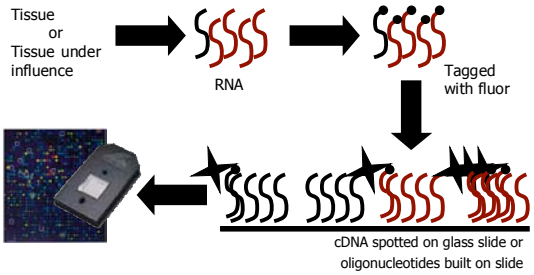
Making sense of microarrays

Isaac S Kohane

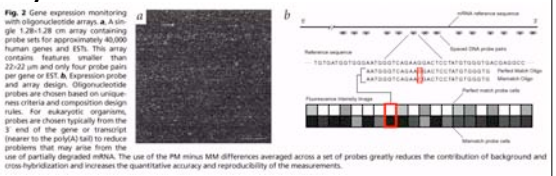
RNA expression detection chips



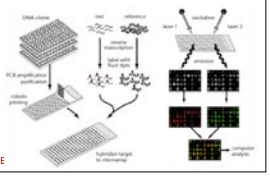
Schena M, et al. Proc Natl Acad Sci USA; 93: 10614 (1996).
Entire issue. Nature Genetics, 21: supplement (Jan 1999).

Two Biochip Technologies

Affymetrix



P. Brown / Stanford



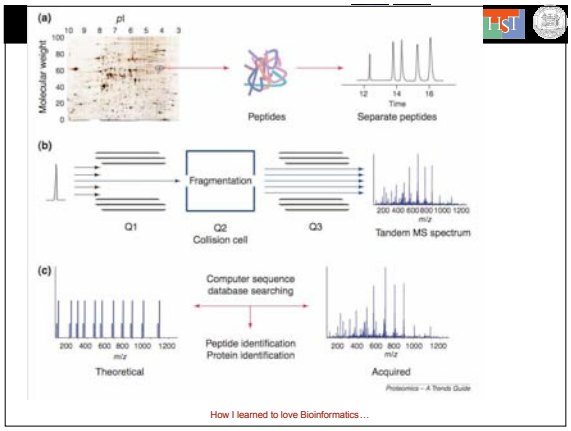
What is a microarray?

- Low cost. The cost should be such that at least hundreds of samples be measurable within a typical NIH RO1 budget
- Commodity level workflow. The microarray should be commoditized such that a routine set of procedures requiring no scientific judgment can be performed using standard equipment to obtain the needed measurement.
- Automation. The process of data acquisition should be completely automated so that after the biomaterial whether it is protein, RNA or DNA is loaded into the analytic pipeline, most of the steps are fully automated and those that are not automated can be done by a non-specialized technician.
- Form factor. The equipment required should easily fit into a standard laboratory bench format and not require its own room.

What is a microarray? (II)

- Translational friendliness. A clinical investigator does not have to understand molecular biology techniques in order to be able to provide the necessary materials for the acquisition of the microarray data.
- Identifiability. All items identified by the microarray technology whether they are proteins or RNA species should be automatically identified against standard reference nomenclatures.
- High Throughput. Hundred of patient samples can be processed within days.
- Commodity level priced infrastructure. The technology equipment and budget should be available to most biological and clinical investigational laboratories.
- Massively parallel measurements of the relevant analytes. That is, the members of transcriptome, the members of the proteome, the metabolome or any other comprehensive measure of molecular physiology.

Can we build microarrays for proteomics?

