10

Microarray Analysis in Drug Discovery and Biomarker Identification

Yushi Liu[1] and Joseph S. Verducci[2]
[1]*Lovelace Respiratory Research Institute*
[2]*The Ohio State University*
USA

1. Introduction

1.1 Motivation of microarray development

Microarrays play important roles in Medicinal Chemistry and Drug Discovery. In the Pre-microarray era, scientists used to study one gene at a time. This approach is costly and time consuming. Quite often, many genes that interact with each other would be ignored. Therefore, the discovery of candidate drug targets is challenging, requiring the rapid development of techniques to identify the difference genomic profiling in disease and normal conditions, which will facilitate the understanding of the disease mechanism and the development of potential drugs for disease treatment.

1.2 Examples of microarray in biomarker identification

Microarray has been successfully applied to the comparison of genomic profiling for human tissues. One advantages of microarray is that it can find some potential drug targets which have been ignored previously. The example for this is the study by Heller *et al* in rheumatoid tissues. (Heller et al, 1997) They found around 100 genes known to be involved in inflammation. (Heller et al, 1997) However, additional genes such as interleukin-6 and matrix metallo-elastase are also found to be overexpressed remarkably, which is not anticipated *a priori*, since matrix metallo-elastase is thought to be distributed only within alveolar macrophages and placental cells.(Debouck and Goodfellow, 1999). Beside human, microarray has also been successfully applied to model organisms such as mouse. (Debouck and Goodfellow, 1999) Animal models play important roles in discovering therapeutic targets and potential drug development. Although the genome for the animals does not agree completely with the human genome, they are more easily to be manipulated. By careful design of the experiments, the treatment effect can be seen more clearly with less noisy background. Moreover, genes can be either knocked down or overexpressed to study the influence on phenotype. People used to use techniques such as differential display PCR to discover genes that are differentially expressed in the animal models and achieved some success. (Wang et al, 1995) However, this technique is much slower compared with microarray.

Quite often, drugs can bind to specific targets within cells and potentially influence different pathways.(Windle & Guiseppi-Elie, 2003) The genes that are differentially expressed

between drug-treated and untreated conditions are typically known as biomarkers. Such biomarkers not only help to identify patients at risk, but they also may lead to breakthroughs in understanding the mechanism for different diseases.(Ko et al., 2005)An example is the review by Chen that summarized the recent microarray research in biomarker identification in atherosclerosis and in-stent stenosis.(Chen et al., 2011)

1.3 Diagnosis using microarray

Microarray has greatly advanced the biology field and the biomarkers identified can be further used to form classifiers for prediction in clinical studies. For example, Gulob *et al* used the gene expression pattern from microarrays to classify acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL) without other information. (Gulob et al, 1999) Therefore, this field draws the attention not only from biologists but also statisticians and bioinformaticians. Through their collaborative efforts, there are many successful instances. We will introduce some of them in section two later.

The rapid development of this technique also resulted in several FDA approved test. AmpliChip CYP450 test is a clinical test to find specific genetic variation of two cytochrome P450 genes CYP2D6 and CYP2C19 genes including deletion and duplications. (de Leon, 2006) These two enzymes account for the variability of drug metabolism for each patient and offers enriched information for the doctors during prescription of psychiatric drugs. (de Leon, 2006) CYP2D6 can be divided into four categories: Poor Metabolizer, Intermediate Metabolizer, Extensive Metabolizer, and Ultrarapid Metabolizer. (de Leon, 2006) Similarly, for CYP2C19, only two categories are found: Poor Metabolizer and Extensive Metabolizer. (de Leon, 2006) The assay works as follows: First, the gene is amplified by PCR and then the amplified product will be fragmented and labelled. Subsequently, these fragments will be hybridized to the microarray chip and the chip is scanned for further analysis. (de Leon, 2006) For further information of this FDA approved test, please see the website at http://molecular .roche.com/assays/Pages/AmpliChipCYP450Test.aspx.

Another FDA approved diagnosis test is MammaPrint to assess the risk of breast tumor and this will help to decide the effectiveness of chemotherapy on the patients. (van't Veer et al, 2002) The assay uses the fresh tissue to study the Amsterdam 70-gene breast cancer gene signature by microarray analysis. (van't Veer et al, 2002) Readers interested in this test can also obtain more information about the MINDACT trial (Microarray In Node negative and 1-3 positive lymph node Disease may Avoid Chemo Therapy) in the paper by Cardoso et al.(Cardoso et al, 2008)

In general, identification of biomarkers by microarray greatly speeds the progress of research by enabling the simultaneous monitoring of the expression of thousands of genes. However, there are many potential pitfalls in analyzing the output from these arrays. (Verducci, et al., 2006) Due to importance of proper analysis, we will give a brief introduction to the statistical methodology underlying proper analysis.

2. Mechanisms and processing of microarrays

Medicinal chemistry has increasingly employed microarrays to identify both key target genes and gene networks that can regulate the effectiveness of drugs. The basic scheme is

illustrated in Figure 1.1. Two cell strains (one is drug treated and one is non-treated) are harvested and the whole RNA from each strain is then extracted. This is followed by reverse transcription of RNA and the resulted cDNA is labelled with either of the two fluorescence dyes (Cys-3 or Cys-5). Then the mixed cDNA from both samples is hybridized to the probesets on the microarray chip. The probesets are the small oligonucleotides that have the complementary sequence of the cDNA attached to the array at each spot. After intensive washing, the intensity from the fluorescence of the dye labelled on cDNA at each spot is measured and recorded. These data will be used for further analysis.

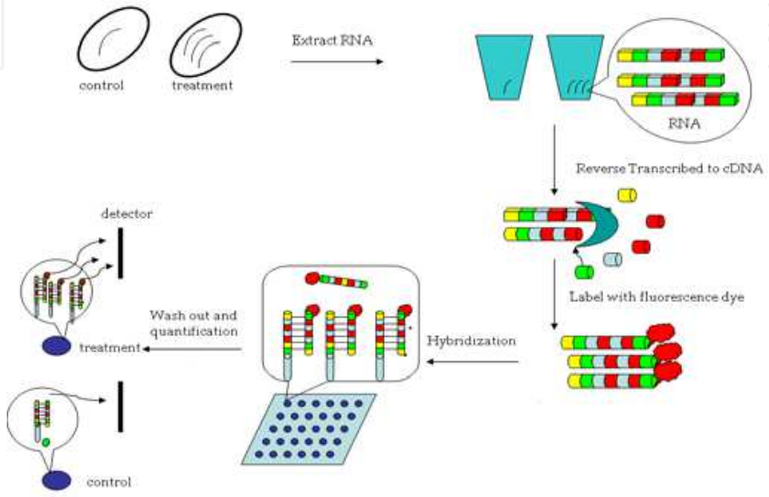


Fig. 1. Illustration of the microarray process. RNA is extracted from treated (treatment) and untreated (control) cell lines, followed by reverse transcription. The reversely transcribed cDNA is then labelled with the fluorescence dye and hybridized to the probes containing complementary fragments. After unbounded cDNA is washed away, the binding at each probe is then quantified based on the fluorescence intensity of the bounded cDNA.

The above method is referred to as a two channel array because a mixture of cDNA from two treatments is measured directly. In contrast, a one channel array will only have cDNA from one sample to be hybridized to the probesets. In this case the fluorescence intensity from each sample is measured separately. For either type of array, processing and analyzing the data present both statistical and biological challenges. Fortunately, many such approaches have been integrated in the freely distrubuted statistical software R (http://www.r-project.org/) and the software Bioconductor (http://www. bioconductor. org/). Typically the data processing step includes four steps: image analysis, quality assessment, pre-processing and statistical inference.(Tibshirani et al., 2005)

2.1 Image analysis and quality assessment

In the image analysis step, each spot is quantified and then converted to intensity afterwards. The method of quantification depends on the brand of the arrays. Quality assessment is usually performed at two levels: array level and probe level. On the array level, fingerprint smudges or washed out corners, are generally recognized. Other problems such as defects of the array, errors in RNA extraction also belong in this category. One common criterion is that if the percentage of spots without any signal is higher than 30%,

the expression array will fail in the quality control step.(Tibshirani et al., 2005) Poor quality at probe level will include errors like faulty printing, uneven distribution, contamination, or poor level of signal to noise ratio. In addition, several parameters can be used to determine the quality of the array: uniformity, which is minimal variation in pixel intensity within a spot; and the brightness, which is the foreground to background ratio.(Tibshirani et al., 2005) Normally, researchers will ignore the spots of poor quality in subsequent analysis.

