DE GRUYTER DOI 10.1515/sagmb-2013-0005 — Statistical Applications in Genetics and Molecular Biology 2013; 12(2): 263–283

Irene S.L. Zeng*, Thomas Lumley, Kathy Ruggiero, Martin Middleditch, See-Tarn Woon and Ralph A.H. Stewart

Two optimization strategies of multi-stage design in clinical proteomic studies

Abstract: We evaluated statistical approaches to facilitate and improve multi-stage designs for clinical proteomic studies which plan to transit from laboratory discovery to clinical utility. To find the design with the greatest expected number of true discoveries under constraints on cost and false discovery, the operating characteristics of the multi-stage study were optimized as a function of sample sizes and nominal type-I error rates at each stage. A nested simulated annealing algorithm was used to find the best solution in the bounded spaces constructed by multiple design parameters. This approach is demonstrated to be feasible and lead to efficient designs. The use of biological grouping information in the study design was also investigated using synthetic datasets based on a cardiac proteomic study, and an actual dataset from a clinical immunology proteomic study. When different protein patterns presented, performance improved when the grouping was informative, with little loss in performance when the grouping was uninformative.

Keywords: optimization of multi-stage design; clinical proteomic study design; simulated annealing; biological grouping information assisting design.

***Corresponding author: Irene S.L. Zeng**, Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand, e-mail: zeng@stat.auckland.ac.nz
Thomas Lumley and Kathy Ruggiero: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
Martin Middleditch: Centre for Genomics and Proteomics, School of Biological Sciences and Maurice Wilkins Centre for Molecular Biodiscovery, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
See-Tarn Woon: Lab PLUS, Virology and Immunology, LabPLUS, PO Box 110031, Auckland City Hospital, Auckland 1148, New Zealand
Ralph A.H. Stewart: School of Medicine, University of Auckland and Green Lane Cardiac Service, PO Box 110031, Auckland City Hospital, Auckland 1148, New Zealand

Introduction and motivation

Most laboratory-based biomarker discoveries do not reach clinical use. One reason may be the lack of connection between laboratory and clinical proteomics studies, so that laboratory selections and the clinical validation of the protein markers are separate processes in study design (Patterson et al., 2010). In addition, there is a risk that false discoveries are introduced by technical artifacts with different proteomic platforms. In 2007, the National Cancer Institute (NCI) suggested a three-stage workflow to link laboratory discovery to clinical utility in proteomic studies (National Cancer Institute, 2007). The stages are: (1) unbiased discovery using tens of samples, followed by (2) targeted verification using hundreds of samples, and finally (3) clinical validation using thousands of samples. The whole process integrates knowledge on proteomic infrastructure, systematic study design and health economics. It thus requires a systematic design to optimize the number of discoveries under constraints of cost and false discovery.

Multi-stage design in gene-association and proteomic studies

In genetic association studies, Satagopan and Elston (2003) proposed a two-stage design excluding markers with little evidence of association in the first stage of the study and selecting only promising markers for the

second stage. They used Monte Carlo grid search to obtain combinations of one-stage design parameters, i.e., power and type I error, and then applied numerical integration to find the solution of the two-stage design parameters that minimized the study cost subject to an overall type I error rate and a statistical power. Wang et al. (2006) expanded the two-stage approach from candidate-gene to genome-wide scale. Zuo et al. (2008) proposed an optimal resource allocation that maximized the overall power for a fixed total cost. They also investigated the impact of genotyping errors. They derived the joint distribution of the test statistic in the first and second stage, and converted the objective function to a mixed integer nonlinear programming problem (MINLP) with only two parameters under a series of constraints. Skol et al. (2007) described a similar approach to Zou et al. but using a different joint test statistic. Moerkerke and Goetghebeur (2008) added that genetic markers should be selected and ranked in order of evidence that balanced false positive rate and false negative rate at the first stage. At the second stage, more samples are selected and data from both stages combined. They proposed a gain function using the weighted sum of the false positive and false negative rates as the objective function for maximization.

Originally multi-stage designs were suggested for gene association studies for which budgetary considerations needed to be balanced against statistical power. However, the falling cost of genotyping and the economies of scale available from off-the-shelf SNP chips made these designs less useful (Spencer et al., 2009). In a study for genome-wide interaction analysis (GWIA), Steffens (2010) also argued against the adoption of a two-stage strategy and suggested that multi-stage screening will prevent the detection of pure epistatic effects.

In contrast, proteomic studies still have substantial per-protein marginal costs, especially in the final stage, so that a multistage design is substantially more affordable. The multistage design also provides built-in technical validation, with protein abundance being measured using different assays at each stage. The multistage design may still be weak for assessing pure interaction without main effects, but this is not currently a major focus of proteomic research.

Similarities and differences between multi-stage gene association studies and multi-stage clinical proteomic studies

A proteomic study using a systems-biology approach to identify disease-related proteins has similarities to a multi-stage gene-disease association study. It starts with systematic identification and screening of hundreds or thousands of proteins. It then uses a targeted candidate quantification approach to verify and/or validate the findings in the same or a separate group of subjects. The decision on the proteins selected at the identification stage for further study is as vital as that in the screening stage of a genome-wide association study (GWAS). The optimization problem in a multi-stage proteomic study also has similar parameters to a multi-stage GWAS. A common problem in both gene and protein association studies is to search for a design that maximizes power with an acceptable false positive rate and cost, or which minimizes cost with fixed power and false positive rate.

However, there are important differences between proteins and genes relevant to association studies: the number of proteins measurable with current technologies is much less than the number of genes. Proteins are highly changeable and have a wider dynamic range: their abundance in a cell ranges from <500 to 2×10^7 copy numbers (Beck et al., 2011). The most abundant plasma proteins such as albumin and IgE are usually not disease-specific, or of primary interest. Depletion of high-abundance proteins can reduce the problem, but large differences in abundance remain after the depletion.

Current biotechnologies allow identification of thousands of proteins. To achieve an unbiased discovery, NCI advocated a second technical verification using a candidate-based platform. Thus, false discovery due to random error and/or technical artifacts needs to be considered in the design. The statistical method used to adjust for multiple tests in gene association studies may not be the optimal solution for proteomic screening. With the upcoming advances in the biotechnologies, tailor-made study design for large-scale protein research is, therefore, a timely objective.

The potential value of using biological information in protein groups

Bioinformatics profiling information is often used to enrich the design of proteomic studies. Proteins are commonly studied in groups defined by function, structure, and localization (Greenbaum et al., 2003). For instance, a therapy or drug may target the proteins in the same disease related pathway. Hence, it is useful for biologists to study each molecule (protein, metabolite) with others that belong to the same signaling pathway (Meani et al., 2009) or biological function. For example, Hoorn et al. (2005) and Chornoguz et al. (2010) successfully identified proteins and their related pathways or networks that associated with disease or a physiological intervention. Hoorn suggested that the combination of pathway analysis and proteomic analysis both facilitated the interpretation of proteins' relationships and made it possible to identify low abundant proteins which otherwise would escape from the proteomic analysis. Meani et al. (2009) considered the understanding of protein signaling pathways in diseased and normal tissues to be the first step in cancer molecule characterization and personalized therapy. An optimal design using pathway or protein network information may increase the likelihood of candidate proteins of an important group being selected from stage I, and thereby improve biologically and clinically relevant discoveries.

Objectives of the proposed study

We investigate optimal designs under the NCI three-stage workflow, and explore extra options that utilize biological information of proteins via bioinformatics approaches or pathway analysis to enrich the study design. The approaches proposed for genetic association studies are expanded, focusing on validation of the discovery via candidate-based platforms. A range of design problems is investigated, starting from the simplest scenario that proteins are selected separately to the comprehensive option of utilizing protein grouping information, but without consideration of the correlation structure across groups. Our main intention is to provide different options with computing algorithms to achieve robust designs when research resources are constrained.

Statistical strategies in the three-stage design

This section describes optimization strategies for a three-stage study from discovery to verification, and validation using different or the same platforms for different independent samples. In the discovery phase, peptides are identified systematically via mass spectrometry (MS) or 2D gel. The discovered peptides are used to identify and quantify proteins through peptide sequence database searching and bioinformatics software (e.g., ProteinPilot™).

In the second verification stage, multiple-reaction monitoring (MRM) mass spectrometry is applied to verify the changes in abundance that were observed for multiple proteins in the discovery phase. Since the 1990s, MRM-based assays have emerged as an alternative candidate approach to enzyme-linked immunosorbent assays (ELISAs). These mass spectrometry assays eliminate the cost of producing a large number of new immunoassays at an early stage of research, allowing the development of antibodies to be deferred until the final stage.

In the third and final stage, new antibodies and immunoassays are developed and used in larger samples of patients for validation. Multiplex ELISA is one type commonly used in clinical laboratories. A novel alternative is the new mass spectrometry-based quantification for candidate peptides smaller than 10 kDa (Anderson, 2005).