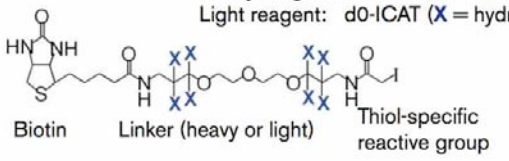


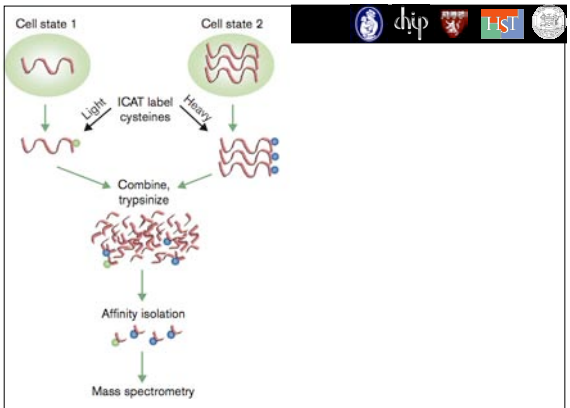
Quantitative Proteomics

ICAT reagents

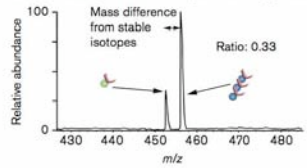
Heavy reagent: d8-ICAT (X = deuteriu

Light reagent: d0-ICAT (X = hydroge

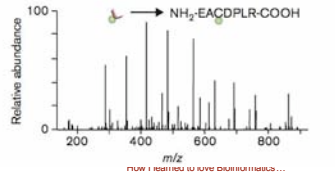




Quantitate relative protein levels by measuring peak ratios



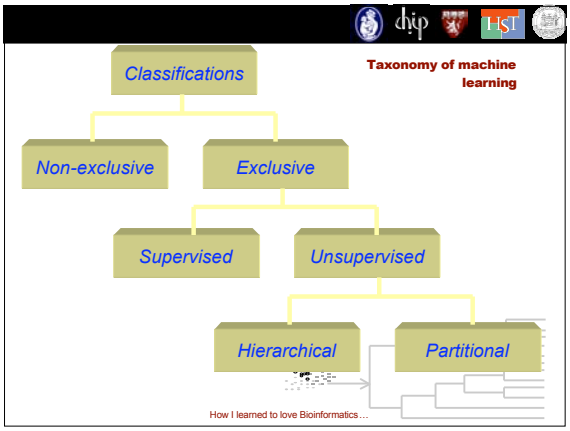
Identify peptide by sequence information (MS/MS scan)

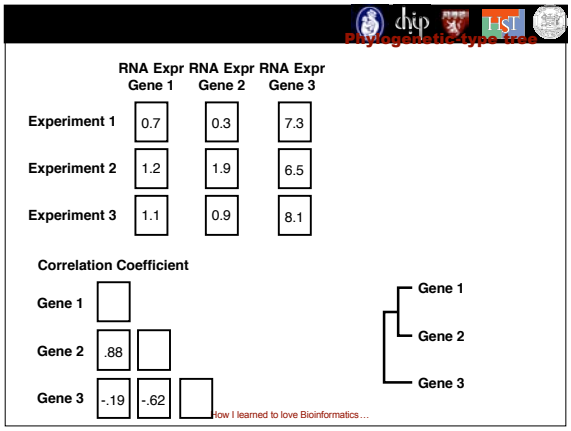


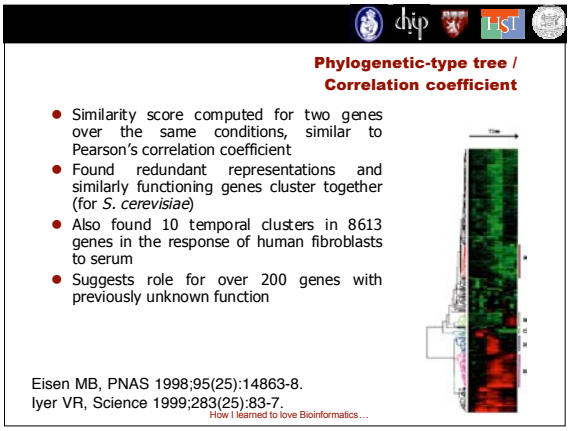
So, do we have proteomic microarrays?

