Can we interpret biological data by computational means?

Ju Han Kim, M.D., Ph.D.
juhan@snu.ac.kr

Integrative biochip informatics

- Exploration: mapping and clustering
- Pattern recognition: finding gene – drug assoc.
- Prognostic subgroup prediction
- Interpretation
  - ✓ GRIP: automatic probe annotation
  - ✓ Xperanto: databasing expression
  - ✓ BioCANDI: database-integrated analysis tool
  - ✓ ChromoViz: visualizing regional regulation
  - ✓ GOTree and GOChase: GO-based annotation
  - ✓ ArrayXPath: pathway-based interpretation
  - ✓ PathMeSH: disease - gene - pathway
- BioEMR: towards real-time clinical trials

Integrative biochip informatics

- Exploration: mapping and clustering
- Pattern recognition: finding gene – drug assoc.
- Prognostic subgroup prediction
- Interpretation
  - ✓ GRIP: automatic probe annotation
  - ✓ Xperanto: databasing expression
  - ✓ BioCANDI: database-integrated analysis tool
  - ✓ ChromoViz: visualizing regional regulation
  - ✓ GOTree and GOChase: GO-based annotation
  - ✓ ArrayXPath: pathway-based interpretation
  - ✓ PathMeSH: disease - gene - pathway
- BioEMR: towards real-time clinical trials

Multi-dimensional Scaling

Table 1 Flying Mileages Between 10 American Cities

Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington, DC	
0	587	1212	701	1936	604	748	2139	2182	543	Atlanta
587	0	920	940	1745	1188	713	1858	1737	597	Chicago
1212	920	0	879	831	1726	1631	949	1021	1494	Denver
701	940	879	0	1374	968	1420	1645	1891	1220	Houston
1936	1745	831	1374	0	2339	2451	347	959	2300	Los Angeles
604	1188	1726	968	2339	0	1092	2594	2734	923	Miami
748	713	1631	1420	2451	1092	0	2571	2408	205	New York
2139	1858	949	1645	347	2594	2571	0	678	2442	San F
2182	1737	1021	1891	959	2734	2408	678	0	2329	Seattl
543	597	1494	1220	2300	923	205	2442	2329	0	Wash

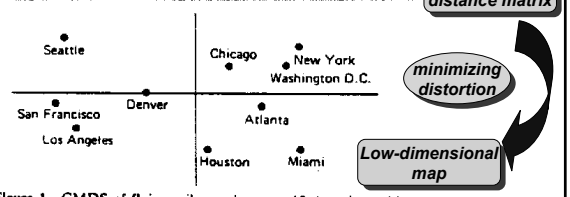


Figure 1 CMDS of flying mileages between 10 American cities.

Young, 1985. Encyclopedia of Statistical Sciences

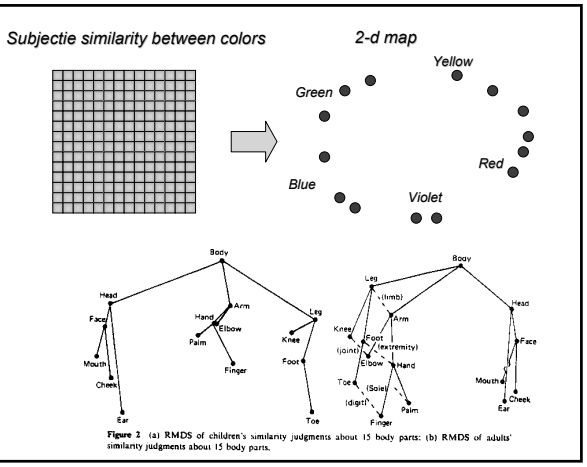
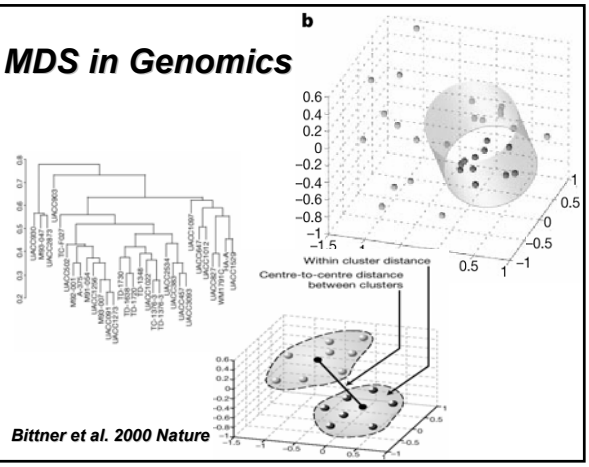


Figure 2 (a) RMDS of children's similarity judgments about 15 body parts; (b) RMDS of adults' similarity judgments about 15 body parts.