2.2 Pre-processing

Two types of errors usually happen at this stage: (1) systematic error, which influences all measurements within one microarray chip with similar effect -- this error may be corrected by estimation; and (2) random error that cannot be explained or corrected, which is typically known as noise. Such errors are totally stochastic and have different influence on different probes.(Tibshirani et al., 2005) Typically, the pre-processing stage contains three steps: background correction, normalization and summarization. For the widely used Affymetrix chips, many Bioconductor routines are available in R for pre-processing. These require creation of an *AffyBatch* object based on raw Affymetrix data (in a .cel file). The first step is the background adjustment. In this step, one tends to subtract the control intensity from the treatment, to 'denoise' the intensity. However, direct subtraction of uncertain quantities can increase the level of noise and possibly result in negative intensity values for certain spots. Various methods to circumvent these problems are available as *method* parameters in the *bg.correct* function in *R*:

a. *RMA method*: which is based on the assumption that the observed signal is a mixture of Gaussian background noise (N) with mean μ variance σ^2 and exponential signal (S) with mean *a.* Thus the fluorescence intensity *O* we observe is the addition of the signal and noise. Assuming the above, $E(S|O)$, which is the conditional expectation of the signal based on the observed intensity will be used as the background corrected values. However, the disadvantages for this are: only the PM(perfect match) values are used and MM(mismatch) values, which contains useful information for background noise are discarded.(Tibshirani et al., 2005) and the results may not be robust if there are gross deviations from the model assumptions. These assumptions may be checked visually via different plotting methods.

b. *MAS 5.0 method*: due to the above disadvantages, *RMA* may not produce optimal result. Therefore, *MAS 5.0* is sometimes used instead. Here, the whole array can be partitioned as k rectangular grids.(Tibshirani et al, 2005) The probeset, with lowest intensity for the grid, is used as the noise value to calculate the background corrected intensity within a particular rectangle. The intensities of these probes are further adjusted according to the weighted average of the background intensity of all grids according to the following formula:

$$W_{k(x,y)}=1/(d^2_{k(x,y)}+S_0) \quad (1)$$

In the above formula, the weight is determined by the Euclidean distance from (x,y) to the centroid of the space k and the smoothing coefficient represented by d^2_k and S_0, respectively.(Tibshirani et al, 2005) Irizarry *et al*. (2003) compare *RMA* and *MAS 5.0* in detail.

c. *Ideal mismatch*: Neither of the above methods uses mismatch information. Although direct subtraction of the mismatch intensity from the perfect match intensity creates the

problems of added noise and negative intensity, *ideal mismatch* adjusts the observed mismatch so that it will never be higher than the PM intensities. The detailed formula of this is available in (Tibshirani et al., 2005) for further reading.

The next step is normalization of scores across different microarrays so that they can be compared fairly with each other. A variety of methods available in the *normalize* function of Bioconductor will be introduced*:*

a. *Scaling normalization*: All the arrays are normalized using the same selected baseline. This is almost identical to fitting linear regression without the intercept.(Tibshirani et al., 2005)

b. *Non-linear transformations*: Although linear regression is simple and easy to implement, in microarray study, the relationship may be more complicated and thus non-linear methods are developed including include cross-validated splines and loess smoothers.(Yang et al., 2001) The "invariantset" method developed by Li and Wong is very robust and is thus recommended.(Li & Wong, 2001) First an "invariantset" is identified. This gene set is composed of non-differentially expressed genes (sometimes called "household function genes) across the arrays and the expression values (or the rankings) of these genes can be used to construct the baseline for normalization (Li & Wong, 2001) However one challenge for this method is the identification of the "invariantset", which may not be available *a priori*.

c. *Quantile normalization*: The purpose of this method is to adjust the empirical distribution on all arrays so that they could be the same. The algorithm in *R* works as follows: First the columns of expression data matrix X are properly ranked (dimension $p \times n$, *p:* number of the genes on the array; n: the number of the arrays). Suppose v is the p-dimensional vector of row means of the sorted data matrix and V is the $p \times n$ matrix whose columns are all equal to v, sort each column of V by the inverse permutation. The obtained matrix is then quantile normalized.(Tibshirani et al., 2005) The basis for this method is that the total energy that cells exert for gene expression remains fairly constant, although the choice of which genes get expressed may differ widely.

d. *Cyclic loess normalization*: An MA plot is used for this normalization procedure: M (which stands for "multiple") is the difference of two log intensities, while A is the average of the two log intensities. Subsequently, a loess curve is fitted for the MA plot and M is predicted by this curve.(Tibshirani et al., 2005) Each intensity value is adjusted based on the difference between the real and predicted M value. The process is iterated until all the arrays or probesets converge. However, the drawback of this method is that it is computationally expensive and time consuming.(Tibshirani et al., 2005) In *R*, the above two steps can be integrated. It has advantages like using all the information across arrays for normalization, and is thus, theoretically, more reliable. The "*vsn*" package in *R* is a representative and can perform the above two steps seamlessly.(Tibshirani et al., 2005)

The final step of preprocessing step is the summation, which is trying to integrate intensity values from multiple probes of a particular gene and obtain its expression value. The R routines *expresso* and *threestep* offer great flexibility in deciding how much to weight each probe. (Tibshirani et al., 2005) Summation completes the pre-processing step, and we are now ready to begin proper analysis.

2.3 Statistical methods for biomarker identification in microarray analysis

2.3.1 Introduction to basic microarray analysis

Since microarray can be viewed as high-dimensional dataset with fewer replicates. Traditional variable selection procedures like stepwise selection cannot identify the biomarkers effectively; modifications or new procedures are developed to accommodate this.

a. *Shrinkage Methods*: One particular drawback from stepwise selection of genes that distinguish treatment from control is its poor performance when the variables (gene expression levels) are highly correlated. However, this is exactly what happens on for microarray data since many genes on the array typically are involved in the same pathway. This inspired the development of the shrinkage methods, which can be viewed as constrained optimization. One advantage of shrinkage methods is they are more continuous than the subset selection and do not exhibit high variance. (Hastie et al., 2001) Theoretically, shrinkage methods do not minimize the residual sum of squares; instead, they impose a penalty on the residual sum of squares. Nowadays, different forms of penalty are proposed and some of the most commonly used ones are introduced here.

Ridge Regression: Ridge regression introduces a penalty on the size of the coefficients, thus leading to the shrinkage of the regression coefficients.(Hastie et al., 2001) Mathematically, $\tilde{\beta}^{ridge}$ solves the following:

$$\tilde{\beta}^{ridge} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\} \tag{2}$$

Here, λ (the regularization parameter) is greater than or equal to zero and controls the amount of shrinkage towards zero. When the regularization parameter is zero, this approach is converted back to ordinary least square (OLS) estimation. The penalized formulation (2) has an equivalent formulation in terms of constrained optimization, which may be achieved using convex programming methods:

$$\begin{aligned}\tilde{\beta}^{ridge} &= \arg\min_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ &\sum_{j=1}^{p}\beta_j^2 \le s\end{aligned} \tag{3}$$

Ridge estimation is suitable for situations when many correlated variables are present in the model.(Hastie et al., 2001) In these cases, the least squares estimator may be poorly determined, since the large positive coefficients may cancel out the negative coefficients on the correlated variables.(Hastie et al., 2001) Ridge regression can effectively prevent this from happening. As the unique solution to (2), the ridge estimator has explicit form:

$$\tilde{\beta}^{ridge} = (X^tX + \lambda I)^{-1}X^tY \tag{4}$$

where I is the identity matrix. Hence, adding a positive constant to the diagonal of X^tX allows a singular matrix to be inverted, effectively reducing the dependencies among the estimated coefficients. This was the original motivation. (Hoerl & Kennard, 1970)

Lasso Regression: Lasso regression is similar to ridge regression, simply replacing the L_2 norm ridge penalty in (2) by an L_1 norm penalty. The lasso form of (3) becomes

$$\tilde{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \sum_{j=1}^{p} |\beta_j| \le t \tag{5}$$

Changing from an L_2 to an L_1 penalty results in a estimator that is nonlinear in y.(Hastie et al., 2001) In this case, nonlinear programming is needed to get the lasso solution iteratively.(Hastie et al., 2001) The algorithms of the least angle regression (LARS) can be simply modified to implement the lasso.(Bradley et al., 2004) A special feature of the lasso is the "sparseness" of the solution: some of the coefficients become exactly zero the constraint *t* becomes sufficiently small. If t is large enough, then no shrinkage is performed. For the case of orthonormal columns of X, the lasso has a simple form in terms of the OLS coefficients and the penalty γ:

$$sign(\hat{\beta}_j)(|\hat{\beta}_j| - \gamma)_+ \tag{6}$$

Bridge Regression: Both ridge and lasso regressions are very popular. They can be generalized to bridge regression to achieve the some of the benefits of both.(Ildiko & Friedman, 1993) In bridge regression, people try to find β that satisfies the following, where $1 \le q \le 2$:

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \sum_{j=1}^{p} |\beta_j|^q \le s \tag{7}$$

SCAD Regression: Despite the good theoretical properties lasso/ridge regression, these penalties do not achieve the following desired properties simultaneously: unbiasedness (the estimator is close to the true parameter when the true parameter is large), sparseness (irrelevant predictors are automatically removed), and continuity (estimator is continuous, preventing the instability of hard thresh holding estimators).(Fan & Li, 2001) The new penalty is defined for a parameter θ as

$$p'_{\lambda}(\theta) = \lambda \left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} \tag{8}$$

This penalty, known as the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan & Li, 2001) helps to improve the properties of L_1 and hard thresh holding penalties such

