High-throughput biotechnologies had improved rapidly within last decade; on the other hand, the statistical methods, for analysing the data generated from these technologies, are falling further behind. One of the most important statistical theories during the process of data analysis is *experimental design*, which was initially proposed by Fisher (1935). Experimental design is a set of procedures that outlines an experiment, which allows us to determine how the analysis can be done. Most of the experiments involves these high-throughput biotechnologies have a two-phase structure, when the responses of experimental units to treatments cannot be measured directly in a single experiment. Subsequent processing (Phase 2) of the initial (Phase 1) experiment is necessary in order for the measurements to be made. The main focuses of this thesis is proteomics experiments, which is the study of proteins. The Phase 1 experiment involves the organisms that are to be perturbed by the experimental conditions of interest. Since the abundance of proteins cannot be measured directly from the organisms, the Phase 2 experiment uses multiplexing techniques such as iTRAQ peptide labelling, coupled with liquid chromatography-mass spectrometry (LC-MS), to measure the abundance of proteins in samples extracted from the organisms in the Phase 1 experiment. These two steps of experimentation are also known as *two-phase experiments*.

The purpose of this chapter is to describe how the methods surrounding the two-phase experiments have evolved over the last few decades. This will aid in the understanding of the two-phase experiment which can also be employed as a part of important idea in the designing two-phase experiment.

**The introduction of two-phase experiment by McIntyre**

Two-phase experiment was first introduced by McIntyre in 1955, where he pointed out several important aspects. The samples from the first phase experiment should be able to evaluate separately from the Phase 2 experiment, this means the analysis of variance (ANOVA) table can be constructed without considering the second phase experiment. In addition, it is essential to have replication in the first phase experiment. Since the treatment groups are normally assigned in the Phase 1 experiment and the statistical test is based on comparing between different treatment groups. Hence, the replication allows us to estimate the experimental error in the first phase experiment which is essential to perform the statistical test. On the other hand, the replicates for the second phase are only important when assessing the measurement errors of the technologies or the methods that applied. The main objective in the theory of two-phase experiment is to link the second phase design to the first phase design by considering all sources of variation from both phases of experiment.

McIntyre then discussed five different two-phase experiment examples with several different design combinations. However, his main focus was on the last example which came from a real experiment that investigated the effects of four light treatments on the synthesis of tobacco mosaic viruses in the tobacco leaves. The first phase experiment consists of eight plants and four leaf positions in two four-by-four Latin squares, then the treatment types are assigned within the plots of each Latin square. Hence, there are 32 observations from the phase one experiment. The second phase experiment consists of four four-by-four Greaco-Latin square designs. The light treatments were estimated by injecting sap from each of the first phase leaves into half-leaves of specific assay plants for the second phase experiment. This generates a total of 128. To analysis this study, McIntyre produced an analysis of variance table explaining the sources of variation introduced from overall two-phase experiment. However, the procedure of how each variance components in the table derived was not mentioned.

In the rest of his paper, McIntyre showed two important concepts. Firstly, he used the light treatments experiment on the different design structures mentioned in the first four examples. He demonstrated the different two-phase design combinations can induce different error variances, in addition, the design used in the real experiment proved to be more efficient than the rest. The second concept is about the replication in the design; McIntyre stated that by doubling the replication of the light treatment experiment, it increases the information of estimated error variance by 19%.

**Curnow’s weighted estimate**

Four years later, Curnow (1957) revisits the theory of two-phase experiment and, in particular, McIntyre’s last example. McIntyre used Latin square analysis to estimate the treatment effect which is equivalent to taking the un-weighted estimate the treatment effects. However, the light treatment experiment applied the incomplete block design; hence the treatment information is located in both “inter-block” and “intra-block” strata. If the accuracies of the two treatment estimates are different and can be estimated without excessive error, then Curnow believed that weighting the estimates according to their accuracies would be a better method. In the light treatment experiment, the treatment effects are relatively small compared to the random error from between half leaves and whole leaves of the test plants; hence, it is not required to use the weighted estimate for the treatment effects. To estimate the variance components, Curnow also believed using the weighted estimates by combing the intra-block and inter-block can give a more accurate result. However, the results from Curnow’s calculation still generated similar results to McIntyre’s for this light treatment experiment. Nevertheless, Curnow’s idea, with the weighted estimates, has given a first stepping stone on the developing a better analysis procedure for the two-phase experiment.

