# Chapter 6 Discussion

The primary purpose of this thesis is to develop a method for the computer generation of optimal designs of two-phase multiplex proteomics experiments. The generation of optimal designs of two- phase experiments uses a combination of theory to define objective functions and computing, to improve the simulated annealing (SA) algorithm. Since the optimal designs are computer generated, there is no restriction on the design parameters (of the Phase 1 experiment) and the end-user does not need to be an expert in designing experiments, and can use this tool to generate these designs. The first part of this thesis applies the method of information decomposition the design of any single- and two-phase experiment, and automates the construction of theoretical ANOVA tables. For the single-phase experiment, the decomposition method is straightforward, as once the strata are defined based on the block structure, the treatment structure is then decomposed within each stratum. In a two-phase experiment, however, decomposition begins with the strata corresponding to the block structure in the Phase 2 experiment, followed by decomposition of the treatment structure into the strata corresponding to the Phase 1 experiment block structure. The procedure for the Phase 1 block-information decomposition is undertaken by regarding the Phase 1 block

factors just as we would treatment factors.

The method of information decomposition applied to designs of any single- and two-phase exper- iments is implemented in an newly developed R package called **infoDecompuTE** which is available on the Comprehensive R Archive Network (CRAN). This R Package allows the user to automate the construction of theoretical ANOVA tables to enable fast assessment of the attributes of designs. These attributes are the degrees of freedom (DF), expected mean squares (EMS), with the variance components and fixed effects components, and the treatment average efficiency factor for every

source of variation.

For researchers who have no R experience, the Shiny application of **infoDecompuTE** package is being hosted at the following link: <https://kcha193.shinyapps.io/infoDecompuTE_Shiny/>. There are three type of outputs that can be generated from this Shiny application: 1) output from the R console as a text file, 2) latex code as a text file, and 3) latex compiled portable document format file.

The second part of this thesis describes a computational approach for finding optimal designs for Phase 2 proteomics experiments using MudPIT-iTRAQTM technologies. Chapter 3 has the Phase 1 experiment arranged in a completely randomised design (CRD). The objective function is constructed aiming to minimise the confounding between Phase 1 Experiment units and Treatment effects with Phase 2 Run and Tag effects. The information matrix is constructed with an orthogonal projection matrix which projects ***y*** onto the Within Runs and Tags vector subspace, by assuming that Tag effects are random.

A three-criterion objective function is derived for generating the optimal design with three properties:

* + 1. information of the Phase 1 Experimental Units is maximised in the Within Runs stratum, based on A-optimal criteria,
    2. treatment information is maximised in the Between Experimental Units Within Runs stratum, based on A-optimal criteria, and
    3. DF of the Treatment effects must still be intact in the Between Experimental Units Within Runs stratum.

The modified nested SA algorithm presented consists of two further improvements. The first improvement is applying the swapping method to only two of the experimental units of the Phase 1 experiment instead of the observational units. The second improvement is the three-stage swapping procedure, which divides a single large search space into three smaller search spaces, swapping the experiment units: 1) within the same runs, 2) within the same tags, and 3) not within the same runs and tags. These improvements are aimed to at speeding up the process in optimising the objective function and then obtaining the optimal design.

Chapter 4 extends the concept to finding the optimal design when the Phase 1 experiment is arranged in blocks, more specifically, a randomised complete block design (RCBD), or a balanced

incomplete block design (BIBD). Having this additional Block factor from the Phase 1 experiment requires us to adjust the objective function to have another criterion in maximising the Residual DF in the Between Plots Within Blocks Within Runs stratum. In addition, instead of having a single equation combining these four criteria with some weights, we optimise this new four-criterion objective function with three incremental steps:

1. The first step is to locate designs where the Phase 1 Plots average efficiency factor in the Within Blocks Within Runs and Tags vector subspace equals 1, and the DF associated with Treatment effects in the Between Plots Within Blocks Within Runs stratum are intact.
2. Then from among the designs located in the first step, the second step uses the modified nested SA algorithm to find optimal designs where the Residual DF in the Between Plots Within Blocks Within Runs stratum are maximised.
3. From among the designs found in the second step, the third step is to find the optimal design where the treatment average efficiency factor in the Between Plots Within Blocks Within Runs and Tags vector subspace is maximised.