$p_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$. The SCAD penalty can also be viewed as a quadratic spline function, with knots at λ and aλ.(Fan & Li, 2001) Also, it does not extremely penalize large values of θ and the solution is continuous.(Fan & Li, 2001) Fan gave the solution to SCAD in the context of wavelets.(Fan, 1997) The SCAD estimator has the following form:

$$\begin{aligned} &\hat{\theta} = sign(z)(|z| - \lambda)_+ \ \ when \ \ |z| \le 2\lambda \\ &\hat{\theta} = \{(a-1)z - sign(z)a\lambda\} / (a-2) \ \ when \ \ 2\lambda < |z| \le a\lambda \\ &\hat{\theta} = z \ \ when \ \ |z| > a\lambda \end{aligned} \tag{9}$$

One feature of this penalty is the oracle property: when some variables are not present in the true model, these corresponding coefficients tend to be estimated as zero when the sample size gets large.(Fan & Li, 2001) Asymptotically, it provides as good estimation of the coefficients as if the underlying model were known beforehand.(Fan & Li, 2001) Selection and estimation of the variable coefficients is automatic and simultaneous.(Fan & Li, 2001) Among the many instances where SCAD has been applied to microarrays for biomarker selection, the study by Wang *et al*. (2007) successfully discovered 71 potential transcriptional factors (TF) in the cell cycle of yeast. These included 19 out of 21 known and experimentally verified TFs related to the cell cycle. Additional TFs showed periodic transcriptional effects and thus were biologically important and worth further study.(Wang et al., 2007)

b. *Methods Involving Derived Inputs: Principal Components Regression:* When a large number of correlated inputs (e.g. potential biomarkers) are available, instead of keeping all these inputs in a regression model, it may be beneficial to consider just a few linear combinations of them. A logically justifiable choice for the coefficients used in the linear combination is the normalized vector *a* that gives the largest sample variance of all possible normalized linear combinations of the input variables. (Hastie et al., 2001) This is called the first principal component. The first *p* principal components are found sequentially, with each successive component maximizing input variation subject to being orthogonal to previous ones. When all the principal components are used, the method becomes the usual least square estimation. However, when fewer principal components are used, this method is similar to ridge regression.(Hastie et al., 2001) As an example, Tan *et al*. (2005) used total principal component regression to classify the tumors into different categories.

Partial Least Squares: In contrast to principal component regression, which uses linear combination only of the input variables, partial least square(PLS) also allows for some information from the response variable y in the linear combinations.(Hastie et al., 2001) The algorithm is as follows: Assume y is centered and each x_j is properly standardized. PLS first regresses y on each x_j to obtain the corresponding coefficient $\hat{\varphi}_{1j}$. Subsequently, we can define the first partial least square direction as $z_1 = \sum \varphi_{1j} x_j$. (Hastie et al., 2001) Then *Y* is regressed on z_1 to obtain the corresponding coefficients which is followed by orthogonalizing $x_1 \ldots x_p$ in reference to z_1.(Hastie et al., 2001) This process is repeated until the desired number of directions is reached. As with principal component regression, using all *p* directions results in the usual least square estimation.(Hastie et al., 2001) Differently from principal component regression, PLS seeks input that is in the direction of high variation and high correlation with *y*.(Hastie et al., 2001) In addition, when the inputs are

orthogonal to each other, the PLS will coincide with the least square estimates after the 1st step, setting the coefficients to be zero for the subsequent steps.

Because of its good properties with small sample sizes and many predictors, PLS has been applied to high-dimensional genomic data.(Boulesteix & Strimmer, 2006; Aaroe, *et al*., 2010) Moreover, PLS regression can also be used to impute missing data. For example, Bras and Menezes (2006) impute missing values using PLS regression with all the genes as predictors. Huang *et al.* (2004) use a penalized version of PLS, which removes genes with poor power of prediction, in order to predict LVAD (left mechanical ventricular assist device) support time. In this case, the shrinkage parameter and the number of latent components are obtained using cross-validation. After proper shrinkage, some genes have coefficient zero, thus removing them from the model.(Huang et al., 2004) This reduces the complexity of the model, and serves as an example for combining both shrinkage and PLS.

c. *Bayes Variable Selection Methods:* Bayesian approaches use knowledge across genes for further inference.(Nott et al., 2007) For example, George and Foster (2000) adopted a binomial prior for the number of the differentially expressed genes (i.e. biomarkers) and a normal prior for their corresponding coefficients, assuming known and constant variance parameter. With informed choice of the hyperparameters, the authors ranked the genes according to the posterior probability that the gene belonged to the differentially expressed gene set or not. Interestingly, the gene ranking agreed with the ranking obtained by other criteria such as AIC (Akaike, 1973) or BIC (Schwartz, 1979; Nott *et al*., 2007) When the variance parameters were unknown, different priors were assumed for effects of the genes and variance of the genes. Lonnstedt and Speed used a normal prior for effects of the differentially expressed genes and an inverted gamma prior for the corresponding variance.(Lonnstedt & Speed., 2002) Thus, after choosing the hyperparameters properly, people derived an explicit expression for the log odds of differentially expressed genes, which was known as B-statistic. (Lonnstedt & Speed, 2002) In contrast, Nott *et al* considered a double tailed exponential prior for effects of the differentially expressed genes and an inverted gamma for the corresponding variance.(Nott et al., 2007) The motivation was that double tailed exponential was heavier on the tails and was related to lasso.(Nott et al., 2007) The proposed linear model was as follows:

$$M_{gj}=\mu_g+\varepsilon_{gj} \tag{10}$$

M_{gj} is the expression value of gene g for array j. μ_g is the gene specific mean expression value. ε_{gj} is $N(0, \sigma_g^2)$ where σ_g^2 is the gene specific variance. All the σ_g^2 are independent. Except the situation where we have infinite sample size, we can only conclude gene g is differentially expressed when $|\mu_g|>k$ and gene g is not differentially expressed when $|\mu_g|\leq k$.(Nott et al., 2007) As a predefined cutoff value, k depends on the purpose of the experiment and other conditions. Correspondingly, B-statistic is then defined as follows to explore whether gene g belongs to the set of differentially expressed gene or not:

$$B(k)=\log\frac{\Pr\left(|\mu_g|>k|M\right)}{\Pr\left(|\mu_g|\leq k|M\right)} \tag{11}$$

The above represents the log odds ratio of the posterior probability given the data M. To implement this hierarchical Bayes procedure, calculation of the posterior probability is

necessary. A modified version of MCMC (Markov Chain Monte Carlo) algorithm was proposed, and details of the computation of the above *B*-statistic when the prior distributions are given is discussed by Nott et al. (2007). When *k* is chosen to be 0 and a proper prior allows μ_g to be exactly 0, an explicit form of B(0) can be developed accordingly.(Lonnstedt & Speed., 2002) People usually use this statistic to rank genes instead of making inferences.(Nott et al., 2007; Lonnstedt & Speed., 2002)

d. Hypotheses Testing as a Variable Selection Methods: The t-statistic has already been widely used for hypothesis testing for a long time. Therefore, people try to apply this to microarray study to select appropriately biomarkers on microarray. However, the traditional t-statistic will not work in this situation due to the following reasons: First, hundreds or thousands hypotheses are being tested simultaneously; therefore, the multiple comparison issue exists. However, Bonferroni correction is too conservative and sometimes no gene can pass this vigorous criterion. Thus, suitable adjustment methods need to be developed to further control the overall error. Second, during the microarray analysis, large outliers are frequently observed, and they tend to drive the t-statistic to be large. Similarly, due to large number of tested genes and small number of replicates, the estimated variance for each gene is usually small, which tends to drive the t-statistic to be large.

To meet the first challenge, a new method known as false discovery rates (FDRs) has been proposed. (Benjamini & Hochberg, 1995; Tusher et al., 2001) False discovery rate is defined as the expected proportion of type I error using the available decision rule.(Benjamini & Hochberg, 1995) This method, readily available in *R*, is especially useful for microarray study, since it is easy to compute and not as overly conservative as is Bonferroni adjustment.

The second challenge requires a robust modification to the current version of t-statistic. One of them is known as *ad hoc* modification, which defines the modification by the data. Efron's 90% rule is in this category.(Efron et al., 2000) The modification is to add a constant term to the denominator to prevent the variance in the t-statistic from being too small.(Efron et al., 2001) The constant a_0 is defined as the 90th percentile of all the standard errors of the genes.(Efron et al., 2000) Thus the ordinary t-statistic has the following format:

$$S_g = M_{g.} / (s_g + a_0) \tag{12}$$

$M_{g.}$ denotes the average expression value for gene *g*, and s_g denotes the corresponding standard deviation. Another example belonging to *ad hoc* modification is the SAM (Significance Analysis of Microarrays).(Tusher et al., 2001) For each gene, the SAM method assigns a score relative based on the changes in expression relative to the standard deviation for the repeated experiments. For genes with scores higher than certain thresholds, a permutation distribution is used to estimate the FDR.(Tusher et al., 2001) This method may be viewed as an empirical Bayes procedure, simply adding a constant each set of genes levels when estimating individual variances. This avoids difficulties when variances are computed from a small number of observations for each gene.(Tusher et al., 2001) This method showed great improvement in gene identification both FDR-wise and fold-wise in terms of the human cell response to the ionizing radiation.(Tusher et al., 2001) Despite of its robustness to individual outliers, the use of this *ad hoc* modification is still limited, and it is challenging to derive and study its theoretical properties.