MDS in Genomics



Bittner et al. 2000 Nature

Temporal patterns revealed by MDS

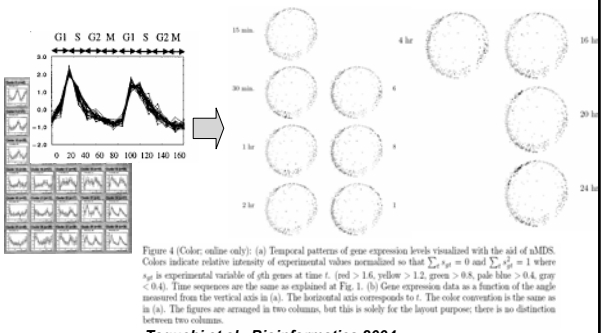
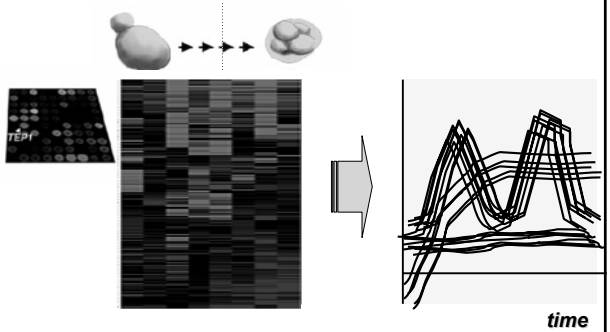


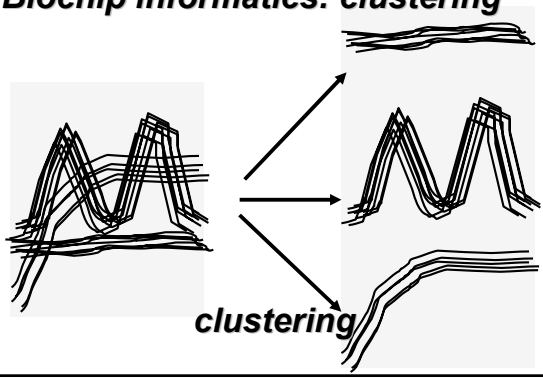
Figure 4 (Color: online only): (a) Temporal patterns of gene expression levels visualized with the aid of nMDS. Colors indicate relative intensity of experimental values normalized so that $\sum_t x_{gt} = 0$ and $\sum_t x_{gt}^2 = 1$ where x_{gt} is experimental variable of gth genes at time t. (red > 1.6, yellow > 1.2, green > 0.8, pale blue > 0.4, gray < 0.4). Time sequences are the same as explained at Fig. 1. (b) Gene expression data as a function of the angle measured from the vertical axis in (a). The horizontal axis corresponds to t. The color convention is the same as in (a). The figures are arranged in two columns, but this is solely for the layout purpose; there is no distinction between two columns.

Taguchi et al., Bioinformatics 2004

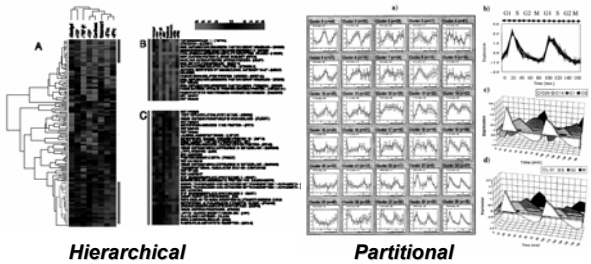
Biochip informatics: clustering



Biochip informatics: clustering



Hierarchical & Partitional Clustering

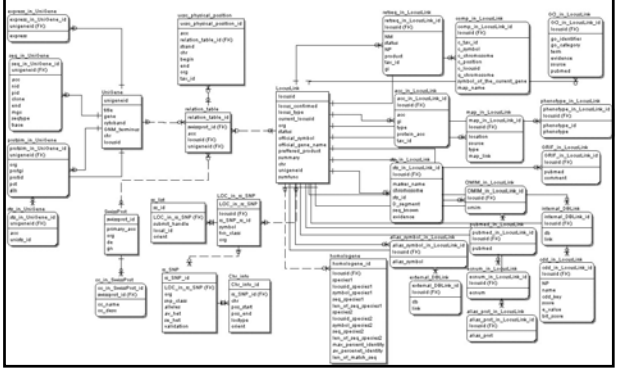


GRIP

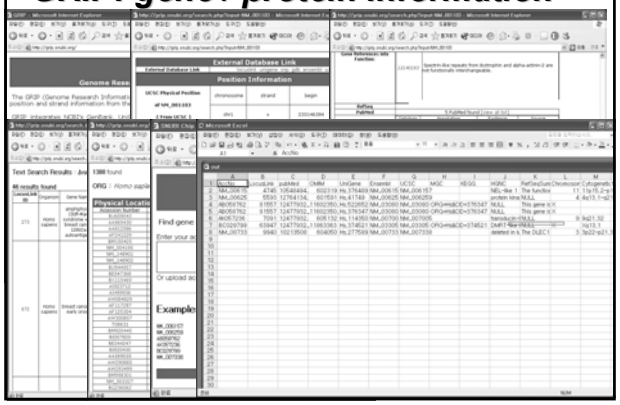
Genome Research Informatics Pipeline

Integrating biological databases

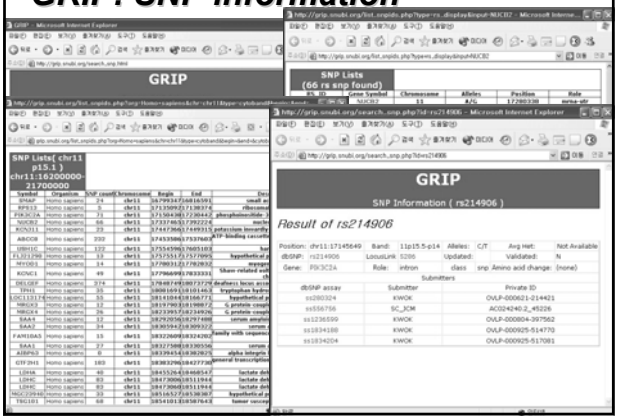
Part of GRIP integration



GRIP: gene / protein information



GRIP: SNP information



Xperanto

Expressionist's Esperanto in XML

Databasing expression with standards

nature

19 April 2001 Volume 410 Issue no 6831

Free and public expression

After a slow start, progress towards developing public repositories for gene expression data is poised to accelerate. For the many biologists working with DNA microarrays, that should be welcome news.