**Further development on the two-phase experiment theory**

Since the initial introduction of two-phase experiments by McIntyre (1955) and subsequent investigation by Curnow, there was no attention on this theory until a publication by Brien in 1983. There, he proposed a set of steps for determining the experimental structures and rules for deriving the analysis of variance tables, focusing on the two-phase experiment. To describe his method briefly, first step is to identify the factors associated the experimental and observational units, then group these factors into different sets, called *tiers*. Using the Wilkinson-Rogers notation, the structure formulae can be written which shows that the relationship between the factors can be determined in each tier, i.e. crossing or nesting relationships. For the two-phase experiment, there will be at least three tiers of factors, two set of block factors and one set of treatment factors. Hence, two phase experiments also known as *multi-tiered* experiment. To derive the analysis of variance table, first step is to expand the structure formulae for each tier to obtain the linear model, then examining for any confounding of any terms between different tiers. The presence of confounding can be determined by study the contrasts that associated specific factors. For example, if a term from a lower tier is confounded to the another terms from a higher tier, then the terms from the lower tier should be within the stratum of the terms from the higher tier in the analysis of variance table. For a more detail description of these procedures see Brien (1983).

Wood et al (1988) emphasised the importance of studying the non-orthogonal block structure which forms the basis of the difficulties encountered in understanding and analysing two-phase experiments. The idea of using the efficiency factors, introduced initially by Yates (1963), was presented to compute the treatment estimators and generalized least squares estimators. In the case of two-phase experiment, there are two sets of block factors from each phase which can be non-orthogonal to each other; therefore, another efficiency factor is present between these two block structures. Wood et al (1988) showed how to incorporate these efficiency factors of the block and treatment structures, and present a set of formulae that computes the treatment estimators and generalized least squares estimators.

In practice, the variance components, from the non-orthogonal block design of two-phase experiments cannot be estimated unless using some computationally intensive methods. Instead, Wood et al noted the idea of using restricted maximum likelihood (REML), described by Patterson and Thompson (1971), to obtain these variance parameter estimates.

Brien and Payne (1999) described a viticultural experiment which has a two-phase structure. They then described an algorithm consists of fitting terms in the model by a series of sweeping operations for analysing the two-phase experiment. There are two types of sweeping operations originally discussed by Wilkinson (1970) and Payne and Wilkinson (1977). The pivotal sweep involves a pivot in which the effects are placed into a unit length vector. This is required for the terms in the first structure formula before sweeps for the terms in the second structure formula which they are confounded. For the two-phase experiments, another pivot sweep is required, since there are always three structure formulae present. The second type of sweep is the reanalysis sweep; this sweeps for the terms from the previous structure to which they are not orthogonal and generally performs after the pivotal sweep. These sweeping sequences allow us to compute the sum of square of the analysis of variance table in the experiment.

**Examples with plant experiments**

A series of plant breeding experiments using two-phase experiment were published since then. Willis et al (2000) investigated the causes of non-indigenous invader plant grows taller and whether it is associated with the genetic level or environmental level. The first phase experiment involves the collection of seeds and the second phase includes the experimental layout of the seedling in the field. The author used the analysis of covariance (ANCOVA) to analyse the data. Due to the missing observations in the experiment, the residual maximum likelihood analysis was also conducted to investigate whether the ANCOVA results were biased by the missing data. The results are both methods were shown to be very similar. This study found a little evidence that increased plant size is a genetically determined characteristic of invasive plants.

Smith et al (2001) investigated the genetic mapping of milling yield in wheat using the two-phase experiment. The first phase is a field experiment which the plant materials consist of double haploid line from two mapping populations in a randomised complete block design. One of two populations was divided into two subsets, and the analysis was performed on each of three populations. The grain sample from the field plots were processed for the subsequent second phase. The authors used resolvable incomplete block design and neighbour balanced design for two of three different sets of samples. The author discussed how the linear model and ANOVA table were built. REML was also used in this study to estimate the variance parameters. In their discussion, authors stated further research is required to provide an efficient arrangement of the sample in the field and in the laboratory with adequate replication in both phases.

Cullis et al (2003) presented a barely malting quality experiment. In their study, there were three different sets of data available from three different field trails. The first two were completed in 2000 and used un-replicated grid plot design with a single plot for each doubled haploid line and multiple plots of Arapiles and Franklin arranged in a systematic grids throughout the trial. The trial in 2001 was designed as randomised complete block design with neighbour balance. Each trial was tested separately in the laboratory for the second phase experiment using a randomised block design. These samples were further processed in order to obtain the traits of interest. This phase involves analysing samples in batches which is a nested block design. Hence, this study consists of three phases: the field trial, malting phase and the analysis of starch enzymes in batches. The author followed the methods described in Brien’s (1983) paper to classify the linear mixed model, structure formulae and ANOVA table for this three-phase experiment. REML was again used to estimate the variance components. The result showed the presence of both field spatial and laboratory variation in the experiment. In addition, the authors also confirmed that accounting this variation in both the design and analysis can substantially change the result. In the discussion, the author emphases a need for developing standard designing software that generates an efficient two-phase design which account for both field spatial and laboratory variation.

