Lecture 18 Gene Expressing Analysis

Tuesday, March 20, 2001
Lectured by: Anne Condon Notes taken by: Fengguang Song

Refs and papers

[1] Ben-Dor et al. Tissue Classification with Gene Expression Profiles, Recomb 2000.
[2] Dolub et a1. Molecular Classfication of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531-537, 1999.
[3] M.J. van der Laan and J. Bryan. Gene Expression Analysis with the Parameter Bootrap, Biostatistics, 2001

1. Overview

Genes are expressed in two steps: DNA → mRNA → proteins. Regulation mechanisms in the cell control which genes are expressed under given conditions.

Measuring mRNA levels can provide a detailed molecular view of gene expression.

Gene expression analysis sheds insight on:

- Gene function
- Interaction between genes
- Disease diagnosis
- Disease treatment monitoring

How?

- Extract mRNA samples from cell.
- Thousands of genes may be co-expressed. Ideally, you'd like to know sequence information for all mRNAs in samples; also expression level need a technology that can "process" many strands in parallel.

2. DNA microarrays

At least two different types arrays:

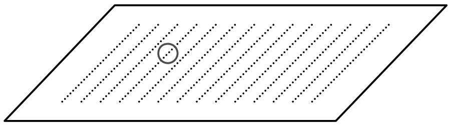
2.1 cDNA arrays

The 'c' represents "complementary to" mRNA.
cDNA:

- Determined from databases
- From a wet library

mRNAs come from two sources: test and reference, which are fluourescently tagged. Say, test---green tags, reference---red tags. The flourescently tagged mRNA pools from both test and reference are washed over the surface, allowing one to determine relative amount of transcript present in the pool for each of he two cell types.

Current technology can generate arrays with over 10,000 cDNA's per square cm.



Attached to the array are cDNA sequences.

Expression level at a given spot is (test fluorescent signals) / (reference fluorescent signals).

There is a company called Synteni developing this technique, which is also a “do it yourself” web site.

2.2 Oligonucleotide microarrays

Photolithography and solid phase synthesis are used to synthesize oligos on surface. Oligo probes are 20-30 based long. And there is no reference on samples.

Affymetrix system: >=300,000 oligos can be synthesized on a 1.28 x 1.28 cm array. Affymetrix sells chips for Arabidopsis, mouse, rat, and human genes. The cost is 500-2k per chip, which can be used only once. Also cost of $175k for setup!

3. Example: molecular classification of cancer

Acute leukemia: ALL and AML types

Motivation: gene expression analysis distinguishes between these two types of genes, more generally, can be discovery as well as diagnosis of cancers.

Cancer classification:

- Class prediction: assignment of particular sample to already defined target types (classes).
- Class discovery: will be taught next lecture.

3.1 Class prediction

To develop the method, Golub dt al. obtained 38 bone marrow samples (27 ALL, 11 AML). RNA from samples was hybridized to oligo arrays containing 6,817 genes. For each (sample, gene) pair, an expression level was measured.

3.1.1 Gene selection

(1) Sort the 6,817 genes by degree of correlation with class distinction.

c: 38-vector (111…11,000…0). ‘1’ represents ALL, ‘0’ represents AML.

For each g: expression level of 38-vector. For each of 38 samples, normalize the log of gene expression level.

Correlation

$$p(g,c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) - \sigma_2(g)}$$

μ: mean of the normalized gene expression levels for gene g taken over class 1 (ALL) samples.
σ: standard deviation.
Large values of |p(g,c)| indicates strong correlation.

(2) Compare result with corresponding sorted correlation information using same gene data but λ permutations of class distinction vector c (400 permutations).
For p(g,c) – 0.3, 1% of λ neighborhoods contain as many points as the observed neighborhood around.

(3) Select most informative genes.
25 genes with largest p(g,c) values;
25 genes with smallest p(g,c) values.

3.1.2 Class predictor
Given new sample x, find (normalized) gene expression levels for all of the 50 selected genes.

Weighted voting scheme: each gene votes for ALL or AML, depending on whether its expression level Xi (the normalized log expression level) is closer to μ(AML) or μ(ALL).

Magnitude of vote is w_g , v_g :
w_g is |p(g,c)|
v_g is $| Xi - \frac{\mu_1(g) + \mu_2(g)}{2} |$
Class with higher total vote is the predicted class, if
$\frac{Vwin - Vlose}{Vwin + Vlose} > 0.7$

3.1.3 Evaluation of class predicton

- Leave–one–out cross-validation (LOOCV) removes one sample, generates predictor based on 37 samples, tests on removed sample.
- 34 new independent samples. Strong predictions were obtained for 29 out of 34 samples.

4. General observations

- Metric for measuring relationships between genes
  a) Pearson correlation measure
  b) Euclidian distance
- Single gene predictors: statistical justification of weighted voting.

-----Euclidian distance

Suppose samples are drawn randomly from large population.

Idealized model: for each gene and class, we have distribution function that maps samples from class to expression levels.

Suppose we knew this distribution. Let x be expression level of sample measured. The vote of a gene can be determined from

$$Vg = \ln \frac{prob(x \mid sample \in C1)}{prob(x \mid sample \in C2)}$$

Assume it is a normal distribution, then

$$Vg = (x - \frac{\mu 1(g) + \mu 2(g)}{2})(\frac{\mu 1(g) - \mu 2(g)}{\sigma^2})$$

5. Select subset of informative genes

The estimate approaches S(μ,σ) when n/log m →infinity, where μ is the number of samples and σ is the number of genes. Selected subset is based on estimates μ1, μ2, σ1, σ2.