With a single format for gene expression data, databases should be able to 'talk' to one another and exchange data. The existence of a standard language should also spur development of software tools to query the databases, and to manage and display gene expression data.

nature

26 September 2002 Volume 419 Issue no 6905

Microarray standards at last

Not a moment too soon, the microarray community has issued guidelines that will make their data much more useful and accessible. *Nature* and the Nature research journals will respond accordingly.

You read a paper with a fascinating conclusion about the expression of several genes. You decide to use some of the same experiments on your system of choice. But when you wade through hundreds of pages of supplementary information, you find that crucial details needed for replication are missing.

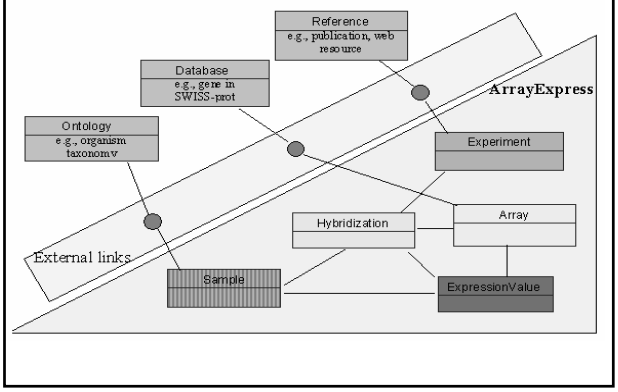
Welcome to the exciting but frustrating world of DNA microarray research. Microarrays are plastic or glass chips spotted with tiny amounts of thousands of probes, used to query the activity levels of that many genes in any tissue or organism at one time. Variables in every step of the experiment often make cross-paper comparison virtually impossible. Microarray papers also pose a considerable strain on the refereeing process; the vast amounts of data mean that critical review is a monumental task.

Yet referees sometimes feel they are not given enough details, leading cautious reviewers to think that they must reanalyse the primary data set. In other cases, the primary data provided are in proprietary software and so are impossible to comment on. Many journals allowed authors to put the huge data files on their own websites for the review process, until it became clear that unscrupulous authors compromised the anonymity of referees by tracking who had visited the website.

For authors, the proposal provides a checklist of variables that should be included in every microarray publication, at http://www.mged.org/Workgroups/MIAME/miame_checklist.html. This checklist, with all variables completed, would be supplied as supplementary information at the time of submission. The MGED group suggests that journals require submission of microarray data to either of two databases emerging as the main public repositories: GEO (www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (www.ebi.ac.uk/arrayexpress).

Harried editors can rejoice that, at last, the community is taming the unruly beast that is microarray information. Therefore, all submissions to *Nature* and the Nature family of journals received on or after 1 December containing new microarray experiments must include the unzipping of five compact disks to the editor. These disks should include necessary information compliant with the MIAME standard. The information must be supplied in a format that could be read by widely available software packages. Data integral to the paper's conclusions should be submitted to the ArrayExpress or GEO databases, with accession numbers where available, supplied at or before acceptance for publication.

How much data should authors provide to the community?



Xperanto: *Expressionist's Esperanto in XML*

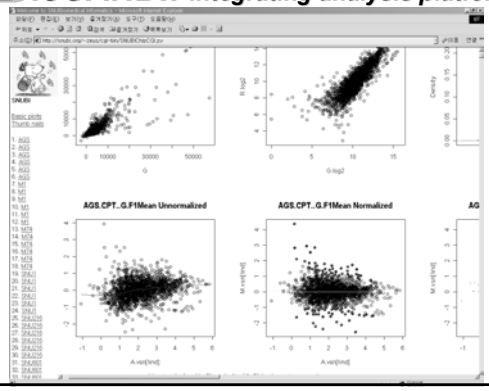
Park et al., Genomics and Informatics 2005

