BACHELOR

Multiple testing and an application on a data set

Thonnard, M.

Award date:
2005

Multiple Testing and an Application on a Data Set
Bachelorproject

Eindhoven University of Technology

M. Thonnard (0540518)
Department of Mathematics & Computer Sciences

16th November 2005

Contents

Chapter 1

Introduction

If we consider 2 random diseases, it can be seen that they have many differences, namely effects, treatments, and many other things. If we consider 2 diseases from the same family, for example normal glucose tolerance and type 2 diabetes mellitus, then the differences are maybe not that clear. The main question in this case is: 'Are there differences between the genes of 2 persons, who have these diseases?' In this report, we will try to find a method to research if there are differences between the genes of persons with these diseases.

Section 2 deals with the experiment and also gene expression is explained. Section 3 contains the model with 1 pair of male and 1 group of genes. Section 4 contains the model with 1 pair of male and all groups. Section 5 explains the model with n pairs of males and all groups. Section 6 deals with the multiple testing theory and the question: 'Can PROC MULTTEST of SAS handle the model with 17 pairs of males and all groups?' And finally, section 7 contains the conclusion.

Chapter 2

Experiment

In this experiment 43 age-matched males are used, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 18 with type 2 diabetes mellitus (DM2). For each male, gene expression, which is explained in section 2.1, is measured for 22283 genes, which covers a large part of the human genome. The genes are divided in 150 groups, which represent biological pathways, and each gene may occur in more than one group. This is done because relevant differences are subtle at the level of individual genes. Only 9120 genes occur in one of the 150 groups. Therefore, we only consider these genes. DNA microarrays are used to identify gene expression changes. These measurements are put in a table.

In the remainder of this report we only consider the diseases normal glucose tolerance (NGT) and type 2 diabetes mellitus (DM2). We want to discover whether differences between these diseases are present. We pair the data to do so, because we want to apply these data to the cDNA data, which are usually paired. We take arbitrarily 2 males, each with a different disease, to compare the 2 diseases, such that we have 17 different paired data (norm1_ngt,..., norm17_ngt). First of all, we take the maximum of 3 and the logarithm of the data, because the data may become too little. Then we normalize the data and finally we take the difference between the data of NGT and these of DM2. These new data are the response data. The 150 groups are the regressors. If the gene occurs in the group, we label it with a 1, otherwise with 0. In the following sections, we consider 3 different models for these data.

2.1 Gene expression

A gene is a little piece of the DNA chain. Genes are codes for proteins. A gene expresses on the moment that an organism commands to do this. This message is send to the gene for this protein in the nucleus. This piece of gene is copied to an other molecule, namely RNA. This is called messenger RNA (mRNA). The mRNA goes from the nucleus to the cell and his information is used to make the proteins. Gene expression of one particular gene is the number of that gene in the RNA sample.

DNA microarrays are perfectly suited for comparing gene expression in different populations of cells. A flash animation of the DNA microarray methodology can be seen at the website [1].

Chapter 3

1 pair and 1 group

In the first model we consider one pair of males and one regressor, namely one group. That is to say, we want to find out whether there is a significant difference between the 2 diseases for these 2 males in a particular group. If we fix group i= I, then the model is

$$y_{Ij} = \alpha_I x_{Ij} + \varepsilon_{Ij},$$

where $\varepsilon_{Ij} \sim N(0, \sigma_I^2)$ and x_{Ij} represents group I. This model has the regression form. The model does not need an intercept, because the data are normalized. If we fit an intercept, then the main effect is stored in the intercept. Because we are interested in the effect of all separate groups, we do not use an intercept. In this case, x_{Ij} is 1 if gene j occurs in group I, otherwise it is 0. Therefore we can rewrite the model. Say, group i contains n_i genes, thus group I contains n_I genes. The model becomes

$$y_{Ij} = \alpha_I + \varepsilon_{Ij}, \qquad \text{j} = 1,...,n_I,$$

where $\varepsilon_{Ij} \sim N(0, \sigma_I^2)$. This is the ANOVA formulation.

