Data Mining Whole-Genome Expression Profiling

Internship Presentation
Computational Biosciences
Arizona State University

Advisors: Dr. Seth Dobrin
Dr. Dietrich Stephan
Intern: Shruti Lal
Schizophrenia:

- Schizophrenia is a chronic and debilitating mental illness.
- The illness affects approximately 1% of the population worldwide.
- It most commonly appears early 20’s in males and early 30’s in females.
- There is often no cure for the illness.
**Signs and Symptoms:**

**Negative signs**

- Dulled emotions (lack of expression), inappropriate display of emotions (laughing while expressing terrifying images) and a change in speech (speaking in a dull monotone)
- Trouble functioning at work or in social situations
- Difficulty with personal hygiene
Positive signs

- Delusions — personal beliefs not based in reality, such as paranoia that you're being persecuted or conspired against
- Hallucinations — sensing things that don't exist, such as imaginary voices
Affected parts of Brain

Source: Dr. Seth Dobrin
Causes:

- The cause of schizophrenia has not yet been identified
- Genetic factors also play a major role
- Chemical or subtle structural abnormalities in the brain may contribute to causing this illness
- Schizophrenia may exist alone or in combination with other psychiatric or medical conditions
Treatments

- There is no cure
- Medications can relieve 80% of symptoms
- Antipsychotic medications (neuroleptics) are the cornerstone of treatment
- Individual therapy, Family therapy and Rehabilitation
Microarray Chips

Types of Microarray:

- Two color (Agilent)
  - DNA molecules made as a copy of mRNA
  - 60-mers

- Oligo based (Affymetrix)
  - Oligomer short polymer
  - Also called GeneChips
  - 25-mers
Microarray Data

Source: Affymetrix/Agilent Website

Affymetrix

Agilent
cDNA technology

Source: Agilent Website
Affymetrix technology

Source: Affymetrix Website
Continued…

Human Genome U133A GeneChip® Array

(1) Probe Array

(2) Probe Set

(3) Probe Pair

Each Probe Pair contains 11 Probe Pairs (PM:MM) of different probes.

(4) Probe Cell

Each Probe Cell contains ~40x10^6 copies of a specific probe complementary to genetic information of interest. probe — single stranded, sense, fluorescently labeled oligonucleotide (25 mers).

The Human Genome U133 A GeneChip® array represents more than 22,000 full-length genes and EST clusters.

Source: Affymetrix Website
Sample Data

- SczD Samples: 21
- Control Samples: 19

- Affymetrix HG-U133 Plus 2.0 with 54,675 genes
- The postmortem brain samples were donated by Stanley Medical Research Institute
# Tgen’s Data