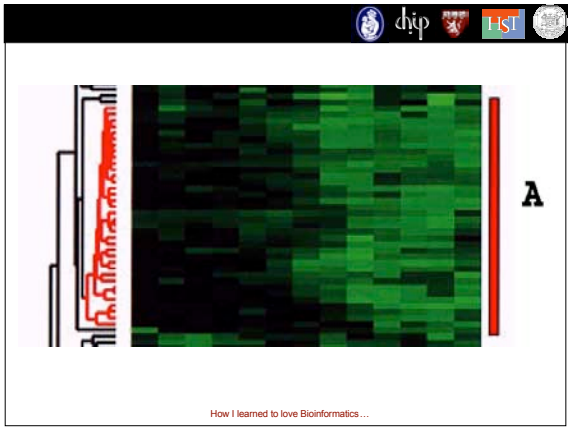
Making sense of all the data

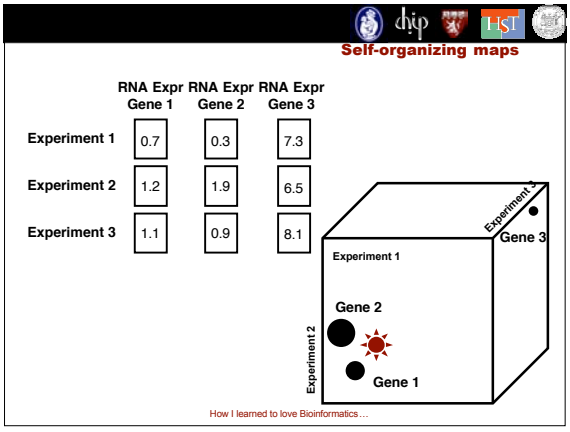
- The underlying dogma
- What data may be included in the data sets?
  - ✓ Beyond genomic data
  - ✓ Multiple scales and non-numeric measures


Taxonomy of machine learning
Classifications
Non-exclusive
Exclusive
Supervised
Unsupervised
Hierarchical
Partitional
How I learned to love Bioinformatics…


Phylogenetic-type tree
RNA Expr Gene 1
RNA Expr Gene 2
RNA Expr Gene 3
Experiment 1
0.7
0.3
7.3
Experiment 2
1.2
1.9
6.5
Experiment 3
1.1
0.9
8.1
Correlation Coefficient
Gene 1
Gene 2
.88
Gene 3
-.19
-.62
Gene 1
Gene 2
Gene 3
How I learned to love Bioinformatics…


Phylogenetic-type tree / Correlation coefficient
Similarity score computed for two genes over the same conditions, similar to Pearson's correlation coefficient
Found redundant representations and similarly functioning genes cluster together (for S. cerevisiae)
Also found 10 temporal clusters in 8613 genes in the response of human fibroblasts to serum
Suggests role for over 200 genes with previously unknown function
Eisen MB, PNAS 1998;95(25):14863-8.
Iyer VR, Science 1999;283(25):83-7.
How I learned to love Bioinformatics…

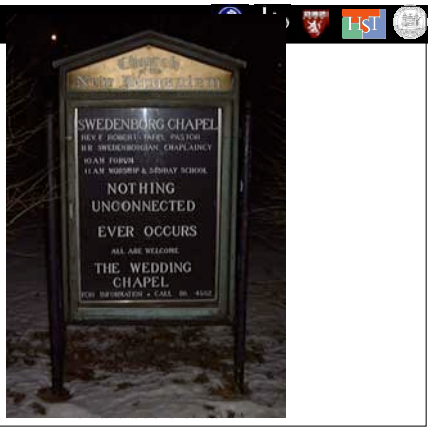

A
How I learned to love Bioinformatics…


Self-organizing maps
RNA Expr Gene 1
RNA Expr Gene 2
RNA Expr Gene 3
Experiment 1
0.7
0.3
7.3
Experiment 2
1.2
1.9
6.5
Experiment 3
1.1
0.9
8.1
Experiment 3
Gene 3
Experiment 1
Gene 2
Experiment 2
Gene 1
How I learned to love Bioinformatics…

Relevance Networks

- Several algorithms have already been developed for knowledge discovery and data-mining of RNA expression data sets
- We are interested in finding networks of genes that are functionally clustered with little or no *a priori* knowledge (unsupervised learning)
- Relevance Networks are an approach to analyze these data sets
- Previously validated in the clinical laboratory result domain

Children's Hospital, Patent Pending.
Butte AJ, Kohane IS, Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks, Symposia AMIA, 1999.



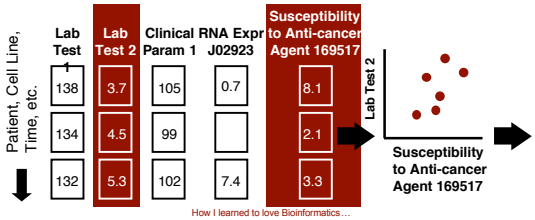
Construction of Relevance Networks 1

- Patients and cell lines are analyzed as cases
- Clinical parameters, laboratory tests, RNA expression, and susceptibility to anti-cancer agents are all example features of those cases

Patient, Cell Line, Time, etc.	Lab Test 1	Lab Test 2	Clinical Param 1	RNA Expr J02923	Susceptibility to Anti-cancer Agent 169517
1	138	3.7	105	0.7	8.1
134	4.5	99		2.1	
132	5.3	102	7.4	3.3	

