

# Basic Statistical references for the Design and Analysis of Gene-Chips Experiments

Chiara Sabatti, etc

November 15, 2000

## Contents

<b>1</b>	<b>Data and data organization</b>	<b>2</b>
1.1	Data matrix . . . . .	2
1.2	Variables or observations? . . . . .	3
1.3	Dimension . . . . .	4
1.4	References . . . . .	4
<b>2</b>	<b>Design</b>	<b>4</b>
2.1	Reference . . . . .	6
<b>3</b>	<b>Measurements</b>	<b>6</b>
3.1	References . . . . .	6
<b>4</b>	<b>Curse of Dimensionality?</b>	<b>6</b>
4.1	Model Selection . . . . .	7
4.2	A New Asymptotics . . . . .	8
4.3	Dimensionality Reduction . . . . .	8
4.4	References . . . . .	8
<b>5</b>	<b>Graphical Representation</b>	<b>8</b>
5.1	References . . . . .	9
<b>6</b>	<b>Shrinking, regularization and Bayesian statistics</b>	<b>9</b>
6.1	References . . . . .	9
<b>7</b>	<b>Exploratory and confirmatory data analysis: the Bootstrap</b>	<b>10</b>
7.1	References . . . . .	10
<b>8</b>	<b>Some statistical techniques and their use in microarrays</b>	<b>10</b>
8.1	ANOVA . . . . .	11
8.2	Discriminant Analysis / Classification . . . . .	11
8.2.1	Reference . . . . .	11
8.3	Multidimensional Scaling . . . . .	12

8.3.1	Reference . . . . .	12
8.4	Clustering . . . . .	12
8.4.1	References . . . . .	13
<b>9</b>	<b>Other study design with array technology</b>	<b>13</b>
<b>10</b>	<b>Functional Genome Project?</b>	<b>13</b>
10.0.2	References . . . . .	13

## 1 Data and data organization

An accurate description of the nature of a microarray experiment can be found at We are here concerned with the formal description of the data derived by an array experiment that is more useful for the purpose of statistical reasoning. As a general rule, we will consider cDNA microarrays, but the content of what follows can be easily adapted to the Affimetrix array. By describing the gene-chip data with traditional statistical language, we will be able to identify a series of specific issues that we will discuss in following sections.

### 1.1 Data matrix

In a first approximation, the data from gene-chips experiments can be summarized in a matrix where each row corresponds to a specific spot on all the arrays and each column corresponds to a different “variety” that involves hybridization of a new array with the cDNA coming from a different cell line. Here are some possible sources of the cell lines whose cDNA is hybridized with the arrays: cell lines of different tumors, of different tissues, or cell lines of the same tissue in different individuals, identical cell lines under different shocks, or in different moments of the cell cycle. The data itself consist in a measurement of the (relative) expression level registered for a spot in correspondence to a particular hybridization. Formally, then, we have the matrix

$$\mathbf{X} = \begin{matrix} & X_{11} & X_{12} & \cdots & X_{1p} \\ X = & X_{21} & X_{22} & \ddots & X_{2p} \\ & \vdots & \vdots & \ddots & \vdots \\ & X_{n1} & X_{n2} & \cdots & X_{np} \end{matrix}$$

where the element  $X_{ij}$  is the measured expression for spot  $i$  in hybridization  $j$ . In the literature, there is a tendency to identify each row with a gene and each column with a variable corresponding to an experimental condition (ex. tumor type or cell cycle phase). In the following we will also use this same convention, initially, however, we wanted to emphasize the correspondence between rows and spots and columns and experiments. This is to take into account of the possible repetitions: a same gene can be spotted in multiple locations of the same array (so that we have more rows than the total number of gene considered) and

the same cDNA can be hybridized to different arrays (so that we have more columns than the total number of conditions considered). Indeed, a microarray experiment includes a number of other variables that we are not considering for the moment and that we will discuss in the sections Design and Measurements: taking into account or not these other variables is appropriate or not according to the nature of the question addressed.

For the purpose of general introduction, unless differently specified we will assume that there is one row per gene and one column per different experimental condition.