<table>
<thead>
<tr>
<th>Sample</th>
<th>Life time alcohol use</th>
<th>Life time drug use</th>
<th>Suicide Status</th>
<th>Smoking at TOD</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-041_Signal</td>
<td>HEAVY</td>
<td>HEAVY</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>43</td>
</tr>
<tr>
<td>A-094_Signal</td>
<td>HEAVY</td>
<td>HEAVY</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>42</td>
</tr>
<tr>
<td>A-092_Signal</td>
<td>HEAVY</td>
<td>HEAVY</td>
<td>YES</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>42</td>
</tr>
<tr>
<td>A-017_Signal</td>
<td>HEAVY</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>53</td>
</tr>
<tr>
<td>A-100_Signal</td>
<td>HEAVY</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>59</td>
</tr>
<tr>
<td>A-045_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>35</td>
</tr>
<tr>
<td>A-075_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>54</td>
</tr>
<tr>
<td>A-079_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>NO</td>
<td>Male</td>
<td>41</td>
</tr>
<tr>
<td>A-087_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>YES</td>
<td>UNKNOWN</td>
<td>Female</td>
<td>44</td>
</tr>
<tr>
<td>A-044_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>YES</td>
<td>NO</td>
<td>Male</td>
<td>39</td>
</tr>
<tr>
<td>A-082_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>33</td>
</tr>
<tr>
<td>A-039_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>47</td>
</tr>
<tr>
<td>A-081_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>32</td>
</tr>
<tr>
<td>A-101_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>52</td>
</tr>
<tr>
<td>A-066_Signal</td>
<td>HEAVY</td>
<td>MODERATE</td>
<td>YES</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>45</td>
</tr>
<tr>
<td>A-068_Signal</td>
<td>LITTLE</td>
<td>MODERATE</td>
<td>YES</td>
<td>UNKNOWN</td>
<td>Female</td>
<td>36</td>
</tr>
<tr>
<td>A-097_Signal</td>
<td>MODERATE</td>
<td>MODERATE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>46</td>
</tr>
<tr>
<td>A-055_Signal</td>
<td>MODERATE</td>
<td>SOCIAL</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>47</td>
</tr>
<tr>
<td>A-011_Signal</td>
<td>MODERATE</td>
<td>SOCIAL</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>51</td>
</tr>
<tr>
<td>A-088_Signal</td>
<td>SOCIAL</td>
<td>SOCIAL</td>
<td>YES</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>43</td>
</tr>
<tr>
<td>A-093_Signal</td>
<td>SOCIAL</td>
<td>UNKNOWN</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>47</td>
</tr>
<tr>
<td>B-095_Signal</td>
<td>SOCIAL</td>
<td>SOCIAL</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>31</td>
</tr>
<tr>
<td>B-077_Signal</td>
<td>LITTLE</td>
<td>SOCIAL</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Female</td>
<td>33</td>
</tr>
<tr>
<td>B-049_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>34</td>
</tr>
<tr>
<td>B-053_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Female</td>
<td>34</td>
</tr>
<tr>
<td>B-043_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>35</td>
</tr>
<tr>
<td>B-033_Signal</td>
<td>MODERATE</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Female</td>
<td>38</td>
</tr>
<tr>
<td>B-090_Signal</td>
<td>HEAVY</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>30</td>
</tr>
<tr>
<td>B-060_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>40</td>
</tr>
<tr>
<td>B-052_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>NO</td>
<td>Male</td>
<td>45</td>
</tr>
<tr>
<td>B-051_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>46</td>
</tr>
<tr>
<td>B-054_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>NO</td>
<td>Male</td>
<td>47</td>
</tr>
<tr>
<td>B-04_Signal</td>
<td>LITTLE</td>
<td>SOCIAL</td>
<td>NO</td>
<td>NO</td>
<td>Male</td>
<td>48</td>
</tr>
<tr>
<td>B-079_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>49</td>
</tr>
<tr>
<td>B-096_Signal</td>
<td>SOCIAL</td>
<td>SOCIAL</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>48</td>
</tr>
<tr>
<td>B-071_Signal</td>
<td>LITTLE</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Female</td>
<td>49</td>
</tr>
<tr>
<td>B-080_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>50</td>
</tr>
<tr>
<td>B-091_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>51</td>
</tr>
<tr>
<td>B-024_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>UNKNOWN</td>
<td>Male</td>
<td>53</td>
</tr>
<tr>
<td>B-70_Signal</td>
<td>SOCIAL</td>
<td>LITTLE</td>
<td>NO</td>
<td>YES</td>
<td>Male</td>
<td>55</td>
</tr>
</tbody>
</table>
Steps of a Microarray Experiment

1. Biological Question
2. Experimental Design
3. RNA Extraction
4. Hybridization
5. Scanning
6. Image Quantification
7. Data Analysis
8. Biological verification and interpretation
Biological Problem

- Is a gene or set of genes differentially expressed in the Schizophrenia from Control?
- How many of such genes expressed?
- How to identify these genes?

“Detect all genes that are differently expressed between Experiment and Control”
Overview of Microarray Data Analysis

- Visualization
  - MA plots
  - Box plots
- Correcting for systematic errors in data
  - Normalization
- Detecting differently expressed genes
  - Statistical analysis
- Further analysis and visualization of data
  - Principle Component Analysis
  - Clustering
  - Gene Ontology
Normalization

The purpose of normalization is to minimize the variations in the gene expression levels of the samples.

Normalization is a particular type of pre-processing step that extracts or enhances meaningful data characteristics.
Normalization methods for Affymetrix data or GeneChip

- MAS 4.0/5.0: Quantile, Lowess
- RMA: Quantile (uses only PM)
- MBEI/dChip: Model-based Expression Index
Normalization

Box Plots:

Before

After

Source: R (AffyImGUI)
Continued...