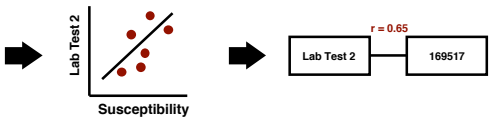
Construction of Relevance Networks 2

- For all pairs of features, we take overlapping values over the cases and make a scatter plot of values



Construction of Relevance Networks 3

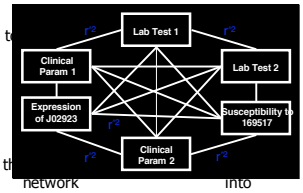
- Perform a pairwise comparison between all features
- For each scatter plot, we fit a linear model and stored
  - ✓ Correlation coefficient r



- Every feature is completely connected to every other feature by a linear model of varying quality

Construction of Relevance Networks 4

- $r^{\wedge}2 = r^2 * r / abs(r)$
- Choose $r^{\wedge}2$ network
- Drop links under threshold
- Breaks connected islands where are stronger than threshold



- Islands are what we call "relevance networks"
- Display graphically, with thick lines representing strongest links

NCI 60 data set

- NCI 60 is set of 60 cancer cell lines against which the National Cancer Institute has tested over 50,000 chemicals to find anti-cancer agents
- NCI recently provided us with consistent, validated data for 5,084 agents, representing the concentration of each agent needed to cause 50% growth inhibition compared to control (GI50) for each cell line in NCI 60
- Data was provided as $-\log_{10}(\text{GI50})$
  - ✓ Lower number means higher GI50, or less sensitivity to an agent
- Unfortunately, few of these anti-cancer agents have documented mechanisms of action (or even a name listed)

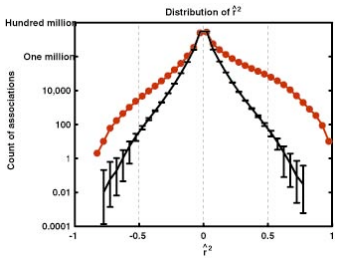
Weinstein, et al. Science; 258: 447 (1992).
http://www.dtp.nci.gov

RNA expression data set

- RNA expression in NCI 68 cell lines was determined using Affymetrix HU6000 arrays
  - ✓ 5,223 known genes
  - ✓ 1,193 expressed sequence tags
- The RNA expression data set and Anti-cancer susceptibility data set were merged, using the 60 cell lines the two tables had in common

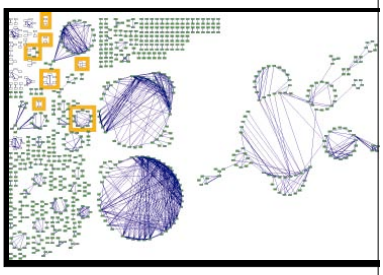
Distribution of r^2

- 11,692 features
- 68,345,586 total associations
  - ✓ 22 M between genes
  - ✓ 12 M between agents
  - ✓ 33 M between a gene and an agent



Genes and Anti-Cancer Agents

- Threshold r^2 was 0.8
- 202 networks
- 834 features out of 11,692 (7.1%)
- 1,222 links out of 68,345,586 (.0018%)
- Only one link between a gene and anti-cancer agent



Taxonomy of Network Links

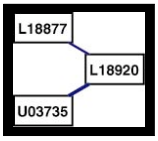
- Identity / Synonomy
- Functional Similarity
- Derivation
- Biological Relationship

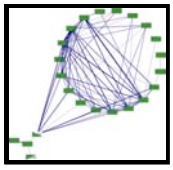
Identity / Synonomy Gene Networks

- 15 networks demonstrated synonymy associations
  - ✓ Most linked endogenous or spiked controls
- 5 / 15 were linked genes with multiple GenBank entries
  - ✓ L10838 and D28423: SRP20
  - ✓ M19267 and Z24727: Tropomyosin alpha chain
  - ✓ X17567 and X52979: snRNP B
  - ✓ U08021 and U51010: Nicotinamide N-methyltransferase
  - ✓ M14199 and U43901: Laminin receptor precursor and laminin receptor mRNA

Functional Similarity Networks

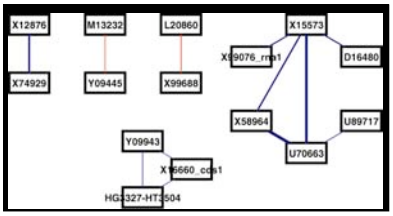
- Melanoma-associated antigens 2, 3 and 12
- Caldesmon 1 and alternative splicing products 3 and 4
- Two sequences from MHC class I (D32129 and X12432)
- 3088 (chlorambucil), 344007 (piperazine alkylator), 34462 (uracil nitrogen mustard), 48034 (aziridine), 6396 (thio-tepa), and 9706 (triethylenemelamine) are all alkylating agents
- 243928 (ethanesulfonamide derivative) and 249992 (amsacrine) are both active against topoisomerase II





