出國報告（出國類別：進修）

美國系統生物研究中心進修生物資訊及系統醫學報告

服務機關：台北榮總
姓名職稱：張家銘 主治醫師
派赴國家：美國
出國期間：102-12-21 to 103-6-19
報告日期：103-7-19

摘要

目前定序技術已經可以在數日之內完成人類基因體三十億個鹼基定序工作，而且預測定序的價格將降至 1000 美金，所以可以預期基因體資訊將逐漸普及化，並大量運用於臨床上，此資訊對人類疾病預測、診斷及個人化醫療有相當重要的價值。但是這基因體資料的數量龐大，總類及結構複雜，基因與基因之間交互作用形成複雜的生物網路，分析這些資料需要用到複雜的資訊科學技術，國內從事相關研究的人不多，這是目前最亟需解決的關鍵點。此次出國學習生物資訊及系統生物醫學技術，這兩種技術正是分析這些資訊方法。研究過程中並進一步發展基因組醫學(geneset-based medicine)模式，此技術可將基因體資料模式化並以電腦分析，達成疾病分類及預測的目的。

關鍵字: 系統生物、生物資訊、系統醫學、全基因體定序

目次

一、 目的

人類完整的基因體序列已經於西元 2000 年完成解碼，之後進入後基因體時代。而最新高速基因定序技術，例如 next generation sequencing 的發展，使得近年來全基因體定序的速度以對數倍速成長，但價格卻下降至一般人可以負擔的水準，預計在 2014 年後全基因體定序的價格將下降至 1000 元美金。後基因體時代努力的方向放在解讀這些基因資訊，並進一步用於預測及治療疾病上。將來發展的趨勢，將是近一步利用生物資訊的方法分析這些基因體的訊息，並根據每個人不同基因序列訊息設計因人而異的治療方式，達成個人化醫療的目的。所以當務之急，是學習分析及運用這些基因資訊的方法。因為基因體的表現受到基因序列、 基因與基因、基因與環境之間相交互作用的調控形成複雜的網路系統，所以目前研究基因體都由以統合研究為架構的系統生物的方向著手。美國系統生物學研究中心(Institute for Systems Biology, ISB)設在西雅圖，是專門發展系統生物研究及個人化醫療的研究機構，它始建於 2000 年，其組織架構是究採取跨學科研究，整合生物學、遺傳學、資訊科學、化學、工程、數學、免疫學、健康和環境的系統生物學不同領域的專家。研究方向著重在基因，蛋白質、其他分子及環境如何控制生物體的機制。這整合性的研究架構將有助於發展個人化醫療、預防及治療疾病，是目前發展系統生物的重鎮。

二、 過程

在 ISB 學習的過程如下 ：

I. 學習程式語言

在進入 ISB 後，首先學習並熟悉電腦程式的撰寫。因為基因體的資料龐大，無法單以人腦及肉眼分析，所以需要借助電腦的計算。主要學習的程式語言為 R 及 Matlab。R 是目前廣泛使用的數學統計計算軟體，因為免費及開放源碼，擁有龐大的支援套件。Matlab 是 ISB 主要使用的程式語言，主要用於工程數學計算，模式分析及模擬。之外也學習 Python。Python 是功能相當完整的高階程式語言，其用途不止限於數學計算，

對於文字分析、網路運用都有強大的計算能力。

II. 學習全基因序列分析軟體

主要使用 Annovar 及 snpEff，這兩者的用途是基因體序列的詮釋(annotation)。基因詮釋是分析基因體序列過程中對臨床醫師而言最重要的部分，Annotation 指在發現基因序列異常點之後，進一步分析此序列異常對蛋白質的影響，及預測其嚴重程度。有了這些資料，就可以針對單基因疾病對病患解釋序列異常處及其影響。因為 Annovar 及 snpEff 都屬於 Linux 系統作業軟體，所以我也同時學習 Linux 作業系統的使用。