The proposed statistical methods for the optimization of multi-stage studies assume that a known set of p_1 proteins are discovered from the stage I process from which a subset of p_2 of these are then selected using a statistical significance threshold based on information from either individual proteins or both individual and

groups of proteins. A subset p_3 of these proteins is then selected based on a second selection criterion at the verification stage. Finally, these p_3 candidate proteins are validated at the last stage. The sample sizes of the second stage, n_2, and the last stage, n_3, and the stage-wise false positive rates (i.e., type I error) are selected to maximize the power of discovery under the constrained study cost and the overall number of false positives.

Before any proof-of-concept pilot experiment, prediction of the number of discoveries at stage I is difficult because of its dependence on various uncontrollable factors, such as the performance of the mass spectrometer, types of biological tissues and other technical artifacts. Given the limited prior information on stage I design parameters, the stage I sample size is not included in the objective function for optimization. We choose to start the optimization from the selection of p_2 from p_1 discovered proteins so that the optimal solution is not influenced by the number of discoveries at stage I. The stage I discoveries will also provide information (i.e. means and standard deviations) for the design parameters to be used in the optimization.

We demonstrate the optimization in the context of studies involving paired samples at each stage, such as 1:1 matched case-control or before-after intervention studies. This method can be generalized to parallel group studies, with or without paired samples. In the paired sample design, the analytical units will be the log-transformed relative intensities. The detectable mean differences between paired samples are determined based on either prior information and/or clinically or biologically relevant differences. The prior information can be obtained from the literature or prior experiments; it is not limited to the stage I discovery study. The standard deviations can be estimated from the stage I discovery study and/or obtained from prior experiments. In the computations for seeking the optimal design solution, the means and standard deviations of the differences are assumed to be constant across stages.

The optimization assumes the budget is fixed. The assay costs at stages II and III, the cost of recruitment and the stage I sample size are known. A solution of stage I/II nominal false positive rates (decision thresholds) and stage II/III sample sizes is derived to maximize the number of discoveries at the final stage. The following sections describe two algorithms for the optimization with and without biological grouping information.

The simplest scenario: proteins are selected individually

In the simplest scenario, selection is carried out independently for each protein, based on single-protein test statistics. Student's paired sample *t*-test is used to assess the differences in the log-transformed fold change between paired samples. p_2 proteins are selected from p_1 proteins based on p_1 individual tests at stage I. p_3 proteins are selected based on p_2 individual tests at stage II and finally p_3 protein candidates are validated at the final stage based on the individual tests.

Using Simulated Annealing (SA) to seek optimized solution in the multi-stage design: the algorithm SA-a

The proposed method maximizes the expected number of proteins with true effects discovered from a three-stage study under a cost constraint. The expected number of true effects is derived from an objective function which has four design parameters: the stage I type I error rate, α_1, the stage II type I error rate, α_2, the sample size at stage II, n_2, and the sample size at stage III, n_3. The values of these parameters were divided into small intervals within defined ranges (i.e. α_1 ranged between 0.005 and 0.50 with interval size 0.025; α_2 ranged between 0.005 and 0.25 with interval size 0.025; n_2 ranged between 100 and 1000 with interval size 10; n_3 ranged between 100 and 5000 with interval size 100). The combinations of knots at these intervals form the solution space of the objective function in the optimization.

Simulated annealing (SA) is used to determine the optimal design parameters in stages II and III for a specified sample size and number of proteins at the first stage. It is a stochastic optimization method that does not require the objective function to be smooth, and is capable of finding global optima even in problems where many local optima exist (Nikolaev and Jacobson, 2010). In the current problem, lack of smoothness and multiple optima result from the constraint and using Monte Carlo averages to approximate the

expected number of discoveries. In contrast to 'hill-climbing' approaches that attempt to find a higher value of the objective function at each iteration, and so cannot escape a local minimum, SA will sometimes step down. At each iteration, the current solution is compared to the next candidate solution. A superior solution will be accepted with 100% probability; an inferior solution will be accepted with a probability based on the current "temperature" which is a predefined constant number decreasing as the algorithm progresses.

Definition of the SA-a algorithm

The solution space, Ω, bounded by the acceptable limit of each design parameter. Let the vector of design parameters $\omega = (n_2, n_3, \alpha_1, \alpha_2)$ be a solution in Ω, where n_2 and n_3 are the stage II and III sample sizes, respectively, and α_1 and α_2 are the stage I and II type I error rates, respectively. Ω contains all the possible combinations of these parameters which are categorized by small intervals within their bounded ranges.

Objective function

Let $f(\omega)$: $\Omega \rightarrow \mathbb{R}$ be the objective function of the solution space, where f is the expected number of proteins that are discovered at stage III associating with the disease being investigated. It is in the range of 0, 1,..., p_1, where p_1 is the number of proteins discovered in stage I and being considered for inclusion in stages II and III of the study.

The proposal neighborhood selection function

The proposed neighborhoods are constructed by M arbitrarily bounded and possibly overlapping solution subspaces, Ω_i (i=1, 2,..., M). The Ω_i are formed by firstly selecting a point ω_i (the centre of Ω_i) according to either a uniform or Beta distributed jumping length from the previous centre point ω_{i-1}, and secondly selecting a uniformly distributed radius R_i with probability 0.5 for each direction from the selected center ω_i. Each candidate point can then be assigned within each Ω_i, according to a uniform distributed probability.

This nested SA starts with a uniformly random assignment of a solution ω in the radius R_1 bounded neighborhood Ω_1, and then a local SA with k iterations is used to seek the global minimum of Ω_1. After the first local SA, a new address is assigned as the centre of the next solution subset Ω_2 and the second local SA is repeated. This procedure repeats for up to M subsets; the solution from each local SA will be updated if it is better than the previous one

The temperature cooling schedule

The logarithmic cooling schedule is defined as,

$$T_k = \frac{temp}{\log\left(\left\lfloor \frac{t-1}{t_{\max}} \right\rfloor \times t_{\max} + \exp(1)\right)},$$

where t is the current iteration, $temp$ is the starting temperature for the cooling scheme and $t_{\max}$ is the number of function evaluations at each temperature (Belisle, 1992).

The acceptance probability

The Metropolis function, i.e.

$$\begin{cases} \exp\left(-\dfrac{f(\omega')-f(\omega)}{T_k}\right), & f(\omega')-f(\omega)>0 \\ 1, & f(\omega')-f(\omega)\le 0 \end{cases},$$

is used to derive the acceptance probability.

The objective function for SA-a

Let pr_i be the probability of protein i being discovered at stage III (i=1... p_1, where p_1 is the number of proteins selected from stage I). The objective function is then given by

$$f(\alpha_1,\alpha_2,n_2,n_3)=E\left(\sum_{i=1}^{p_1} pr_i\right)=\sum_{i=1}^{p_1} E(pr_i),$$

where α_1 and α_2 are the significance levels at stages I and II, respectively, and n_2 and n_3 are the sample sizes at stages II and III, respectively.

Now, let $c_1=Pt^{-1}(1-\alpha_1/2,df_1)$, $c_2=Pt^{-1}(1-\alpha_2/2,df_2)$ and $c_3=Pt^{-1}(0.975,df_3)$ be the t quantiles corresponding to the type I error rates at stage I, II and III respectively, where Pt^{-1} is the quantile function for Student's t-distribution, and df_1, df_2 and df_3 are the corresponding degrees of freedom at stages I, II and III, respectively.

Let $\beta_{1,i}$, $\beta_{2,i}$ and $\beta_{3,i}$ denote the paired t-test type II error rates at stages I, II and III, respectively, for protein i. It follows that $(1-\beta_{j,i})$ is the power at each corresponding stage, j (j=I, II, III). The expected number of true discoveries (power) is expressed as a function of the cumulative density of t-statistics for the ith protein at each stage, i.e.

$$E(pr_i)=(1-\beta_{1,i})\,(1-\beta_{2,i})\,(1-\beta_{3,i}),$$

where

$$\beta_{1,i}=P\left(\frac{\bar{x}_i-\theta_0}{\delta_i/\sqrt{n_1}}<c_1\right)=P\left(\frac{\bar{x}_i-\theta_i}{\delta_i/\sqrt{n_1}}<c_1+\frac{\theta_0}{\delta_i/\sqrt{n_1}}-\frac{\theta_i}{\delta_i/\sqrt{n_1}}\right)=P\left(T_{df_1}<c_1+\frac{\theta_0-\theta_i}{\delta_i/\sqrt{n_1}}\right)$$

and n_1 represents the known stage I sample size. If θ_0=0, this simplifies to $1-\beta_{1,i}=1-Pt\left(c_1-\dfrac{\theta_i}{\delta_i/\sqrt{n_1}}\right)$, where θ_i is the absolute difference between the matched diseased and normal groups under the alternative hypothesis for protein i and Pt is the cumulative paired sample Student's t-distribution function. Analogously, the objective functions for $1-\beta_{2,i}$ and $1-\beta_{3,i}$ are given by

$$f(\alpha_1,\alpha_2,n_2,n_3)=\sum_{i=1}^{m}\left(1-Pt\left(c_1-\frac{\theta_i}{\delta_i/\sqrt{n_1}}\right)\right)\left(1-Pt\left(c_2-\frac{\theta_i}{\delta_i/\sqrt{n_2}}\right)\right)\left(1-Pt\left(c_3-\frac{\theta_i}{\delta_i/\sqrt{n_3}}\right)\right).$$

The cost function is defined as $n_2\times p_2\times t_2+n_3\times p_3\times t_3+(n_2+n_3)\times R$, where t_2 and t_3 are the assay costs and p_2 and p_3 are the numbers of proteins being tested at stages II and III, respectively, and R is the recruiting cost. This cost function is used in the following simulation study; it may vary based on different cost structures.

The actual objective function of SA-a computes the expected number of positive findings by using the Monte Carlo average of 1000 simulations. Additionally, technical differences between the stage I and stage II assays can be simulated by multiplying each θ_i by a random "technical artifact" adjustment, λ_i, in the Stage I calculations. Our simulations below incorporate this adjustment.

Comparison of nested neighborhood selection with single-step selection

Instead of using single-step SA, algorithm SA-a employs a nested-search strategy on subsets of the solution space determined by both the jumping length from one centre to another and the radius of the search space. Comparing the single-step method with the nested-search method, the latter constructs a local structure

of the global search surface. This strategy is shown to be more efficient with shorter computation time and without losing effectiveness in finding a good solution. In a case study to identify the global solution of a function with known maxima under inequality constraints, the computing time of using the single-step search was about twice of that using the nested search. The discovery rate for the known maximum from 100 experiments using 10,000 iterations in the global search was 54%. Compared to an equivalent nested-search of 100 subsets × 100 iterations, the discovery rates were 64%, 58% and 97% for uniform, $\beta(\alpha=4, \beta=6)$-, and $\beta(\alpha=4, \beta=20)$-distributed jumping lengths, respectively. When the global search used 100,000 iterations and, equivalently, 100 subsets of 1000 iterations in the nested-search, the discovery rate of the known global maximum from 100 experiments were all 100%.

The convergence of SA-a can be proved by theorem 1 of both Belisle (1992) and Hajek (1988). Belisle's theorem 1 is a special case of Hajek's result in which the state space is discrete and finite. SA-a is defined over subsets of $\mathbb{R}^d$, with a temperature scheme converging in probability to 0. Its transition probability from one candidate to another is positive. When M (the number of subsets) is sufficiently large, it can naturally deduce that SA-a converges in probability to the global minimum of the bounded space Ω.

An enrichment design: using protein group information and protein selection by group and individual

Under this more complex scenario, proteins are analyzed in biological groups. Selection of proteins at stages I and II is based on the combined criteria of group and individual hypothesis tests. A protein is selected if the single-protein test statistic exceeds the threshold of a corresponding type I error rate for the t-test or if the group test statistic exceeds the threshold of a corresponding type I error rate for the Hotelling's T-test. The validation/selection of proteins at the final stage is only based on t-tests for the individual proteins.

The following paragraph describes a simulated annealing algorithm SA-b (Table 1.), which is used to optimize and simulate the three-stage design when grouping information for each discovered protein is available in a paired sample study. Utilizing the additional grouping information, nested simulated annealing with Beta-distributed jumping lengths is used to find the optimal design solution. The selection criteria combine decision thresholds of Hotelling's T-squared statistics for the groups and the t-statistics for the individual proteins.

Apart from using grouping information, compared to SA-a, several improvements have also been made in SA-b. The first is to convert the inequality cost constraint into an equality cost constraint by using a series of slack terms (Nocedal and Wright, 1999). The second is the reduction in the dimension of the design problem by using the fact that the cost constraint will always bind. Instead of searching the entire interval of the stage III sample size n_3, now n_3 is derived from the current cost constraint and other chosen design parameters from the early stages. Because the cost function is monotonic with all the design parameters, this change reduces the computing time used to search those n_3s with inferior solutions. The third improvement is to add an overall false-positive constraint in the algorithm.

Definition of the simulated annealing algorithm SA-b using grouping information

The solution space Ω bounded by the acceptable limit of each design parameter

Let the vector of design parameters $\omega = (n_2, n_3, \alpha_{t_1}, \alpha_{t_2}, \alpha_{f_1}, \alpha_{f_2})$ be a solution in Ω, where n_2 and n_3 are the Stage II and III sample sizes, α_{t_1} and α_{t_2} are the stage I and stage II type I error rates for the individual tests and α_{f_1} and α_{f_2} are the type I error rates for the group tests. Ω contains all the possible combinations of these parameters categorized into small intervals within the bounded ranges.

Objective function

Let $f(\omega)$: $\Omega \to \mathbb{R}$ be the objective function of the solution space, where f is the expected number of proteins detected at stage III. The expected number of detected proteins with true effects is subject to first- and

Table 1 The contents of the SA-b algorithm.

Step 1. Assign study parameters: cost constraint, "technical artifact" adjustment vector λ, mean difference and its standard deviation for each protein, and cost functions for stages II and III.
Step 2. Initialize number of iterations, simulated annealing parameters and solution.
Step 3. Initialize the sequences of slack term, S_i, for the cost constraint; *i* ranges from 1 to J.
Step 4. While the number of iterations <*M*, repeat the following steps:
4.1 Randomly select an address as the centre of the local search neighborhood using a uniformly or Beta distributed jumping length.
4.2 Activate simulated annealing for the local search with *k* iterations. The simulated annealing local search algorithm contains three functions: 1. the objective function, which uses Monte Carlo simulation to calculate the expected number of detected positives at the final stage; 2. the proposal neighborhood function, which determines the next searching subset of new candidate points; and 3. the cost-sample size function that calculates the stage III sample size according to the inequality cost constraint, slack term S_i, cost functions and the currently chosen design parameters.
4.3 Compare the local maximum with the best solution from the past. If the current solution is better, then replace the previous best solution with the current one.
4.4 Start next neighborhood search and repeat Step 3.
4.5 Repeat Step 2 using the next slack term S_{i+1}.

second-stage type I error rates of the group Hotelling's *T*-tests, the individual *t*-tests, second-stage sample size and third-stage sample size. In the optimization, this objective function is constrained by: 1) cost and 2) the number of false positives. The selection criteria of the multi-stage design are:

– Stage I: (group test p-value $<\alpha_{f_1}$) or (individual test p-value $<\alpha_{t_1}$ and group test p-value <0.05), i.e.

$$\left(T^2>F^{-1}_{df(p_1),df(n_1-p_1)}(1-\alpha_{f_1})\right)\cup\left(T>Pt^{-1}_{df(n_1)}(1-\alpha_{t_1}/2)\cap T^2>F^{-1}_{df(p_1)df(n_1-p_1)}(0.95)\right)$$

– Stage II: (group test p-value $<\alpha_{f_2}$ and individual test p-value <0.05) or (individual test p-value $<\alpha_{t_2}$), i.e.

$$\left(T^2>F^{-1}_{df(p_2),df(n_2-p_2)}(1-\alpha_{f_2})\cap Pt^{-1}_{df(n_2)}(0.975)\right)\cup\left(T>Pt^{-1}_{df(n_2)}(1-\alpha_{t_2}/2)\right)$$

– Stage III: $T>Pt^{-1}_{df(n_2)}(0.975)$

In the above, α_{t_1} and α_{t_2} are the significance levels of individual tests at stages I and II; α_{f_1} and α_{f_2} are the significance levels of the group tests at stages I and II; T^2 is the *F*- distributed Hotelling's *T*-squared statistic with degrees of freedom determined by the number of proteins and the sample size at each stage; *T* is the Student's *t*-statistic; F^{-1} is the quantile function for the *F*-statistic.

The configuration of the objective function is described in Section 2.2.2.

A similar cost function as described in 2.1.2 is defined as $n_2\times p_2\times t_2+n_3\times p_3\times t_3+(n_2+n_3)\times R-S$, where t_2 and t_3 denote the assay costs at stages II and III, respectively, *R* is the recruiting cost, *S* is the slack term of the total budget, and p_2 and p_3 are the numbers of proteins being tested at stages II and III, respectively.

The false-discovery constraint controls the expected number of false discoveries and is defined as $m\times 2Pt(c_1)\times 2Pt(c_2)\times 2Pt(c_3)$, where *m* represents the total number of proteins with true effects.

The actual objective function in SA-b computes the expected number of proteins with true effects by using the Monte Carlo average of 1000 simulations with adjustment for technical artifacts. To utilize the grouping information and according to requirements from the subject area, the first-stage criterion is set to select groups with a changeable significance level, and proteins with a changeable significance level but belonging to groups significant at the fixed 0.05 level during the optimization. The second-stage criterion is set to select proteins with a changeable significance level, and proteins significant at 0.05 levels but belonging to groups with a changeable significance level. The third-stage selection is based only on the individual tests being significant at the 0.05 levels.

In SA-b, the proposal neighborhood selection function, temperature cooling schedule, and acceptance probability are set to be the same as those of SA-a.

Use of analytical approximation to compute the analytical objective function for SA-b

In SA-b, using the Monte Carlo average to estimate the expected number of true discoveries prolongs the optimization process. To simplify the optimization, we investigated using an approximated analytical function to replace the Monte Carlo average. The expected number of true discoveries is given by

$$\sum_{i=1}^{p}(1-\beta_{1,i})(1-\beta_{2,i})(1-\beta_{3,i}),$$

where $\beta_{1,i}$, $\beta_{2,i}$ and $\beta_{3,i}$ represent the nominal type II error rates at stages I, II and III, respectively, for the *i*th protein. Under the selection criteria for this multi-stage design utilizing the protein group information, described in Section 2.2.1, the analytical function for the type II error, $\beta_{1,i}$, of the *i*th protein at stage I is equivalent to the probability that *the group containing the* i*th protein is not selected at the current group test decision threshold* (*event* A), and *either the* i*th protein is not selected at the current individual test decision threshold* (*event* B) or *the group is not selected at the 0.05 level* (*event* C).

The probability of the *i*th protein not being selected at stage I is, therefore, be expressed as $pr(A\cap(B\cup C))$, and can be expanded to

$$Pr((A\cap B)\cup(A\cap C))=pr(B)\times(A|B)+pr(A\cap C)-pr(B)\times pr((A\cap C)|B)$$

Analytically, $\beta_{1,i}$ is a function of the cumulative density function of the *t*-statistic and the cumulative density function of the group Hotelling's *T*-squared statistic which is *F* distributed after the transformation and is conditional on the individual *t*-statistic for each protein. It can be decomposed as follows.

Let $pr(B)$ denote the probability that the *i*th protein is not selected at the current *t*-test threshold. It can be expressed as $pr(B)=Pt(t<c_1+t_i)$, described in 2.1.2, where c_1 is the threshold for the corresponding type I error of the *t*-test; and t_i is the *t*-statistic for the *i*th protein. Now, let $pr(A|B)$ denote the probability that the group containing the *i*th protein is not selected at the current group test decision threshold, given that the *i*th protein is not selected at the current *t*-test threshold. This can be expressed as $pr(A|B)=F\left(T_i^2<d_1\middle|t<c_1+t_i\right)$, where T_i^2 represents the scaled *F* distributed Hotelling's *T*-squared statistic of the group containing the *i*th protein; and d_1 represents the *F*-statistic for p-value < the decision threshold of Hotelling's *T*-test for the group.

$pr(A\cap C)$ is the probability that the group of *i*th protein is not being selected under the combination of the group test statistic thresholds ($d_{0.05}$ and d_1) and can be expressed as $pr(A\cap C)=F\left(T_i^2<\min\left(d_1,d_{0.05}\right)\right)$, where $d_{0.05}$ represents the *F* statistic for p-value <0.05 in the group test. Finally, let $pr((A\cap C)|B)$ denote the conditional probability of $A\cap C$ given the *i*th protein is not selected.

The conditional cumulative *F* density, defined as $pr(A|B)=F\left(T_i^2<d_1\middle|t<c_1+t_i\right)$, is equivalent to the marginal distribution of the cumulative *F* density with respect to the *t*-statistic for the *i*th protein, i.e.

$$F\left(T_i^2<d_1|t<c_1+t_i\right)=\int_{-\infty}^{c_1+t_i}F\left(T_i^2<d_1\right)\times pt(t)\,dt, \qquad (a)$$

where $pt(t)$ represents the density function of the *t*-statistic, and Hotelling's *T*-squared statistic is given by

$$T_i^2=\frac{\lambda_i\left(X_i-u_i\right)^T\left(X_i-u_i\right)}{S_i},$$

where

$$\lambda_i=\frac{n_1-p_{1,i}}{p_{1,i}\left(n_1-1\right)}$$

denotes the scale factor which transforms Hotelling's *T*-squared statistic into an *F*-statistic; X_i and u_i are the observed and null-hypothesis means for all proteins in the group containing the *i*th protein; S_i is the variance-covariance matrix of this group, n_1 is the stage I sample size, and $p_{1,i}$ is number of proteins included in stage I for the group to which the *i*th protein belongs.

The integrand in equation (*a*) is approximated by $F\left(\tilde{T}_i^2<d_1-\lambda\times t_{1,i}^2\right)\times pt(t)$, i.e.

$$F\left(T_i^2 < d_1 | t < c_1 + t_i\right) \approx \int_{-\infty}^{c_1 + t_{1,i}} F\left(\tilde{T}_i^2 < d_1 - \lambda_i t_{1,i}^2\right) \times pt(t)\, dt,$$

where the group test, Hotelling's T-squared statistic T_i^2, is approximated by the sum of $\tilde{T}_i^2$ and the t-statistic for the ith protein (i.e. $T_i^2 \approx \tilde{T}_i^2 + \lambda_i t_i^2$), and $\tilde{T}_i^2$ is T_i^2 excluding the mean effect of the ith protein (i.e. $\tilde{T}_i^2 = S^{-1}\lambda_i \left(X_{-i} - u_{-i}\right)^T \left(X_{-i} - u_{-i}\right)$).

Finally, we approximate $pr((A \cap C)|B)$ by $pr(A \cap C)$, which would be exact if B were independent of A and C.

A similar approximation is also applied to the stage II nominal type II error $\beta_{2,i} = \Pr(B \cap (A \cup D))$, where D denotes the event that *a protein is not selected at the 0.05 significance level*. $\beta_{2,i}$ is expanded as

$$\Pr((B \cap A) \cup (B \cap D)) = pr(B) \times pr(A|B) + pr(B \cap D) - pr(B \cap D) \times pr(A|(B \cap D)).$$

The approximated analytical objective function for SA-b was implemented in several synthesis datasets for comparing with its Monte Carlo simulated function. The computing times of using the analytical approximation were shown to be between 20–100 times faster than using the Monte Carlo average in SA-b. The design parameters and solutions were also shown to be similar to the results utilizing the Monte Carlo simulated objective function. More discussions are provided in the following immunology case study.

Case studies

An immunology study

The lymphocyte proteome was analyzed in 17 Common Variable Immunity Deficient (CVID) patients and 34 normal controls. CIVD, also known as acquired hypogammaglobulinemia, is the most common primary immunodeficiency disorder encountered in clinical practice (Park et al., 2008). CVID patients have low levels of immunoglobulin G, A and M; and also are susceptible to recurrent infections because of their inability to produce antibodies. Much of the past research has focused on deciphering the genetic basis of CVID (Park et al., 2008). However, the genetic causes of this disease are complex and still not fully understood. We hypothesize that proteomic characterization of CVID cases (beyond the gross immunoglobulin deficiencies) will be an alternative approach to reveal genetic causes and mechanisms. This study aimed to identify proteins with differentiated expression in CVID patients compared to the matched controls.

Patients and controls were matched by age group, ethnicity and gender. All patients and controls are Caucasian. Lymphocytes were isolated from blood using Ni-NTA agarose (Invitrogen) in an accredited lab (Lab PLUS, Auckland City Hospital). The proteome of the lymphocytes cell lysates were then analyzed and quantified by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) at the Center for Genomics and Proteomics, University of Auckland. Samples were all analyzed using the tagged proteomics technique iTRAQ, where eight samples were allocated as one batch of the multiplex assay. A reproducibility pilot study was performed before the main discovery study. Since the reproducibility of the experiments was shown to be satisfactory, the main study was performed. This mass spectrometry-based approach identified peptides from 289 proteins and provided the relative quantification for each peptide. The log-transformed relative quantity of the peptide was used to derive the natural log transformed protein ratio for patients and normal controls in the hierarchical multi-level mixed effect model. The proteins were grouped into 20 rudimentary classes according to their biological function by the Biochemist (STW) in Table 2, while the overlapped functions of some proteins were not presented.

Grouping of proteins

The 289 proteins discovered were grouped according to their functions: namely immunity, metabolic, tumor, protein synthesis/degradation, nuclear metabolic, cell migration, ER membrane, protein structure, signaling

Table 2 The functional group and observed fold-changes in their relative intensity between the matched normal controls and patients on the natural log scale.

Proteins	Estimate	StdErr	Group
EZR cDNA F	2.3702	0.5794	Immunity
CTSG Cathe	2.6498	0.6883	Immunity
VIM Viment	0.4954	0.2677	Immunity
MNDA Myelo	0.5187	0.2532	Immunity
PSMA6 28 k	0.3144	0.1672	Immunity
MPO Isofor	0.3874	0.2368	Immunity
IL4R IL4R n	1.0728	0.5822	Immunity
CALR Calre	0.2572	0.1735	Immunity
MIR1244-3;	−1.1069	0.755	Immunity
S100A8 Pro	0.5802	0.4038	Immunity
DSG2 Desmo	0.8601	0.5558	Immunity
LTF Unchar	0.6448	0.463	Immunity
DEFA1;DEFA1	0.4371	0.3296	Immunity
YBX1 Prote	2.0715	0.6459	Immunity
LCP1 Plast	0.3222	0.2527	Immunity
CORO1A Cor	0.3327	0.267	Immunity
PPIA Pepti	0.2315	0.1871	Immunity
S100A9 Pro	0.3585	0.2926	Immunity
MSN Moesi	0.2576	0.2173	Immunity
PSME2 Unch	0.2518	0.217	Immunity
MIF Macroph	0.6192	0.4975	Immunity
UBA1 Ubiqu	2.5894	0.6904	Metabolic
GLRX Gluta	0.7093	0.3601	Metabolic
LYZ Lysozy	0.3312	0.1898	Metabolic
CA1 Unchar	0.3841	0.2144	Metabolic
PGLS 6-pho	1.4955	0.9774	Metabolic
HNRNPK cDN	0.5699	0.3119	Tumor
PKM2 Pyruv	0.3182	0.3106	Tumor
HSPA5 cDNA	0.6351	0.1413	Protein synthesis
AARS cDNA	−1.2433	0.436	Protein synthesis
RPS10-NUDT	−0.7353	0.2703	Protein synthesis
RPS5 40S r	0.3428	0.2061	Protein synthesis
HNRNPA2B1 I	0.6766	0.3045	Nuclear metabolic
HNRNPK cDN	0.5699	0.3119	Nuclear metabolic
APRT Adeni	0.4482	0.2525	Nuclear metabolic
MNDA Myelo	0.5187	0.2532	Nuclear metabolic
S100A4 Unc	0.341	0.2102	Nuclear metabolic
BANF1 Barr	0.3083	0.1815	Nuclear metabolic
ANP32A;ANP	0.3749	0.2375	Nuclear metabolic
HNRNPC cDN	0.6791	0.4071	Nuclear metabolic
LGALS1 Gale	1.1107	0.7363	Nuclear metabolic
VCP Transi	0.4181	0.3348	Nuclear metabolic
TUBA1B Unc	0.6446	0.2011	Cell migration
FCHO2 89 k	0.6863	0.2114	ER membrane
TMSB4X TMS	0.5501	0.2172	Protein structure
KRT9 Kerat	0.8742	0.3978	Protein structure
PRKAR1A cAM	0.6009	0.1081	Signaling
HSPA5 cDNA	0.6351	0.1413	Signaling
ANXA4 cDNA	0.9266	0.3574	Signaling
S100A11 Pr	0.4635	0.2458	Signaling
HSP90AA1 I	0.5218	0.2874	Signaling
RAC1 Isofo	0.5155	0.3021	Signaling

function, mitochondrial, blood protein, DNA repair/structure, trafficking/secretory, inflammation, apoptosis, autoantibody, ER/membrane, angiogenesis, transcription, neuro protection and redox. Nine groups were noted to contain protein candidates with significant fold changes between CVID patients and normal controls. Fifty-two proteins in total were needed to be considered for inclusion in the stage II verification study. We used the SA-a, SA-b, and the SA-b with analytical approximation to identify the optimal solutions for the second and third stages of this study.

Demonstration of code and results for three-stage design using SA-a, SA-b and approximation for SA-b

The cost structure used in this study is different to that described in 2.2.1. At stage II, the cost per protein for peptide synthesis is $280 and per biological sample for proteomic analysis is $1015. At stage III, the cost is assumed to be $200 per protein per biological sample for laboratory analysis. The recruitment cost is set to be $100 per biological sample. The assay cost functions in the R language for stages II and III are defined as

```
assaycost2=function(n,p){280*p+1015*n}
```

and

```
assaycost3=function(p) (200*p),
```

respectively, where p is the number of proteins selected at the nominal stage and n is the sample size.

The programs were run in the computer clusters of NeSI: http://www.nesi.org.nz/, where each program was assigned to 4GB memory within a cluster.

The codes used in the R function to utilize group information and analytical approximations are:

```
> optim.two.stage.appr(budget=6e6, protein=protein, N1=30,
  artifact=rep(1,52),iter.number=10,assaycost2.function=assaycost2,
  assaycost3.function=assaycost3,recruit=100,a1.t.min=0.01,a1.t.max=0.25,
  a1.f.min=0.01,a1.f.max=0.25,a1.step=0.025,a2.t.min=0.01,a2.t.max=0.05,
  a2.f.min=0.05, a2.f.max=0.05, a2.step=0.025, n2.min=100, n2.max=1000,
  n2.step=100, n3.min=100, n3.max=1000, n3.step=100)
```

The approximation programs had run times within an hour. The group program SA-b had run times between 7 and 10 h for the five proteins examples and between 20 and 30 h for the 52 proteins examples.

Different budget ranges determined by the known health funding agents were tested for this case study. Three different budgets with the solutions are presented for the verification/validations of the top five proteins of interest, and the targeted 52 proteins in Table 3. Considering the relatively low prevalence of CVID, all budgets were assessed by different ranges of stage II sample size. To verify and validate the top five proteins, using ranges of 30–100 and 100–1000 for the stage II sample size are demonstrated to be feasible. The 1.2 million dollar budget was shown to be insufficient for a sample size between 100 and 1000 at stage II, 1000–5000 at stage III, but is sufficient for a sample size in the range of 30–100 at stage II and 100–1000 at stage III.

The solutions using the approximated analytical objective functions are shown to be in a similar range to the results from their Monte Carlo simulated objective functions, except for the first and third scenarios presented in Table 3. In these two scenarios, while both solutions from the analytical function achieve the same numbers of discoveries as their Monte Carlo simulation, they contain two smaller stage II sample sizes as the design parameters and thus results in different resource allocations. This discrepancy indicates the existence of multiple global optimal solutions for the objective function.

Despite the similar results (sample sizes and costs) from using and not using grouping information in this study, due to the large fold changes for all included proteins, proteins' functional group information is still considered essential for biologists to assess the discoveries and assist in the decision making in the protein selections from stage I.

Table 3 The optimal design parameters for a given budget using three different algorithms for the multi-stage CVID proteomic study.

Objectives	Method		
	SA-a	SA-b	SA-b, with analytical approximation
Full discovery of 52 proteins Cost =\$6×10^6 n_2=100–1000	pt_1, pt_2, n_2, n_3: 0.10,0.04,500,517 Cost stage II: 572,060 Cost stage III: 5,426,940 Time: 12.7 h	pt_1, pf_1, pt_2, pf_2, n_2, n_3: 0.22,0.10,0.04,0.05,500,517 Cost stage II: 572,060 Cost stage III: 5,426,940 Time: 20.0 h	pt_1, pf_1, pt_2, pf_2, n_2, $n_{3\ (100–1000)}$: 0.22, 0.22, 0.01, 0.05,365,100 Monte Carlo objective function used to derive n_{3*}: 532 Cost stage II: 421,535 Cost stage III: 5,577,465 Time: 56 min
Full discovery of 52 proteins Cost =\$1.2×10^6 n_2=30–100	pt_1, pt_2, n_2, n_3: 0.18,0.01,86,104 Cost stage II: 110,450 Cost stage III: 1,088,550 Time: 11.7 h	pt_1, pf_1, pt_2, pf_2, n_2, n_3: 0.11, 0.18, 0.01, 0.05,86,104 Cost stage II: 110,450 Cost stage III: 1,088,550 Time: 19.0 h	pt_1, pf_1, pt_2, pf_2, n_2, $n_{3\ (100–1000)}$: 0.04,0.15,0.01,0.05,90,100 Monte Carlo objective function used to derive n_{3*}: 104 Cost stage II: 114,910 Cost stage III: 1,084,090 Time: 53 min
Discovery of five *most* interesting proteins Cost =\$5×10^5 n_2=100–1000	pt_1, pt_2, n_2, n_3: 0.20,0.01,330, 118 Cost stage II: 369,350 Cost stage III: 129,650 Time: 3.2 h	pt_1, pf_1, pt_2, pf_2, n_2, n_3: 0.05,0.20,0.01,0.05,330,118 Cost stage II: 369,350 Cost stage III: 129,650 Time:7.0 h	pt_1, pf_1, pt_2, pf_2, n_2, $n_{3\ (100–1000)}$: 0.02,0.04,0.01,0.05,100,200 Monte Carlo objective function used to derive n_{3*}: 351 Cost stage II: 112,900 Cost stage III: 386,100 Time: 5 min
Discovery of five *most* interesting proteins Cost =\$5×10^5 n_2=30–100	pt_1, pt_2, n_2, n_3: 0.01,0.01,60,392 Cost stage II: 68,300 Cost stage III: 430,700 Time: 3.3 h	pt_1, pf_1, pt_2, pf_2, n_2, n_3: 0.06,0.01,0.01,0.05, 60, 392 Cost stage II: 68,300 Cost stage III: 430,700 Time: 8.3 h	pt_1, pf_1, pt_2, pf_2, n_2, $n_{3\ (100–1000)}$: 0.04,0.01,0.01,0.05,74,100 Monte Carlo objective function used to derive n_{3*}: 378 Cost stage II: 83,910 Cost stage III: 4,150,90 Time: 7 min

[a]Stage III sample size, n_3, was re-derived using the Monte Carlo simulated objective function. The solution from this Monte Carlo simulated function assumes the study used up the budget minus the slack term.
[b]The stage I sample size equals to the number of controls in this study. It needs to be greater than the number of proteins each group. The minimal stage II sample size also needs to be greater than the number of proteins in each group.

Using simulated protein datasets

Data

To assess the performance of the SA-b algorithm and to investigate the factors that are associated with the efficiency of the program, we simulated different protein patterns from synthetic datasets that were generated from a cardiac proteomic study (Zeng et al., 2009). The cardiac proteomic study collected coronary plasma blood samples of eight ischemic patients before and after an angioplasty procedure, and used LC-MS/MS with iTRAQ™ labeling to discover and quantify the proteins. The simulated datasets were created by using mean differences and ranges of variances in the relative quantity on the log scale between these two time points. Different patterns were simulated by setting the mean difference to zero or by doubling the variances of some proteins. The factors being investigated included the grouping property of proteins, number of proteins with non-zero mean differences, variations in the protein effect, and budgets. The grouping property focuses on the co-regulation of proteins in the same biological functional group, which are believed *a priori* to act in concert with one another.

Table 4a Optimal design for a given budget in scenario: dataset comprises of 50 identified proteins of interest, and of which 44 proteins with true effects distribute in seven of the 10 protein groups (proteins with true effects are clustering within groups; each group has more than one protein with true effect-informative grouping).

Budget	**$1 million**		**$5 million**		**$10 million**	
	SA-b	**SA-a**	**SA-b**	**SA-a**	**SA-b**	**SA-a**
Expected number of true effects for the final optimized solution	No acceptable solution for a full discovery		40.8	40.5	41.4	41.1
Design parameters of the optimized solution	NA		pt_1, pf_1, pt_2, pf_2, n_2, n_3	pt_1, pt_2, n_2, n_3	pt_1, pf_1, pt_2, pf_2, n_2, n_3	pt_1, pt_2, n_2, n_3
			0.10,0.25,0.01,0.05,100,138	0.25,0.01,100,156	0.135,0.225,0.05,0.05,200,274	0.25,0.05,200,306
Results (budget allocation, false negative rates, discovery rate, and computing time) for the current optimized solution						
False negative rates for different proteins of true effects	NA		Protein no. 100: 0.5% Protein no. 104: 0.2% Protein no. 137: 15.2% Protein no. 139: 9.3% Protein no. 142: 0.1% Protein no. 144: 0.6% Protein no. 146: 3% Protein no. 148: 0% Protein no. 149: 0.2% Protein no.s 105,121,145: 100%	Protein no. 100: 3.7% Protein no. 104: 1.7% Protein no.137: 26.3% Protein no. 139: 16.5% Protein no. 142: 0.1% Protein no. 144: 0.9% Protein no. 146: 4.3% Protein no. 148: 0% Protein no. 149: 0.3% Protein no.s 105,121,145: 100%	Protein no. 100: 0% Protein no. 104: 0% Protein no. 137: 0.1% Protein no. 139: 0.1% Protein no. 142: 0% Protein no. 144:0% Protein no. 146: 0% Protein no. 148: 0% Protein no. 149: 0% Protein no.s 105,121,145: 100%	Protein no. 100: 0.9% Protein no. 104: 0.2% Protein no. 137: 5.8% Protein no. 139: 4% Protein no. 142: 0% Protein no. 144: 0.2% Protein no. 146: 1.3% Protein no. 148: 0% Protein no. 149: 0.1% Protein no. 105,121,145: 100%
Discovery rates for different proteins of true effects	NA		100% for others (excluding proteins recorded above)			
Costs at stage II and III	NA	NA	Stage II: 3,727,840 Stage III: 1,271,160	Stage II: 3,585,760 Stage III: 1,413,240	Stage II: 7,437,120 Stage III: 2,561,880	Stage II: 7,171,520 Stage III: 2,827,480
Computation time	172,876 s	69,787 s	171,847 s	69,530 s	174,224 s	97,464 s

Table 4b Optimal design for a given budget in scenario: dataset comprises of 50 identified proteins of interest, and of which 18 proteins with true effects distribute in 18 of the 18 protein groups (only one protein has true effect in each group- non informative grouping).

Budget	**$1 million**		**$5 million**		**$10 million**	
	SA-b	**SA-a**	**SA-b**	**SA-a**	**SA-b**	**SA-a**
Expected number of true effects	No acceptable solution for a full discovery		16.9	16.8	16.9	17.0
Design parameters of the optimized solution	NA		$pt_1, pf_1, pt_2, pf_2, n_2, n_3$	pt_1, pt_2, n_2, n_3	$pt_1, pf_1, pt_2, pf_2, n_2, n_3$	pt_1, pt_2, n_2, n_3
			0.25,0.25,0.05,0.05,100,411	0.11,0.01,200,345	0.25,0.25,0.05,0.05,200,820	0.25,0.05,200,1288
Results (budget allocation, false negative rates, discovery rate, and computing time) for the current optimized solution						
False negative rates for different proteins of true effects	NA		Protein no. 100: 4.1%	Protein no. 100: 4.5%	Protein no. 100: 3.3%	Protein no. 100: 1.2%
			Protein no. 137: 12.2%	Protein no. 137: 11.1%	Protein no. 137: 8.5%	Protein no. 137: 4.7%
			Protein no. 142: 0.1%	Protein no. 142: 0.1%	Protein no. 142: 0.1%	Protein no. 142: 0%
			Protein no. 105: 100%	Protein no. 105: 100%	Protein no. 105: 100%	Protein no. 105:100%
Discovery rates for different proteins of true effects	NA		100% for others (excluding proteins recorded above)			
Costs at stage II and III	NA	NA	Stage II: 3,124,000 Stage III: 1,875,000	Stage II: 3,492,000 Stage III: 1,507,000	Stage II: 6,248,000 Stage III: 3,751,000	Stage II: 4,242,880 Stage III: 5,756,120
Computation time	219,787 s	54,065 s	225,338 s	71,590 s	226,110 s	76,367 s

Table 4c Optimal design for a given budget in scenario: dataset comprises of 50 identified proteins of interest, and of which 44 proteins with true effects distribute in seven of the 10 protein groups (proteins with true effects are clustering within groups; each group has more than one protein with true effect-informative grouping).

Budget	**$1 million**		**$5 million**		**$10 million**	
	SA-b	**SA-a**	**SA-b**	**SA-a**	**SA-b**	**SA-a**
Expected number of true effects	No acceptable solution for a full discovery	5.9	6.0	6.0	6.0	6.0
Design parameters of the optimized solution	NA	pt_1, pt_2, n_2, n_3 0.01,0.01,100,176	pt_1, pf_1, pt_2, pf_2, n_2, n_3 0.10,0.25,0.01,0.05,100,1464	pt_1, pt_2, n_2, n_3 0.20,0.01,200,1091	pt_1, pf_1, pt_2, pf_2, n_2, n_3 0.10,0.25,0.01,0.05,100,3687	pt_1, pt_2, n_2, n_3 0.09,0.04,472,2585
Results (budget allocation, false negative rates, discovery rate, and computing time) for the current optimized solution						
False negative rates for different proteins of true effects		Protein no. 144: 1% Protein no. 145: 1% Protein no. 146: 1% Protein no. 147: 1.2% Protein no. 148: 10% Protein no. 149: 0.3%	Protein no. 149: 0.1%	Protein no. 149: 0.1%	Protein no. 149: 0.1%	Protein no. 149: 0.1%
Discovery rates for different proteins of true effects	NA	100% for others (excluding proteins recorded above)				
Costs at stage II and III	NA	Stage II: $580,000 Stage III: $385,349	Stage II: $1,300,000 Stage II: $3,220,114	Stage II: $2,760,000 Stage III: $2,398,333	Stage II: $1,300,000 Stage III: $8,109,918	Stage II: $4,248,000 Stage III: $5,686,083
Computation time	111,032 s	39,858 s	106,202 s	39,553 s	111,032 s	39,858 s

– The above program used ranges of stage I *t*-test p-value (pt_1) between 0.01 and 0.25 with step size 0.025; stage II *t*-test p-value (pt_2) between 0.01 and 0.05 with step size 0.025; stage I *F* test p-value (pf_1) between 0.01 and 0.25 with step 0.025; stage II *F* test p-value (pf_2) 0.01 and 0.05 with step 0.025; n_2 from 100 to 1000 with step 100; false positive rate <0.01. The final stage used *t*-test with >85% power at 0.05 significance level.

– Table summarized results used 9×1000 Hybrid Simulated Annealing search; all results were verified by 19×1000 SA search. The technical artifact λ is set to be `(1, 1, 0.8, repeat (1, 45 times), 0.9, 0.8)`. The assay cost is set to (NZ$800, NZ$200) with recruitment cost of NZ$1000.00 and slack term cost of NZ$1000.00.

Each synthetic dataset comprises 50 proteins of which 44, 18 or six have non-zero mean difference, which we will refer to as "true effects". These are either clustered in a few groups or scattered across different groups, with some proteins either in overlapping or non-overlapping groups.

The expected number of discovered true effects (true positives, power) is affected by multiple factors. These factors include cost, significance thresholds at stages I and II, sample sizes at stage II and III, and the effect size (mean difference/standard deviation) of each protein. Results from some of these datasets are shown in Tables 4a–c.

Results using SA-b for a multi-stage design in different simulated protein datasets

Computation time and number of true effects

The simulations were implemented using computer clusters with 16 CPUs of 1 GB per CPU. Computation time is shown to increase with the number of true effects.

Budget, numbers of true effects and design parameters

In the simulated data of 44 true effects among 50 proteins (Table 4a), the budget of $10 million results in 90% discovery. In the simulation with 18 (Table 4b) or six (Table 4c) true effects among the 50 proteins of interest, $5 million is sufficient for 95% discovery in the 18 true-effects scenario and 100% discovery for the six true-effects scenario. The budget of $1 million achieves 100% discovery in the six true-effects scenario. All simulations use the same stage I sample size of 60 and the same cost function as described in sections 2.2.1 and 2.2.2 and footnotes of Table 4c.

In scenarios where the $10 million budget cannot achieve 100% discovery of all true effects, we note that the optimal stage I *F*-test decision threshold for selection is close to the upper bound of the parameter space. This phenomenon indicates that, the default 0.05 threshold would be far from optimal given the small sample size at stage I and the budget constraint. Both of the decision thresholds for the stage I *F*- and *t*-test are >0.05. Conversely, in Table 4c, where a $1 million budget can achieve a 100% discovery for the five true effects, the optimal stage I *t*-test decision threshold is smaller than 0.05.

The relations between the cost ratio of stage III-to-stage-II and the p-value of the stage I individual *t*-test, the cost ratio of stage III-to-stage-II and the p-value of the stage I group test were investigated using the 44 true effects data. The stage II sample size was fixed at 100, and the budget at $5 million. The p-values of the stage I *t*-tests were set between 0.001 and 0.25, and the p-values of the stage I *F*-tests were set between 0.01 and 0.25. When using SA-a, the cost ratio is shown to decrease with a higher p-value threshold for the *t*-test (Figure 1b). When using SA-b, although a similar relation between the cost ratio and the p-value threshold for the groups' *F*-tests is observed, the p-value of the *t*-test does not influence the cost ratio within the same band of the *F*-test p-values.

Effect size and number of detectable true effects

In the synthesized datasets, there are several proteins with extremely small effect sizes that cannot be detected. The detection of these proteins are hindered by the sample size and significance thresholds at stages I and II. Under the unconstrained optimization, 100% discovery was achieved for the case of 44 true effects with a second stage sample size of 670 and third stage sample size of 2800 when the stage I individual test p-value <0.36 and the second stage individual test p-value <0.16, given that the stage I sample size was 60. When there are no multiple stage selections, a sample size of 4751 can detect the protein with the smallest effect size (mean difference=0.1, standard deviation=2.3) with 85% statistical power and 5% type I error rate. This indicates that the detection of proteins with small effect size may be restricted using the systematic approach due to the step-wise type I error rate control and the constrained monetary resource in a proteome-wide study.

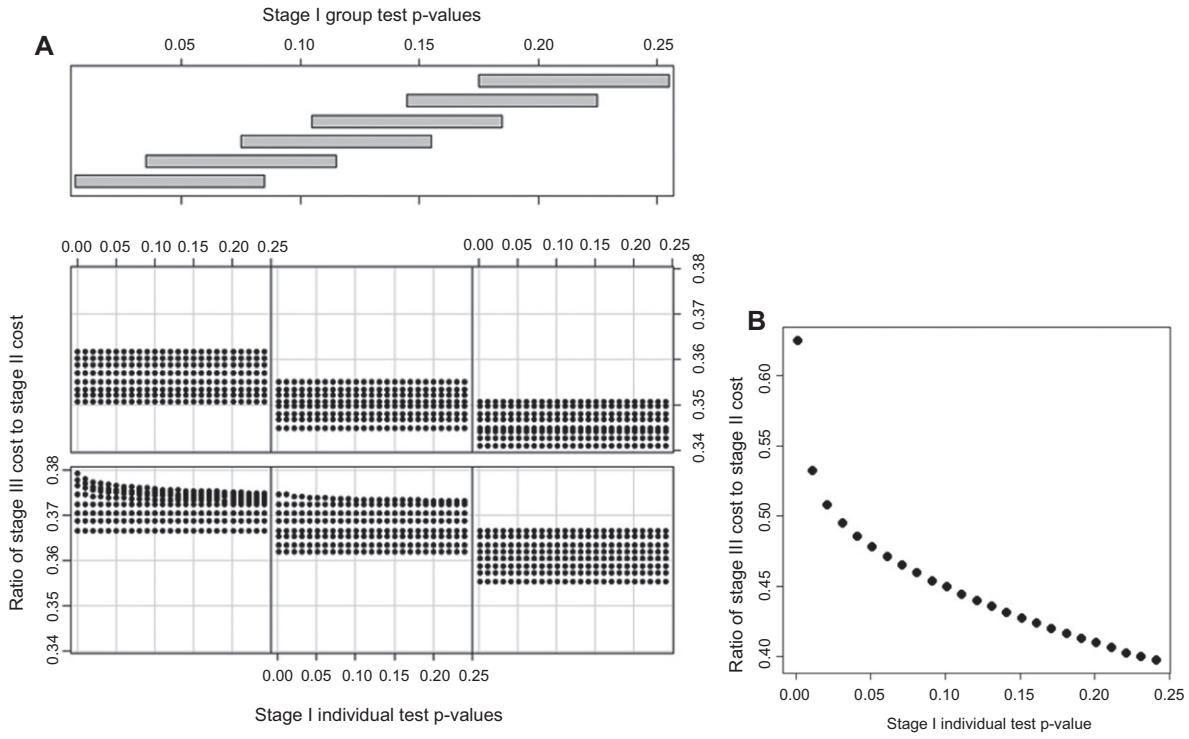

Figure 1 Associations between cost ratios and test decision thresholds in scenarios of using vs. not using biological group information.
(A) The six graphs represent the associations between cost ratios and stage I *t*-test, when the group *F* test p-values are in different ranges. The six graphs arrange in a descending order of the group test p-values, starting from the bottom left corner to the upper right corner. The protein dataset has 44 true effects among 50 proteins and is the same one to that used in Table 4a. (B) The graph represents the association between cost ratios and stage I *t*-test p-values with a same range as that in Figure 2a. The protein dataset has 44 true effects among 50 proteins discovered at stage I and is the same one to that used in Table 4a.

Convergence

SA-b is restricted to a smaller solution space, in which only those n_3 meeting the cost constraint are included. Thus, the convergence of algorithm SA-b is better than SA-a when the same number of iterations is applied.

Overlapping groupings

When utilizing biological information, a protein may belong to several functional groups (Whitford, 2005). It is known that there are overlapping protein complexes sharing several proteins within biological networks. For example, in the TNF/NF-κB signaling pathway, proteins p100, 1KKa, 1KKb and 1KKc are shared by several functional groups in this pathway (Zotenko et al., 2006). When utilizing SA-b, the overlapping proteins can be included in the group statistic for every group to which they belong.

A comparison between using grouping information and not using grouping information

Simulations using different synthetic protein datasets were conducted and used to investigate the influence of different protein patterns in optimizations using SA-a (without group information) and SA-b (with group information). When the budget is under a tight constraint and the grouping is informative, SA-b results in

Table 5 Different scenarios to use SA-a and SA-b.

When to use SA-a	When to use SA-b
1. There is a small number of proteins that are of interest (i.e., <5)	1. There is a large number of proteins that are of interest
2. The fixed budget will be more than sufficient for the verification/validation of all proteins of interests	2. There is informative group information (i.e., some proteins have a large effect size and are clustered in the same group)
3. All proteins in the same group have a large effect size	3. A number of proteins of interest have small effect size and cluster with proteins of large effect size in the same group
4. All proteins belong to a single group	

more proteins being selected from stage I given that the number of proteins in each group is less than the sample size. SA-a results in fewer proteins being selected from stages I, but larger sample size in stage III.

Comparison of protein discovery rates between SA-a and SA-b within the same ranges of design parameters shows that, SA-b has more favorable results in the protein-wise discovery when there is informative grouping. Informative grouping information increases the individual protein discovery rate and the average number of true discoveries. Uninformative grouping information does not make a meaningful difference to the discovery rate and cost allocation. The benefit of using grouping information is greater when the budget is under a tight constraint for detecting a large number of true effects. Under this condition, SA-b tends to allocate more resources to verifying more proteins at stage II. With respect to CPU running time, SA-b uses about twice to three times more system time than SA-a.

Table 5 provides scenarios of when to use SA-b and SA-a. Since SA-b with the analytical approximation runs much faster than the other two methods, it should be used firstly to assess whether a fixed budget will yield a good design solution to verify/validate the proteins of interest.

Discussion

Proteomic techniques used to investigate large numbers of proteins simultaneously are comparable to genomic platforms used to investigate gene-disease associations, and have similar challenges in experimental design and data analysis (Greef et al., 2007). In this paper, we used simulated annealing to simultaneously optimize the design for a multi-stage proteomic study comprising discovery, verification and clinical validation phases, taking into account the resource constraints for maximizing the number of true discoveries.

We investigated two different strategies for the design of a multi-stage clinical proteomic study, and recommend considering biological grouping information in the optimization of the design. Multi-stage designs are cost-effective because non-promising candidates can be eliminated after the first stage, leaving only promising candidates to be validated in later stages. While, with the falling cost of genotyping, multi-stage designs are no longer commonly used in genome-wide association studies, they remain appealing for proteomic studies given the substantial per-protein cost of clinical validation. As suggested by the NCI, verification using a candidate-based platform and validation in large-scale clinical samples will improve the discoveries of disease related proteins and their final translation to utilization. A systematic approach to design optimization allows resources to be allocated efficiently across the different stages of the study. Further, using integrated biological information enriches the design for laboratory discovery and clinical application and thereby optimizes the solution. From simulations of different protein datasets in the current paper, we discovered that using protein grouping information improves the optimization results when the grouping information is informative.

We also found that a structured two-step search was more efficient than a one-step global search and that using a Beta distribution for jump lengths in the two-step search further improved the speed.

A design based only on individual-protein tests could be optimized more easily because the objective function is smooth and can be calculated analytically, but individual-protein tests do not make full use of available biological information. Using a combination of individual-protein and group tests gives an objective function that has no simple analytical form, and for this reason Monte Carlo estimation and simulated annealing is necessary.

An important limitation of the current group algorithm is that Monte Carlo estimation prolongs the computing time required for the optimization process. However, the computations that form the main computing load can be easily parallelized, and the code made more efficient by using a faster programming language. Greater gains are also shown to be achievable from an analytical function to approximate the objective function. The current algorithms are conditional on the stage I discovery design parameters (sample size and number of discoveries). This limitation reflects a common problem in the funding process that many biomedical researchers currently face. Before significant funding can be sought for a multiple-phase study, pilot data from a stage I discovery is often needed as proof-of-concept; the stage I sample size is, therefore, determined by the available funds at this pilot stage. In general, the pilot study has a small available budget. As recommended in the current practice, the stage I sample size is in the range of 10–100. However, some of our simulations showed that a larger stage I sample size (>100) leads to a smaller cost allocation in the stage II verification, and increase the statistical power at stage I. This suggests that a bigger range of sample size at stage I may need to be considered in some cases. This will be one topic of our future research.

The software

The R functions optim.two.stage.single (SA-a), optim.two.stage.group (SA-b) and optim.two.stage.app (SA-b using analytical approximation) performing the methods described in this paper are contained in the R package proteomicdesign 2.0. This package is available from the CRAN website: http://www.r-project.org. The R functions have been assessed and tested on multiple synthetic datasets (parts of these results were shown in the paper), and an actual case study dataset at the desktop and the computer cluster.

Acknowledgements: The authors wish to express their gratitude to Sharon Browning (Department of Statistics, University of Washington, Seattle), Rohan Ameratunga and Wikke Koopman (Lab PLUS, Auckland District Health Board, New Zealand), Patrick Gladding (North Shore Hospital, Auckland, New Zealand) and Jocelyne Benatar (Cardiac Vascular Research Unit, Auckland City Hospital) for useful discussions and involvement in the cardiac and immunology proteomics studies. The authors wish to acknowledge the financial support from Green Lane Research and Education Trust and A+ Charitable Trust for the cardiac and immunology proteomic studies. The authors also wish to acknowledge the contribution of the NeSI high-performance computing facilities and the staff at the Centre for eResearch at the University of Auckland (Gene Soudlenkov and Sina Masoud-Ansari). New Zealand's national facilities are provided by the New Zealand eScience Infrastructure (NeSI) and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation and Employment's Infrastructure programs (URL: http://www.nesi.org.nz). The authors also acknowledge the associate editor and two anonymous reviewers for their constructive comments.

References

Anderson, L. (2005) "Candidate-based proteomics in the search for biomarkers of cardiovascular disease," J. Physiol., 563(1), 23–60.

Beck, M., A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg and R. Adbersold (2011) "The quantitative proteome of a human cell line," Mol. Syst. Biol., 7, 549, 1–8.

Belisle, C. J. P. (1992) "Convergence theorems for a class of simulated annealing algorithm on R^d," J. Appl. Probab., 29, 885–895.

Chornoguz, O., L. Grmai, P. Sinha, K. A. Artemenko, R. A. Zubarev and S. Ostrand-Rosenberg (2010) "Proteomic pathway analysis reveals inflammation increases myeloid-derived suppressor cell resistance to apoptosis," Mol. Cell. Proteomics, 10(3), 1–9.

Greef, J. V. D., S. Martin, P. Juhasz, A. Adourian, T. Plasterer, E. R. Verheij and R. N. McBurney (2007) "The art and practice of systems biology in medicine: mapping patterns of relationship," J. Proteome Res., 6, 1540–1558.

Greenbaum, D., C. Colangelo, K. Williams and M. Gerstein (2003) "Comparing protein abundance and mRNA expression levels on a genomic scale," Genome Biol., 4(9), 117.1–117.8.

Hajek, B. (1988) "Cooling schedules for optimal annealing," Math. Opera. Res., 13, 311–329.

Hoorn, E. J., J. D. Hoffert and M. A. Knepper (2005) "Combined proteomics and pathways analysis of collecting duct reveals a protein regulatory network activated in vasopressin escape," J. Am. Soc. Nephrol., 16(10), 2852–2863.

Meani, F., S. Pecorelli, L. Liotta and E. F. Petricoin (2009) "Clinical application of proteomics in ovarian cancer prevention and treatment," Mol. Diagn. Ther., 13(5), 297–311.

Moerkerke, B. and E. Goetghebeur (2008) "Optimal screening for promising genes in 2-stage designs," Biostatistics, 9(4), 700–714.

National Cancer Institue. (2007). Building the Foundation for Clinical Cancer Proteomics Clinical proteomic technologies for cancer 2007 Annual Report. Retrieved from http://proteomics.cancer.gov/.

Nikolaev, A. G., and S. H. Jacobson (2010) Simulated annealing. In: J.-Y. P. M. Gendreau (Ed.), Handbook of Metaheuristics. New York: Springer.

Nocedal, J., and S. J.Wright (1999) Numerical optimization, New York: Springer.

Park, M. A., L. T. Li, J. B. Hagan, D. E. Maddox and R. S. Abraham (2008) "Common variable immunodeficiency: a new look at an old disease," Lancet, 372, 489–502.

Patterson, S. D., J. E. V. Eyk and R. E. Banks (2010) "Report from the Wellcome Trust/EBI "Perspectives in Clinical Proteomics" retreat- a strategy to implement next-generation proteomic analyses to the clinic for patient benifit: pathway translation," Proteomics Clin. Appl., 4, 883–887.

Satagopan, J. M. and R. C. Elston (2003) "Optimal two-stage Genotyping in population-based association studies," Genet. Epidemiol., 25(2), 149–157.

Skol, A. D., L. J. Scott, G. R. Abecasis and M. Boehnke (2007) "Optimal designs for two-stage genome-wide association studies," Genet. Epidemiol., 31(7), 776–788.

Spencer, C. C. A., Z. Su, P. Donnelly and J. Marchini (2009) "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," PLoS Genetics, 5(5), e1000477.

Steffens, B. (2010) "Feasible and successful: Genome-wide interaction analysis involving all 1.9×1011 pair-wise interaction tests," Hum. Heredity, 69(4), 268–284.

Wang, H., D. C. Thomas, I. Pe'er and D. O. Stram (2006) "Optimal two-stage genotyping designs for genome-wide association scans," Genet. Epidemiol., 30, 356–368.

Whitford, D. (2005). An introduction to protein structure and function. In Proteins structure and function, USA: John Wiley & Sons, Ltd.

Zeng, I. S. L., S. R. Browning, P. Gladding, M. Jullig, M. Middleditch and R. A. H. Stewart (2009) "A multi-feature reproducibility assessment of mass spectral data in clinical proteomic studies," Clin Proteomic., 5, 170–177.

Zotenko, E., K. S. Guimarães, R. Jothi and T. M. Przytycka (2006) "Decomposition of overlapping protein complexes: A graph theoretical method for analyzing static and dynamic protein associations," Algorithms Mol. Biol., 1(7), 1–11.

Zuo, Y., G. Zou, J. Wang, H. Zhao and H. Liang (2008) "Optimal two-stage design for case-control association analysis incorporating genotyping errors," Ann. Hum. Genet., 72, 375–387.