Biological Associations

- Keratin 8 linked to keratin 18
- Glycoprotein Ib beta linked to *psd*
- Others were not easily explained using the medical literature

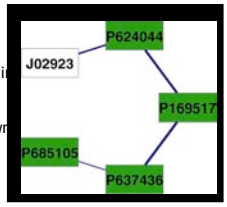


Derivative Networks

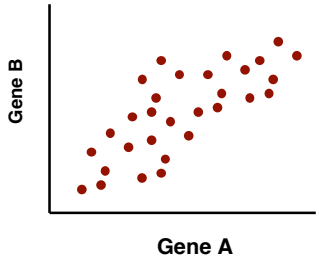
- 112167 (19-norlanosta-1) was derived from 112166 (curcurbitacin)
- 274555 (gold derivative) was derived from 306388
- 295500, 374028, 606497, 606499, 610456, 610457, and 610459 are all camptothecin derivatives
- 302325 and 382034 are pyrimidinediamine derivatives
- 351710, 352299, and 627505 are methylellipticinium derivatives
- 603071 and 629971 are 9-aminocamptothecin derivatives
- 634785 and 634786 are 4-piperidinone derivatives

Genes and Anti-Cancer Agents

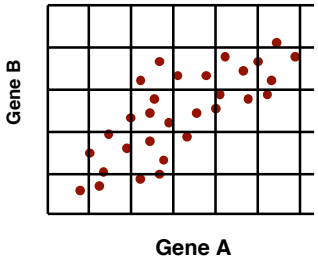
- Elevated levels of J02923 (lymphocyte cytosolic protein-1, LCP1, L-plastin, pp65) is associated with increased sensitivity to 624044
- Agent 624044 is 4-Thiazolidinecarboxylic acid, 3-[[6-[2-oxo-2-(phenylthio)ethyl]-3-cyclohexen-1-yl]acetyl]-2 thioxo-, methyl ester, [1R-[1a(R*),6a]]- (9CI))
- LCP1 is an actin-binding protein involved in leukocyte adhesion
- A role for LCP1 has been previously postulated
- Low level expression thought to occur i cancer cell lines
- Other thiazolidine derivatives are known tumor cell growth



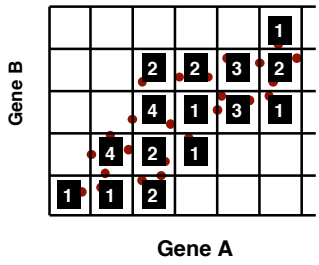
Mutual Information



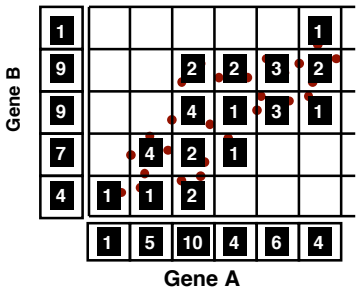
Mutual Information



Mutual Information



Mutual Information



Mutual Information (A,B) =
Entropy (A) + Entropy (B) – Entropy (A and B)

Global Regulatory Pathways Using Relevance Networks

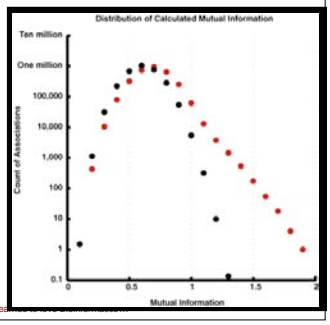
- Stanford provides a yeast RNA expression data set for public use
- Data set of 2,467 open reading frames (ORF) measured under 79 conditions
- Several experiments with various time points in
  - ✓ Diauxic shift
  - ✓ Mitotic cell division cycle
  - ✓ Sporulation
  - ✓ Temperature shocks
  - ✓ Reducing shocks

Eisen MB, Proc Natl Acad Sci U S A 1998;95(25):14863-8.

Distribution of Mutual Information

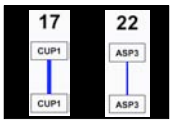
- 2,467 features
- 3,041,811 associations

Butte AJ, Kohane IS, Using Correlations To Discover Large Scale Functional Relationships Between RNA Expression and Anti-Cancer Agent Susceptibility, Pacific Symposium on Biocomputing, 2000.