## 1.2 Variables or observations?

In most statistical frameworks there are two separate notions of variables and observations (columns and rows of the data matrix). In the data matrix described above, such distinction is not immediate. In fact, gene-arrays are an example of what Art Owen calls “transposable data”. If the goal of the analysis is to formulate a classification rule for tumor types on the base of the expression levels of different genes, the variables are the genes and the different experiments are different observations. However, if the question is to identify groups of genes that have similar regulatory mechanisms, the variables are the different experiments and the genes are the observations. Indeed, the gene-array context helps under-scoring the fact that the labels of “variable” or “observation” are more dependent on the kind of interrogation of the data that one intend to carry on rather than on an essential characteristics of the object. Furthermore, it should be noticed that in many cases “covariates” are available either for the rows or the columns of the data matrix or both. For example, for each gene, we may have sequence information of the regulatory region upstream the coding one, information that we may want to relate to the expression pattern of the gene (see ). On the other hand, for each cell line, we may have information with regard to the particular tumor type it belongs to, the survival time of the affected individual to whom it belonged, the effect of a particular therapy (see ). This underscore the fact that a priori there is not a particular “direction” in which the analysis should take place.

If we take the approach of the “transposable data”, one should somehow work both on rows and columns to gain information in both directions. This is particularly true for exploratory analysis whose goal is to find a convenient graphical representation of the data. Examples are Hartigan, double clustering, gene shaving, plaid, paf. We will discuss some of these in the sections graphical representation and clustering. In general, it is a statistics prejudice (quite well backed up, by the way) that you should formulate a question before trying to answer it looking at data. If this is the case, in most situations it will be clear what should be used as variables and classical statistical approaches will help with the analysis. If you then change the question and want to use the same data set, there may be the need of some thinking about how much you can use your data with out “fishing” for answers, but this will be discussed in the section tests. Here are the example of some questions that one may want to

answer by using the above data-matrix and that call for a specific use of rows and columns and that we discuss in the following:

1. Can I diagnose tumor type based on gene expression levels?
2. Can I predict survival time using gene expression levels?
3. Can I identify distances between different tumors based on the expression levels of various genes?
4. Can I identify genes that are coregulated using the values of their expression in different moments of the cell cycle and/or under different shocks, in presence or absence of sequence information in their promoter region?
5. Here is a new gene that has a give expression profile: who are his buddies?

### 1.3 Dimension

As in traditional notation, our data matrix  $X$  has  $n$  rows and  $p$  columns. Typically,  $n$  is in the order of thousands while  $p$  is in the order of tens. Great setting if we consider the columns as variables, but quite different from the classical statistical paradigm if we consider the rows as variables, or if we want to investigate both rows and columns. In general, one has to come to terms with the fact that gene-chip data are high-dimensional, the usual kind of asymptotic will not apply and the number of variables can be way bigger than the number of observation. This require some additional steps, but there are indeed some precise answer that statistics has to offer. We will discuss them in the section Curse of Dimensionality?

### 1.4 References

- Lazzeroni, L. and Owen, A.B. "Plaid Models for Gene Expression Data" <http://www-stat.stanford.edu/~owen/reports/>

## 2 Design

“To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.” (R.A. Fisher, Indian Statistical Congress, Sankhya, ca 1938)

Fisher started working as a statistician in 1919 at the Rothamsted Experimental Station. This research center in England was conducting a series of experiments to measure the effects of different fertilizers on various crops (and still is, see <http://www.iacr.bbsrc.ac.uk/res/treshome.html>). It is this context that much of the modern statistical theory was born and that he studied “The

design of experiments”. The context of gene arrays (where different DNA fragments are spotted in a plane and groups of them are “treated” with different hybridizations) is actually very similar to the one in which Fisher was working.