Smith et al (2006) published a simulation study based on a plant breading and genetic mapping experiment to show the benefits of using the two-phase experimental design. The authors proposed a general linear mixed model that removes any restriction concerning orthogonality of block structure. A new design principle was proposed namely p/q-rep design, where p is the proportion of Phase 1 replicates to the genotypes and q is the proportion of the Phase 2 replicates to the Phase 1 replicates. This study shows that p=q=0.1 was better than the experiment without replication. The study also shows using phase 2 replication always had higher realized genetic gain that those did not, which conclude the Phase 2 replication is essential for the genetic mapping experiment. There was not any spatial correlation was shown in the phase 1 experiment, which indicating replication in the first phase is not as essential.

**Moving toward a high-throughput biotechnology**

Jarrett and Ruggiero (2008) looked at the two-phase experiment for gene expression two-colour microarray. Initially, they discussed the different orders of fitting the factors to the model can affect the structure of ANOVA table and the way to analyse the treatment effects i.e. computing the F-ratio. Two different types of designs were investigated, these were multiple dye-swap and alternating loop designs. The results shows the multiple dye-swap design is a more robust method, because there are always have at least (r-1) DF for estimating the variance for treatment comparisons. REML was again shown to be a better approach to estimate the variance components. The effectiveness of the estimate was assessed using the effective degrees of freedom computed by the first two moments to those of an approximating chi-square distribution.

**Work by Brien and Bailey (2006, 2009, 2010)**

More recently, Brien and Bailey have done more extensive work on the two-phase experiments. Brien and Bailey (2006) categorised six types of multiple randomisations that involves with the two-phase experiments. Brien and Bailey (2009, 2010) are then establishing the decomposition with respect to each of the six types of multiple randomisations.

These six different types of multiple randomisations are composed, coincident, independent, double, randomised-inclusive and un-randomised-inclusive multiple randomisations across three tiers. The first tier contains the block factors of the Phase 2 experiment, the second tier contains the block factors of the Phase 1 experiment and the third tier contains the treatment factors of the overall two-phase experiment.

The design is assumed to be structure balanced.

For the composed randomisation, the factors of tier 3 are randomised to tier 2 factors and the factors of tier 2 are randomised to the factors of tier 1. These two randomisations are independent to each other. Given that the design is assumed to be structure balanced, the decomposition of the factors between the two tiers can be achieved in any order.

For the randomised-inclusive randomisation, the factors of tier 3 are first randomised to tier 2 factors. The combined tiers 3 and 2 factors are then randomised to the tier 1 factors. The pseudofactors can be used to combine the factors of tiers 3 and 2. Similar to the composed randomisation, under the structure balanced design, the decomposition of the factors between the two tiers can be achieved in any order.

For the un-randomised-inclusive randomisation, the factors of tier 2 factors are first randomized to tier 1 factors. The tier 3 factors are then randomized to the combination of tier 1 and 2 factors. Hence, the pseudofactors can be used to combine the factors of tiers 1 and 2. For the decomposition procedure, the factors of tier 1 are first decomposed by the factors of tier 2. The factors of tier 3 are then decomposed to the result of the first decomposition. If the factors of tier 3 are orthogonal to the first decomposition, then the decomposition described is the same as the combination of the first decomposition and the factor of tier 1 can decomposed directly by the factors of tier 3.

Independent and coincident randomisations are similar where the factors from tier 2 and 3 are randomised to the factors of tier 1 in two separate randomisations. The factors in the first tier, which involve in randomisation, can be the same or different set of factors for distinguishing in between the independent and coincident randomisations. The decomposition procedure for independent randomisations is combination of the factors of tier 1 decomposed by the factors of tier 2 and the factors of tier 1 decomposed by the factors of tier 3. For the coincident randomisations is similar to the un-randomised-inclusive randomisation in terms of randomisation, depending on the relationship between the factors of tier 1 and 3, the combination of the factors of tier 1 decomposed by the factors of tier 2 and the factors of tier 1 decomposed by the factors of tier 3 can be used. Otherwise, the left to right decomposition can also be applied if

The last multiple randomisation is double randomisation which is when the factors from the first tier are randomised to the factors from the tier 2 and 3, i.e. opposite in direction to the coincident and independent randomisations. The decomposition procedure is also left to right.