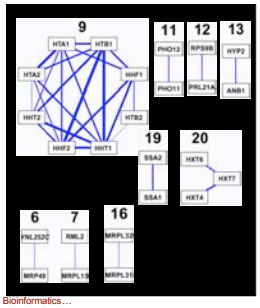
Synonomy Networks

- Threshold MI was 1.2
- 22 networks
- 199 features out of 2,467 (8.1%)
- Two networks with synonomy associations
  - ✓ Copper mettalothionein
  - ✓ L-aspariginase II



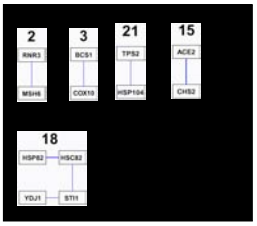
Similar Function Networks

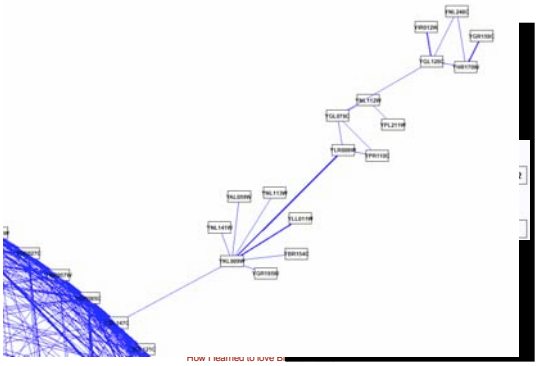
- Nine networks with associations linking genes with similar functions
  - ✓ Histones
  - ✓ Acid phosphatases
  - ✓ Ribosomal proteins
  - ✓ Translation initiation
  - ✓ 70 kDa heat shock proteins
  - ✓ Hexose transporters
  - ✓ Mitochondrial ribosomal proteins



Biological Pathway Networks

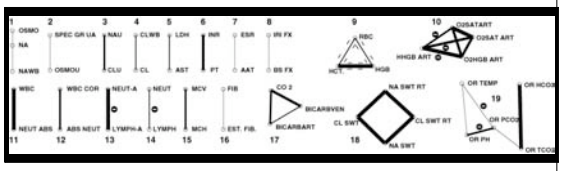
- Five networks with associations linking genes known to be related in the same biological pathway
  - ✓ Base pair mismatch repair
  - ✓ Cytochrome complex assembly
  - ✓ Trehalose-6-phosphate and chaparone
  - ✓ Chitinase expression regulator and chitin synthase II
  - ✓ Isoforms of chaperones





Another domain: Clinical Laboratory

- Matrix of 642 lab tests with 28,566 overlapping results
- Threshold r^2 was 0.6, n was 50

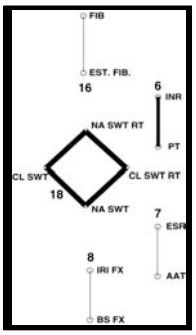


- 48 lab tests out of 642 (7%), 36 links out of 205,761 (.017%)

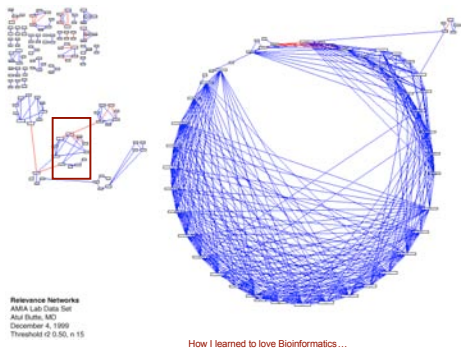
Butte AJ, Kohane IS, Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks, Symposia AMIA, 1999.