Intensity Plots:

Source: R (AffyImGUI)
Continued…

M-A Plots:

Source: R (AffyImGUI)
Differential Gene Finding

Methods:

- Student t-test
- Fold Change
- FDR (False Discovery Rate)
Gene Expression

Ho $= \overline{X}$

Ha $\neq \overline{X}$

$\overline{X}_S$

$\overline{X}_C$

Ho $\rightarrow \overline{X}_S = \overline{X}_C$

Ha $\rightarrow \overline{X}_S \neq \overline{X}_C$

Ho = null hypothesis

Ha = alternate hypothesis
Student t-test

- Ho $\Rightarrow \overline{X}_s = \overline{X}_c$ and Ha $\Rightarrow \overline{X}_s \neq \overline{X}_c$

- $t = \frac{\overline{X}_s - \overline{X}_c}{\sqrt{\frac{S^2_s + S^2_c}{n_s + n_c}}}$

- If $Pr(-t < X < t) < 1 - \alpha$, accept Ho otherwise reject Ho and accept Ha.
Student t-test (Example)
Excel Plug-Ins: T-Tests

- Age: 30-40 / 40-50 / 50-60
- Gender: Male / Female
- Smoking: Yes / No
- Alcohol Use: Little / Moderate / Heavy / Social
- Drug Use: Little / Moderate / Heavy / Social
- Suicidal Status: Yes / No
Smoking/Non-smoking (SCZD / Control)

Source: MS Excel

Only 5 genes P-Value < 0.0001 for Non-Smoking
Age (Sczd/Control)

Source: MS Excel

Age 30-40 P-Value < 0.005 (55 Genes)

Age 40-50 P-value < 0.005 (250 Genes)

Age 50-60 P-Value < 0.005 (75 Genes)
Discussion (t-test)

- Smoking
  - The results of t-test showed that there are more numbers of genes found significantly regulated in samples who smoked during their life time.

- Age
  - The group 40-50 showed more significant genes relative to 30-40 and 50-60 age groups.
Fold Change

- $r = \text{target} / \text{reference}$
- If $r_i > r_{th}$ then Gene $g_i$ is differentially expressed
Fold Change Results

Fold change = 1.0
Genes found = 344

Source: MS Excel
Fold Change Results

Fold change = 1.0
Genes found = 344
Discussion (fold change)

- Genes selected by FC process can be consistently compared between experiments and enables the user to globally extract information with a high degree of confidence.

- Since it uses constant threshold for all genes, it will introduce false positives at the low end and reduce true positives at high end.
Genes and Arrays Expression

- Hierarchical Clustering
- K-means Clustering
Biological Question

- Can genes be grouped on the basis of similar transcription behavior?
- Can genes be grouped on the basis of similar cellular functions?
- Can arrays be grouped on the basis of biological similarities?
- Can data be adjusted in some way to enhance certain relationship?
- What distance measure should be used to group together related genes?
Types of Hierarchical Clustering

- Agglomerative (bottom-up)
  - Most commonly used approach
  - Begins with singletons and merge them until root is achieved
- Divisive (top-down)
  - Opposite of agglomerative
  - Begins with root and recursively partition the root until singleton sets are achieved
Hierarchical Clustering

Dendrogram
Types of Linkage in Hierarchical Clustering

- Single Linkage
- Complete Linkage
- Centroid Linkage
- Average Linkage
Hierarchical Clustering Results

Gender: SCZD/Control

Source: Cluster/Java TreeView

Average
Centroid
Complete
Single

Coloring from green to red reflect low to high expression.
Continued...

Suicide:

Source: Cluster/Java TreeView

View Status
Select Node to view annotation.

Array Tree

Array Names

Usage Hints
Click to select node – use arrow keys to nav

GeneAnnotation
Text View

GeneTree

Global Pixel View

Zoom Pixels

LOC389123
NCC27005
C14orf25
KIAA1193
PTPN2
LOC339229
C14orf145
HLA-DQA1
TAF15
ZNF605
FLJ20313
ID2
KRTAP1–3
TSFAN-1
CHD3
FLJ36144
FLJ25143
DRG1
IMF-3
EXO1
EFG8
Continued…

Smoking: SCZD/Control

Source: Cluster/Java TreeView
Continued…

Source: Cluster/Java TreeView

Smoking Gene Tree:
Continued...

Source: Cluster/Java TreeView
Continued...