Now we apply this model to the data. We choose arbitrarily a response variable, norm1_ngt, and we fix this variable. We consider if the model is significant for several groups, which are also arbitrarily chosen. First of all, we choose group 1. Then the model becomes

$$y_{1j} = \alpha_1 + \varepsilon_{1j}, \qquad \text{j} = 1,...,n_1,$$

where $\varepsilon_{1j} \sim N(0, \sigma_1^2)$. If we put these data in SAS, then we get this output.

Source	DF	Sum of Squares	Mean Square	F value	p-value
Model	1	0.05951	0.05951	0.09	0.7633
Error	9119	5983.55151	0.65616		
Uncorrected Total	9120	5983.61102			

Table 3.1: Analysis-of-Variance table of pair 1 and group 1.

The parameter estimate of α_1 is 0.01277. The p-value is 0.7633, which means that there is no significant difference between the 2 diseases in group 1. The normality assumption of the error can also be checked with a normal probability plot of the residuals (figure 3.1). We

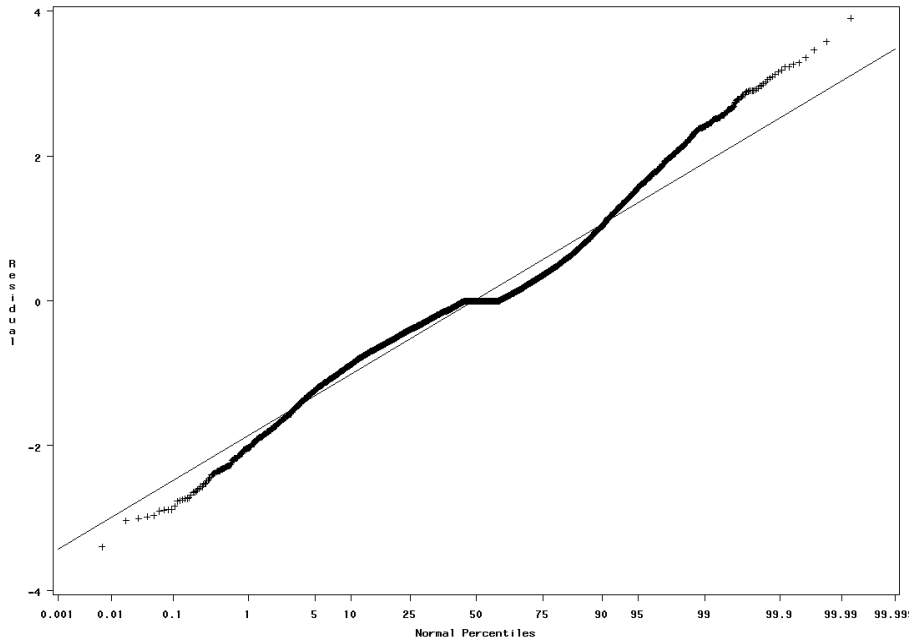

Figure 3.1: Normal probability plot of residuals.

cannot say that the normality assumption is not satisfied.

Secondly, we choose the seventh group. Then the model becomes

$$y_{7j} = \alpha_7 + \varepsilon_{7j}, \qquad \text{j} = 1,...,n_7,$$

where $\varepsilon_{7j} \sim N(0, \sigma_7^2)$. If we put these data in SAS, then we get this output.

Source	DF	Sum of Squares	Mean Square	F value	p-value
Model	1	5.20374	5.20374	7.94	0.0049
Error	9119	5978.40728	0.65560		
Uncorrected Total	9120	5983.61102			

Table 3.2: Analysis-of-Variance table of pair 1 and group 7.

The parameter estimate of α_7 is 0.14040. The p-value of this model is 0.0049 and we choose the confidence level, α, to be 0.05. We can also see in the normal probability plot that the normality assumption of the error is satisfied. We can see that this model is significant and that means that there is a difference between the 2 diseases in group 7. This does not mean that for each pair of males this model is significant. If we consider pair 16 (norm16_ngt), then the p-value is 0.5009, such that this model is not significant.

In the next chapter, we go one step further and we take all the groups into the model.