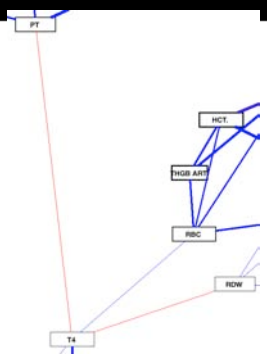
Taxonomy of links

- Identity or synonymy
  - ✓ Serum fibrinogen and estimated fibrinogen
- Mathematical
  - ✓ Prothrombin time and International Normalized Ratio
- Physiologic
  - ✓ Sodium and chloride in sweat
- Pathologic
  - ✓ Erythrocyte sedimentation rate and alpha-1 antitrypsin
- Causal
  - ✓ Serum blood sugar and serum insulin level



Dropping the thresholds





Possible novel links

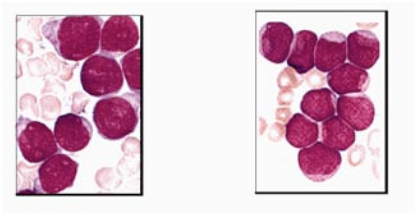
Supervised learning

- Several techniques available
- Decision trees
  - ✓ Y. Sun et al.
- Support Vector Machines
  - ✓ M. P. S. Brown *et al.*, *Proc Natl Acad Sci U S A* **97**, 262-267 (2000).
- Neural Networks
  - ✓ C. Wu, S. Shivakumar, *Nucleic Acids Research* **22**, 4291-4299 (1994).

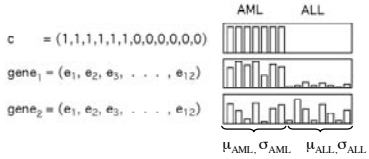
Leukemia Classification

Morphology does not distinguish leukemias very well

Acute lymphoblastic leukemia (ALL) Acute Myelogenous Leukemia (AML)



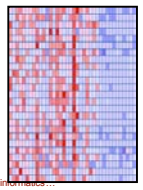
Courtesy P. Tamayo



Sort genes according to measure of correlation (distance or class separation):

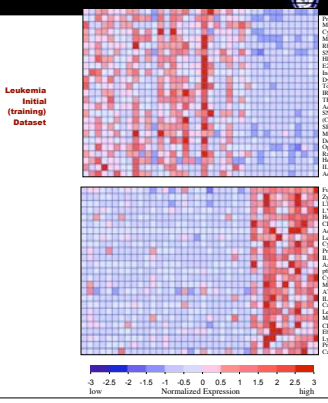
Euclidean = $\Sigma_i (g_i - c_i)^2$

Signal-to-noise = $|\mu_{AML} - \mu_{ALL}| / (\sigma_{AML} + \sigma_{ALL})$



Higher correlation

Best Marker Genes

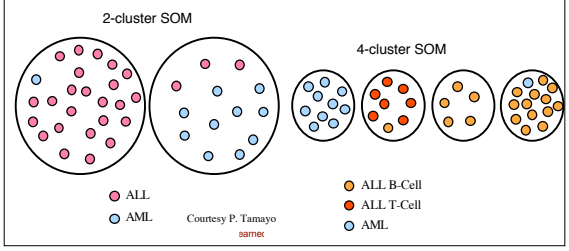


Leukemia AML, ALL B- and T-Cell Classes

38 Sample collection:

11 AML + 19 ALL B-Cell + 8 ALL T-Cell

Discovered Classes:

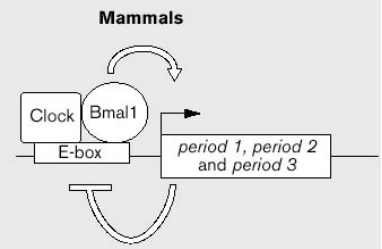


Courtesy P. Tamayo

Resources

- MeV - TIGR
- Genecluster 2.0 (Broad Institute)

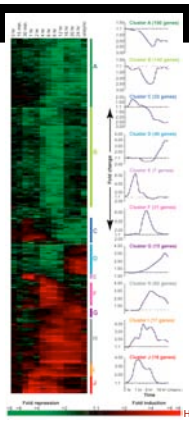
Physiology of dynamics of biological processes



- Elegant mechanism for the internal clock discovered in several species.

How do we translate this knowledge into cellular physiology on a genomic scale?

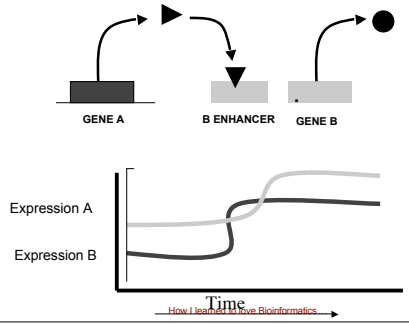
Transcriptional response of fibroblasts to Serum



- Gene expression of fibroblasts across time
- Clusters of characteristic time signatures
  - ✓ Transcriptional protein and signal transduction pathways (c-FOS, Jun B, Mitogen Activated protein) (cluster A,B)
  - ✓ Cell-cycle-specific: Cyclin A, Cyclin B1, cdc28 kinase (cluster D)
- Can we sort out individual gene interactions?