BioCANDI

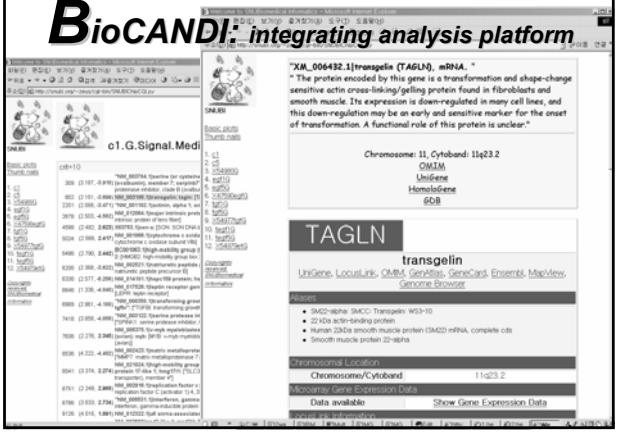
BioChip Analysis N Data Integration

Database-backed microarray analysis

BioCANDI: *integrating analysis platform*



BioCANDI: *integrating analysis platform*

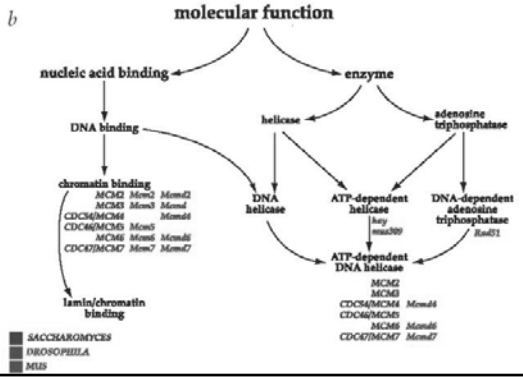


BioCANDI: *integrating analysis platform*

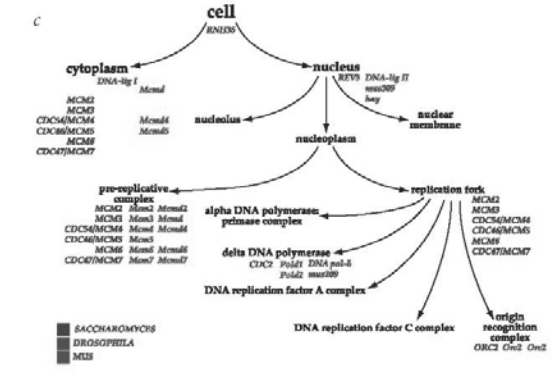
GOChase and GOTree

Interpreting gene expression clusters

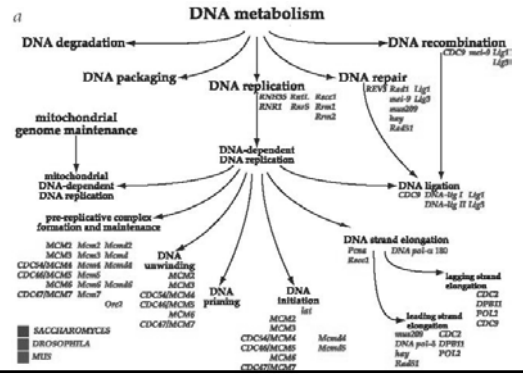
Gene Ontology



Gene Ontology

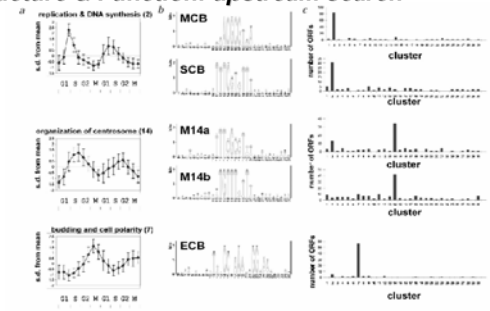


Gene Ontology



Upstream(Motif / Regulon) search

Structure & Function: upstream search



GOA-based interpretation of clusters

Table I: Comparison of GOTM with related software*

	FatiGO	Onto-Express	GOSurfer	GoMiner	GOTM
Interface/OS	Web	Web	Windows	Windows/Mac	Web
Input Identifier	Unigene ID, Gene symbol, Swiss-Prot ID, Ensembl ID, GenBank ID	GenBank ID, Affymetrix probe set ID, Unigene ID	Affymetrix probe set ID, LocusID, Unigene ID	HUGO gene names	LocusID, Gene symbol, Affymetrix probe set ID, UnigeneID, Swiss-Prot ID, Ensembl ID
Multi-level analysis	No	Yes	Yes	Yes	Yes
Visualization of classification	Bar chart, Table, Fixed tree	Bar chart, Table	Fixed tree	Expandable tree, DAG	Expandable tree, Bar chart, Fixed tree
Statistical Analysis	Fisher's exact test	Binomial test, χ^2 test, Fisher's exact test	χ^2 test	Fisher's exact test	Hypergeometric test
Correction for multiple tests	Yes	Yes	No	No	No
Visualization of enriched GO categories	Bar chart, Table	Bar chart, Table	Highlight in the full GOTree	Highlight in the full DAG	Sub-tree and DAG of enriched GO categories; Highlight in the full GOTree and bar chart
Availability	Free	Partially free	Free	Free	Free

Gene Ontology based interpretations

Every error that has a beginning has an end.

GOChase: correcting errors from gene ontology-based annotations for gene products

Yu Rang Park, Chan Hee Park, Ju Han Kim

Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea

The Gene Ontology (GO) is a controlled biological vocabulary that provides three structured networks of terms to describe biological processes,cellular components, and molecular functions. Many databases of gene products are annotated using the GO vocabularies. We found that some GO-updating operations are not easily traceable by the current biological databases and GO browsers. Consequently, numerous annotation errors arise and are propagated throughout biological databases and GO-based highlevel analyses. GOChase is a set of web-based utilities to detect and correct the errors in GO-based annotations.