Accordingly, a penalized likelihood version of t-statistic has been proposed and implemented. For example, Wu (2005) proposed another modified t-statistics, taking advantage of both SAM and lasso methods. The method works as follows: assume the linear regression situation,

$$x_{ij}=\beta_0+ y_j+\varepsilon_j \tag{13}$$

where x_{ij} represents the expression of gene i on array j; y_j is the indicator, whether the j^{th} sample belongs to the control or treatment group. A *t*-statistic or *F*-statistic can be developed.(Wu, 2005) The test statistic involves ordinary between/within group sum of squares, both of which can be penalized like in lasso regression.(Wu, 2005) The test statistic in this scenario can be derived as:

$$t_i^* = sign\left(\bar{x}_{i1}-\bar{x}_{i2}\right)\frac{\left(\left|\bar{x}_{i1}-\bar{x}_{i2}\right|-\lambda\right)_+}{\sqrt{n/(n_1n_2)s_i^2+\lambda^2/(n-2)}} \tag{14}$$

$$F_i^* = \frac{\left(\left|\bar{x}_{i1}-\bar{x}_{i2}\right|^2-\lambda^2\right)_+}{n/(n_1n_2)s_i^2+\lambda^2/(n-2)} \tag{15}$$

Tusher's SAM statistic is as follows:

$$d_i = \frac{\bar{x}_{i1}-\bar{x}_{i2}}{s_0+s_1\sqrt{n/(n_1n_2)}} \tag{16}$$

Comparing the formulas (14—16), we can see that the penalized *F* or *t* statistic can be viewed as a special version of SAM, since the term $\lambda^2/$ (n-2) coincides with the constant s_0 in SAM statistic, which helps to stabilize the variance.(Wu, 2005) Furthermore, Wu showed that FDR can be calculated by permutation and then a cutoff can be put on the test statistic.(Wu, 2005) What makes the penalized SAM statistic superior to the ordinary SAM statistic is that penalized SAM statistic is derived rigorously from the situation of linear model and thus easier to develop its theoretical properties.(Wu, 2005) Through applications, this statistic also shown good performance.(Wu, 2005)

Another modified *t*-statistic is refined for a statistical model assuming both multiplicative and additive errors.(Ideker et al., 2000) The parameters within the model are subsequently estimated using maximum likelihood method with all the observations.(Ideker et al., 2000) Subsequently, a traditional maximum likelihood ratio test for each individual gene is carried out to identify the significance of the intensities.(Ideker et al., 2000) In some examples, this method can be shown superior to the simple fold approach.(Ideker et al., 2000) However, this method is naïve and has potential limitation as follows: first, the author does not use any multiple comparison adjustment techniques when performing multiple tests on thousands of genes simultaneously; this may be corrected by introducing the traditional FDR. Second, the author used chi-square as the distribution of -2*ln(likelihood ratio), which may not hold for small sample size.(Ideker et al., 2000) A more suitable distribution needs to be derived accordingly.

One more example of a "modified" t-statistic is derived from a Bayesian approach, which has become popular in statistics. For example, the B-statistic, the log odd ratio of the

posterior probability can be viewed as a Bayesian version of t-statistic.(Lonnstedt & Speed, 2002) Another example of "moderated" *t*-statistic is proposed by Smyth.(Smyh, 2004) It assumes a scale inverse chi square prior for the variances of the genes.(Smyh, 2004) Additionally, the parameters can be estimated using Bayesian method and the 'moderated' t statistic is obtained by substituting the corresponding variance with their estimate.(Smyh, 2004) Cui *et al* propose another modified t-statistic using similar approach.(Cui et al., 2005) First, they performed a simulation from a chi-square distribution, whose degrees of freedom depends on the sample size to estimate the variance. Then they derive a bias-corrected Stein estimator on the log scale.(Cui et al., 2005) Thus, this estimator is more robust since the shrinkage in the variance makes the estimator of variance more robust.

As we can see, the main drawback of "moderated t" is that it depends on a particular type of distribution. When the distribution assumption is not satisfied, these estimators will be inefficient and often lead to false inferences. This inspires the birth of the distribution free 'shrinkage t' statistic.(Opgen-Rhein & Strimer, 2007) The main idea behind this is shrinking the empirical variance of each gene towards the common median of all the variance.(Opgen-Rhein & Strimer, 2007) For each group, the ordinary variance is replaced by the corresponding shrinkage variance in the test statistic:

$$t_k^* = \frac{\bar{x}_{k1} - \bar{x}_{k2}}{\sqrt{\frac{v_{k1}^*}{n_1} + \frac{v_{k2}^*}{n_2}}} \tag{17}$$

For the above formula, n_1 and n_2 are sample size for group 1 and 2, respectively. v^* stands for the corresponding shrinkage variance estimator. Here, each empirical variance is shrinking towards the median, which is shown to be more efficient or robust than shrinking towards the mean or zero.(Opgen-Rhein & Strimer, 2007) We can also view the "shrinkage t" as a combination of a standard t statistic and the fold change statistic.(Opgen-Rhein & Strimer, 2007) Another feature is that the "shrinkage t" belongs to the James-Stein estimator, not relying on any explicit prior distribution assumption and its theoretical property will be easily derived. Furthermore, this method is computationally efficient and the corresponding gene ranking is consistent with other tests.(Opgen-Rhein & Strimer, 2007)

Shrunken centroid method and SCOOP: From a different point of view, Tibshirani developed the shrunken centroid method for biomarker identification.(Davies & Bromage, 2002; Tibshirani et al., 2005) For each gene within each group (i.e. treatment group or control group), the overall mean and the group means are calculated. The group means are shrunk toward the overall mean iteratively for each gene.(Davies & Bromage, 2002; Tibshirani et al., 2005) The shrunken values are used to rank the genes and the cutoff is chosen by cross-validation. (Davies & Bromage, 2002; Tibshirani et al., 2005) The shrunken values can be also used to form a classifier and the authors used this method to classify the cancer conditions. (Davies & Bromage, 2002; Tibshirani et al., 2005) Despite the successes this method has achieved, it has one potential drawback: information about correlation among genes is distorted or lost during successive shrinkage, and, therefore, the identified genes may appear falsely to be independent of each other. Based on this method, Liu *et al.* (2009) developed an improved version of shrinkage centroid method: SCOOP (**S**hrunken **C**entroid **O**rthogonal **O**rdering **P**rojection) to extend to the cases with correlation variables. Instead of

shrinking along the natural axes (Tibshirani et al., 2005), which ignores the potential linkage between variables, SCOOP rotates the axis and shrinks the group means in the direction preserving the least correlation information of variables. The algorithm of SCOOP is as follows: With the input of group information and the gene expression values from microarrays, two matrices are further identified: Between Epoch Covariance Matrix, containing all the variation between different groups, and Within Epoch Covariance Matrix, containing the variation originating from replication. Then, the eigenvalues and eigenvectors for both Between Epoch Covariance Matrix and Within Epoch Covariance Matrix are calculated by spectral decomposition. Since we have small number of samples and large number of variables (i.e. the genes) for microarray studies, both Between Epoch Covariance Matrix and Within Epoch Covariance Matrix are going to be highly singular. The union of the eigenvectors of the Between Epoch Covariance Matrix and Within Epoch Covariance Matrix with nonzero eigenvalues will form the basis functions of the new space (known as the Augmented Discriminant Space). For each gene, the group mean expression is shrunk towards the overall mean along the direction orthogonal to the Augmented Discriminant Space until the group means coincide, at which point that gene is eliminated from consideration. The amount of shrinkage needed for each gene is considered as its measure of importance. The above algorithm is carried out individually for each gene, producing a ranking of genes according to the importance measure. SCOOP has been successfully applied to identify biomarkers responsible for female rainbow trout reproductive cycle.(Liu, 2009, 2011)

2.3.2 Introduction to basic microarray time course analysis

Due to the decreasing cost of microarrays, their use in time course analysis has become ever more popular. The corresponding analysis is more challenging statistically than the two sample microarray situation. The time course may be longitudinal (where the mRNA samples for different time points are taken from the same individual), or cross-sectional (where the mRNA samples are extracted from different individual).(Tai & Speed, 2005) As a result, gene expression tends to be correlated for the longitudinal study or a design used for the cross-sectional study using a common reference. In addition, usually only 5-10 time points are available. Therefore, the traditional time series model cannot deal with such small series. This will require the development of new methods for analysis.