Iyer Science 1999

Dynamics

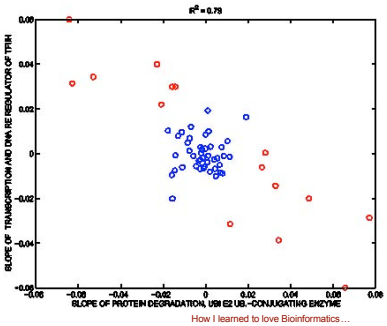


Point by point slopes

- Different networks identified than in static networks (why?)
- Artifactual high correlations near rest

$$\text{Slope}(n,n+1) = \frac{(\text{expression\_level}_{n+1} - \text{expression\_level}_{n})}{(\text{time}_{n+1} - \text{time}_{n})}$$

Simple slope analysis



Negative correlation between transcription and protein degradation

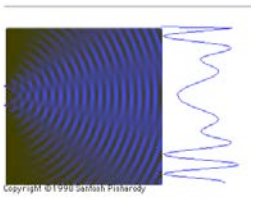
Had very low pairwise association in static correlation of entropic space

Borrowing from signal processing techniques

- Examining time-series in the frequency domain

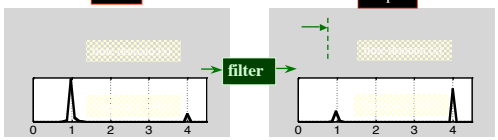
Reconstructing the source waveforms

Ripples in Water



Linear Approach

Schematic input → filter → output



At each frequency, the relationship between oscillations of in input and output are quantified by:

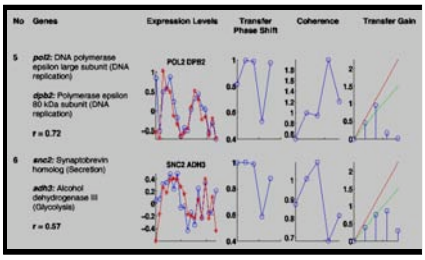
Transfer gain: amplitude modulation from input to output.

Transfer phase: time lead or time lag from input to output.

Coherence: linearity of relationship between signals.

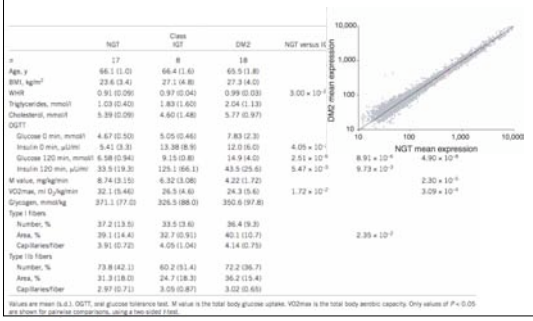
Signal Processing Finds Gene Regulation

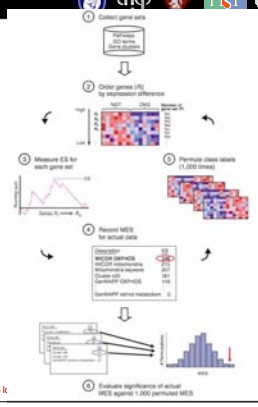
- Given enough compute-power, digital signal processing can be successfully applied to all possible pairs of genes

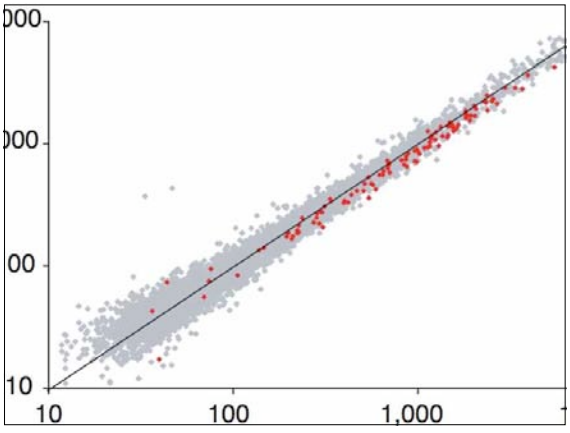


Butte AJ, et al. Comparing the Similarity of Time-Series Gene Expression Using Signal Processing Metrics, Submitted.

Using Knowledge To Extract Information







How I learned to love Bioinformatics…

How I learned to love Bioinformatics…