Availability: http://www.snubi.org/software/GOChase/
Contact: juhan@snu.ac.kr

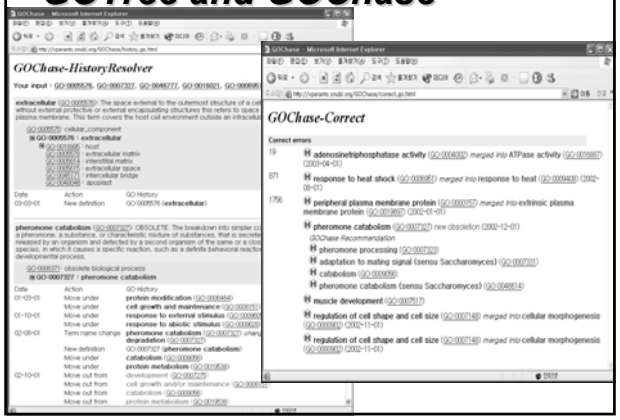
Bioinformatics 2004 doi:10.1093/bioinformatics/bth106

GOChase:

Park et al., Bioinformatics 2004

The Gene Ontology (GO) is a rapidly growing hierarchy of controlled vocabularies for describing biological processes, cellular components, and molecular functions

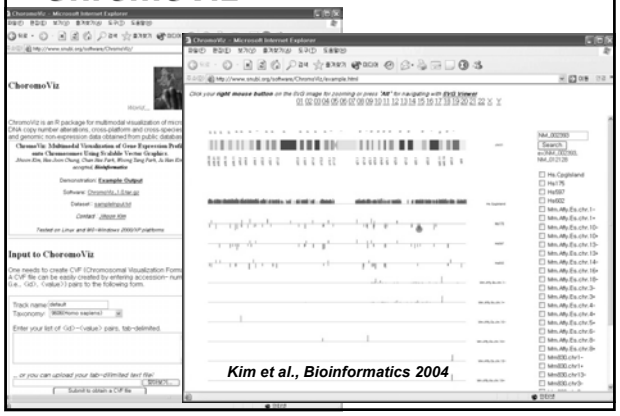
GOTree and GOChase



ArrayXPath and ChromoViz

Interpreting gene expression clusters

ChromoViz



Kim et al., Bioinformatics 2004

Pathway map with expression prof.

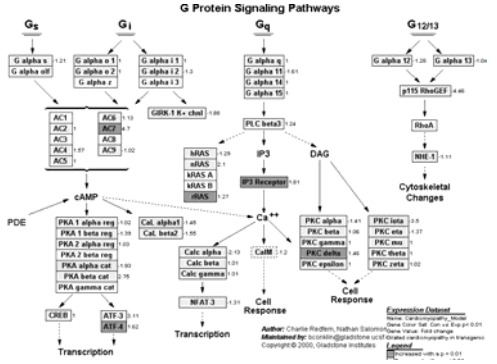
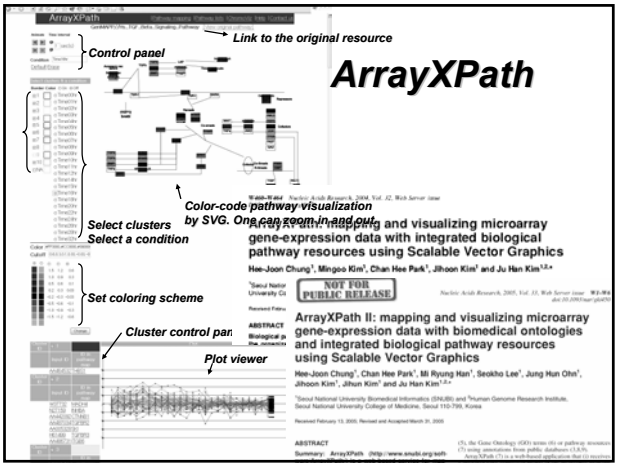


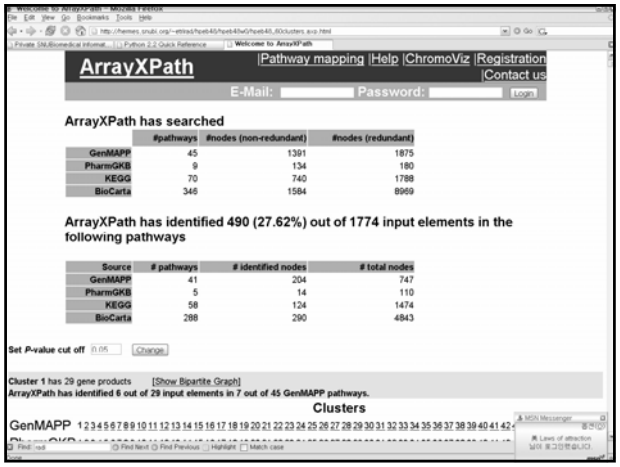
Table 1. Distribution of pathway–node identifiers among the major pathway resources

