PIMA:建立一個 cDNA 微陣列和寡核苷酸微陣列的整合平台網站

PIMA:Building a web-based Platform for Integration of cDNA and oligo Microarray Data

陳彥臣、陳宇瑄、邱泓文、李元綺
Yen-Chen Chen, Lillian Y. Chen, Hung-Wen Chiu, Yuan-Chii Gladys Lee

臺北醫學大學醫學資訊研究所
Graduate Institute of Medical Informatics, Taipei Medical University

摘要

目前網際網路上，有許多公開免費的微陣列數據資料庫可供人研究。因技術不同分為 cDNA 和寡核苷酸微陣列，整合兩種微陣列的數據，可增加對實驗結果的可信度，得到更完整的資訊。PIMA 是一個線上的微陣列數據整合平台，提供 cDNA 微陣列和寡核苷酸微陣列基因表現的對照資料，利用 LocusLink 為 Identifier 建立微陣列的資料建立關聯性，自動化定期更新資料表，幫助我們減少對微陣列數據作資料探勘的繁雜步驟。

關鍵字：微陣列、寡核苷酸、整合平台、資料探勘、PIMA, cDNA, affymetrix.

Abstract

Nowadays, many microarray databases can be accessed freely for research on the Internet. Since microarray can be categorized into either cDNA or oligo microarray, it is important to integrate the two different types of datasets in order to increase the reliability of the experimental results. PIMA is a newly developed microarray integration platform that will provide comprehensive gene expression information on both cDNA and oligonucleotide microarray via LocusLink as the identifier to build microarray relationship. Automatically update of the microarray data information will also reduce the workload in performing microarray data mining process.

Keyword: *Microarray, cDNA, oligo, affymetrix, data mining, integration.*

Introduction:

微陣列生物晶片(Microarray)技術，是種能同時測量數千個，甚至上萬個基因表現的技術，不但能增加效率，更降低了系統誤差的可能性。因此用微陣列生物晶片(Microarray)來測量各物種的基因表現，是近幾年來十分熱門的工具。尤其是在研究與疾病相關的議題上，微陣列生物晶片的大量使用，對於疾病的分類、診斷、預測皆有極大的貢獻。但是由於使用微陣列生物晶片(Microarray)的成本較高，非一般實驗室可輕鬆負擔，而且 Microarray 的實驗數據有很高的重複研究與利用的價值，因此有許多機構致力於 Microarray Data 的收集管理與制訂標準化的的資料交換格式(MIAME: Minimum information abut a microarray experiment）[1-2] ，目前公認的是以 The Microarray Gene Expression Data (MGED) Society 所制訂的 MAGE-ML (Micrarray Gene Expression-Markup Language)格式來存放 Microarray Data 為共同的標準，並且公開散佈在網際網路上，例如：EBI 的 ArrayExpress[3-4] 、 Stanford 的 SMD(Stanford Microarray Database)[5] 、 NCBI 的 GEO(Gene Expression Omnibus) [6]、密西根大學的 ONCOMINE [7] . 耶魯大學的 YMD [8] 等。

Microarray Gene Expression 測量技術，本身又分為兩種技術，一種是 cDNA Microarray ，普遍為一般 Home made 所採用，另一種是 oligonucleiotide Microarray 由 Affymetrix Corp.[9] 所研發，兩種技術的原理相同，但差別在晶片上的序列長短不同，計算表現量的方式也不同，各有優點。因此，如果能把

這兩種技術所得到的結果拿來一起比較，當兩種 Microarray 同時得到一致的基因表現的結果一致時，我們則可以更加確定該實驗的正確性；反之，兩種 Microarray 的結果不同時，我們則要考慮是否採用這個數據，進而幫助我們有效地縮小尋找重要基因的範籌。基於這個想法，我們收集並整理了類似想法的文獻資料，其中最具代表性的是 Lee 等人(2003)的研究[10]，他們利用兩種 Microarrays， cDNA and oligonucleiotide， 同時對 NCI-60 cancel cells 進行研究，將兩種 Microarrays 的結果，利用減去平均值的方法做 Normalization 之後，再做 hierarchical clustering 觀察兩種 Microarray 實驗所得到的 dendrogram pattern 是否相同，並計算兩者之間的相關係數，結果證明兩種 Microarray 的實驗數據可以得到相似的結果，具有高度相關性。