**Most recent paper by Brien et al.**

The most recent paper by Brien et al. (2011) discuss a systemic approach in designing the Phase 2 experiment taking into account of the Phase 1 experiment. The first principle is to formulate the skeleton ANOVA table using this factor-allocation diagram. The factor-allocation diagram is the randomisation diagram as discussed in Brien and Bailey (2006). In this paper, the author presents a list of rules for calculating the expected mean squares of the ANOVA table. The Chapter 2 of this thesis will present an R package which can automatically produce a skeleton of ANOVA table. The author also address some of the fundamentals in the designing the experiments. These are replication, to measure the random error, randomisation, to avoid systematic biases, and blocking to reduce the variation among experimental units.

Brien et al. (2011) states the replication of the Phase 1 experiment is only required where there is uncontrolled variation in the Phase 2 experiment. For the cases of MudPIT-iTRAQ experiment, the variation between MudPIT runs is known to be large; hence the replication of the Phase 1 experiment is required. The treatment effects of the Phase 2 experiment for the MudPIT-iTRAQ experiment is just tag effect which generally is not the main interest for estimation of the biologists. Thus, in Chapter 4 in finding the optimal design where the Phase 1 experiment is randomised block design, we will show that by confounding the block factor of the Phase 1 experiment with the tag effects can maximised the degrees of freedom in estimating the variance of the treatment effects.

The block factor of the Phase 1 experiment with the highest variation but without treatment information should be confounded with the block factor of the Phase 2 experiment with the highest variation. This situation can be observed in Chapter 4 in finding the optimal design where the Phase 1 experiment is randomised block design. The block factor of the Phase 1 experiment can also be assigned in such way that the block is confounded more with run.

If a given block factor of the Phase 1 experiment contain treatment information, this block factor should be confounded with the block factor of the Phase 2 experiment with the smallest variation.

The treatment should be assigned with the random effects of Phase 2 experiment which consist of the smallest variation. The treatment can be confounded with multiple random effects of Phase 2 experiment which is caused by the non-orthogonal design. However, this paper only focuses on the orthogonal design. In the chapter where searching for the optimal design, there are some cases with the non-orthogonal design will be described.

Pseudofactors can be used to group together a set of levels of a factor and to be randomised for the other factor. This will allows us to keep track of all factors in the experiments or to procedure structure-balanced designs. However, the use of pesudofactor can cause the given factor split which may result in the degradation in the estimation.

If the randomisation procedure is complicated, the factor-allocation diagram can also be complicated. Thus, the author proposed to start with a simple composed randomisation method before using a more complicated randomisation procedure. In addition, all of the Phase 1 factors should always allocate to the Phase 2 factor, and randomise them when it is possible.

Most importantly, plan both phases of the experiment before the biologists commencing any of them. However, typically the biologists would have already done the Phase 1 experiment before consulting in the strategies of designing the Phase 2 experiment.

**Overview of the thesis**

An R package enables the researcher to produce the theoretical ANOVA table of any two-phase experiments quickly. This theoretical ANOVA table consists of the degrees of freedom for each associated mean squares as well as the coefficients of the variance components in every stratum. In addition, the package can also generate the treatment average efficiency factors for every treatment effect to indicate the amount of information that is presented across strata. The theoretical ANOVA tables are used extensively in this thesis to examine the properties of the design given, especially for a complicated design such as two-phase experimental design.

For a given set of design parameters, there are often many ways to allocate the samples collected from the Phase 1 experiment to the iTRAQ labels of the Phase 2 experiment. The following two chapters show the developments in the method for finding the optimal two-phase experiments. The objective function defined can identify the best allocation in terms of allowing a valid test for the treatment effects with the highest average efficiency factor. An improved version of simulated annealing algorithm is then presented for optimising the objective function defined.

The method is focusing in find the optimal design for the two-phase MudPIT-iTRAQ experiment, where the Phase 1 experiment is arranged in completely randomised design and the Phase 2 experiment is arranged in randomised block design. The subsequent chapter will modify the objective function to accommodate for the cases where the Phase 1 experiment is either arranged in randomised block design or balanced incomplete block design.

Finally, the method in computing the variance component and the effective degrees of freedom is presented. The effective degrees of freedom will further help in comparing the design from knowing the variance component estimates, i.e. between runs or between animals.