Furthermore, two different types of confounding schemes were investigated, where Phase 1 Block effects are intentionally confounded with Tag effects, and where Phase 1 Block effects are intention- ally confounded with Run effects. In general, designs where Phase 1 Block effects are intentionally confounded with Tag effects are shown to have higher Residual DF in the Between Plots Within Blocks Within Runs stratum, because some DF associated with Tag effects are now estimated in the Between Block stratum.

From optimal designs found, if the Phase 1 experiment is arranged in a CRD with fewer than 16 animals (experimental units), it is more preferable to use the four-plex system instead of the eight-plex system, due to the two extra DF available in the Between Animals Within Runs stratum. However, where more Phase 1 animals (experimental units) are used, the degrees of confounding between Animal effects and Run effects increases in the Phase 2 experiment; thus, it becomes preferable to use the eight-plex system over the four-plex system. If the Phase 1 experiment is arranged in Blocks, in general, the four-plex system should still be used where there are fewer than 16 animals (experimental units). However, there is no clear cut-off number of experimental units at which the eight-plex system become better than the four-plex system. This is because having the additional Block component can generate designs with higher Residual DF where Blocks effects

that confounded with Tag effects.

The main purpose of Chapters 3 and 4 is to describe the development of an automated process for finding the optimal design for a wide range of two-phase multiplexing experiments. Even though the main examples are comprise of four- and eight-plex experiments, the methods presented are more general and can be applied to all two-phase designs. This allows researchers using these technologies to design their experiments without requiring expert knowledge in experimental design. In addition, having this tool available also allows the consulting statisticians to present a quick solution to their client. A set of optimal designs that was found is presented in Appendices [C](#_bookmark302), [E](#_bookmark304) and [G](#_bookmark306) and their property is presented as tables in Appendices [D](#_bookmark303), [F](#_bookmark305) and [H](#_bookmark307).

The last part of the thesis is showing how to estimate the VCs (variance components) using a REML (restricted maximum likelihood) where the Phase 2 Run effects are assumed to be random. We then show how to approximate the EDF (effective degrees of freedom), which indicates how well we estimate the variances of Treatment effects, i.e. the residual MS of the stratum associated with the experimental unit. A design with higher EDF provides a better estimate of the variance of Treatment effects. However, the REML method described here does not improve the approximation of the EDF from the optimal designs found in Chapters 3 and 4. This is due to these optimal designs having the property where the Phase 1 experimental units to the Phase 2 Blocks are always balanced, which ensures that we always have a valid F-test for testing Treatment effects. Thus, these optimal designs are robust to the VCs estimation procedure.

### Further line of work

#### Shiny application of generating optimal designs of Phase 2 exper- iments

Scientists are very adaptive at using these technologies, and they even have a good intuitive sense of needing to design their experiments to protect against unwanted systematic sources of variation. The introduction of labelling technologies in multiplexing for the “omics” experiments is evidence of this.

For finding optimal designs in Chapters 3 and 4, we give a set of designs found as presented in Appendices [C](#_bookmark302), [E](#_bookmark304) and [G](#_bookmark306). Researchers can select the design for the two-phase experiment which matches the design parameters that they have applied.

Some R functions on the optimisation algorithm have been written, which will be published as a publicly available package on CRAN. Furthermore, we will also turn this R package into a Shiny application, so that it is are easily accessible to end-users. These end-users would be researchers from a wide range of scientific disciplines. Thus, even the scientists who are unfamiliar with R will feel comfortable using it. Our design methods will become publicly available to researchers with a user-friendly interface.

#### Effective degrees of freedom versus treatment efficiency factor

In Chapter 5, there is an example of a the Phase 1 experiment involving *ν* = 8 treatments assigned to *na* = 16 animals. We compare the theoretical ANOVA from the designs of the Phase 2 experiment using four-plex and eight-plex in Tables [6.1](#_bookmark289) and [6.2](#_bookmark290).

In Table [6.1](#_bookmark289), where the Phase 2 experiment uses the four-plex system, there are 3 DF associated with the Treatment effects estimated in the Between Runs stratum, with a treatment efficiency factors of 0*.*3. Thus, the Run effects are assumed to be fixed, because we cannot recover the extra information on Between Animals VC, *σ*2,from the MS in the Between Animals Between Run stratum for estimating the variance of the Treatment effects. Hence, the EDF of the Between Animal Within Run stratum in this case are always 4 DF. As in Table [6.1](#_bookmark289) where the Phase 2 experiment uses the eight-plex system, confounding occurs between Treatment and Tag effects, with Tag effects containing 0*.*3 of the treatment information. Since the Run effects are assumed as random, we can recover the extra information from the Between Animals Between Runs stratum for estimating the variance of Treatment effects, thus the EDF can be as high as 5 DF.