因此，為了要將兩種 Microarray 的基因表現能對在一起比較，他們使用 Bussey 等人在 2003 年所發展的 MatchMiner［11］這個工具，用來將 cDNA Microarray 每個 clone ID 所代表的 Gene 和 oligo Microarray 每個 Affy ID 所代表的 Gene 作對應，使兩種 Microarray 的數據能做比較。但是在實際上操作時，發現到有許多的 spots 無法對應。例如：Affymetrix chip 中 ID 為 1004_at 的 spot 應該要對應到 UniGene Hs.113816 這個基因，但是在 matchminer 搜尋的結果卻是無資料，事實上卻是有資料可以對應起來的，原因在於 matchminer 只建立一個簡單的基因對應雛形，無定期更新對應資料。使用其他類似功能的網站如 GeneLynx［12］，也有相同類似的問題，像是資料不完整，及沒有定期更新資料庫。因此我們發展一套 Web-Based Microarray Data Integration Platform - PIMA，用來整合研究異質或同質的 Microarray Data。

Materials and Methods:

首先我們收集各個資料庫的資料，利用 Perl/CGI, Java, Microsoft VBA program 撰寫了程式來擷取並整合了多個資料庫的資料到 MySQL Database，其中包含 Swissprot, GeneOntology, Affymetrix ID 和 NCBI 的 LocusLink 、 Unigene 及 OMIM 等，所收集的資料來源列表如（表一）。在 Bussey 等人所發展的工具中，他們所使用的方法是以 UniGene［13］為主建立 cDNA clone ID 和 oligo Affy ID 的對應關係。

表一：資料來源表

Identifier Name	Source	HyperLink
LocusLink	NCBI	http://www.ncbi.nlm.nih.gov/LocusLink/
UniGene	NCBI	http://www.ncbi.nlm.nih.gov/UniGene/
GeneBank accession number	NCBI	http://www.ncbi.nlm.nih.gov/GenBank/
cDNA Image Clone ID	LLNL	http://image.llnl.gov/
Affymetrix Probe Set ID	Affymetrix	http://www.affymetrix.com/
SAGE tag	SAGENET	http://www.sagenet.org/
Gene Ontology	AmiGO	http://www.godatabase.org/
Gene Symbol	HUGO	http://www.gene.ucl.ac.uk/hugo/
Swiss-prot	Swissprot	http://ca.expasy.org/sprot/

UniGene 是是根據複雜的 GeneBank，整理分析簡化而來，專門收集 non-redundant set 的基因來源的 clusters 數據。每一個 UniGene Cluster 包含代表單一基因的序列和相關的資訊，例如表達此基因的組織類型和圖譜定位資訊。

除了已被註解過序列以外，成千上萬的 Unknown EST 也被收錄在內。因此，這些收集的資源可以作為發現新基因的來源。現在，許多實驗室研究人員已經利用 UniGene 進行大規模的基因表達圖譜分析。所有屬於同一個基因的 splicing variants 放置在同一個集合中，由於新的序列資料不斷的加入和每星期的不斷更新，因此在 UniGene 中的 resulting clusters 也不斷的被合併，或新產生。由於 Clusters 的不穩定變化，所以使用 UniGene Cluster ID 作為 Identifier 是不明智的，因此我們採用 LocusLink 來解決 Identifier 的問題。

LocusLink［14］提供一個 single query interface 來找到某一個 genetic loci 的 sequence 和 descriptive information。資料的有系統化的正式基因名稱 (gene name)，簡稱 (symbol)，別名 (aliases)，Reference sequence，GenBank accession number，表型 (phenotypes)，EC numbers，OMIM numbers，UniGene clusters，同源(homology)，map locations，Gene Ontolgoy information 和相關的網站資訊。因此我們以 LocusLink 為主要的 Identifier，收集各個資料庫的資料，撰寫 Microsoft VBA program 抽取出我們需要的資料欄位和對應資料，最後將所有的資料統合，整理出 LocusLink, UniGene, cDNA clone ID, Affy

ID, SAGE tag, Gene Ontology, Gene Symbol, Swiss-prot etc. 彼此間的對應關係（ 表二 ），所有的基因註解的資料和對應關係都儲存在 MySQL Database 中。

我們利用表二之對應關係，以 LocusLink IDs 作為主要的 Identifier 建立資料表之間的關聯，使所有基因註解的資料能夠對應轉換，並且撰寫數個 perl program 定期自動去取得各個相關資料庫中的資料，存放到我們的資料庫中， PIMA 使用 Perl/CGI 和 Java Serve Page 結合 Apache Web Server 而建立，作業系統使用 Linux System Kernel 2.6，整個系統架構流程如（ 圖一 ）所示。

Results:

我們所發展的 cDNA 微陣列和寡核苷酸微陣列的整合平台網站 PIMA，網址 http://bio.tmu.edu.tw/PIMA/，主要的功能有：

1. 利用 LocusLink 為 Identifier 來 query signal array 中所有的基因註解資料，使用者可以選擇已存在於資料庫中的 Array 或是自行上傳 Array data，Array type 可以是自訂的或是商用的 Array chip。
2. Query 兩個 Array 以上基因的交集、聯集、差集的基因表現和註解的資料，可以是 cDNA-to-cDNA, cDNA-to-Affy 或是 Affy-to-Affy 互相比較。
3. Query array 上使用者有興趣的基因，並且和 KEGG Pathway Database 連接。
4. Query 各種基因資料的對應，如 LocusLink, UniGene, Gene Symbol, OMIM etc.
5. 可依照 LocusLink, UniGene, Affy ID 等對 Query 後的結果整理排序。
6. 自動化定期更新基因註解資料。

Discussion:

NCI panel of 60 human tumor cell lines

NCI-60 panel［15］是 National Cancer Institute 在 1990 年時所建立的，裡面包含了 60 種 cancer cell lines。到現在為止，NCI-60 cancer lines 已經被用來觀測超過十萬個和抑制細胞成長相關的化合物實驗。這些生化實驗提供每個化合物在六十個 cancer cell lines 中的實驗結果，提供了藥物學豐富的數據資訊。這些豐富的資訊，幫助了解化合物在分子層級上的特性，在 cell lines 中評估 DNA, RNA, Protein, Functional 和 Pharmacological levels 。因此，由於 NCI-60 cell lines 比其他任何的關於細胞集合資料庫的資料豐富，他們的資料庫已經成為許多實驗室中重要的資源。

Comparison of the capability with MatchMiner

PIMA 主要的目的和 MatchMiner 相同，是要幫助我們能整合分析不同的 Microarray Data。主要的差異在於 PIMA，是以 LocusLink 當主要的 Identifier，不同於 MatchMiner 使用 UniGene 為 Identifier，使用 LocusLink 有較好的規則。而且不只提供不同 Microarray 之間 match 基因完整的基因註解，也可找出不同 array 之間基因的交集、聯集和差集，以及刪除目前已知的 HouseKeeping Genes 資料，並且與 KEGG Pathway database 結合分類，提供 OMIM 的相關連結，減少使用者整理及搜尋資料的步驟。

Future direction

目前 PIMA 資料庫中只有 Human 的資料，原因是我們希望 focus 在人類疾病相關的基因上面，未來也許會增加其它物種的同源資料。此外，我們希望能加入線上分析的功能，能將經過 PIMA 整理過後的資料，直接線上做統計分析。

Conclusion:

在 PIMA 建立完成之後，透過 PIMA 來處理我們要分析的 Microarray Data，大幅減少我們所要花費的時間，原本需要分別到各個資料庫去搜尋的工作，皆可在 PIMA 之中輕鬆完成。

Reference：

1. Brazma A., Robinson A., Cameron G., Ashburner M. (2000). One-stop shop for microarray data. Nature 403, 699-700.

2. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P. et al. ,Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 29, 365-71..
3. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M et al. ArrayExpress: a public database of gene expression data at EBI. C R Biol. 2003 Oct-Nov;326(10-11):1075-8.
4. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J,. et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2003 Jan 1;31(1):68-71. [http://www.ebi.ac.uk/arrayexpress]
5. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, et al. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. 2003 Jan 1;31(1):94-6.
6. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res., 30, 207–210.
7. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform.Neoplasia. 2004 Jan-Feb;6(1):1-6.
8. Cheung,K.H., White,K., Hager,J., Gerstein,M., Reinke,V., . et al. YMD: a Microarray database for large-scale gene expression analysis. Proceedings of the American Medical Informatics Association 2002 Symposium, Hanley and Belfus, Inc., San Antonio, TX, pp. 140–144.
9. Affymetrix [http://www.affymetrix.com/]
10. Lee JK, Bussey KJ, Gwadry FG, Reinhold W, Riddick G, Pelletier SL,.. et al..Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells.Genome Biol. 2003 ; 4(12):R82. Epub 2003 Nov 25.
11. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC.. et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers.Genome Biol. 2003;4(4):R27. Epub 2003 Mar 25.

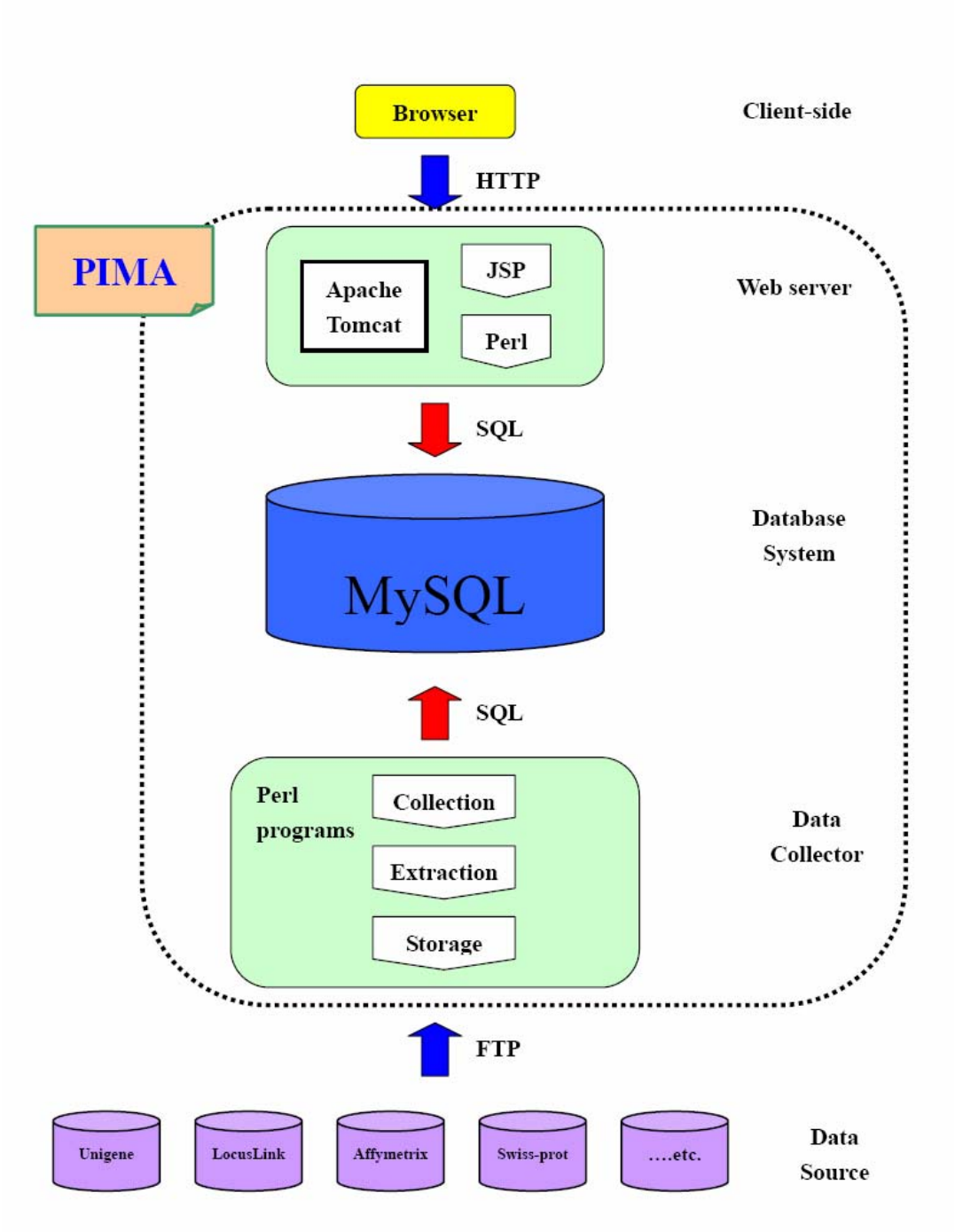
    [http://discover.nci.nih.gov/matchminer]
12. Lenhard B, Hayes WS, Wasserman WW. GeneLynx: a gene-centric portal to the human genome. Genome Res. 2001 Dec;11(12): 2151-7.
13. UniGene. [http://www.ncbi.nlm.nih.gov/UniGene/]
14. LocusLink.[http://www.ncbi.nlm.nih.gov/LocusLink/]
15. NCI panel of 60 human tumor cell lines.

    [http://dtp.nci.nih.gov/]

表二：對應關係表

	LocusLink	UniGene	Affy ID	SAGE Tage	GO NO.	OMIM	Symbol	Image Clone ID	Swissprot
LocusLink									
UniGene	M : M								
Affy ID	M : 1	M : 1							
SAGE Tage	M : 1	M : 1	M : M						
GO NO.	M : 1	M : 1	1 : M	1 : M					
OMIM	M : 1	M : 1	M : 1	M : 1	M : 1				
Symbol	1 : 1	1 : M	1 : M	1 : M	1 : M	1 : 1			
Image Clone ID	M : 1	M : 1	M : M	M : M	M : 1	M : 1	M : 1		
Swissprot	M : M	M : M	M : 1	M : 1	1 : M	1 : 1	1 : M	1 : 1	

Row v.s Column
一對一 1 : 1
一對多 1 : M
多對多 M : M
多對一 M : 1



圖一：PIMA 系統架構圖

PIMA 的 Data Collector 會自動定期去下載各個相關資料庫的資料包含 Swissprot, GeneOntology, Affymetrix ID 和 NCBI 的 LocusLink 、 Unigene 及 OMIM，經過整理之後存放到 MySQL 資料庫中。使用者利用瀏覽器對 PIMA 進行資料搜尋的動作，PIMA 使用 JSP,Perl program,MySQL Database System 建置。