本內蛋白質的資訊。生物資訊學被廣泛的應用於這些數據的分析。對於前者，生物資訊學所面臨的問題與RNA微陣列數據分析中遇到的問題相似；對於後者，生物資訊學將所獲得的大量質譜數據與通過已知蛋白質資料庫預測的數據進行比較，並使用複雜的**統計學**方法進一步分析。



(六)比較基因組學

比較基因組學的核心課題是識別和建立不同生物體的基因或其他基因組特徵的聯繫。這些跨基因組的聯繫使得我們能夠跟蹤基因組的進化過程。基因組進化由多個不同層次的事件完成。在最低的層次，單個核苷酸上發生了點變異。在較高的層次，大的染色體片段經歷了**複製**、**橫向遷移**、**逆轉**、**調換**、**刪除**和**插入**等過程。在最高的層次上，整個基因組會經歷**雜交**、**倍交**、**內共生**等變異，並迅速產生新的物種。



(七)基因表達分析

- 多種生物學技術可以用於測量基因的表達，如DNA**微陣列**、**表達序列標籤**（expressed sequence tag），**基因表達連續分析**（serial analysis of gene expression）、**大規模平行信號測序**（massively parallel signature sequencing），多元原位雜交法（multiplexed in-situ hybridization）等。上列所有方法均**嚴重依賴於環境**並會產生大量高雜訊的數據。生物資訊學致力於發展一套統計學工具以從中提取有用的資訊。



(八)調控分析

調控是指**當細胞受到外信號刺激之後其內發生的一系列反應過程的總和**。生物資訊學技術被用於分析調控的各個步驟。例如，**基因表達的數據**可用於分析基因調控。對於一個單細胞生物體，我們可以用**生物晶片技術**觀察受到不同外界刺激、處於細胞周期不同狀態的細胞，並將採集到的數據利用**聚類演算法**分析，以尋找表達相似的基因或樣本。該結果可用於多種深入的分析。

(九)生物系統模擬

系統生物學通過對細胞子系統的計算機模擬，分析這些細胞過程間的複雜聯繫，並將分析結果可視化。例如：**人工生命**就通過計算機模擬簡單的生物形式，以幫助人類了解進化過程。



(十)藥物研發方面的應用

生物資訊學可用於**藥物靶標**基因的發現和驗證。有許多資料庫可用來獲得在不同組織在正常／疾病狀態下基因表達的差異，通過搜索這些資料庫，可以得到候選基因作為藥物靶標，特異性地針對某一種疾病。另外，還可根據蛋白質功能區和三維結構的預測來對藥物靶標進行鑒定，以便早期了解所研究蛋白的屬性，預測它是否適用於**藥物作用**。

【藥物靶標】

藥物靶標是指體內具有藥效功能並能被藥物作用的生物大分子，如某些蛋白質和核酸等生物大分子。那些編碼靶標蛋白的基因也被稱為靶標基因。事先確定靶向特定疾病有關的靶標分子是現代新藥開發的基礎。



軟體工具

BLAST（基本局部比對搜尋工具）和FASTA是目前使用得最為頻繁的兩套資料庫搜索程序。它們的功能相近，都是把*用戶提交的一個核酸序列或蛋白質序列拿去與指定的資料庫中的全部序列作比較*。一般認為，BLAST運行速度*快*，對蛋白質序列的搜尋更為有效。FASTA運行較*慢*，對核酸序列更為敏感。也有基於網頁交互的軟體如STING，用於結構生物資訊學的分析。



臺灣的生物資訊

臺灣的生物資訊是一門跨領域且具整合性之新學科領域，它是綜合**生物學科領域**與**資訊學科領域領域**所發展出來的。由於分子生物學快速進展和基因組技術的研發，產生了許多生物相關的大量資料，這些資料需要電腦協助儲存、分析及搜尋。另一方面，資訊科技近年來突飛猛進，藉由資訊科技將生物科技的相關資料作系統性的整理儲存，進而經由資料掘礦（data mining）技術將已知生物間的基因或蛋白序列資料作比較分析，可演繹出彼此功能間的關係。最終的目標是發現新的生物認知，進而建立生物系統的概念，以辨識生物學的各項準則。

我國生物資訊現階段的主要發展方向，分為：

(1)**基因體醫學、蛋白體學、功能基因體學、比較基因體學**：包括**人類遺傳疾病基因定位方法研發、生物晶片設計、蛋白體的研究、反應路徑的研究、比較基因體學的研究、基因體醫學之研發與應用**等。

(2)發展生物**資訊分析工具以及應用軟體**：包括**演算法、生物資訊知識管理系統、基因體醫學方面之分析工具**及**應用軟體**等。

(3)建立及**維護生物資訊**各類資料庫：包括**建立國人重要疾病之基因醫藥資料庫、臺灣遺傳資訊資料庫、生物資訊加值資料庫**等各類型整合性資料庫。

心得

生物資訊為我們生活增添許多便利，不管是在任何方面都有許多貢獻。☺

生物資訊在我們的生活中息息相關～～它可以用在醫療、食品、農產品等有許許多多的用途。

我覺得生物資訊學的發展有助於現有網路生物資源及生物資訊之取得與利用，對未來的發展有很大的幫助。

資料來源

※維基百科：

http://zh.wikipedia.org/wiki/%E7%94%9F%E7%89%A9%E8%B3%87%E8%A8%8A#.E5.9C.A8.E8.8D.AF.E7.89.A9.E7.A0.94.E5.8F.91.E6.96.B9.E9.9D.A2.E7.9A.84.E5.BA.94.E7.94.A8

※臺灣大百科全書：

http://taiwanpedia.culture.tw/web/content?ID=26329



The End of The Report

Thanks for your watching!



