# Chapter 6 Discussion

The primary purpose of this thesis is developing for the computer generation of optimal designs of two-phase multiplex proteomics experiments. The generation of the optimal design of Phase 2 experiment uses combination of theory, to define objective functions, and computing, to improve the simulated annealing (SA) algorithm. Since the optimal design are computer generated, there is no restriction on the design parameters (of the Phase 1 experiment) and the end-users does not need to be an expert in designing experiments can use this tool to generate these designs.

The first component of this thesis is applying the method of information decomposition on de- signs of any single- and two-phase experiments to automate the construction of theoretical ANOVA tables. For the single-phase experiment, the decomposition method is straightforward as once the strata are define based on the block structure, the treatment structure is then decomposed within each stratum. In a two-phase experiment, however, the decomposition begins with the strata cor- responding to the block structure in the Phase 2 experiment, followed by the decomposition into the strata corresponding to the Phase 1 experiment block structure. The procedure for the Phase 1 block-information decomposition by regarding the Phase 1 block factors just as we would treatment factors.

The method of information decomposition on designs of any single- and two-phase experiments was implemented in an 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 design. 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 has 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 (pdf) 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 was constructed aiming to minisied the confounding between the Phase 1 Experiment units and Treatment effects with the Phase 2 Run and Tag effects. The information matrix was constructed by assuming the Tag effects as the random, i.e. orthogonal projection matrix which projects ***y*** into the Within Runs and Tags vector subspace.

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

The modified nest SA algorithm presented consists of two further improvements: The first improvement is applying the swapping method to only the two 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: this involves swapping the experiment units: 1) within the same runs, 2) within the same tags, and 3) not within the same runs and tags. These improvement is aiming to speed up the process in optimising the objective function and then obtaining the optimal design.

Chapter 4 extends the concept of finding the optimal design having the Phase 1 experiment arranged in a blocks, more specifically, a randomised complete block design (RCBD) and a balanced incomplete block design (BIBD). Having this additional Block factor from the Phase 1 experiment

required 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 have 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 which Phase 1 Plots average efficiency factor in the Within Blocks Within Runs and Tags vector subspace equal to 1 and DF associated with treatment effects in the Between Plots Within Blocks Within Runs stratum must be 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 is maximised.
3. From among the designs found in the second step, the third step is to find the optimal design where treatment average efficiency factor in the Between Plots Within Blocks Within Runs and Tags vector subspace is maximised.

Furthermore, two different types of the confounding schemes where the Phase 1 Block effects are intentionally confounded with Tag effects and Phase 1 Block effects are intentionally confounded with Run effects were investigated. In general, designs where Phase 1 Block effects are intentionally confounded with the Tag effects has 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 less than 16 animals (experimental units), it is more preferred 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, when more Phase 1 animals (experimental units) are used, the degrees of confounded between the Animal effects and Run effects increases in the Phase 2 experiment. Thus, it becomes more preferable to use 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 when less than 16 animals (experimental units) are used. However, there is no a clear cut-off number of experimental units, that the eight-plex system become better than the four-plex system. This is because having the an additional Block component can generate designs with higher Residual DF where the Blocks effects that confounded with the Tag effects.

The main purposes of Chapter 3 and 4 is to have an automated process of finding the optimal design for wide range of two-phase multiplexing experiments, which is very important and useful for researchers using these technologies to design their experiments without requiring the expert knowledge in experimental design. In addition, the consulting statisticians can also present an quick solution to their client having this tool available for them. A set of optimal designs that was found are present in Appendices [C](#_bookmark301), [E](#_bookmark303) and [G](#_bookmark305) with their properties presented as tables in Appendices [D](#_bookmark302), [F](#_bookmark304) and [H](#_bookmark306).

The last facet of the thesis was showing how to estimate the VCs (variance components) with 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 the Treatment effects, i.e. the residual MS of the Between Experiemtnal units stratum. A design with a higher EDF provides a better estimates of the variance of Treatment effects, as well as the valid F-test of the Treatment effects. However. REML method described here did not improve the approximation of the EDF from the optimal designs found in Chapters 3 and 4. This is due to these optimal designs have 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 for the Treatment effects. Thus, these optimal designs are robust to the VCs estimation procedure.

### Further line of work

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 have given a set of designs that was found as presented in Appendices [C](#_bookmark301), [E](#_bookmark303) and [G](#_bookmark305). The researchers can used these design for the two-phase experiments which match the design parameter that they have applied.

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

interface.

In Chapter 5, there is an example where the Phase 1 experiment involving *ν* = 8 treatments assigned to *na* = 16 animals. Comparing between the theoretical ANOVA from the designs of the Phase 2 experiment using four-plex and eight-plex in Tables [6.1](#_bookmark288) and [6.2](#_bookmark289).

In Table [6.1](#_bookmark288) 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 treatment efficiency factors of 0*.*3. Thus, the Run effects are assumed to be fixed, because we cannot recover the extra informa- tion 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 is always 4 DF. As in Table [6.1](#_bookmark288) where the Phase 2 experiment uses the eight-plex system, the confounding is occurred between the Treatment and Tag effects, with the Tag effects containing 0*.*3 of the treatment information. We can recovered 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*

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

An additional work can be done in comparing between recovering the treatment information across runs or recovering the extra DF in EDF to get a better estimate of the variance. To achieve this, a future work would mean performing more extensive simulation studies to understand which of these two designs is preferable and under which circumstances. These circumstances are not the just with different ranges of values of VCs, as well as different range of the value in the fixed effects

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

for the simulation study.

One of the issues that arise with high-throughput multiplexing experiments is that of missing data. For a single protein, there are various way in which the missing values can arise in a iTRAQTM proteomics experiment. One way which we most interested in is when a unique peptide, which only belongs to a specific protein, was 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, thus, this protein will be consider as missing for one entire run of the Phase 2 experiment. This can become problematic in the analysis stage, as the design is likely become unbalanced due to unequal replication of treatment group or the experimetnal units of Phase 1 experiment.

For example conisder an example of Phase 2 experiment with Phase 1 experiment consists of *ν* = 4 treatments assigned to *na* = 12 animals. Each animal is then further split to *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](#_bookmark290).

The theoretical ANOVA of the full design in Table [6.3](#_bookmark290) is presented in Table [6.4](#_bookmark291). 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, the Treatment effects can be estimated with 0*.*96 amount of the treatment information with 5 Residual DF for estimating the variance of the Treatment effects. In addition, there is a valid F-test for comparing between the treatments, because the coefficients of VCs are the same for the Treatment and Residual EMS in the Between Animals Within Runs

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* |

stratum.

Table 6.4: Theoretical ANOVA table of design in Table [6.3](#_bookmark290).

###### 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*σ*2 + 6*θγ* + 0*.*67*θτ* | 1 | 0*.*1111 |
| Treatment  Residual | 3 *σ*2 + 2*σ*2 + 5*.*76*θτ*  5 *σ*2 + 2*σ*2  *a* |  | 0*.*96 |
| Tag | 2 *σ*2 + 6*θγ* | 1 |  |
| Residual | 7 *σ*2 | | |

*a a*

Within Animals

If a given protein is not detected in Run 6, there are four observation are missing for the Phase 2 experiment. The theoretical ANOVA is presented Table [6.5](#_bookmark292), which shows the total DF is reduce to 19 DF. The Residual DF in the Between Animals Within Runs stratum is reduced to 3 DF, which is 2 lesser 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 fo the treatment information is also reduced from

0*.*96 to 0*.*8.

If a given protein is not detected in Runs 5 and 6, we left with 16 observation for the Phase 2 experiment. The theoretical ANOVA of the new design is presented in Table [6.6](#_bookmark293). The Residual DF in the Between Animals Within Runs stratum is reduced to 2 DF, which is 3 lesser than the full

Table 6.5: Theoretical ANOVA table of design in Table [6.3](#_bookmark290) 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

design. There is a valid F-test for the treatment effects, with Treatment effects are fully estimable in the Between Animals Within Runs stratum. This is due to the how to structure our initial design with 2-by-2 arrays system. Hence, when the last two runs of the experiment become missing, we basically lost 1 biological replicates, ie. there are now 8 animals from the Phase 1 experiments, so 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](#_bookmark290) has 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 type of missingness will results in designs that has no valid F-test for the treatment effects or difficult in estimating the VCs from the theoretical ANOVA.

Table 6.6: Theoretical ANOVA table of design in Table [6.3](#_bookmark290) 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

Further simulation study can be done to explore what happens to the properties of designs

considered in Chapters 3 and 4 with different pattern of missingness. We can investigate how the design can start to breakdown like what we have observed in Table [6.5](#_bookmark292), where there is only one run of experiment is missing. We can further examine for any alternative designs which have more desirable properties in terms of their robustness for downstream statistical analyses when we have missing values.

An alternative approach is to construct an imputation model under a Bayesian multivariate and multilevel inference framework (cite Irene’s paper). This model uses the information from the experimental factors such as the physical properties of the peptides, the effects from iTRAQTM tags and MudPIT runs 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 distribution of missing values. The resultant posterior distribution of these parameters, including parameters of interest are therefore estimated utilizing both the pattern of the missingness and information for missing values. We can incorporate this framework on how to better design the Phase 2 experiment, which will enable us to impute reliable value for the final analysis.

### Future research directions

Another multiplexing technology which was becoming popular only a few years ago is *Next-Generation Sequencing* (NGS) of DNA. This multi-plexing technology can be carried out by adding unique in- dex sequences, namely *barcodes*, on the both end each DNA or RNA fragment. Therefore, different barcodes are attached to different biological samples, allowing NGS instrument to sequence multiple samples simultaneously. The abundance levels of the sequences are then measured based on the number of barcodes presented for each sample. These barcodes are very similar to the iTRAQTM tags when measuring protein abundances. Thus, the methods of optimal designs described in this thesis also apply to this technology.

We can currently obtain a kit with 96 barcodes, meaning that we can quantify up to 96 samples at the same time. However, using more barcodes is not always ideal, because the intensity deceases as more barcodes are used. Hence, deciding on number of barcode is more practical than theoretical. Let’s consider a Phase 1 experiment arranged in a CRD with *ν* = 8 treatments assigned to *na* = 48 animals, the sample from each animal are split to *ns* = 2 sub-samples, which gives us a total of *n* = 96 sub-samples to be measured using the NGS technology. If a researcher decided to uses all 96

barcode with just one run of the experiment, then the Treatment effects are completely confounded with the Tag effects. Using the objective function and SA algorithm derived in this thesis, we can quickly generate four optimal designs with multiple runs, where all have valid F-test for the treatment effects, with different number of barcodes uses in the Phase 2 experiment. The Residual DF and the treatment average efficiency factors of these four designs are presented in Table [6.7](#_bookmark295). This shows the best option is to use 8 runs of the experiment with 12 barcodes, which generated the highest Residual DF and the treatment average efficiency factors in the Between Animals Within Runs stratum. However, given that each run of experiment cost about five thousand dollars, it may be ideal to advice the researcher to use 4 runs with 24 barcodes, because there were not a lot of improvement compare to using 8 runs of the experiment with 12 barcodes. Therefore, more work can be done in examining the efficiency of using different number 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 runs and barcodes for Next-Generation Sequencing technology

|  |  |  |  |
| --- | --- | --- | --- |
| Number of runs | 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 response. The method in this thesis assume that the responses, which has been log transformed, are normally distributed. Thus, all of the designs we have generated assume unit-treatment additivity. Having a counts as response violate this assumption, thus, another further work can be on how to obtain optimal designs of two-phase experiments with counts as response.