Making sense of microarrays

Isaac S Kohane

Two Basic Steps

Affymetrix

P. Brown / Stanford

RNA expression detection chips

- Tissue
- Tissue under influence
- RNA
- Tagged with fluor
- cDNA spotted on glass slide or oligonucleotides built on slide

Entire issue, Nature Genetics, 21: supplement (Jan 1999).

What is a microarray?

- Low cost. The cost should be such that at least hundreds of samples can be measurable within a typical NIH ROI 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?
How I learned to love Bioinformatics

Quantitative Proteomics

ICAT reagents
Heavy reagent: d8-ICAT (X = deuterium)
Light reagent: d0-ICAT (X = hydrogen)

Biotin Linker (heavy or light) Thiol-specific reactive group

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
How I learned to love Bioinformatics

Classifications

- Non-exclusive
- Exclusive

Supervised

Unsupervised

Hierarchical

Partitional

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


Self-organizing maps
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.

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.

<table>
<thead>
<tr>
<th>Lab Test</th>
<th>Lab Test 2</th>
<th>Param 1 J02923</th>
<th>Susceptibility to Anti-cancer Agent 169517</th>
</tr>
</thead>
<tbody>
<tr>
<td>134</td>
<td>3.7</td>
<td>105</td>
<td>0.7</td>
</tr>
<tr>
<td>134</td>
<td>4.6</td>
<td>69</td>
<td>2.1</td>
</tr>
<tr>
<td>134</td>
<td>5.3</td>
<td>102</td>
<td>7.4</td>
</tr>
<tr>
<td>134</td>
<td>5.3</td>
<td>102</td>
<td>7.4</td>
</tr>
</tbody>
</table>

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 $r$

Construction of Relevance Networks 4

- $r^2 = r^2 + \ln(\text{abs}(r))$
- Choose $r^2$ network
- Drop under threshold links
- Breaks connected islands
- Islands are what we call "relevance networks"
- Display graphically, with thick lines representing strongest links.
How I learned to love Bioinformatics

**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 $-\log_{10} (GI50)$.
- 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).

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 (0.0018%)
- Only one link between a gene and anti-cancer agent.

**Identity / Synonymy 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: Troponycin alpha chain
  - X17567 and X19297: v-raf8B
  - 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 I and alternative splicing products 3 and 4
- Two sequences from MHC class I (D32129 and X12432)
- 3088 (chlorambucil), 344007 (piperazine alkylation), 34462 (uracil nitrogen mustard), 48034 (azadine), 6396 (bix-tesp), and 9705 (trimethylene melamine) are all alkylating agents
- 243928 (ethanesulfonamide derivative) and 249992 (amazine) are both active against topoisomerase II

Biological Associations
- Keratin 8 linked to keratin 18
- Glycoprotein Ib beta linked to pad
- Others were not easily explained using the medical literature

Derivative Networks
- 112167 (19-nortestoster-1) was derived from 112166 (curcubitacin)
- 274555 (gold derivative) was derived from 306388
- 295500, 374028, 606497, 606499, 610456, 610457, and 610459 are all camptothecin derivatives
- 302325 and 380304 are pyrimidine derivative amides
- 351710, 35299, and 627505 are methyldiisocinum derivatives
- 603071 and 629971 are 9-amino camptothecin derivatives
- 634785 and 634786 are 4-piperidine derivatives

Genes
- Elevated levels of J02923 (lymphocyte cytosolic protein-1, LCPI, L-plastin, p90) is associated with increased sensitivity to 62404H
- Agent 624044 is 4'-thiazolidine-carboxylic acid, 3-[(6-(2-oxo-2'-phenylethyl)-1-yl)acetyl]-2-thioxo, methyl ester, [1R-[1a(R*)]-6a]- (9CI)
- LCP1 is an actin-binding protein
- A role for LCP1 has been previously postulated
- Low level expression occur to cancer cell lines
- Other thiazolidine derivatives are known to tumor cell growth

Mutual Information
- Gene B
- Gene A
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


Distribution of Mutual Information

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


Synonym Networks

- Threshold MI was 1.2
- 22 networks
- 199 features out of 2,467 (8.1%)
- Two networks with synonym associations
  - Copper metallothionein
  - L-asparaginase II
How I learned to love Bioinformatics

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 chaperone
- 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 r² was 0.6, n was 50

- 48 lab tests out of 642 (7%), 36 links out of 205,761 (0.017%)


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
- Support Vector Machines
- Neural Networks

**Leukemia Classification**

Morphology does not distinguish leukemias very well

- Acute lymphoblastic leukemia (ALL)
- Acute myelogenous leukemia (AML)

**Formulae**

Euclidean distance:

\[ d \left( \mathbf{x}, \mathbf{y} \right) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

Signal-to-noise:

\[ \text{Signal-to-noise} = \frac{\left| \mu_{\text{AML}} - \mu_{\text{ALL}} \right|}{\sqrt{\frac{\sigma_{\text{AML}}^2 + \sigma_{\text{ALL}}^2}{2}}} \]

**Resources**

- MeV - TIGR
- Genecluster 2.0 (Broad Institute)
How I learned to love Bioinformatics

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?

Dynamics

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

Point by point slopes

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

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

At each frequency, the relationship between oscillations of 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 computer power, digital signal processing can be successfully applied to all possible pairs of genes.

How I learned to love Bioinformatics...