Data Analysis of 3 Types of Salmonella: Entiriditis, Newport, and Typhimurium

Dmitry Devetyarov, Zhiyuan Luo, Muna Anjum (VLA),
Nick Coldham (VLA), Alex Gammerman

January 2008

1 Microarray Data Description

The microarray data consists of preprocessed Salmonella strains of 3 classes:

- Newport
- Entiriditis
- Typhimurium

In addition, we are provided with 7 *controls* - the strains on which the array was based originally. These are DT104, SL1344, Lt2 (3 Typhimurium strains), CT18 (Typhi strain that provided the majority of genes of microrray), PT4 (Entiriditis strain), S.gallinarum, S.bongori.

Each strain is represented by at least two replicates. The breakdown of the dataset into 3 classes is shown in Table 1.

Table 1: The breakdown of data into classes.

Class	Number of strains	Overall number of replicates
Entiriditis	24	51
Newport	14	33
Typhimurium	20	40
Controls	7	14
In total (3 classes + controls)	65	138

Strains of all classes were normalised together (which is a crucial difference from the data available before).

Initially, microarrays comprised 5592 genes (features). After elimination of genes with missing data ('No Data' fields), 4736 genes were left.

2 Data Analysis

2.1 Replicate Variability

The variability of replicates of samples has been verified. The coefficients of variation (CV) have been calculated for replicates of each sample and for each gene.

CV appeared to be quite high:

- mean value across all genes and samples = 20.8%
- maximum value for a gene and for sample = 159.4%

Nevertheless, all further analysis was carried out with values averaged across replicates of the same strain.

2.2 Discrimination between Classes

The aim of data analysis is to separate three classes of Salmonella strains: Entiriditis, Newport, and Typhimurium.

2.2.1 Heat Map

Figure 1 represents a heat map for three types of Salmonella (24 Entiriditis strains, 10 Newport strains, 20 Typhimurium strains) and 4736 genes. The figure demonstrates visually clear discrimination. It can be seen from the heat map that for each pair of classes there are groups of genes, each of which separates these two classes (that is, for any of these genes you can find the threshold that separates these two classes). Meanwhile, it is not visually clear whether there is a gene that separates all three classes from each other.

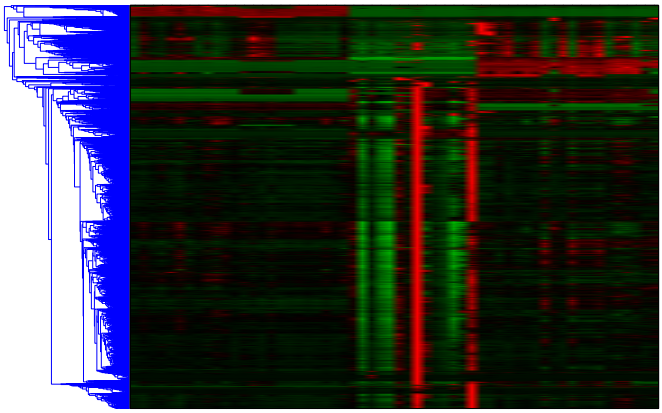

Figure 1: The heat map of 24 Entiriditis strains, 10 Newport strains and 20 Typhimurium strains.

2.2.2 One-against-one Separation

For each pair of classes and for each gene, we found the threshold that provides the smallest error rate when predicting using the cutoff model with this threshold. As a result, for each pair of classes we found a big group of genes, each of which separates these two classes (that is, for any of these genes you can find the threshold that separates these two classes):

- Entiriditis vs Newport: 218 genes with names starting with SBG, SDT, SEN, SG, STM, STY, PSLT;
- Newport vs Typhimurium: 145 genes with names starting with AF, SBG, SDT, SEN, SG, STM, STY, PSLT;
- Entiriditis vs Typhimurium: 336 genes with names starting with SBG, SEN, SG, STM, STY.

2.2.3 Discrimination between Three Classes

The analysis showed that there is the only gene (STM2087) that separates any pair of classes, that is, can separate all three classes. The distribution of STM2087 values on the logarithm scale is given in Figure 2. The figure shows that the values of these gene have the following relationship for different classes of Salmonella:

$$\text{Newport} < \text{Entiriditis} < \text{Typhimurium}$$

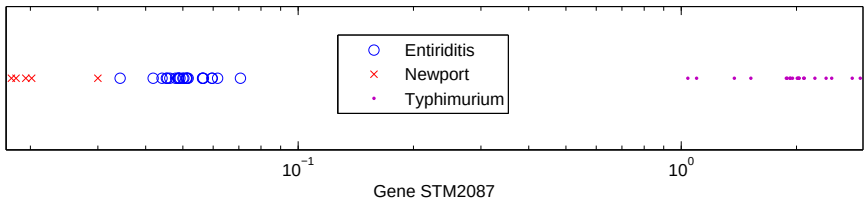

Figure 2: The distribution of STM2087 values for three classes on the logarithm scale.

Thus, all three classes are easily separable by one gene, and no further analysis with application of complicated machine learning algorithms is needed.

2.2.4 Kruskal-Wallis p-values

We calculated Kruskal-Wallis p-values that reject the null hypothesis that three classes of samples are drawn from populations with the same distribution. The minimum p-value 1.5063e-011 corresponds to the same gene STM2087. Top 20 Kruskal-Wallis p-values are listed in Table 2. 558 genes (out of 4736) have Kruskal-Wallis p-values less than 10^{-6}.

This proves that there are lot of genes that discriminate two classes out of three. However, these p-values cannot prove or disprove the ability to separate all three classes.

Table 2: Top 20 Kruskal-Wallis p-values for a single gene.

Gene Name	K-W p-value
STM2087	1.51E-11
STM2234	2.24E-11
SEN2615	3.15E-11
STM2767	3.25E-11
STY2296	3.48E-11
SEN1384	4.33E-11
STM2232	4.81E-11
STM2743	4.81E-11
STM4497	5.19E-11
STM4496	6.55E-11
STM2745	6.62E-11
STM2741	7.15E-11
SEN1996	7.45E-11
SG4353	7.67E-11
SEN1991	8.40E-11
STM2230	8.53E-11
STM2940	8.82E-11
SEN1752	9.37E-11
SG1833	9.77E-11
SG2007	9.82E-11