Chapter 4

1 pair and all groups

This model deals with one pair of males and all the regressors, namely 150 groups. With this model we want to research, if there is a significant difference between the 2 diseases for these 2 males. If we fit a model and it is significant, then we can not be sure whether it is a individual effect or not. The model is

$$y_j = \sum_{i=0}^{150} \alpha_i x_{ij} + \varepsilon_j,$$

where $\varepsilon_j \sim N(0, \sigma^2)$ and x_{ij} is an indicator variable. This model has no ANOVA form, because one gene may correspond to more then one group.

If we fit this model, we can see that 4 groups are a linear combination of other groups. They are non-identifiable. Therefore, we delete these 4 groups, namely group 70, group 84, group 118 and group 144.

We choose arbitrarily a response variable, one of the 17 pairs of males. Pair 1 is chosen. The SAS-output is given in table 4.1.

Source	DF	Sum of Squares	Mean Square	F value	p-value
Model	146	123.90124	0.84864	1.30	0.0092
Error	8974	5859.70978	0.65297		
Uncorrected Total	9120	5983.61102			

Table 4.1: Analysis-of-Variance table of pair 1 and all groups.

The model is significant, because of the p-value, which means that disease 1 differs from disease 2 for this pair of males. If we test the hypothesis

$$H_0 : \alpha_i = 0$$

against

$$H_1 : \alpha_i \neq 0,$$

then we come to the conclusion that only α_4, α_7, α_{24}, α_{72}, α_{121}, α_{130} and α_{137} are significant at the confidence level of 0.05. The parameter estimates of the significant α's are in table 4.2.

α_4	0.19803
α_7	0.13683
α_{24}	0.12749
α_{72}	1.27192
α_{121}	-0.95331
α_{130}	2.39005
α_{137}	1.11788

Table 4.2: Parameter estimates of significant α's.

For all the pairs of males is the model significant. For 15 pairs of males the model is significant at the confidence level of 0.05 with a p-value < 0.0001 and for one pair is the p-value 0.0394. For the third pair, 7 of the α_i's are significant. For the other pairs, more α_i's are significant. For example, for the seventeenth pair, 42 of the α_i's are significant.

Because in this model we look at one pair, the effect can be individual, if the model is significant. We want to find out whether the effect depends on the individu or on the difference between the diseases. This can be done by looking at more than one pair.

Chapter 5

n pairs and all groups

The next step is to find a model for more pairs of males, that is to say, n pairs of males are the response variable. All the groups are the regressors. This is the most interesting model, because we can give a more general answer. If the model is significant, then the chance is little that there is an individual effect.

Because there are n measurements at the same gene, we need to consider a stochastic variable, β. We assume also two things. Firstly, two measurements at different genes, but in the same pair, are assumed to be uncorrelated. Secondly, two measurements of the same gene, but in different pairs of males are assumed to be correlated. Thus, the model becomes

$$y_{jk} = \sum_{i=1}^{150} \alpha_i x_{ij} + \beta_j + \varepsilon_{jk}, \qquad \text{k=1,...,n,}$$

where $\varepsilon_{jk} \sim N(0, \sigma^2)$ i.i.d., $\beta_j \sim N(0, \tau^2)$ i.i.d. and x_{ij} is an indicator variable.

It can easily be seen that the assumptions are satisfied. For the first assumption, the covariance is

$$\begin{aligned} Cov[y_{jk}, y_{lk}] &= E[y_{jk}y_{lk}] - E[y_{jk}]E[y_{lk}] \\ &= E[(\beta_j + \varepsilon_{jk})(\beta_l + \varepsilon_{lk})] \\ &= E[\beta_j\beta_l] + E[\beta_j\varepsilon_{lk}] + E[\beta_l\varepsilon_{jk}] + E[\varepsilon_{jk}\varepsilon_{lk}] \\ &= 0. \end{aligned} \tag{5.1}$$

For the second assumption, the covariance is

$$\begin{aligned} Cov[y_{jk}, y_{jl}] &= E[y_{jk}y_{jl}] - E[y_{jk}]E[y_{jl}] \\ &= E[(\beta_j + \varepsilon_{jk})(\beta_j + \varepsilon_{jl})] \\ &= E[\beta_j^2] + E[\beta_j\varepsilon_{jk}] + E[\beta_j\varepsilon_{jl}] + E[\varepsilon_{jk}\varepsilon_{jl}] \\ &= Var[\beta_j] - (E[\beta_j])^2 \\ &= Var[\beta_j] \\ &= \tau^2. \end{aligned} \tag{5.2}$$

5.1 2 pairs

In this section we put 2 pairs of males in the model. Pair 1 and pair 2 are chosen. This model is put in SAS. Table 5.1 gives the significant groups.