In the terminology of experimental design, each outcome (observation of the variable of interest, which is the expression level in our case) is measured in correspondence of a set of levels of different factors, that may influence it. For example, in the case of arrays, these are some factors present in the experiment:

- dye
- position of the measured spot on the array (peripheral vs central, or top vs bottom, or left vs right)
- pen which created the spot (in case there is more than one pen for array)
- hybridization experiment (every experiment can result in a different level of hybridization)
- cell line hybridized
- gene that is represented in the spot

Some of these factors are of interest and the experimenter specifically wants to study how the outcome depends on their variation. Some other correspond to inevitable variability in the experimental setting and the researcher mainly wants to control for their presence in interpreting the results. In the typical microarray experiment, cell line and gene are the effects one wants to estimate and the remaining are factors one wants to control for. However the situation might be different when a laboratory is setting up the protocols and the machinery to do microarray experiments: then dye, position, pen, array are the effects one wants to measure to identify the procedure that minimize their variability.

To try to control for both dye and hybridization experiment, researcher in cDNA array technology often hybridize the same array to the cDNA from two cell lines colored with a different dye. The ratio of the expression levels in the two dyes is measured: this is supposed to correct for the different amount of hybridization from one array to the other. To correct for the dye effect the values are then renormalized, so that their logs have mean zero.

This is a particular way of addressing the problem. There are, however, others that statistical design of experiment can suggest and that may be more effective. In particular, notice that in the setting described above, dye and cell line effect are completely confounded. Is it really necessary for this to be the case?

Already in the literature, different methods are explored: for example it is often recommended for two arrays to be hybridized to the same couple of cell lines, switching the dye across experiments. This is only one example of the type of suggestions that could come from design of experiments; an other easy example can involve multiple spotting of the same gene on an array.

In general, the following are few basic principles of the design of experiments. **Replication** is the repetition of the basic experiment and is necessary

to estimate the error and obtain a precise estimate of the effects of interest. **Randomization** allows to control for effects of extraneous factor that are not modeled. **Blocking** identifies subsets of the data that are more homogeneous.

## 2.1 Reference

- Kerr and Churchill(2000), Experimental design for gene expression microarrays<http://www.jax.org/research/churchill/pubs/index.html>

## 3 Measurements

How one measures the outcome of the experiment depends mainly on the available technology. However, there is often room for choices and it is important to keep an eye on the way in which data will be analyzed to optimize the results. In general, the way in which measurements are taken determines the nature of the error: because statistical analysis is all about recovering the true signal from the error, this is obviously very important.

Measuring the outcome of a microarray experiment is a complex matter; one step of the process involves transforming a two-dimensional picture in an array of numbers. The image segmentation technique used here, for example, can bias the results in different directions, so that it is important to discuss it with the statistician.

If the measurements are taken as ratio, it is often convenient to take the Log of these numbers, and actually to truncate their values. If a dye-renormalization is done to the data, this should be discussed with the statistician also.

### 3.1 References

- ScanAlyze<http://rana.lbl.gov/>
- Always Logs<http://www.stat.Berkeley.EDU/users/terry/zarray/Html/log.html>
- Terry Speed Group<http://www.stat.Berkeley.EDU/users/terry/zarray/Html/>
- Wing Wong Group<http://biosun1.harvard.edu/complab/>

## 4 Curse of Dimensionality?

We now consider the situation when data has been collected and is in the form described in section 1. One of the interesting features of microarray experiments is the fact that they gather information on a large number of genes (six, ten thousand). If we are considering, in the statistical lingo, genes as variables, this means that our observations are in a 6000-dimensional space and  $p \gg n$ . The expression “curse of dimensionality” is due to Bellman and in statistics it relates to the fact that the convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow. In terms

of microarrays, this means that, a priori, we need an “enormous” amount of observations (hybridizations to different cell lines) to obtain a “good” estimate of a function of the genes (that identifies, for example, which genes have altered expression patterns in a specific tumor type). A pretty dim scenario.

Fortunately, however, there are “blessings” associated with dimensionality (see Donoho). One of them has to do with what mathematicians call “concentration of measure”. To try to get a commercial slogan out of it, we could say that in many cases, there are really “few things that matter” and that the function will be constant on most of the space. This opens up the possibility of doing statistics in a meaningful and novel way.

## 4.1 Model Selection

Suppose we are interested in classifying our  $n$  cell lines in two groups (cancer vs non cancer) based on the expression levels of  $p$  genes. If  $p$  is bigger than  $n$  we can certainly find a rule based on  $n$  genes that classifies our data perfectly. The problem is, however, that when the expression levels from a new cell line will be observed, our rule will do extremely poorly in predicting the cancer status of the cell line. This is because our rule was constructed “ad hoc” for our data set. To make sure that the classification rule we produce is meaningful for cell lines whose expression pattern has not been observed yet, we need to select a classification rule that is based on few genes. In this way, we can hope to identify relations between genes and tumor types that are real and not merely present by accident in the particular data set in hand. “For this reason, statisticians have, for a long time, considered model selection by searching among subsets of possible explanatory variables [genes], trying to find just a few variables among the many which adequately explain the dependent variable [tumor type]. The history of this approach goes back to the early 1960’s when computers began to be used for data analysis and automatic variable selection become a distinct possibility.” (from Donoho)

There is an important warning, however, implicit in any variable selection procedure, and particularly serious when the number of variables is really large: searching long enough and among numerous enough variables will find a pattern, even if all the variables are noise. How do we protect ourselves from this phenomena?

Typically, variable selection amounts to the search of a set of variables using which we can construct a model that minimize an error criterion. To take into account that the more variables are included in the model, the higher the variability of the prediction, a penalized form of the criterion is minimized:

$$\min \text{ERR}(\text{model}) + c(\text{Model Complexity}),$$

where  $c$  is a constant and the model complexity is roughly the number of variables in the model.

In situations where the number of variables to search among is really big ( $p \gg n$ , as in microarrays), proposals have appear in the literature that suggest  $c = \text{cost} \log(p)$ , to take into account of the search effect.

## 4.2 A New Asymptotics

An other way of doing statistics in high dimension is to change the way of doing asymptotics. It is indeed realistic to consider situations where  $d$  goes to infinity together with  $n$ . A successful example of this is the work of Johnstone on principal components (that could be applied to microarray data) and, in a strictly microarray context, of Mark Van Der Laan.

## 4.3 Dimensionality Reduction

A simpler, but sometimes very effective, way of dealing with high-dimensional data is to reduce the number of dimensions, by eliminating some coordinates that seem irrelevant. In the case of microarray data this can often be done effectively and simply, by eliminating from consideration all those genes whose expression value doesn't vary across hybridization experiments. There are a variety of threshold rules that can be employed.

The statistic literature contains a whole set of techniques to identify the “relevant” coordinates of a dataset (see principal components analysis, independent component, SIR, etc. ). Some of these may be applicable to microarrays, even though certainly not in a blind fashion.

## 4.4 References

- Donoho, ”High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality”<http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html>
- M.J. van der Laan, J. Bryan (1999), Gene Expression Analysis with the Parametric Bootstrap. <http://www.stat.Berkeley.EDU/users/laan/papers/geneexpr11.ps>

# 5 Graphical Representation

In the previous section we considered some problems associated with the estimation of models from microarray data. Model fitting is a classical statistical approach and arguably the more instructive one, when appropriate and successful.

However, we may be looking at a process that we do not understand enough to attempt to model and, for the time being, we may mainly want to “look at the data”. Even this task is not very simple and requires some thinking. Tufte has written some beautiful books about visual display of quantitative information. The challenge is to produce visual appealing displays that uses our brain processing ability to convey the maximal amount of information. The fact that we “process” images has to be taken into account also to make sure that the display doesn't induce the observer to believe in patterns and effects that are not present in the data.

Microarray data analysts have taken ample advantage of color to display the numerical values of a matrix in an appealing way. Reordering of genes

and varieties has also been used in these plots. Particular attention should be paid to the method used for the reordering. Patches of uniform color are visually appealing, but they are effective in representing the data only when the associations between genes and varieties that they suggest are real.

There are various methods to achieve a reordering of the data, some of which will be mentioned in the section devoted to clustering, others are referred to in the references. There are entire journals devoted to graphical methods of statistical analysis, so we really aren't aiming to be exhaustive. A particularly useful tool for graphical analysis of data sets is Xgobi.

## 5.1 References

- Tufte, Edward R. *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1989.
- GAP <http://gap.stat.sinica.edu.tw/>
- Xgobi <http://www.public.iastate.edu/~arcview-xgobi/>

## 6 Shrinking, regularization and Bayesian statistics

One of the most important discoveries of statistics in the last 50 years has been the importance of “shrinking”, regularizing observations from noisy sources. This can be justified using a Bayesian perspective or from a decision-theoretic point of view. The current applications of this methodology in microarray analysis are connected to the necessity of converting the noisy expression measurements to an estimate of the true expression level.

### 6.1 References

- M.A. Newton, C.M. Kendzioriski, C.S. Richmond, F.R. Blattner, K.W. Tsui, 2000. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*. To appear. (Also Technical Report 139, Department of Biostatistics and Medical Informatics.) <http://www.stat.wisc.edu/newton/papers/publications/>
- S. Dudoit, Y.H. Yang, M. J. Callow and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. August 2000. <http://www.stat.Berkeley.EDU/users/terry/zarray/Html/papersindex.html>

## 7 Exploratory and confirmatory data analysis: the Bootstrap

Tuckey initially suggested the distinction between exploratory and confirmatory data analysis. The first consisting in “finding patterns” in data, the second one in attempting to validate them, making sure that the perceived association are “real” and not due to random chance. Because of the vast amount of search involved in the analysis of microarray data it is often quite difficult to use appropriate validation methods. The bootstrap, an “assumption free” technique can be particularly useful.

### 7.1 References

- M.J. van der Laan, J. Bryan (1999), Gene Expression Analysis with the Parametric Bootstrap. <http://www.stat.Berkeley.EDU/users/laan/papers/geneexpr11.ps>
- Kerr and Churchill(2000), Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. <http://www.jax.org/research/churchill/pubs/index>

## 8 Some statistical techniques and their use in microarrays

The goal of this section is to clarify which statistical technique is best used to address some question that have appeared in the microarray literature. We start proposing a non-exhaustive table that suggests a technique for different questions and then we give a brief description of the techniques with references.

**I have here observation from cell lines of different colon cancers: which genes are differentially expressed in the the various types?**  
Classification with genes as variables and cell lines as observations is a possibility.

**I have expression levels of genes in different experiments (heat shock etc) which genes behave similarly?** Clustering of the genes using experiments as variables.

**I have expression levels of genes in different moments of the cell cycle: which genes behave similarly?** One can try clustering again, but there is a clear dependency among the variables (times in the cell cycle) and one should be careful with the definition of distances.

Also, one may discretize the stages of the cell cycle and use classification to answer the question: when I see which genes on (off) I can be sure the cell is about to divide?

**I have observations on different tumors in different organs: few colon cells, few lungs cells etc. I believe that some of these tumors are**

**closer together (same biological mechanism) than others: how can I find an ordering of the tumors?** Multidimensional scaling could help. Or clustering. In both case, though, one should use the available information to keep close together tumors that are known to be the same.

## 8.1 ANOVA

ANOVA is the classical statistical technique to analyze the outcome of experiments. It can be very valuable in microarray study to assess what are the sources of variation in the experiment and what portion of the variance does each of them explain.

## 8.2 Discriminant Analysis / Classification

Suppose that one has a vector of labels  $y$  associated to each experiment: examples are tumor types in an experiments that analyzes different tumor cells; tumor status in an experiments that compares tumor cells with normal ones; phase of the cell cycle in an experiment comparing the same cell lines in different moments; type of shocks that the cell underwent. Then, an interesting question is: which genes are most useful to discriminate between the various values of  $y$ ? How do I formulate such a classification rule? This help us identifying genes that may be involved in the process that underlies the classification (for example, I could identify a set of genes that when turned on clearly signal that the cell is in meiosis and then learn that these genes are implicated in meiosis) and is helpful to predict the value of  $y$  for new cell lines of which we just know the expression pattern. This is particularly interesting, for example, in the case of tumor cells. The current criteria to classify a tumor type in a given organ may require observing the response of the cancer to different treatments. It would be useful, to be able to classify the tumor type of a new patient by looking at the gene expression pattern of his cancer cell. If this is done successfully, the patient could immediately receive the best treatment.

Discriminant analysis and /or classification are classical problems considered in statistics and a variety of tools are available. Fisher linear or quadratic discriminant, generalized linear models, trees, are examples of basic techniques. Boosting and Bagging are methods to improve the precision of given classifiers. Cross validation is an important tool to assess the performance of any given discriminant rule.

Because of the special nature of array experiments, one may additionally need to resort to some of the techniques we discussed in the section curses of dimensionality. What we intend to stress here is the usefulness of this type of analysis and the opportunity to conduct it when the data are in the form described.

### 8.2.1 Reference

- . Dudoit, J. Fridlyand, and T. P. Speed. Comparison of Discrimination

### 8.3 Multidimensional Scaling

It is often the case that experimenters would like to derive from microarray data an ordering of some variables of interest. It may be that the hybridizations originating the datasets are with cells from different tumors and it is believed that some of these cancer types are closer to each other than others. It is of interest to gather information about such similarity from the expression patterns exhibited by genes from these different cancer cells. There are many ways of answering this question, but if the goal is a reordering of the cell lines in a way that represents the distances among them, multidimensional scaling is one possibility that is worth exploring. If there is previous knowledge of some tumor types, this information should be taken into account in the analysis.

It is possible to obtain a similar reordering from clusters and trees, but these methods will only make sure that observations that are adjacent in the proposed order are close, it will not be possible to attach any meaning to the position of distant observations.

#### 8.3.1 Reference

- 

### 8.4 Clustering

Clustering is the statistical technique that doesn't need advertising in microarray context as it is the single most used instrument of analysis.

It is appropriate to cluster data when the observations are believed to be non-homogeneous (coming from different populations) and there is no precise notion of how they could be separated in homogeneous groups and it is desired to find such groups (that is, if there are labels identifying different classes, there is no need to cluster per se). There are a variety of cluster methods based on different notions of distances between the observations and on different algorithms. In case of microarrays particular attention should be devoted to the definition of distance. For example, suppose that we want to cluster genes on the base of the variation of their expression values during the cell cycle. Then, one needs to decide if two genes that have opposite behavior (one is on while the other is off and vice versa) should be in the same cluster (as they probably have a similar regulatory mechanism) or in different clusters—and decide how to define distance accordingly. Also, what about a gene that behaves exactly like an other, but with a lag?

Inference with clusters is not easy: it is not easy to determine how many clusters are there in the data and how likely is that the assignments of each observation to a given cluster is correct. One may use bootstrap techniques to answer to these questions.

Microarrays are special in that, as we described in the introduction, the statistical notions of “observations” and “variables” are interchangeable. With reference to the data matrix introduced in section 1, one may want to cluster both with respect to row and columns. To do this, one may independent cluster both (as done in Eisen), or cluster jointly. The latter is more complicated: the following section contains references to a method of Hartigan to do so and to two methods explicitly developed for microarray context (see Lazzeroni and Hastie).

#### 8.4.1 References

- Lazzeroni, L. and Owen, A.B. ”Plaid Models for Gene Expression Data”<http://www-stat.stanford.edu/~owen/reports/>
- Trevor Hastie, Robert Tibshirani, Michael Eisen, Pat Brown, Doug Ross, Uwe Scherf, John Weinstein, Ash Alizadeh, Louis Staudt, David Botstein ”Gene Shaving: a New Class of Clustering Methods for Expression Arrays”.<http://www-stat.stanford.edu/~hastie/Papers/>
- Hartigan (1972), “Direct clustering of a data matrix”, JASA

## 9 Other study design with array technology

We are particularly interested in using array technology to help localize the gene responsible for some disease. If you are also interested in questions of this type, please contact us or see the publication section.

## 10 Functional Genome Project?

A special attention and a particular design should be considered when one wants to collect expression data not to answer a specific scientific question, but with the idea of constructing a library of expression values under a standard set of circumstances, that subsequent researchers may consult. One of the most important issues in the construction of such library is the identification of a standard set of relevant conditions under which measure expression levels and the definition of standard experimental procedures. Once such standardized expression levels are constructed, one could use statistically based scoring systems (analogous to BLAST) to identify the proximal genes to any novel entry.

#### 10.0.2 References

BLAST<http://www.ncbi.nlm.nih.gov/BLAST/>