*a*

Additional work can be done in comparing between recovering the treatment information across runs, and recovering the extra DF in EDF to get a better estimate of the variance. To achieve this, it would mean performing more extensive simulation studies to understand which of these two designs would be preferable and under which circumstances. These circumstances are not just different ranges of values of VCs, but also different ranges of values in the fixed effects for the simulation study.

#### Missing values

One of the issues that arise with high-throughput multiplexing experiments is that of missing data. For a single protein, there are various ways in which the missing values can arise in a MudPIT-

Table 6.1: Theoretical ANOVA table from the Phase 1 experiment arranged in CRD with *ν* = 8 and *rb* = 2, and from the Phase 2 experiment using the four-plex system.

###### Source of Variation DF EMS *Eγ Eτ*

Between Runs Between Animals

Treatment 3 *σ*2 + 2*σ*2 + 4*σ*2 + 1*.*2*θτ* 0*.*3

*a r*

Within Animals 4 *σ*2 + 4*σ*2

*r*

Within Run Between Animals

Tag 1 *σ*2 + 2*σ*2 + 8*θγ* 1

*a*

Treatment 7 *σ*2 + 2*σ*2 + 3*.*23*θτ* 0*.*8077 Residual 4 *σ*2 + 2*σ*2

*a*

*a*

Within Animals

Tag 2 *σ*2 + 8*θγ* 1

Residual 10 *σ*2

iTRAQTM proteomics experiment. One way in which we are most interested is when a unique peptide, which only belongs to a specific protein, is simply not found in one run of the experiment, but can be found on the other runs of the experiment. Thus, during the database searching, the bioinformatic software cannot re-construct this specific protein; and this protein would be considered as missing for one entire run of the Phase 2 experiment. This can be problematic in the analysis stage, as the design is likely to become unbalanced due to unequal replication of the treatment group or the experimental units of the Phase 1 experiment.

For example, consider the Phase 2 experiment with the Phase 1 experiment consisting of *ν* = 4 treatments assigned to *na* = 12 animals. Each animal is then further split into *ns* = 2 sub-samples and measured in the Phase 2 MudPIT-iTRAQTM experiment comprising *nr* = 6 runs and *nγ* = 4 tags. An optimal design of Phase 2 experiment is presented in Table [6.3](#_bookmark291).

The theoretical ANOVA of the full design in Table [6.3](#_bookmark291) is presented in Table [6.4](#_bookmark292). The total of 23 DF are partition to 5 DF for Between Runs stratum and 18 DF for Within Runs stratum. In the Between Animals Within Runs stratum, Treatment effects can be estimated with 0*.*96 amount of the treatment information with 5 Residual DF for estimating the variance of Treatment effects. In addition, there is a valid F-test for comparing between treatments, because the coefficients of VCs are the same for the Treatment and Residual EMS in the Between Animals Within Runs stratum. If a given protein is not detected in Run 6, then there are four observation are missing for the Phase 2 experiment. The theoretical ANOVA is presented Table [6.5](#_bookmark293), which shows the total DF

Table 6.2: Theoretical ANOVA table from the Phase 1 experiment arranged in CRD with *ν* = 8 and *rb* = 2 and the Phase 2 experiment using the eight-plex system.

###### Source of Variation DF EMS *Eγ Eτ*

Between Runs

Between Animals 1 *σ*2 + 2*σ*2 + 8*σ*2

*a r*

Within Animals 2 *σ*2 + 8*σ*2

*r*

Within Runs Between Animals

Tag 3 *σ*2 + 2*σ*2 + 4*θγ* + 1*.*2*θτ* 1 0*.*3 Treatment 7 *σ*2 + 2*σ*2 + 3*.*23*θτ* 0*.*8077 Residual 4 *σ*2 + 2*σ*2

*a*

*a*

*a*

Within Animals

Tag 4 *σ*2 + 4*θγ* 1

Residual 10 *σ*2

Table 6.3: Optimal design for Phase 2 experiment showing the allocation of sub-samples from treatments assigned to animals, where the Phase 1 experiment consists of *ν* = 4 treatments assigned to *na* = 12 animals, *ns* = 2 sub-samples are then taken from each animals and measured in the Phase 2 MudPIT-iTRAQTM experiment comprising *nr* = 6 runs and *nγ* = 4 tags.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Run** | 114 | **Tag**  115 116 | | 117 |
| 1 | *Jb* | *Ld* | *Ea* | *Cc* |
| 2 | *Ld* | *Jb* | *Cc* | *Ea* |
| 3 | *Aa* | *Gc* | *Fb* | *Dd* |
| 4 | *Gc* | *Aa* | *Dd* | *Fb* |
| 5 | *Hd* | *Ia* | *Kc* | *Bb* |
| 6 | *Ia* | *Hd* | *Bb* | *Kc* |

Table 6.4: Theoretical ANOVA table of design in Table [6.3](#_bookmark291).

