Clustering and reverse engineering: from genes to the metabolome

Humberto Ortiz Zuazaga

University of Puerto Rico
High Performance Computing facility

September 18, 2009

Outline

1 Introduction

2 Clustering

3 Reverse Engineering

Bioinformatics

“The creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems posed by or inspired from the management and analysis of biological data.” — Wikipedia

Computational biology

The application of computers to the collection, analysis, and presentation of biological information.

Metabolomics

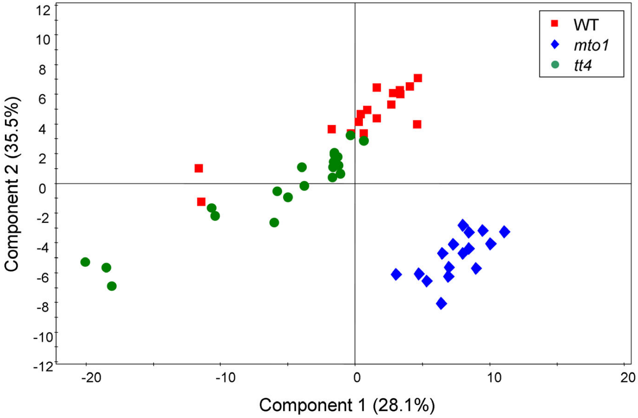
“the chemical profiling of (all) cellular metabolites by their identification and quantification.” [1]

[1] Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in Arabidopsis thaliana. Miyako Kusano, Atsushi Fukushima, Masanori Arita, Pär Jonsson, Thomas Moritz, Makoto Kobayashi, Naomi Hayashi, Takayuki Tohge and Kazuki Saito. BMC Systems Biology 2007, 1:53 doi:10.1186/1752-0509-1-53

Clustering

Dividing the elements of a set into related subsets based on a distance metric among elements.
Question: What other biological problem groups elements based on their ”distance”?

Principal components



Transcriptional clustering

- Microarrays measure abundance of many (all) genes in a sample.
- Microarray analysis makes extensive use of clustering.
- Extensive review in PMID: 11099257

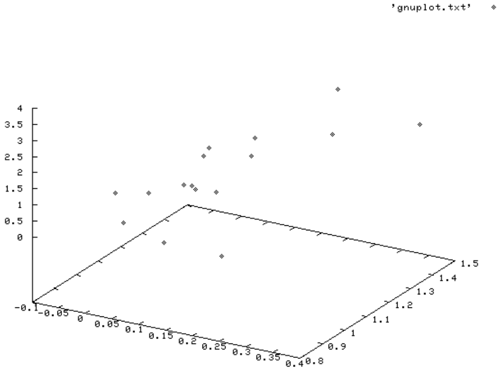
What is a distance metric?

Distance metrics in microarray analysis

- Euclidean distance
- Mutual information
- Coeficient of correlation

Euclidean distance

- according to Euclid's formula for geometric distance
- can generalize to n dimensions



Common clustering techniques

- Hierarchical - Eisen et al
- K Means, Fuzzy K Means
- Self Organizing Maps (SOM) - GENECLUSTER
- Support Vector Machines (SVM)
- clique graphs - Amir Ben-Dor

CLUSTER

- Eisen et al PNAS `http://rana.lbl.gov/papers/Eisen_PNAS_1998.pdf`
- `http://rana.lbl.gov/`
- Free software and manuals (registration required)
- Question: what clustering technique and distance function?

GENECLUSTER

- Tamayo et al PNAS `http://www.pnas.org/cgi/content/abstract/96/6/2907`
- Question: what clustering technique and distance function?

Reverse engineering gene expression networks

Deduce patterns of gene regulation from measured expression data.

Inference Techniques

- Boolean networks
- Mutual information
- Linear networks
- Neural Networks

A Comparison of Genetic Network Models, L.F.A. Wessels, E.P. Van Someren, and M.J.T. Reinders; Pacific Symposium on Biocomputing 6:508-519 (2001).

Boolean networks

- Represent gene levels and stimuli as on or off
- Very simple biological model, simple computational approach

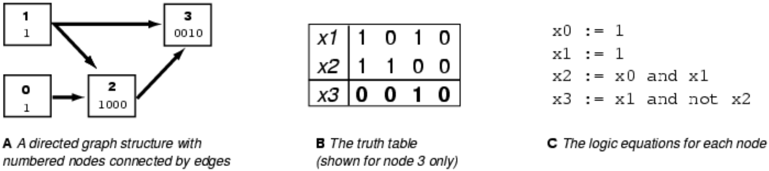
Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design, T.E. Ideker, V. Thorsson, and R.M. Karp; Pacific Symposium on Biocomputing 5:302-313 (2000).

Boolean formulas

True 1, False 0, and ($\wedge$), or ($\vee$), not ($\neg$)

$$\begin{aligned} 1 \wedge 0 &= 0 \\ 1 \wedge 1 &= 1 \\ 1 \vee 0 &= 1 \\ 1 \vee 1 &= 1 \\ \neg 0 &= 1 \\ \neg 1 &= 0 \end{aligned}$$

Boolean network



A *A directed graph structure with numbered nodes connected by edges*

x1	1	0	1	0
x2	1	1	0	0
x3	**0**	**0**	**1**	**0**

B *The truth table (shown for node 3 only)*

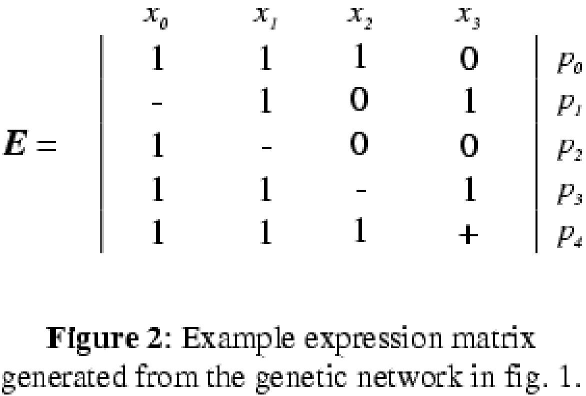
```
x0  := 1
x1  := 1
x2  := x0 and x1
x3  := x1 and not  x2
```

C *The logic equations for each node*

Figure 1: Example of the Boolean steady-state network model

Expression matrix

For a set of genes and a set of perturbation experiments construct an expression matrix as shown:



$$E = \begin{array}{cccc|c} x_0 & x_1 & x_2 & x_3 & \\ \hline 1 & 1 & 1 & 0 & p_0 \\ - & 1 & 0 & 1 & p_1 \\ 1 & - & 0 & 0 & p_2 \\ 1 & 1 & - & 1 & p_3 \\ 1 & 1 & 1 & + & p_4 \end{array}$$

Figure 2: Example expression matrix generated from the genetic network in fig. 1.

Inference Procedures

- From the expression matrix, the *Predictor* generates (possibly several) network hypothesis
- The *Chooser* selects a new perturbation experiment, that would best discriminate between available hypotheses.

Predictor

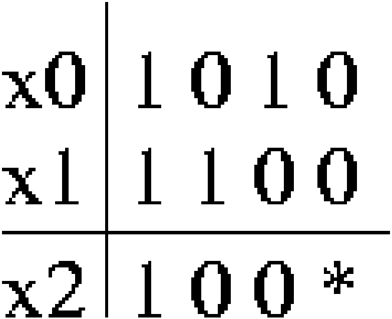
- Look at all pairs of experiments where a given gene differs except where it is forced (-, +).
- Build a multiset of all other genes that also changed between those rows.
- Construct the hitting set, the smallest set of elements such that there is a member of each subset.
- Generate the boolean functions by inspection of the members of the hitting set.

The Predictor in action

- x0 - no changes
- x1 - no changes
- x2 - row pairs, set
  - (0,1) x0, x3
  - (0,2) x1
  - (1,4) x0
  - (2,4) x1, x3
  - hitting set Smin = x0, x1

Generating the boolean functions

- The truth table for x2 can be generated by looking at the values seen for the members of Smin
- The '*' represents an unknown value (x0 and x1 are never 0 in the same experiment)



x0	1	0	1	0
x1	1	1	0	0
x2	1	0	0	*