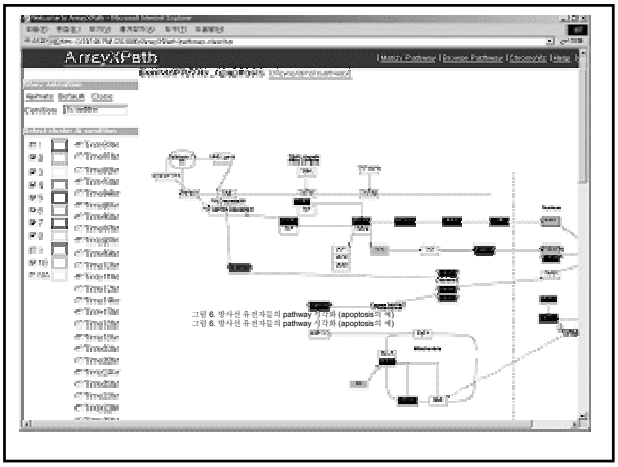
Pathway		Gene / Protein			ID resolution					Metabolite		Embedded pathway		Free text description	
		simple	compl ex	redundant	Total	OGS	LL	SP	UR						
KEGG	70	(256)* (505)* 740	(121)	(637) (469) (1,106)	740	720	20	0	0	1,896	(2,624)	0	(0)	121	(275)
GenMAPP	45	1,454		(1,942)	1,391	1,329	39	23	63	83	(97)	4	(4)	130	(372)
BioCarta	346	1,584		(8,976)	1,584	1,580	4	0	0	0	(0)	50	(141)	18	(53)
PharmGKB	9	134		(189)	134	133	1	0	0	11	(25)	1	(1)	23	(26)
Overall	470	3,025		(12,900)	3,088	2,938	64	23	63	1,990	(2,746)	55	(146)	55	(146)

* There were 21 elements redundant in the simple (256) and composite (505) elements so that 740 unique elements were found in the 70 KEGG human pathways.
Numbers in parentheses are redundant counts. KEGG has 121 composite elements containing 505 identifiable gene products.
OGS: official gene symbol, LL: LocusLink, SP: SwissProt, UR: unresolved

ArrayXPath







Current Molecular Medicine 2005, *5*, 11-21

Pathway and Ontology Analysis: Emerging Approaches Connecting Transcriptome Data and Clinical Endpoints

L. Yue and W.C. Reisdorf

Bioinformatics, GlaxoSmithKline, USA

Abstract: The increasing use of gene expression profiling offers great promise in clinical research into disease biology and its treatment. Along with the ability to measure changing expression levels in thousands of genes at once, comes the challenge of analyzing and interpreting the vast sets of data generated. Analysis tools are evolving rapidly to meet such challenges. The next step is to interpret observed changes in terms of the biological properties or relationships underlying them. One powerful approach is to make associations between the genes that are under investigation and well-known biochemical or signaling pathways, and further to assess the significance of such associations.

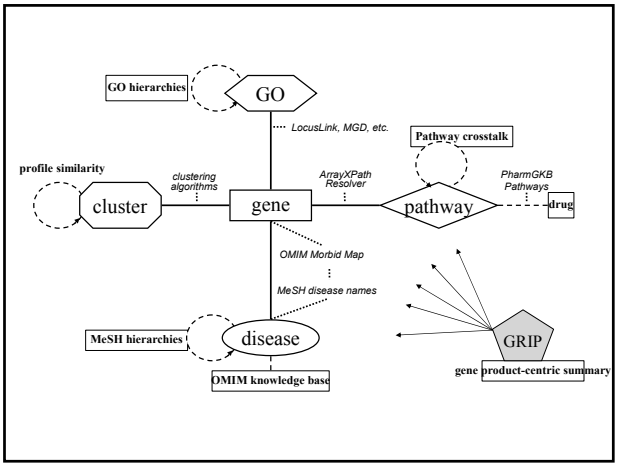
Table 1. Tools used to analyze pathway and biological relationship data.

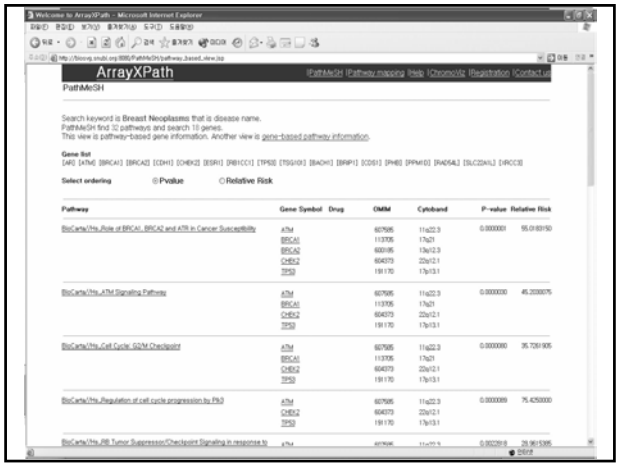
Tool	Pathway/ Ontology	Statistical Methods	Visualization	Refs	URL
ArrayXPath	Pathway	F, M	Y	[38]	http://www.snubi.org/software/ArrayXPath/
Pathway Miner	Pathway	F	Y	[39]	http://www.biorag.org/pathway.html
Knowledge Editor	Both	None	Y	[53]	http://gscope.gsc.riken.go.jp/
EASE	Pathway	O	N	[47]	http://www.DAVID.niaid.nih.gov
GeneMerge	Both	H	N	[52]	http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html
MAPPFinder	Both	O	Y		
DAVID	Both	None	Y		
GFINDer	Both	H, F, C	N		
OntoTools	Both	H, F, C, O	N		
GOSurfer/ChipInfo	Ontology	C, M	Y		
GOFish	Ontology	None	Y		
GOGet/GoView	Ontology	None	Y		
GOTree Machine	Ontology	H	Y		
FatiGO	Ontology	F, M	N		
FuncAssociate	Ontology	F, M	N		