III 學習其他生物資訊相關技術

1. DNA microarray 分析：包括 normalization， background correction，filtering 及分析表現差異基因，並用 GO(gene ontology)及 KEGG pathway 詮釋基因異常。
2. 生物資料庫的使用：包括 GEO， RECON-X， 1000 genomes， GO，reactome, ，KEGG etc。
3. Biological networks (包括 signaling network，regulatory networks，protein-protein interaction 等) ，及其分析計算軟體(cytoscape)。
4. Machine learning：Machine learning(機器學習，包括 kNN、neural network、naive Bayes、ecision trees、regression tree、SVM、SOM、clustering) 是利用電腦運算式，針對輸入的資料，找出其中的規則，將資料分類及預測，是目前資訊科學用於型態辨識重要的工具之一，在此我將 machine learning 運用於疾病的分類(詳見 V)。

IV.學習 metabolic network

我在 ISB 的指導教授是 Nathan Price，他主要專門研究代謝網路。代謝網路主要研究細胞及生物所有代謝物產生路徑及其相關基因及蛋白質，代謝網路是目前所有生物網路中模式化最完全的生物網路，這代表代謝網路可以用數學模式預測生物體代謝物的產生途徑，在正常及在基因異常時的產量產物的變化，例如肝細胞發展成肝癌時，肝細胞代謝物的變化，變化明顯的代謝物所以可以用於發展生物標記；也可以用於，癌就在不同基因發生異常時所發生的代謝變化，進一步預測對生物影響。

V. 研究及發展基因組模式醫學(geneset-based model medicine)

特定生物體的功能大部分是由一群，而非單一基因所負責，例如 KEGG pathways database 中收錄的 pathways 就是其中的代表。所謂基因組模式醫學是指研究在一個疾

病中，是因那些功能異常所引起，而有哪幾組基因的異常造成這些功能的變化，所以是以一組基因，而非單一基因為單位，研究疾病。

早期研究疾病，都針對單一基因研究，但單基因疾病只適用於少數罕見疾病，常見的疾病，例如癌症、糖尿病等，都是多基因疾病，目前醫學研究的重點放在多基因疾病，多基因疾病因為是基因與基因之間交互作用形成，致病機轉相當複雜，研究時需要分析大量的基因體資料。ISB 之前已經發展出 DIRAC (Differential Rank Conservation) 運算式，此運算式可以根據基因組表現排序，並算出基因組規律係數(ranked conservation index) ，此係數代表這個基因組或 pathway 表現正常與否。我進一步修改並利用此運算式，經由分析下載自 NCBI GEO database 數百到數千個案例不同疾病及正常組織之間基因表現，在計算出特定疾病的不同基因組規律係數之後，用機器學習程式(machine learning) 藉由電腦分析、歸類及預測疾病。以肝癌為例，在計算基因表現晶片數萬個基因表現的資料後，預測肝癌的正確率可達 93.4% (見附錄)。這技術將可以用於臨床預測，同時我正在撰寫這研究結果，將發表於論文。

三、 心得

在進入後基因體時代，主要的技術關鍵是如何分析這些龐大的基因體資訊，在分析這些資料時，要考慮的因數相當多，除個單一基因的功能及其表現之外，還必須考慮基因-基因之間的交互作用，而基因體資訊的種類相當複雜，除了序列的資料外，還有基因表現、基因調控、蛋白質表現的資料要同時列入考慮。系統生物醫學的工作即是將這些不同形式的資料統合，同時納入單一模式中加以計算，以求得診斷或預測疾病的效果。這工作主要依靠電腦的運算，經由適當的演算式，可以將基因體資料中隱藏的重要資訊找出，並進一步運用於個人化醫療及臨床上。

四、 建議事項

1. 系統生物及生物資訊都是與數學及程式設計高度相關的學門，所以如果選擇學習這相關的學問，建議出國前要將相關的知識先學好，並達到一定的熟悉程度，以免出國後還要花時間熟悉相關軟體的操作，浪費時間。
2. 人類基因體序列及金因表現的資料數項相當龐大，所以在進行相關計算時，所需的硬體設備要求也相對提高，建議出國時自己攜帶的電腦要注意是否合乎硬體需求。

五 附錄

1. 以基因組模式(gene set based model)分析 698 個肝癌患者及 585 個控制組基因表現數據(DNA microarray), 藉由 DIRAC 統合 4 種不同 DNA microarray 平台(GPL10687, GPL570, GPL96, GPL3921) 計算其基因組規律係數, 經 SOM(self organizing maps, 一種 neural network 機器學習演算式) 重複 10 次分類及預測的結果, 正確率平均為 93.4%。

	Sensitivity	Specificity	Accuracy
1	0.907143	0.972222	0.940141
2	0.943038	0.936508	0.940141
3	0.895105	0.985816	0.940141
4	0.924658	0.927536	0.926056
5	0.899329	0.955556	0.926056
6	0.928571	0.938462	0.933099
7	0.905405	0.933824	0.919014
8	0.910345	0.964029	0.93662
9	0.911392	0.952381	0.929577
10	0.980392	0.916031	0.950704
Average =	**0.920538**	**0.948236**	**0.934155**

2.