Typically, researchers are interested in identifying the genes whose expressions change over time. In the one-sample problem, some genes' patterns vary according to a common pattern. In the two-sample problem, we need to identify genes whose temporal changes differ under two or more biological conditions.(Tai & Speed, 2006)

One popular method typically used is a regression model. As an example, maSigPro belongs to this category and is available in *R*. (www.bioconductor.org) To find significantly different genes for two or more biological conditions, maSigPro first builds a global regression model with different experiment conditions acting as dummy variables.(Conesa et al., 2006) Then the significance of the estimated parameters in the model was tested to assess the significant differences between gene time course profiles.(Conesa et al., 2006)

Another method for microarray time course analysis is via ANOVA and the F-statistic. The classical ANOVA and mix-effect ANOVA models are used for cross-sectional and

longitudinal study, respectively.(Diggle et al., 2002; Neter et al., 1996; Tai & Speed, 2005) For the one-sample problem, time is treated as one factor. Thereafter, the corresponding F-statistic is calculated.(Tai & Speed, 2005) Moreover, this method can be extended to the situation with multiple experiment conditions. Time, experiment condition and potentially their interaction are included for this model. An example of the classical ANOVA in time course study is available in (Wang & Kim, 2003). Also the multiple comparison adjustment for testing error is discussed for this method.(Ge & Speed, 2003)

Later, a robust version of ANOVA approach was proposed by Park *et al*, since it does not require the normality assumption.(Park et al., 2003) Similar to a two-way ANOVA model which includes time, biological conditions and their interaction as factors, genes that are concluded insignificant in this model will be reanalyzed in the same ANOVA model without the interaction term. (Park et al., 2003) Genes that are concluded significant in both models are chosen.(Park et al., 2003) Another modified ANOVA method is the ANOVA-SCA (analysis of variance-simultaneous component analysis), which takes into consideration about the correlation structure of the measured variables.(Nueda et al., 2007) Basically, principle component analysis is used to the estimated parameters of each source of variation in the ANOVA model.(Nueda et al, 2007) One advantage of this method is that it utilizes information from the experiment design and takes into consideration about correlation among the each source of variability associated with experimental factors. To identify the differentially expressed genes, the authors proposed another criterion for ANOVA-SCA: the mixture of leverage and SPE (square prediction error).(Nueda et al., 2007) Leverage quantifies how much a particular gene contributes to the multivariate ANOVA-SCA model, while SPE evaluates the goodness of fit of the model to a particular gene.(Nueda et al., 2007) The potential test statistic is and its p-value are obtained with reference to a weighted χ^2 distribution.(Box, 1954) Nonetheless, the drawback of this method is that it does not use the actual time scale and direct smoothing cannot be applied. Besides, this method cannot be used when the time course points are irregular.

In summary, the ANOVA and the corresponding modified versions offer substantial advantages: they can separate variation due to each different factor, therefore, removing the non-random effects and reducing the potential noise within the data.(Box, 1954) However, there are two innate limitations: First, it assumes independent among different time points ignoring the potential correlation; second, the small number of replicates leads to unstable estimation of gene-specific variance, leading to big value of within time F-statistics even for genes with just small amount of changes. This leads to high false positive rates.(Tai & Speed, 2005) In addition, some differentially expressed genes may have outliers which tend to cause low F-statistic, resulting in false negative rates.(Tai & Speed, 2005) Thus, the idea of moderation is introduced.

To reduce the false positive rate or false negative rate, the gene-specific variance is shrinking towards a common value estimated from the whole gene set, known as moderation.(Tai & Speed, 2005) One example about the application of moderation to microarray time course is performed by Tai and Speed.(Tai & Speed, 2006) They derived the *MB*- and $\tilde{T}^2$-statistic for one-sample or two-sample problem in the scenario of longitudinal microarray time course study, taking into consideration about the correlation across times. In detail, *MB*-statistic is the log 10 of the posterior odds whether the null or alternative hypothesis is true. When the number of replicates is equal for all genes, the *MB*-statistic under the null hypothesis is

supposed to have the expected profile equal to 0 in one-sample case or equal expected profiles in two-sample scenario.(Tai & Speed, 2006) Then the form of *MB*-statistic becomes a monotonic increasing function in $\tilde{T}^2$. $\tilde{T}^2$-statistic is $\tilde{t}'\tilde{t}$ where $\tilde{t}$ is the moderated multivariate t-statistic in the form of $\tilde{t} = n^{1/2}\tilde{S}^{-1/2}\bar{X}$.(Tai& Speed, 2006). $\tilde{S} = \frac{(n-1)S + \nu\Lambda}{n-1+\nu}$

where S represents the gene-specific variance-covariance matrices, $\bar{X}$ is the gene-specific average time course vector. n represents the number of replicates. The other two parameters v and Λ can be estimated from all the genes. Both of the two statistic are derived when we assume independent and identical inverse Wishart priors to the gene-specific covariance matrices.(Tai & Speed, 2006) The advantage of this method comes from the incorporation of the information about the correlation structure, moderation and replication.(Tai & Speed, 2006) In addition, this statistic outperforms the ordinary F-statistic, due to moderation in empirical Bayes framework.(Tai & Speed, 2006) This procedure is shown to be very effective in false positive or negative rate reduction.(Tai & Speed, 2005) Thus, this procedure is incorporated in the Bioconductor "timecourse" package in *R*. (http:/ /www. bioconductor.org/packages/2.3/bioc/vignettes/timecourse/inst/doc/timecourse.pdf). The drawback of this method is modeling each gene independently, ignoring the latent genes pathway network and making no use of the actual time scale.

Another method that is used similar idea to estimate the unstable variance robustly and incorporate correlation in the study is based on the likelihood-based approach. Guo *et al.* develop a test based on the Wald statistic for one-sample longitudinal data.(Guo et al., 2003) This method adds a positive number to each diagonal element in the denominator matrix to incorporate the idea of moderation and stabilize the estimation of the variance.

$$w(i) = [L\hat{\beta}(i)]^T [L\hat{V}_s(i)L^T + \lambda_\omega I_{r\times r}]^{-1} [L\hat{\beta}(i)] \tag{18}$$

In the above formula, L represents a matrix with dimension r X p, $\hat{\beta}$ represents the p X 1 regression parameters estimation and $\hat{V}s$ is the corresponding estimated variance-covariance matrix. λ_ω is an estimated positive scalar to prevent inverting a highly singular matrix.(Guo et al., 2003) However, the limitation of this method is that it is only suitable for one-sample problem and using the asymptotic theory will not be suitable for small number of replicates.

Despite of the popularity of the above method, they all ignore one important fact in time course study: They do not make use of the time points dynamically. This is the reason to introduce B-splines or wavelets to model the gene temporal expression profiles. Natural B-splines are piecewise cubic polynomials, which are smoothly connected at knots. It can describe the complicated gene expression patterns over time, since the linear combination of a series of basis functions can mimic any temporal profiles for genes. Each basis function can be thought as the potential expression pattern locally (i.e. the basis function will be zero outside certain time range). Comparing with methods that do not utilize time scale directly, B-splines have many advantages: reduce the noise, assuming only smoothing changes occur with time; use the actual time taken for the samples, easy to adapt for schedules with irregular time points; As an example, Bar-Joseph *et al.* present an algorithm to characterize the expression pattern of each gene by a continuous curve fitted by B-splines.(Bar-Joseph et

al., 2003) They constrain the spline coefficients of genes within the same class so that their expression profiles can vary similarly. Thus, the gene expression pattern can be viewed dynamically. Comparing with previous methods, the reconstruction of the gene timecourse has 10-15% less error for those points that are not observed.(Bar-Joseph et al., 2003) Another approach proposed by Hong and Li is to solve the two-sample problem with B-splines adaption.(Hong & Li, 2006) In details, to identify biomarkers whose expression profiles are different under multiple biological conditions, linear combinations of basis functions are used to create smooth gene expression timecourse. The Markov chain Monte Carlo EM algorithm (MCEM) can be used to estimate the gene-specific parameters and hyperparameters from the hierarchical model. The genes are chosen using the empirical Bayes log posterior odds and the posterior probability based FDR.(Hong & Li, 2006) As a result, this method outperforms the traditional ANOVA model and is suitable for long time course data. Another example developed by Storey *et al.* and denoted as EDGE (**E**xtraction of **D**ifferential **G**ene **E**xpression) is also widely used for microarray timecourse study.(Storey et al., 2005) It estimates the coefficients of a B-spline function to fit the timecourse for each gene, and test whether all the coefficients are zero or not by an F-statistic. If all the coefficients are zero, the genes are not differentially expressed. Q-value based on false discovery rate(FDR) is calculated for each individual gene to offer a suitable cutoff value.(Storey et al., 2005) This method is an example to combine B-splines with the hypothesis testing, using FDR to control the error rate. Therefore, this method is superior to other methods. However, this method does not use the correlation information between variables (i.e. genes) and needs improvement. Thus SCOOP in combination with B-spline offers an alternative for biomarker identification for microarray timecourse study. (Liu, 2009, 2011)

When the situation of multiple biological condition in microarray timecourse study is encountered, Yuan *et al.* develop a hidden Markov model approach.(Yuan & Kendziorski, 2006) For this method, the authors consider all possibilities of equality and inequality for all the means among the different biological conditions and the expression pattern process is modelled as a Markov chain.(Yuan & Kendziorski, 2006) These biological conditions are referred as states. Thus, the observations are conditionally independent given the state of the chain. In summary, this method can monitor the expression pattern for each gene and the observations at different time points may be dependent on each other. The differentially expressed genes are then selected based on the posterior probabilities of states of interest.(Yuan & Kendziorski, 2006) Moreover, it is suggested that the associated posterior probability is useful to cluster genes.(Yuan & Kendziorski, 2006)