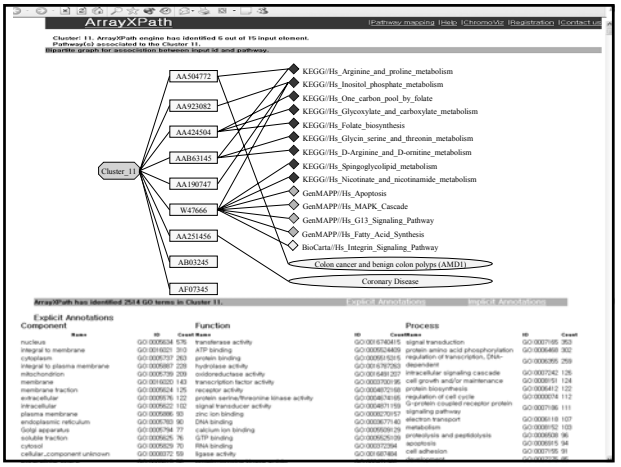
complication when analyzing such large data sets is that one is effectively doing thousands of comparisons of genes and pathways, and one should also include a correction factor for multiple hypothesis testing. Some of the tools listed in the table incorporate such a correction, but a general discussion of the methods and their relevant merits is beyond the scope of this review (and the expertise of the authors).

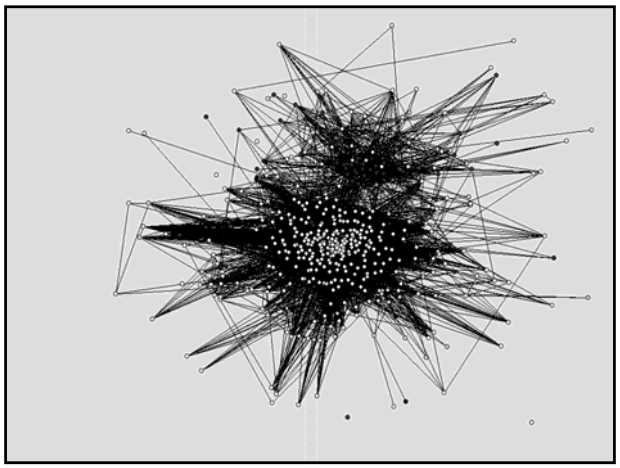
Of the tools primarily focused on pathways, ArrayXPath [38] appears to be the most comprehensive. The tool accepts a range of gene identifiers (Swissprot, GenBank, UniGene, LocusLink, etc.) and includes KEGG, BioCarta and GenMAPP as its data source. ArrayXPath uses Fisher's exact test along with the False Discovery Rate method for multiple hypothesis testing correction, and produces a very nice graphical display of results. Pathway Miner [39] can map the genes from up to four different experiments onto KEGG, BioCarta or GenMAPP pathways, and it can

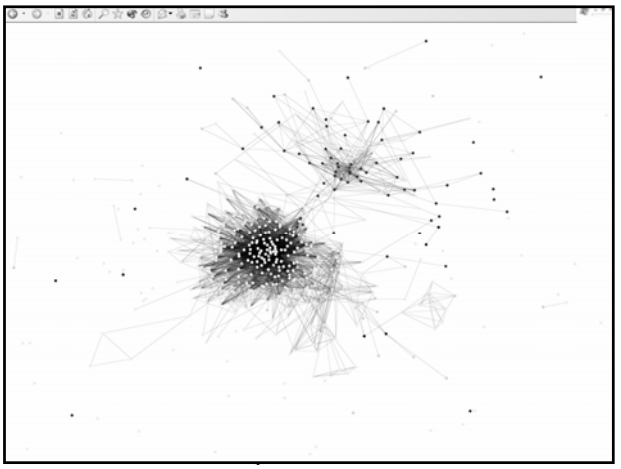
Where to go from here?
Understanding gene expression clusters

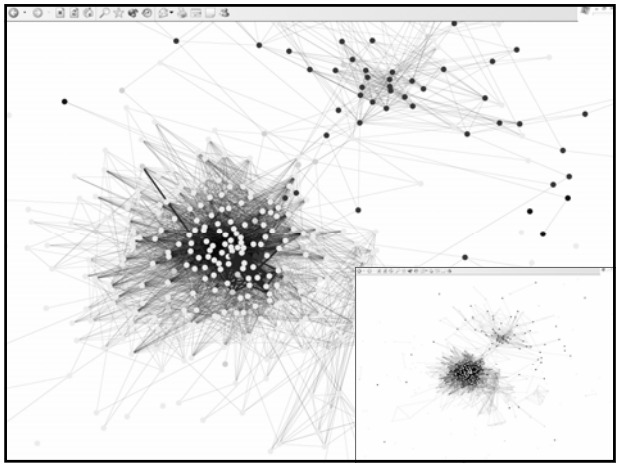


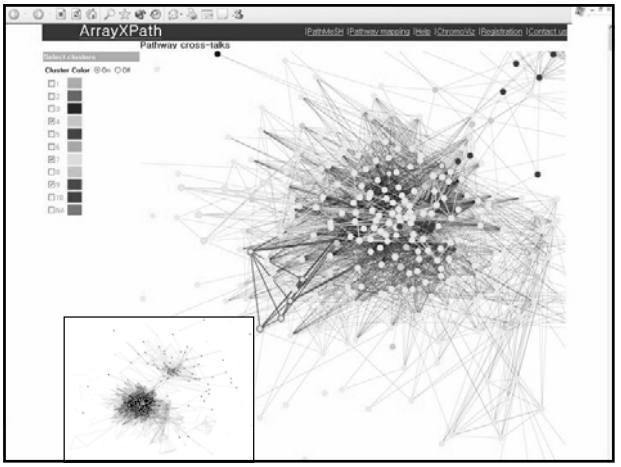

ArrayXPath
PathMeSH
Search keyword is Breast Neoplasms that is disease name.
PathMeSH find 32 pathways and search 18 genes.
This view is pathway-based gene information. Another view is gene-based pathway information.
Gene list
Select ordering
Pvalue
Relative Risk
Pathway
Gene Symbol
Drug
OMIM
Cytoband
P-value
Relative Risk


ArrayXPath
Cluster 11. ArrayXPath engine has identified 6 out of 15 input element.
Pathway(s) associated to the Cluster 11.
Bipartite graph for association between input id and pathway.
Cluster_11
AA504772
AA923082
AA424504
AAB63145
AA190747
W47666
AA251456
AB03245
AF07345
KEGG//Hs_Arginine_and_proline_metabolism
KEGG//Hs_Inositol_phosphate_metabolism
KEGG//Hs_One_carbon_pool_by_folate
KEGG//Hs_Glycoxylate_and_carboxylate_metabolism
KEGG//Hs_Folate_biosynthesis
KEGG//Hs_Glycin_serine_and_threonin_metabolism
KEGG//Hs_D-Arginine_and_D-ornitine_metabolism
KEGG//Hs_Spingoglycolipid_metabolism
KEGG//Hs_Nicotinate_and_nicotinamide_metabolism
GenMAPP//Hs_Apoptosis
GenMAPP//Hs_MAPK_Cascade
GenMAPP//Hs_G13_Signaling_Pathway
GenMAPP//Hs_Fatty_Acid_Synthesis
BioCarta//Hs_Integrin_Signaling_Pathway
Colon cancer and benign colon polyps (AMD1)
Coronary Disease
ArrayXPath has identified 2514 GO terms in Cluster 11.
Explicit Annotations
Component
Function
Process



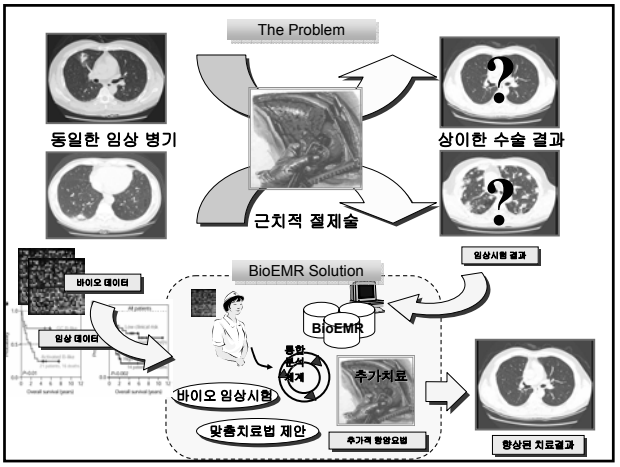


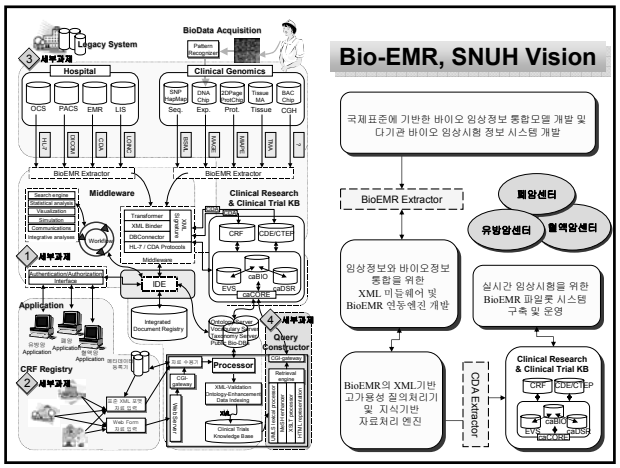



ArrayXPath
Pathway cross-talks
Cluster Color

Integrative biochip informatics

- ***Exploration: mapping and clustering***
- ***Pattern recognition: finding gene – drug assoc.***
- ***Prognostic subgroup prediction***
- ***Interpretation***
  - ✓ ***GRIP: automatic probe annotation***
  - ✓ ***Xperanto: databasing expression***
  - ✓ ***BioCANDI: database-integrated analysis tool***
  - ✓ ***ChromoViz: visualizing regional regulation***
  - ✓ ***GOTree and GOChase: GO-based annotation***
  - ✓ ***ArrayXPath: pathway-based interpretation***
  - ✓ ***PathMeSH: disease - gene - pathway***
- ***BioEMR: towards real-time clinical trials***





Then, isn't it just a tool?

Life

Science
科學

Logos

Classification

- ***Informatical Phenomena***
- ***Molecular biology***
- ***Organic/Bio- Chemistry***
- ***Observation / classification***
- ***created by God***