將上述肝癌基因組模式數據以 GO (gene ontology)詮釋，顯示肝細胞在變成癌症後主要在代謝功能方面出現異常 (Pathway 為有統計顯著意義之 GO terms, adjPvalue 為用 Benjamini–Hochberg procedure 校正後之 p 值)

Pathway	meanDisR	meanConR	_differenc	RejectHo	p_value	goid	adjPvalue
SUGAR_BINDING	0.726934	0.847337	-0.1204	1	1.28E-210	GO:000552	1.81E-207
CARBOHYDRATE_BINDING	0.762928	0.852631	-0.0897	1	7.41E-199	GO:003024	5.24E-196
CHROMOSOME_CONDENSATION	0.659963	0.862365	-0.2024	1	4.67E-198	GO:003026	2.20E-195
MICROTUBULE	0.734091	0.835961	-0.10187	1	1.50E-194	GO:000587	5.30E-192
DNA_MODIFICATION	0.714111	0.891354	-0.17724	1	3.65E-188	GO:000630	1.03E-185
REGULATION_OF_CYCLIN_DEPENDENT_PROTEIN_KINASE	0.784251	0.869054	-0.0848	1	5.02E-187	GO:000007	1.18E-184
CHROMOSOME_SEGREGATION	0.772026	0.865199	-0.09317	1	4.55E-185	GO:000705	9.19E-183
MITOTIC_CELL_CYCLE	0.787274	0.864961	-0.07769	1	3.23E-179	GO:000027	5.71E-177
REGULATION_OF_GENE_EXPRESSION_EPIGENETIC	0.760871	0.844246	-0.08338	1	5.79E-179	GO:004002	9.10E-177
ISOMERASE_ACTIVITY	0.783073	0.863204	-0.08013	1	4.85E-177	GO:001685	6.87E-175
CHROMOSOME	0.809858	0.879164	-0.06931	1	8.93E-173	GO:000569	1.15E-170
MITOTIC_CELL_CYCLE_CHECKPOINT	0.726368	0.866038	-0.13967	1	1.23E-171	GO:000709	1.45E-169
MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BI	0.793152	0.873751	-0.0806	1	4.20E-171	GO:000022	4.57E-169
REGULATION_OF_MITOSIS	0.741523	0.838703	-0.09718	1	1.35E-167	GO:000708	1.28E-165
SPINDLE	0.76834	0.862874	-0.09453	1	1.30E-167	GO:000581	1.28E-165
DEOXYRIBONUCLEASE_ACTIVITY	0.807422	0.895853	-0.08843	1	8.92E-165	GO:000453	7.89E-163
CHROMOSOME_ORGANIZATION_AND_BIOGENESIS	0.810532	0.86189	-0.05136	1	2.95E-164	GO:005127	2.46E-162
CHROMOSOMAL_PART	0.808437	0.881162	-0.07272	1	4.91E-161	GO:004442	3.86E-159
REGULATION_OF_CELL_CYCLE	0.796499	0.859448	-0.06295	1	4.33E-160	GO:005172	3.23E-158
CELL_CYCLE_GO_0007049	0.802837	0.864568	-0.06173	1	1.61E-159	GO:000704	1.14E-157
MICROTUBULE_ORGANIZING_CENTER_PART	0.714557	0.817749	-0.10319	1	3.03E-159	GO:004445	2.04E-157
CELL_CYCLE_CHECKPOINT_GO_0000075	0.77722	0.874319	-0.0971	1	1.49E-157	GO:000007	9.60E-156

3. 如上，進一步將 GO terms 做相關分析， 顯示肝細胞癌代謝功能的異常及其相關基因路徑。