# Abstract

The primary purpose of this thesis is to develop a method for the computer generation of optimal designs of two-phase multiplex proteomics experiments. Two-phase experiment arise when observa- tion cannot be obtained from a experiment (Phase 1), and an extra experiment (Phase 2) is needed to collect the measurements of interest. Such experiments require two different experimental designs for either phases. Proteomics experiments have a two-phase structure, as it starts with organisms being first perturbed by the experimental conditions of interest, i.e. the *Phase 1 experiment*. Since protein abundance cannot be measured directly from the organisms, the organ must be harvested, and the proteins are extracted and then measured in the subsequent laboratory-based experiments,

i.e. the *Phase 2 experiment*. The *multiplexing* technology is introduced that allows the simultaneous analysis of up to eight samples, however it complicates the Phase 2 experimental designs, i.e. how to allocate the samples from the Phase 1 experiment to the Phase 2 experiment.

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. A newly developed R package called **infoDecompuTE** is also developed and available on the Compre-

hensive 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.

The second part of this thesis describes a computational approach for finding optimal designs for Phase 2 proteomics experiments using MudPIT-iTRAQTM technologies. 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. Chapter 3 has the Phase 1 experiment arranged in a completely randomised design. The three-criterion objective function is derived aiming to minimise the confounding between Phase 1 Experiment units and Treatment effects with Phase 2 Run and Tag effects. The concept is further extended when the Phase 1 experiment is arranged in blocks, more specifically, a randomised complete block design, 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. Even though the examples described only involves with four and eight tags, the method of finding optimal designs can also be applied to proteomics multiplexing experiments involving other than four or eight tags. The main purpose is to allow 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 variance components using a restricted maximum likelihood (REML) where the Phase 2 Run effects are assumed to be random. We then show how to approximate the effective degrees of freedom (EDF), 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.