###### Source of Variation DF EMS *Eγ Eτ*

Between Runs

Between Animals 2 *σ*2 + 2*σ*2 + 4*σ*2

*a r*

Within Animals 3 *σ*2 + 4*σ*2

*r*

Within Runs Between Animals

Tag 1 *σ*2 + 2*σa* + 6*θγ* + 0*.*67*θτ* 1 0*.*1111 Treatment 3 *σ*2 + 2*σa* + 5*.*76*θτ* 0*.*96

2

2

Residual 5 *σ*2 + 2*σ*2

*a*

Within Animals

Tag 2 *σ*2 + 6*θγ* 1

Residual 7 *σ*2

are reduce to 19 DF. The Residual DF in the Between Animals Within Runs stratum are also reduced to 3 DF, which is 2 DF less than the full design. However, there is no direct valid F-test for this design, as coefficients of the VCs from the Treatment and Residual EMS are different in the Between Animals Within Runs stratum. In addition, the amount of the treatment information is also reduced from 0*.*96 to 0*.*8.

Table 6.5: Theoretical ANOVA table of design in Table [6.3](#_bookmark291) with Run 6 missing.

###### Source of Variation DF EMS *Eγ Eτ*

Between Run

Between Animals 2 *σ*2 + 1*.*6*σ*2 + 4*σ*2

*a r*

Within Animals 2 *σ*2 + 4*σ*2

*r*

Within Run Between Animals

Tag 3 *σ*2 + 1*.*27*σ*2 + 1*.*36*θγ* + 0*.*43*θτ* 0*.*2727 0*.*0857

*a*

Treatment 3 *σ*2 + 1*.*96*σ*2 + 4*.*23*θτ* 0*.*8471 Residual 3 *σ*2 + 1*.*78*σ*2

*a*

*a*

Within Animals

Tag 2 *σ*2 + 4*θγ* 0*.*8

Residual 4 *σ*2

If a given protein is not detected in Runs 5 and 6, we are left with 16 observations for the Phase 2 experiment. The theoretical ANOVA of the new design is presented in Table [6.6](#_bookmark294). The Residual DF in the Between Animals Within Runs stratum are reduced to 2 DF, which is 3 DF less than the full design. There is a valid F-test for Treatment effects, with Treatment effects being fully estimable in the Between Animals Within Runs stratum. This is due to how we structured our initial designs a 2-run-by-2-tag array system. Hence, if the last two runs of the experiment are missing, we basically lose one biological replicate, i.e. there are now 8 animals from the Phase 1 experiments, so that the allocation of sub-samples of animals and treatments, to be labelled with tags and analysed with runs still has a balanced structure. The optimal design presented in Table [6.3](#_bookmark291) is shown to be robust in dealing with certain type of missingness, i.e. when Runs 1 and 2, or Runs 3 and 4, or Runs 5 and 6 are missing. Other different types of missingness will result in designs that have no valid F-test for treatment effects, or make it difficult to estimate the VCs from the theoretical ANOVA.

Further simulation studies can be done to explore what happens to the properties of the designs considered in Chapters 3 and 4 with different patterns of missingness. We can investigate how the design can start to breakdown as observed in Table [6.5](#_bookmark293), where there is one run of the experiment that is missing. We can further examine any alternative designs which have more desirable properties in

Table 6.6: Theoretical ANOVA table of design in Table [6.3](#_bookmark291) with Runs 5 and 6 missing.

###### Source of Variation DF EMS *Eγ Eτ*

Between Run

Between Animals 1 *σ*2 + 2*σ*2 + 4*σ*2 Within Animals 2 *σ*2 + 4*σ*2

*r*

*a r*

Within Run Between Animals

Tag 1 *σ*2 + 2*σ*2 + 4*θγ* 1

*a*

Treatment 3 *σ*2 + 2*σ*2 + 4*θτ* 1

*a*

Residual 2 *σ*2 + 2*σ*2 Within Animals

*a*

Tag 2 *σ*2 + 4*θγ* 1

Residual 4 *σ*2

terms of their robustness for downstream statistical analyses where we have missing values.

An alternative approach is to construct an imputation model under a Bayesian multivariate and multilevel inference framework ([Irene SL Zeng](#_bookmark333), [2017](#_bookmark333)). This model uses the information from experimental factors, such as the physical properties of the peptides, the effects from iTRAQTM tags and MudPIT runs, along with the clinical factors of each patient to construct a likelihood model. Each parameter in the likelihood model is estimated by using Empirical Bayesian Hamiltonian MC algorithm, which integrates prior information for missingness and the distribution of missing values. The resultant posterior distribution of these parameters, including parameters of interest, are therefore estimated utilizing both the pattern of missingness and information for missing values. We can incorporate this framework into how to better design the Phase 2 experiment, which will enable us to impute reliable values for the final analysis.

### 6.2 More general future research directions

Another multiplexing technology, which started becoming popular only a few years ago is *Next- Generation Sequencing* (NGS). This multi-plexing technology can be carried out by attaching unique index sequences, namely *barcodes*, on the end of each DNA or RNA fragments ([Smith et al.](#_bookmark361), [2010](#_bookmark361)). Therefore, different barcodes are attached to different biological samples, allowing NGS instrument to sequence multiple samples simultaneously. The abundance levels of sequences are then measured based on the number of barcodes present in each sample. These barcodes are very similar to the iTRAQTM tags, where measuring protein abundances. Note that MudPIT runs of the proteomics

experiments are referred to as *lanes* of the NGS experiments. Thus, the methods of optimal designs described in this thesis also apply to this technology.

We can currently obtain a kit with 96 and 384 barcodes, meaning that we can quantify up to 96 or 384 unique samples at the same time ([Smith et al.](#_bookmark361), [2010](#_bookmark361); [Shapland et al.](#_bookmark360), [2015](#_bookmark360)). However, using more barcodes is not always ideal, because as more barcodes are used the number of DNA or RNA sequences for each barcode decreases ([Campbell et al.](#_bookmark322), [2015](#_bookmark322)). Hence, deciding on the number of barcodes is more practical than theoretical.

Let us consider a Phase 1 experiment arranged in a CRD with *ν* = 8 treatments assigned to *na* = 48 animals, and the sample from each animal split into *ns* = 2 sub-samples, which gives us a total of *n* = 96 sub-samples to be measured using the NGS technology. If a researcher decides to use the kit with 96 barcodes for just one lane of the experiment, then the Treatment effects are completely confounded with Tag effects.

Using the objective function and SA algorithm derived in this thesis, we can quickly generate four optimal designs with multiple lanes, where all have a valid F-test for Treatment effects, with different numbers of barcodes used in the Phase 2 experiment. The Residual DF and the treatment average efficiency factors of these four designs are presented in Table [6.7](#_bookmark296). This shows the best option is to use 8 lanes of the experiment with 12 barcodes, which generated the highest Residual DF (32 DF) and the treatment average efficiency factors (0*.*9837) in the Between Animals Within Runs stratum. However, given that each lane of experiment costs about five thousand dollars, it may be ideal to advise the researcher to use 4 lanes with 24 barcodes, because there is not a lot of improvement compared to using 8 lanes of the experiment with 12 barcodes. Therefore, more work can be done in examining the efficiency of using different numbers of barcodes for generating a better optimal design of the Phase 2 experiment.

Table 6.7: Residual DF and treatment average efficiency factors from the optimal design with different number of lanes and barcodes for Next-Generation Sequencing technology

|  |  |  |  |
| --- | --- | --- | --- |
| Number of lanes | Number of barcodes | Residual DF | *Eτ* |
| 2 | 48 | 17 | 0*.*56 |
| 4 | 24 | 28 | 0*.*8532 |
| 8 | 12 | 32 | 0*.*9837 |
| 16 | 6 | 31 | 0*.*9510 |

Finally, the NGS experiment returns counts as the response. The method in this thesis assumes that the response, once log transformed, is normally distributed; thus, all of the designs we have

generated assume unit-treatment additivity. Having a count as the response violates this assump- tion, and so further research could be undertaken on how to obtain optimal designs of the two-phase experiment where the response exhibits a non-normal distribution.

Further research outlined here will help to maximise the benefits of new technologies, such as NGS, while at the same time extending the capabilities of our method, for generating optimal designs, to a wider range of settings in two-phase multiplex proteomics experiments.