Smoking Array Tree:

Source: Cluster/Java TreeView
Discussion about Hierarchical clustering results

- Smoking plays a vital role in Schizophrenia
- Biological question: Is Smoking a cause of Schizophrenia or does Schizophrenia increase the propensity to smoke.
- The reality is literature says 99% Schizophrenics smoke.
- Not much confidence in clustering results due to very few non smoking samples.
**K-means Clustering Algorithm**

- Pick a number (k) of cluster centers
- Assign every gene to its nearest cluster center
- Move each cluster center to the mean of its assigned genes
- Repeat 2-3 until convergence
**k-means Algorithm**

**Step 1**

![Graph showing the expression in condition 1 vs. expression in condition 2](image)
Continued…

Step 6

![Graph showing expression in condition 1 vs. expression in condition 2 with markers k_1, k_2, and k_3.](chart.png)
**K-means Clustering Results**

Source: Cluster/Java TreeView

**Sczd/Control:**

Cluster 1 = 18/19  
Cluster 2 = 20/21

Array: \( k = 2 \)  
Iterations = 100

05/16/05
Linked to hypothyroidism and brain edema following trauma

Multicatalytic proteinase complex

Smoking:
Genes: k = 10
Iterations = 100
Continued...

Suicide:

Arrays: k = 2
Iterations = 100

Suicide sample: 5
Non-suicide sample: 16
Continued…

Gender:

Arrays: k = 4
Genes: k = 4

Source: Cluster/Java TreeView

- Responsible for neural cell protection and cell survival
- Multicatalytic Proteinase complex
- Glycolytic signaling processes
- Molecular function and cellular activities (Kinase activity, nucleic acid binding)
Discussion about k-means clustering results

- It was not surprising that the genes that are expressed together share common functions.
- The extent to which gene expression patterns separate genes into functional categories across a relatively small collection of condition is surprising.
- On the basis of observations, it is possible that many of the genes of other confounding factors share common functions.
- The array clustering based on gender did not produce expected results. This supports the fact that the schizophrenia is not biased towards any gender.
- The array clustering based on suicide also did not produce expected results.
Problems faced

- Excel crashed many times
- FDR ran for 7 consecutive days and finally did not produce any significant gene for $Q = 0.05$ and gave approximately 30,000 significant genes for $Q = 0.9$ and gave zero significant genes for $Q = 0.89$
Conclusion about the project

- Microarray technology offers an opportunity to generate functional data on a genome wide scale.
- The biological variance in the patients recapitulates the complex variance in phenotypes seen in SCZD patients.
- The expression of genes found using statistical tests are similar to the findings in the literature.
- All the explorations provide hypotheses that can be tested in the laboratory using standard biological methods.
- SNP analysis confirms the multitude of genotypes that can cause SCZD.
- Found the previously discovered candidate genes (DISC-1, Syanpsin-II).
Future Work

- Biological interpretation using GO browser
- Biological verification using gene regulatory network models (Bayesian and Boolean networks)

"If gene $X_1$ is active and gene $X_2$ is suppressed, gene $Y$ would be activated"
Acknowledgements

- Dr. Seth Dobrin
- Dr. Phillip Stafford
- Dr. Jeff Touchman
- Dr. Dietrich Stephan
References

- **Normalization**
  - Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application-
  - Exploration, normalization, and summaries of high density oligonucleotide array probe level data Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Terence P. Speed
  - A comparison of normalization methods for high density oligonucleotide array data based on variance and bias - B.M. Bolstad, R.A.Irizarry, M.Astrand and T.P. Speed

- **Differential gene finding**
  - A mixture model for estimating the local false discovery rate in DNA microarray analysis-J.g. Liao, Yong Lin, Zachariah E. Selvanayagam and Weichung Joe Shih
  - Identifying differentially expressed genes using false discovery rate controlling procedures- Anat Reiner, Daniel Yekutieli and Yoav Benjamini
Clustering

- Cluster analysis and display of genome-wide expression patterns - Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, And David Botstein
- Inference from Clustering with Application to Gene-Expression Microarrays - Edwardr. Dougherty, junior barrera, Marcel brun, Seungchn Kim, Robertom. Cesar, Yidong chen, Michael bittner, And Jeffreym. Trent
- Data Clustering: A Review- A.K. Jain, M.N. Murty And P.J. Flynn
- Cluster analysis of gene expression dynamics - Marco F. Ramoni, Paola Sebastiani, and Isaac S. Kohane

- Dr. S. Kim Lecture notes
- Affymetrix and Agilent websites