2.3.3 What is next?

Although the microarray technology has lead to big breakthroughs in biology, there is one innate drawback in this technique: since all the sequence information about genes incorporated into the probes needs to be known *a priori*, the microarray can only obtain fixed and partially information about gene variants within the cell. This limitation requires the development of new techniques to gain the information for all the gene alleles simultaneously. Therefore, the next generation sequencing technique gains popularity and may be consequently lead to more informative microarrays. The first generation sequencing

is accredited to Frederick Sanger in 70s. (Sanger et al, 1977) Before the development of Sanger's chain-terminator method, Maxam and Gilbert used toxic chemicals to modify the bases, inferring the sequence of DNA fragment.(Maxam & Glibert, 1977) Sanger's chain-terminator method gained popularity since the method involved less toxicity. The key of the chain-terminator method is the dideoxynucleotide triphosphates(ddNTPs) to terminate the DNA chain elongation. To sequence a particular DNA fragement, the DNA template, primer, DNA polymerase and deoxynucleotidephosphates(dNTPs) is split into four separate reactions with the addition of only one of the four radioactively or fluorescently labelled dideoxynucleotides (ddATP, ddGTP, ddCTP or ddTTP) in the four reactions.(Sanger et al., 1977) Therefore, during the elongation, ddNTPs are incorporated into some of the strands, leading to DNA fragments that have varying length. These fragments can be separated using gel electrophoresis and the relative position of the band on the gel be used to determine the base identity.(Sanger et al., 1977)

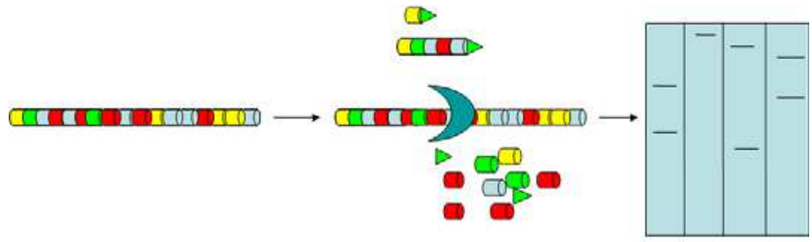


Fig. 2. The Sanger's chain-terminator method. For a fragment of DNA, the sample is split into four reactions containing dNTP, polymerase. Each reaction is supplemented with one type of ddNTP, serving as the chain terminator. In the above figure, we show only one reaction: the dNTP is depicted as tubes and ddCTP is depicted as triangles. The ddCTP terminates the reaction upon the addition of the ddCTP. The other three reactions form similar ladders and the sequences can be detected based on their relative position on the gel after gel electrophoresis.

In recent years, instead of using one fluorescence dye and four reactions, four different fluorescence dyes with unique emission wavelength will be used in a single reaction. Then the dye reader can automatically read the base identity after capillary electrophoresis. Readers interested in this technique can read the user's manual for ABI PRISM® 373 DNA Sequencer manual available at http:/ /www3.appliedbiosystems.com /cms/groups/ mcb_ support/documents/generaldocuments/cms_041831.pdf.

Therefore, the previous sequencing technique is laborious and time consuming. The current biological studies require more efficient ways to sequence. This is the motivation for next generation sequencing development.

The first step for the high-throughput sequencing is to prepare a template. In this step, genomic DNA is randomly split into small pieces to construct fragment template.(Metzker, 2010) When the genomic DNA is first circularized by ligation and then split into small fragments, this is known as mate-pair template, which has advantages over fragment template in alignment.(Metzker, 2010) However, due to the reason that single fluorescence event is hard to detect, the templates need to be amplified. Emulsion PCR by Roche and bridge PCR by illumina are introduced here. The sheared genomic DNA will be ligated with

adaptors containing the same fragment.(Metzker, 2010) This allows the amplication of the DNA fragment using common PCR primer. For emulsion PCR, the water droplet containing the bead-DNA complex, primer, polymerase, and dNTP will be used to perform PCR amplification. Numerous droplets are created by emulsifying the oil-aqueous mixture, allowing all the genomic DNA fragments to be amplified simultaneously.(Metzker, 2010) Another popular amplification method is bridge PCR. Bridge PCR has two steps: initial priming and extension of the template. The genomic DNA fragments with adaptors at both ends will be immobilized and bent over to form a bridge. Subsequently, the DNA molecules will be amplified to form clusters.(Metzker, 2010) Despite the great success, the amplification procedure is time consuming and complicated. Moreover, AT or GC-rich sequences may be biased during the amplification. (Metzker, 2010) Therefore, single-molecule templates technique which involves the immobilization of primer, template, or polymerase has become popular.(Metzker, 2010) The readers interested in this can obtain more details about this technique in Metzker 2010.

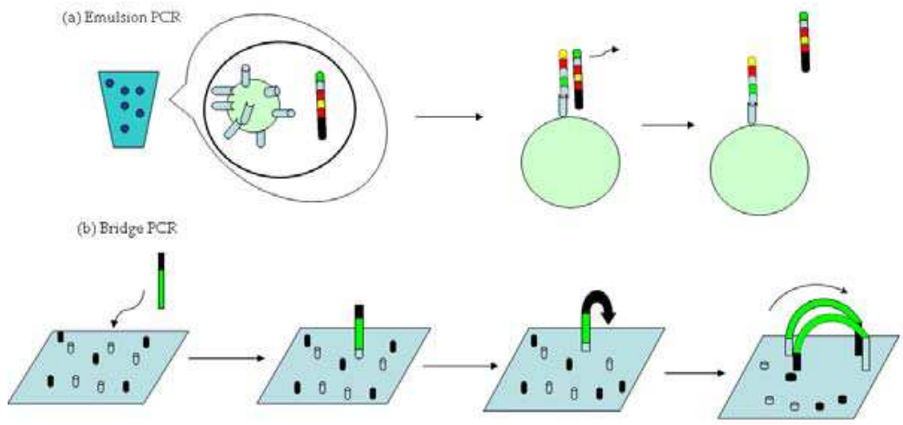


Fig. 3. The illustration of the emulsion PCR and bridge PCR. For emulsion PCR, the aqueous droplets can be created by emulsion in the oil water mixture. Then the template can be amplified with the primers within the bead. In the end, thousands of DNA fragments containing identical sequences to the template will be available within one bead for each aqueous droplet. For bridge PCR, the template is immobilized and bridge amplified to form a cluster.

Sequencing and imaging step follows the above amplification step. The four colour reversible termination method by illumina is introduced here. Right after the template clusters are obtained, the four nucleotides labelled with distinct fluorescence dye will be incorporated according to the template sequence and the elongation step halts upon the addition of fluorescence labelled nucleotide. Upon total internal reflection fluorescence imaging, TCEP (tris(2-carboxyethyl)phosphine) will be used to cleave the fluorescence dye and 3'-inhibitor to allow the next cycle of elongation. This process is iterated until the identities of all the nucleotides are known.(Metzker, 2010) The sequencing process for Roche/454 is called pyrosequencing, which uses a different mechanism for sequencing: following the emulsified PCR, the DNA-amplified beads are loaded into PTP (PicoTiterPlates) wells. Subsequently, this method allows the polymerase to add only one particular type nucleotide with the release of pyrophosphate. The pyrophosphate will then be converted with the emission of light by a

series of reactions.(Ronaghi et al., 1998) This is further recorded by a charge-coupled device camera. The order of the light emission will be used to induce the sequence. This mechanism is totally different from the reversible termination method and it does not require the use of modified dNTP to halt the elongation process.(Metzker, 2010)