Source	DF	SS	Mean Square	F value	p-value
group 4	1	15.79119	15.79119	11.87	0.0006
group 10	1	11.15118	11.15118	8.38	0.0038
group 16	1	7.50967	7.50967	5.65	0.0175
group 17	1	8.95416	8.95416	6.73	0.0095
group 31	1	6.38230	6.38230	4.80	0.0285
group 36	1	9.03690	9.03690	6.79	0.0092
group 63	1	8.68028	8.68028	6.53	0.0107
group 64	1	8.40909	8.40909	6.32	0.0119
group 96	1	7.10606	7.10606	5.34	0.0208
group 102	1	5.68600	5.68600	4.27	0.0387
group 110	1	6.11652	6.11652	4.60	0.0320
group 112	1	5.52938	5.52938	4.16	0.0415
group 129	1	8.73816	8.73816	6.57	0.0104
group 143	1	10.73623	10.73623	8.07	0.0045

Table 5.1: Significant groups at the confidence level 0.05 for 2 pairs of males.

14 groups are significant at the confidence level of 0.05, which means that there is a significant difference between the 2 diseases in these groups of genes. Because this model contains only 2 pairs of mails the differences between the 2 diseases can depend on the individuals. To get a more general answer, the model has to enlarge.

5.2 5 pairs

Now 5 pairs of males are put in the model. Pairs 1, 2, 3, 4 and 5 are chosen. The groups in Table 5.2 are significant at the confidence level 0.05. This model has not much connection with the previous model (2 pairs). Only 6 groups are significant in both models. This model needs more pairs to formulate a more general answer.

5.3 17 pairs

The model with all the pairs of males, namely 17, is the most important. This model gives a more general answer. The groups in Table 5.3 are significant at the confidence level 0.05. Thus, there is a difference between the 2 diseases in these particular groups. It can also be seen that group 147 is the most significant one, which is a confirmation of the real experiment [3]. This model has also not much connection with the model with 5 pairs of males. Only 3 groups are significant in both models. Because there is so little connection between the models, we probably need multiple testing.

Source	DF	SS	Mean Square	F value	p-value
group 4	1	6.31729	6.31729	5.56	0.0183
group 14	1	6.16114	6.16114	5.43	0.0198
group 15	1	4.86986	4.86986	4.29	0.0384
group 36	1	7.70450	7.70450	6.79	0.0092
group 53	1	4.80315	4.80315	4.23	0.0397
group 63	1	16.48745	16.48745	14.52	0.0001
group 64	1	15.79224	15.79224	13.91	0.0002
group 90	1	4.60691	4.60691	4.06	0.0440
group 93	1	5.49688	5.49688	4.84	0.0278
group 96	1	4.70164	4.70164	4.14	0.0419
group 104	1	9.76219	9.76219	8.60	0.0034
group 107	1	6.59481	6.59481	5.81	0.0160
group 112	1	13.14191	13.14191	11.58	0.0007
group 113	1	9.59743	9.59743	8.45	0.0037

Table 5.2: Significant groups at the confidence level 0.05 for 5 pairs of males.

Source	DF	SS	Mean Square	F value	p-value
group 10	1	20.682488	20.682488	19.05	<.0001
group 11	1	13.389161	13.389161	12.33	0.0004
group 12	1	9.381408	9.381408	8.64	0.0033
group 18	1	10.882049	10.882049	10.02	0.0016
group 27	1	7.597493	7.597493	7.00	0.0082
group 28	1	6.075762	6.075762	5.59	0.0180
group 67	1	5.976489	5.976489	5.50	0.0190
group 69	1	4.801250	4.801250	4.42	0.0355
group 78	1	7.472606	7.472606	6.88	0.0087
group 104	1	12.732991	12.732991	11.73	0.0006
group 106	1	5.160548	5.160548	4.75	0.0293
group 112	1	15.080036	15.080036	13.89	0.0002
group 113	1	12.493271	12.493271	11.50	0.0007
group 133	1	8.444442	8.444442	7.78	0.0053
group 136	1	9.326109	9.326109	8.59	0.0034
group 145	1	5.051222	5.051222	4.65	0.0311
group 147	1	23.494119	23.494119	21.63	<.0001
group 148	1	7.600641	7.600641	7.00	0.0082

Table 5.3: Significant groups at the confidence level 0.05 for 17 pairs of males.