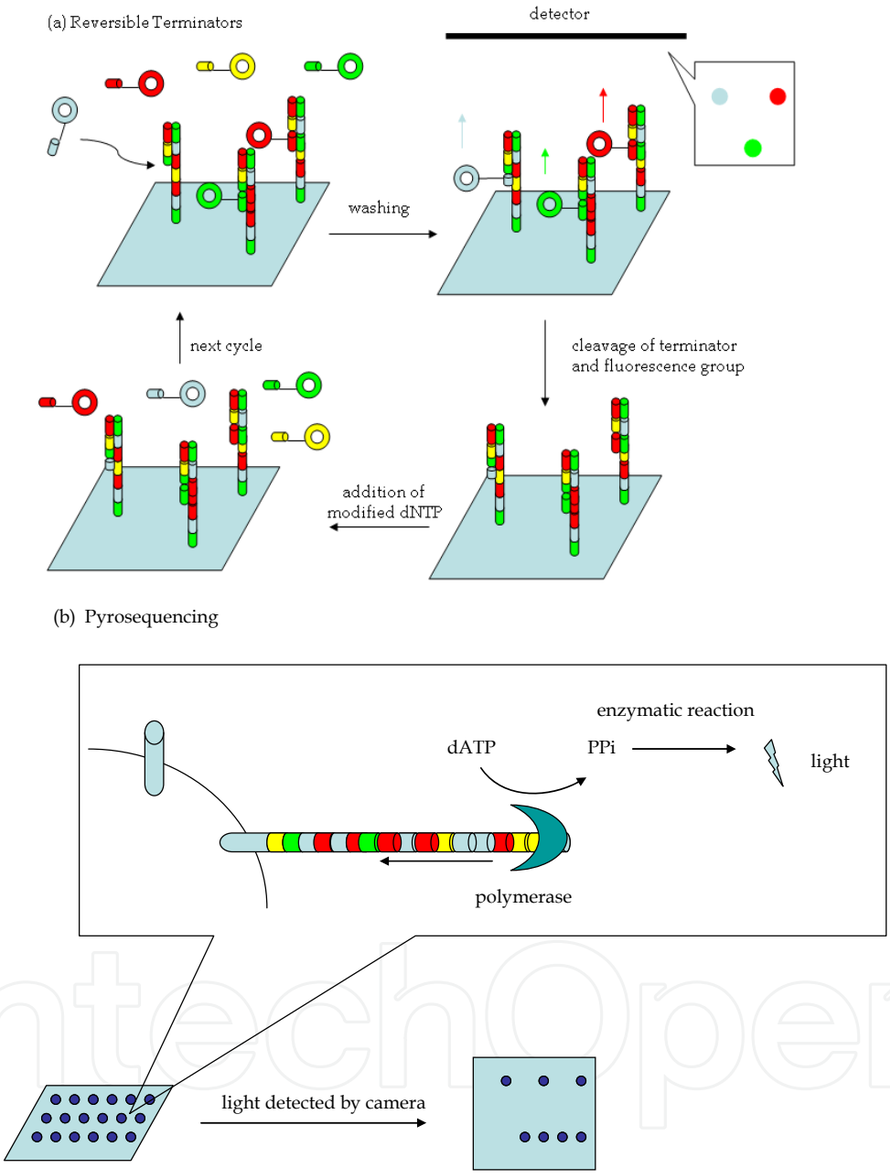


Fig. 4. The illustration of the reverse terminator sequencing and pyrosequencing. For reverse terminator sequencing, different dNTP is labelled with different fluorescence dye. Upon addition of each dNTP, the reaction halts and the fluorescence is the recorded. Then the terminator and fluorescence dye of dNTP is cleaved. Subsequently, the next dNTP is incorporated and the whole process is iterated. For pyrosequencing, one single dNTP flows through with the addition of the nucleotide in the corresponding position. The release of the pyrophosphate will undergo enzymatic reaction to produce light. Therefore, the camera will record which of the fragments has this dNTP at its current position.

The next generation sequencing reads will subsequently need to be aligned to the reference genome or assembled *de novo*.(Chaisson et al., 2009; Pop & Salzberg, 2008; Trapnell & Salzberg, 2009) There are several challenges for the genome assembly and alignment besides cost and effort: First, some reads may not be aligned to reference genome due to the structural variant (e.g. deletion or insertion).(Metzker, 2010) Second, some reads are difficult to align to the highly repetitive regions.(Metzker, 2010) *de novo* assembly will be complicated for large genome although some successes are reported.(Butler *et al*, 2008; Hernandez et al., 2008; Zerbino & Birney, 2008)

Although there are so many challenges, this field is still undergoing rapid development and will play a main role in the personal genome era and personalized medicine field. The gigantic information from the next generation sequencing studies will require the collaboration between biologists, bioinformaticians, and biostatisticians. What we envision are more and more big breakthroughs in the field of life science.

3. Conclusion

In summary, we presented a detailed overview of microarray studies. We introduced the mechanism, the associated statistical analysis, and the potential substitution for microarray-next generation sequencing. Several examples of microarray studies to identify biomarkers are also presented. We hope this chapter can serve as a guide for beginners in the field of biomarker identification and drug discovery.

4. Acknowledgment

The authors thank NCI NIH HHS for the support of Grant R01 CA095568/ and NSF for the support of Grant DMS-0540693. The authors also thank the editor and reviewers for their constructive comments.

5. References

Aarøe, J.;Lindahl, T.; Dumeaux, V.; Sæbø, S.; Tobin, D.; Hagen, N.; Skaane, P.; Lönneborg, A.; Sharma, P.; & Børresen-Dale, A-L. (2010). Gene Expression Profiling of Peripheral Blood Cells for Early Detection of Breast Cancer. *Breast Cancer Research,* Vol. 12, R7.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Second International Symposium on Information Theory*, pp. 267-281.

Bar-Joseph, Z.; Gerber, G.; & Gifford, D. K. (2003). Continuous representations of time-series gene expression data. *Journal of Computational Biology,* Vol. 10, pp. 341-356

Benjamini, Y.; & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. *Journal of The Royal Statistical Society,* Vol. Ser B 57, pp. 289-300.

Boulesteix, A.; & Strimmer, K. (2006). Partial Least Squares: a Versatile Tool for the Analysis of High-dimensional Genomic Data. *Briefings in Bioinformatics,* Vol.8, pp. 32-44.

Box, P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one way classification. *Ann. Math.Stat.,* Vol. 25, pp. 290-302.

Bras, L.P.; & Menezes J.C. (2006). Dealing with Gene Expression Missing Data. *IEE Syst Biol,* Vol. 153, pp. 105-119.

Butler, J.; MacCallum, I.; Kleber, M.; Shlyakhter, I.A.; Belmonte, M.K.; Lander,E.S.; Nusbaum, C.; & Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.,* Vol. 18, pp. 821-829.

Cardoso, F.; Van't Veer, L.; Rutgers, E.; Loi,S.; Mook, S.; & Piccart-Gebhart, M.J. (2008) Clinical application of the 70-gene profile: the MINDACT trial. *J. Clin. Oncol.*, Vol. 26, pp. 729-735.

Chaisson, M.J.; Brinza, D.; & Pevzner, P. A. (2009). De novo fragmentassembly with short mater-paired reads:does the read length matter? *Genome Res,* Vol. 19, pp. 336-346.

Chen, J.; Traci, T. G; Brigitta, C. B.; Li, J.; Spencer, B. K.; Nicolas, C.; Jiang, H.; & Hou, D. (2011). Microarray Applications in Occlusive Vascular Disease. *Cardiovascular & Hematological Agents in Medicinal Chemistry,* Vol.9, pp. 84-94.

Conesa, A.; Nueda, M.; Ferrer, A.; & Talon M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics,* Vol. 22, pp. 1096-1102.

Cui, X.; Hwang, J.T.G.; Qiu, J.; Blades, N.J.; & Churchill G.A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics,* Vol. 6, pp. 59-75.

Davies, B.; & Bromage, N.(2002). The effects of fluctuating seasonal and constant water temperatures on the photoperiodic advancement of reproduction in female rainbow trout, Oncorhynchus mykiss. *Aquaculture,* Vol. 205, pp. 183-200.

Debouck C.; & Goodfellow P.N. (1999). DNA microarrays in drug discovery and development. *Nature*, Vol 21, pp. 48-50.

de Leon J.(2006)AmpliChip CYP450 test: personalized medicine has arrived in psychiatry. *Expert Rev Mol Diagn,* Vol. 6 pp. 277-286.

Diggle,P.J.; Heagerty, P.; Liang, K.-Y.; & Zeger S.L. (2002). *Analysis of Longitudinal Data* (2nd ed.). New York: Oxford University Press.

Efron, B.; Tibishirani, R.; Goss. V.; & Chu, G. (2000). *Microarrays and Their Use in a Comparative Experiment*. *Technical Report*, Vol. 213, Stanford Univeristy, Available from
http://statistics.stanford.edu/~ckirby/techreports/GEN/2000/2000-37B.pdf

Efron, B.; Hastie, T.; Johnstone I.; & Tibshirani, R. (2004). Least Angle Regression. *Ann. Statist.,* Vol.32, pp. 407-499.

Efron, B.; Tibishirani, R.; Storey, J.; & Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association,* Vol. 96, pp. 1151-1160.

Fan J. (1997). Comments on 'Wavelet in Statistics: A Review' by A. Antoniadis. *Journal of the Italian Statistical Association,* Vol. 6, pp. 131-138.

Fan, J.; & Li. R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association,* Vol. 96, pp. 1348-1359.

Ge, Y.; Dudoit, S.; & Speed, T. (2003). Re-sampling based multiple testing for microarray data analysis. *Test,* Vol.12, pp. 1-77.

Gentleman, R.; Carey V.J.; Huber, W.; Irizarry, R.A.; & Dudoit, S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor:*(1st ed.) Springer, New York.

George, E.I.; & Foster, D.P. (2000). Calibration and Emprical Bayes Variable Selection. *Biometrica,* Vol.87, pp. 731-747.

Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; Bloomfield, C. D.;& Lander E. S.(1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, Vol. 286, pp. 531-537.

Guo, X.; Qi, H.; Verfaillie, C.M.; & Pan, W. (2003). Statistical significance analysis of longitudinal gene expression data. *Bioinformatics,* Vol. 19, pp. 1628-1635.

Hastie, T.; Tibshirani, R.; & Friedman, J. (2001). *The Elements of Statistical Learning* (1st ed.). Springer, New York.

Hernandez, D; Francois, P.; Farinelli, L.; Osteras, M.; & Schrenzel, J. (2008). *De novo* bacterial genome sequencing: millinos of very short reads assembled on a desktop computer. *Genome Res,* Vol. 18, pp. 802-809.

Heller, R.A.; Schena, M.; Chai, A.; Shalon, D.; Bedilion, T.; Gilmore, J.; Woolley, D. E.; & Davis, R. W. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci*., Vol. 94, pp.2150-2155.

Hoerl, A.E.; & Kennard, R. (1970). Ridge Regression:Biased Estimation for Nonorthogonal Problems. *Technometrics,* Vol. 12, pp. 55-67.

Hong, F.; & Li, H. (2006). Functional Hierarchical Models for Identifying Genes with Different Time-Course Expression Profiles. *Biometrics,* Vol. 62, 534-544.

Huang, X.; Pan, W.; Park, S.; Han, X.; Miller L. W.; & Hall, J. (2004). Modeling the Relationship between LVAD Support Time and Gene Expression Changes in Human Heart by Penalized Partial Least Squares. *Bioinformatics,* Vol. 20, pp. 888-894.