Chapter 6

Multiple testing

6.1 Introduction

Incorrect conclusions arise easily from extensive data analysis and manipulation. It is not always clear whether bad data or data manipulation cause false conclusions. If 20 aspects of a data set are considered, then it is expected that one result will be significant, even if no effects are real. This is called the multiplicity problem. A single question can become multiple questions, resulting in a large collection of tests. In this situation, multiple testing is particularly appealing. A general solution for the multiplicity problem is the use of resampling methods. The resampling approach displays results as adjusted p-values.

After a theoretical introduction, we will try to apply multiple testing on the problem with 17 pairs of males and all groups. The mean question is 'Can proc multtest of SAS handle this model?'.

6.2 Familywise error rates

In the multiple testing applications, there are many hypotheses and their alternatives, say H_1 vs. H_1', H_2 vs. H_2',..., H_k vs. H_k'. A Simultaneous Test Procedure (STP) is any procedure by which each of these k hypotheses is determined to be accepted or rejected at a level α. STP's are commonly devised to control the Familywise Error Rate (FWE). Two kinds of FWE's are defined:

FWEC = Pr(reject at least one H_i | all H_i are true),

FWEP = Pr(reject at least one H_i, i=j_1,...,j_t | H_{j_1},...,H_{j_t} are true).

A STP is said to control the FWE in the weak sense if $FWEC \leq \alpha$ and in the strong sense if $FWEP \leq \alpha$. It is more desirable that an STP strongly control the FWE.

6.3 Adjusted p-values

In most cases it is possible to compute an adjusted p-value, $\tilde{p_i}$, i=1,...,k, for each test of H_i vs. H_i'. If $\tilde{p_i} \leq FWE = \alpha$, then H_i is rejected. The definition is as follows:

$\tilde{p}_i = \inf\{\alpha \mid H_i \text{ is rejected at FWE}=\alpha\}$.

Thus, the adjusted p-value is the smallest significance level for which one still rejects H_i.

6.4 Statistical resampling

Resampling methods refers to methods in which the observed data are used repeatedly. If it were possible, an experimenter would repeat the experiment. This is what resampling does with a computer, namely the observed variable's values are randomly re-assigned to treatment groups, and the test statistics are recomputed. This is done thousands of times. The original test statistic is considered unusual if it is unusual compared to the resampling distribution. The bootstrap method is one of the resampling methods which is considered in [2]. For example, we have a random sample, Y_1, Y_2,...,Y_n, from a large population with G as its distribution function. The bootstrap is a method for approximating the probability

$$\alpha_n = \Pr(\tfrac{\bar{Y}-\mu(G)}{s/\sqrt{n}} \geq 1.96 \mid G)$$

This probability can be estimated by substituting an estimate $\hat{G}$, the empirical distribution function, for G. The empirical distribution function is

$$\hat{G}(y) = \tfrac{1}{n}[\sharp \text{ of } Y_i\text{'s} \leq \text{y}]$$

After substituting

$$\hat{\alpha_n} = \Pr(\tfrac{\bar{Y}^*-\mu(\hat{G})}{s^*/\sqrt{n}} \geq 1.96 \mid \hat{G})$$

With the distribution of $\bar{Y}^*$ induced by $\hat{G}$ and the distribution of $\bar{Y}$ induced by G. In bootstrap resampling, one creates random variables having distribution $\hat{G}$. Let Y_1^*, Y_2^*,..., Y_n^* be generated as follows:

$$Y_i^* = \begin{cases} Y_1, & \text{with probability} \frac{1}{n} \\ Y_2, & \text{with probability} \frac{1}{n} \\ \ldots & \\ Y_n, & \text{with probability} \frac{1}{n}. \end{cases}$$

Each Y_i^* has as distribution the empirical distribution function $\hat{G}$. A random sample, Y_1, Y_2,...,Y_n, can be obtained *with replacement*. This means that the value of Y_1^* a random selection from the observed data Y_1, Y_2,...,Y_n, say Y_j is. Next Y_j is returned to the collection, and Y_2^* is sampled from the same collection. This continues until all values are selected. Some observed data Y_j can appear more than once.

Another method is the *without replacement* method. In this case the values are not replaced after having been selected, such that each original value Y_j appears only once among the sampled values. This method is not useful for estimating the probability that the t-statistic exceeds 1.96.

The probability $\hat{\alpha}_n$ can be estimated *with replacement* with the following algorithm:

Algorithm 1: A bootstrap probability estimate

0. Initialize a counting variable: COUNT = 0.
1. Generate resampled data Y_1^*, Y_2^*,..., Y_n^*, a *with replacement* sample from the original data Y_1, Y_2,...,Y_n.
2. If

$$\frac{\bar{Y}^* - \bar{Y}}{s^*/\sqrt{n}} \geq 1.96$$

then increment the count variable: COUNT $\leftarrow$ COUNT + 1.
3. Repeat steps 1-2N times, creating N resampled data sets. The estimated value of $\hat{\alpha}_n^{(N)}$ = COUNT/N.

6.5 Bonferroni method

The Bonferroni method is the simplest single-step method, which is a simultaneous test procedure (STP). Stepwise STP's allow different adjustment techniques for different hypotheses, depending upon how the hypotheses are ordered. Mostly, they are ordered according to the size of the p-values. The Bonferroni method rejects an hypothesis H_i when the p-value p_i is less than α/k, where α is the FWE level and k is the number of tests. The Bonferroni single-step adjusted p-value is

$$\tilde{p}_i = \min(kp_i,1).$$

Protection of FWEC motivates the Bonferroni single-step method:

$$\begin{aligned} Pr(\text{reject at least one } \mathrm{H_i} \mid \mathrm{H_0^c}) &= Pr(\min_{1\leq \mathrm{i} \leq \mathrm{k}} \mathrm{P_i} \ \leq \ \alpha/\mathrm{k} \mid \mathrm{H_0^c}), \\ &\leq \sum_{i=1}^{k} Pr(\mathrm{P_i} \ \leq \ \alpha/\mathrm{k} \mid \mathrm{H_0^c}). \end{aligned} \tag{6.1}$$

Assuming that all p-value distributions are $U[0,1]$ under their respective null hypotheses H_{0i}, the upper bound becomes $k(\alpha k) = \alpha$. Since α is an upper bound for (6.1), the Bonferroni method is conservative when the marginal p-value distributions are uniform.

6.6 Free step-down resampling method

For the free step-down resampling method the p-values need to be ordered, $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(k)}$. Let these ordered p-values have indexes r_1, r_2,..., r_k so that $p_{(1)} = p_{r_1}$ and so on. The indexes r_j are fixed. The free step-down adjusted p-values are defined as follows:

$$\begin{aligned} &\tilde{p}_{(1)} = \Pr(min_{l\in\{r_1,r_2,...,r_k\}} \ P_l \leq p_{(1)} \mid H_0^c) \\ &\tilde{p}_{(2)} = \max[\tilde{p}_{(1)}, \Pr(min_{l\in\{r_2,...,r_k\}} \ P_l \leq p_{(2)} \mid H_0^c)] \\ &... \\ &\tilde{p}_{(j)} = \max[\tilde{p}_{(j-1)}, \Pr(min_{l\in\{r_j,...,r_k\}} \ P_l \leq p_{(j)} \mid H_0^c)] \\ &... \\ &\tilde{p}_{(k)} = \max[\tilde{p}_{(k-1)}, \Pr(\ P_{r_k} \leq p_{(k)} \mid H_0^c)] \end{aligned}$$

The following algorithm estimates the free step-down adjusted p-values:

Algorithm 2: Free step-down resampling method

0. Initialize counting variables: $COUNT_i = 0$, i=1,...,k.
1. Generate a vector ($p^*_{r_1}$,..., $p^*_{r_k}$) from the same (or at least, approximately the same) distribution as the original p-values (p_{r_1},..., p_{r_k}) under the complete null hypothesis. Note that the sequence $\{r_j\}$ is fixed throughout the simulation. Thus the $p^*_{r_j}$ will not have the same monotonicity as the original p-values $p_{(j)}$.
2. Define the successive minima:

$$q^*_k = p^*_{r_k}$$
$$q^*_{k-1} = \min(q^*_k, p^*_{r_{k-1}})$$
$$q^*_{k-2} = \min(q^*_{k-1}, p^*_{r_{k-2}})$$
$$\ldots$$
$$q^*_1 = \min(q^*_2, p^*_{r_1})$$

3. If $q^*_i \leq p_{(i)}$, then $COUNT_i \leftarrow COUNT_i + 1$.
4. Repeat 1-3N times. Compute $\tilde{p}^{(N)}_{(i)} = COUNT_i/N$.
5. Enforce monotonicity using successive maximization:

$$\tilde{p}^{(N)}_{(1)} = \tilde{p}^{(N)}_{(1)}$$
$$\tilde{p}^{(N)}_{(2)} \leftarrow \max(\tilde{p}^{(N)}_{(1)}, \tilde{p}^{(N)}_{(2)})$$
$$\ldots$$
$$\tilde{p}^{(N)}_{(k)} \leftarrow \max\ (\tilde{p}^{(N)}_{(k-1)}, \tilde{p}^{(N)}_{(k)})$$

Once monotonicity is enforced the estimates $\tilde{p}^{(N)}_{(j)}$ are reasonable approximations of the actual values $\tilde{p}_{(j)}$, provided N is sufficiently large. N $\geq$ 10,000 is recommended.

6.7 Multiple testing on 17 pairs and all groups

In this section we will try to link multiple testing to the model with 17 pairs of males and all groups. The question is: 'Can PROC MULTTEST of SAS handle this model?' In Section 5.3 we used PROC GLM of SAS to program this model with the 150 groups as the class variables, but incorrect conclusions can arise from this analysis. That is way we need to consider another method, namely PROC MULTTEST. Because PROC MULTTEST does not accept more than one class variable, we need to find another way to analyse the model. The 150 groups need to be reduced to one column. First of all, the 50 largest groups are selected. These 50 groups and the 9120 genes form a 9120x50-matrix. Next, all dependent rows are labeled with the same number and these labels are put in a columnmatrix. This column is the new class variable. If one number appears less than 5 times in the column, then these rows are removed from the column. The rows represent the 9120 genes.

After trying to analyse this model with PROC MULTTEST, we have to conclude that we did not find a method to program this model, because PROC MULTTEST can only compare two groups with each other. PROC MULTTEST can only have a contrast between two things.

Chapter 7

Conclusion

In this report we have tried to discover whether differences between normal glucose tolerance (NGT) and type 2 diabetes mellitus (DM2) are present. The genes are divided in 150 groups and each gene may occur in more than 1 group. DNA microarrays were used to detect gene expression changes in each gene. We also paired the data to become a model.

The first model was the most simple model, namely 1 pair and 1 group. From this model, we can conclude that some are significant and some are not. We have not found a straight line. The next model is more complicated and contains 1 pair of male and all groups. For all the pairs of males is this model significant, but we do not know whether the effect depends on the individu or on the difference between the 2 diseases. That is way we need the most complicated model, namely the model with n pairs of males and all groups. After analyzing different models with PROC GLM of SAS (n=2, n=5, n=17) it can be seen that there is not much connection between these models. We also did not find a straight line. We probably need another method to analyse this model.

Incorrect conclusions can arise from extensive data and that is something we want to avoid. Multiple testing is such a method that can avoid the incorrect conclusions. Multiple testing is based on resampling methods. PROC MULTTEST of SAS can handle these extensive data that need multiple testing. But we did not find a method to program the model with 17 pairs of males and all groups with PROC MULTTEST.

Bibliography

[1] *http://www.bio.davidson.edu/courses/genomics/chips/chip.html*

[2] P.H. Westfall and S.S.Young, *Resampling-based multiple testing, Examples and methods for p-value adjustment*, 1993, John Wiley & Sons, Inc., New York.

[3] Nature Genetics, *PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes*, July 2003, volume 34, number 3, p. 267-273.