Ideker, T.; Thorsson, V.; Siegel, A.F.; & Hood L.E. (2000). Testing for Differentially-Expressed Genes by Maximum-Likelihood Analysis of Microarray Data. *Journal of Computational Biology,* Vol. 7, pp. 805-817.

Ildiko, E. F.; & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics,* Vol. 35, pp. 109-135.

Irizarry,R. A.;. Bolstad, B. M; Collin, F.; Cope, L. M.; Hobbs, B.; & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research,* Vol.31 pp. e15 doi:10.1093/nar/gng015

Ko, D.; Xu, W.; & Windle, B. (2005). Gene Function Classification Using NCI-60 Cell Line Gene Expression Profiles. *Computational Biology and Chemistry,* Vol. 29, pp. 412-419.

Li, C.; & Wong, W. H. (2001). Model-based analysis of oligonucleotides arrays: model validation, design issues and standard error application. *Genome Biology,* Vol. 2, pp. research 0032.

Liu, Y.; Verducci, J.; Nagler, J.; Schultz, I.; Hook, S.; Cracium, G.; Sundling K.; & Hayton, W. (2009). Time Course Analysis of Microarray Data for the Pathway of Reproductive Development in Female Rainbow Trout *Statistical Analysis and Datamining*, Vol. 2, pp. 192-208

Liu, Y. (2011). *Properties of the SCOOP method of selecting gene sets. OhioLink*, Ph.D. dissertation, Dept. of Statistics,The Ohio State University, Columbus.

Lonnstedt, I.; & Speed, T.P. (2002). Replicated Microarray Data. *Statist. Sinica,* Vol. 12, pp. 31-46.

Nott, D. J.; Yu, Z.; Chan, E.; Cotsapas, C.; Cowley, M.J.; Pulvers, J.; Williams, R.; & Little, P. (2007). Hierarchical Bayes variables selection and microarray experiments. *Journal of Multivariate Analysis,* Vol. 98, pp. 852-872.

Nueda, M.J.; Conesa, A.; Westerhuis, J.A.; Hoefsloot, H.C.; Smilde, A.K.; Talón, M.; & Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics,* Vol. 23, pp. 1792-1800.

Maxam, A.M.; & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci,* Vol. 74, pp. 560-564.

Metzker, M.L. (2010). Sequencing technologies--the next generation. *Nature Reviews,* Vol. 11, pp. 31-46.

Neter, J.; Kutner, M.H.; Wasserman, W.; & Nachtsheim, C. (1996). *Applied Linear Statistical Models* (4th ed.). Irwin: McGraw-Hill.

Opgen-Rhein, R.; & Strimer, K. (2007). Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. *Statist. Appl. Genet. Mol.Biol.,* Vol.6, pp. article 9.

Park, T.; Yi, S-G.; Lee, S.; Lee, S. Y.; Yoo, D-H.; Ahn, J-I.; & Lee Y-S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics,* Vol. 19, pp. 694-703.

Pop, M.; Salzberg S.L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet.,* Vol. 24, pp. 142-149.

Ronaghi, M.; Karamohamed, S.; Pettersson, B.; Uhlen, M.; & Nyren, P. (1998). Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.,* Vol. 242, pp. 84-89.

Sanger, F., Nicklen, S.; & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA,* Vol. 74, pp. 5463-5467.

Schwartz, G. (1979). Estimating the Dimension of a Model. *Ann. Statist.,* Vol. 6, pp. 461-464.

Smyh, G.K. (2004). Statistical Application in Genetics and Molecular Biology. *Statist. Appl. Genet. Mol.Biol.,* Vol. 3, pp. article 3.

Storey, J.D.; Xiao, W.; Leek, J.T.; Tompkins, R.G.; & Davis, R.W. (2005). Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci,* Vol. 102, pp. 12837-12842.

Tai, Y.C.; & Speed, T.P. (2005). Statistical Analysis of Microarray Time Course Data. Available from *http://www.ds.unifi.it/StatGen2005/works/day4/speed_latest.pdf.*

Tai, Y.C.; & Speed, T.P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.,* Vol. 34, pp. 3287-2412.

Tan, Y.; Shi, L.; Tong W.; & Wang, C. (2005). Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Research,* Vol. 33, pp. 56-65.

Tibshirani, R.; Hastie, T.; Narasimhan, B.; & Chu, G. (2005). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci,* Vol. 99, pp. 6567-6572.

Trapnell, C.; & Salzberg, S.L. (2009). How to map billions of short reads onto genomes. *Nature Biotech,* Vol 27, pp. 455-457.

Tusher, V.; Tibishirani, R.; & Chu, C. (2001). Significance Analysis of Microarrays Applied o Transcriptional Response to Ionizing Radiation. *Proc. Natl. Acad. Sci,* Vol. 98, pp. 5116-5121.

van't Veer, L.J.;Dai, H.; van de Vijver, M.J.; et al. (2002) Gene expression profiling predits clinical outcome of breast cancer. *Nature*, Vol. 415, pp. 530-536.

Verducci, J., Melfi, V., Lin, S., Roy, S and Sen, C. (2006) Microarray Analysis of Gene Expression: Considerations in Data Mining and Statistical Treatment *Physiological Genomics* 25(3):355-63.

Wang, L.; Chen, G.; & Li, H. (2007). Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data. *Bioinformatics,* Vol. 23, pp. 1486-1494.

Wang, J.; & Kim, S.K. (2003). Global analysis of dauer gene expression in Caenorhabditis elegans. *Development,* Vol. 130, pp. 1621-1634.

Wang, X.; Yue, T.L.; Barone, F.C.; White, R. F.; Clark, R. K.; Willette, R. N.; Sulpizio, A. C.; Aiyar, N. V.; Ruffolo Jr, R. R.; & Feuerstein G. Z. (1995). Discovery of adrenomedullin in rat ischemic cortex and evidence for its role in exacerbating focal brain ischemic damage. *Proc. Natl. Acad. Sci.*, Vol. 92, pp.11480-11484.

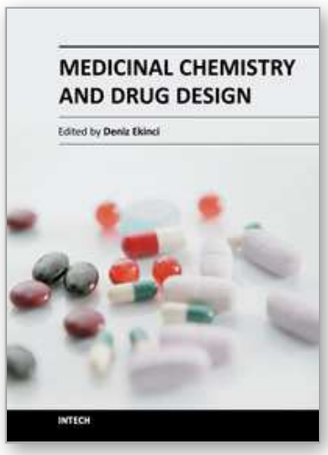
Windle, B. & Guiseppi-Elie, A. (2003). Microarrays and Gene Expression Profiling Applied to Drug Research in: *Burger's Medicinal Chemistry and Drug Discovery* (6th ed.). John Wiley and Sons, Inc. Hoboken, New Jersey.

Wu, B. (2005). Differential Gene Expression Detection Using Penalized Linear Regression Models: the Improved SAM Statistics. *Bioinformatics,* Vol. 21, pp. 1565-1571.

Yang, Y.H.; Dudoit, S.; Luu P.; & Speed, T.P. (2001). *Normalization for cDNA Microarray Data*. San Jose, California. Available from http://www.stat.berkeley.edu/users/terry/zarray/TechReport/589.pdf

Yuan, M.; & Kendziorski, C. (2006). Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions. *Journal of the American Statistical Association,* Vol. 101, pp. 1323-1332.

Zerbino, D.R.; & Velvet, B.E. (2008). algorithms for de novo short read assembly using de Bruijin graphs. *Genome Res.,* Vol. 18, pp. 821-829.



Medicinal Chemistry and Drug Design
Edited by Prof. Deniz Ekinci

ISBN 978-953-51-0513-8
Hard cover, 406 pages
Publisher InTech
Published online 16, May, 2012
Published in print edition May, 2012

Over the recent years, medicinal chemistry has become responsible for explaining interactions of chemical molecules processes such that many scientists in the life sciences from agronomy to medicine are engaged in medicinal research. This book contains an overview focusing on the research area of enzyme inhibitors, molecular aspects of drug metabolism, organic synthesis, prodrug synthesis, in silico studies and chemical compounds used in relevant approaches. The book deals with basic issues and some of the recent developments in medicinal chemistry and drug design. Particular emphasis is devoted to both theoretical and experimental aspect of modern drug design. The primary target audience for the book includes students, researchers, biologists, chemists, chemical engineers and professionals who are interested in associated areas. The textbook is written by international scientists with expertise in chemistry, protein biochemistry, enzymology, molecular biology and genetics many of which are active in biochemical and biomedical research. We hope that the textbook will enhance the knowledge of scientists in the complexities of some medicinal approaches; it will stimulate both professionals and students to dedicate part of their future research in understanding relevant mechanisms and applications of medicinal chemistry and drug design.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yushi Liu and Joseph S. Verducci (2012). Microarray Analysis in Drug Discovery and Biomarker Identification, Medicinal Chemistry and Drug Design, Prof. Deniz Ekinci (Ed.), ISBN: 978-953-51-0513-8, InTech, Available from: http://www.intechopen.com/books/medicinal-chemistry-and-drug-design/microarray-analysis-in-drug-discovery-and-biomarker-identification

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen