Available online at www.sciencedirect.com
ScienceDirect
Journal of Nutritional Biochemistry 21 (2010) 561–572
Journal of Nutritional Biochemistry

REVIEWS: CURRENT TOPICS

Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology☆

Wenjiang J. Fu[a], Arnold J. Stromberg[b], Kert Viele[b], Raymond J. Carroll[c], Guoyao Wu[d,*]

[a]Department of Epidemiology, Michigan State University, East Lansing, MI 48824, USA
[b]Department of Statistics, University of Kentucky, Lexington, KY 40536, USA
[c]Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[d]Faculty of Nutrition and Department of Animal Science, Texas A&M University and Department of Systems Biology and Translational Medicine, Texas A&M College of Medicine, College Station, TX 77843, USA

Received 26 September 2008; received in revised form 10 November 2009; accepted 12 November 2009

Abstract

Over the past 2 decades, there have been revolutionary developments in life science technologies characterized by high throughput, high efficiency, and rapid computation. Nutritionists now have the advanced methodologies for the analysis of DNA, RNA, protein, low-molecular-weight metabolites, as well as access to bioinformatics databases. Statistics, which can be defined as the process of making scientific inferences from data that contain variability, has historically played an integral role in advancing nutritional sciences. Currently, in the era of systems biology, statistics has become an increasingly important tool to quantitatively analyze information about biological macromolecules. This article describes general terms used in statistical analysis of large, complex experimental data. These terms include experimental design, power analysis, sample size calculation, and experimental errors (Type I and II errors) for nutritional studies at population, tissue, cellular, and molecular levels. In addition, we highlighted various sources of experimental variations in studies involving microarray gene expression, real-time polymerase chain reaction, proteomics, and other bioinformatics technologies. Moreover, we provided guidelines for nutritionists and other biomedical scientists to plan and conduct studies and to analyze the complex data. Appropriate statistical analyses are expected to make an important contribution to solving major nutrition-associated problems in humans and animals (including obesity, diabetes, cardiovascular disease, cancer, ageing, and intrauterine growth retardation).

Keywords: Bioinformatics; Nutrition research; Statistical analysis; Systems biology

1. Introduction

Nutrition plays a vital role in health, genetic predisposition, and disease [1–3]. Effective prevention and treatment of metabolic disorders require adequate knowledge about the molecular mechanisms responsible for the actions of nutrients and other dietary components on cell metabolism and function [4–7]. Traditional research in molecular nutrition involves the analysis of expression of one or a very few genes at one time. While this approach has led to important findings on the discoveries of key regulatory pathways for nutrient utilization, it generally requires prior knowledge of genes of interest. There is increasing evidence that most genes do not function in isolation and that dietary nutrients interact to modulate expression of a set of genes and their biological functions [5]. Thus, nutritionists face a challenging task of defining cellular and molecular mechanisms that control the digestion, absorption and metabolism of dietary nutrients. With the recent completion of sequencing of the genomes of many species, including the human [8,9], mouse [10,11], rat [12] and yeast [13], we now have useful tools to identify complex interactions between genes and the diet as an environmental factor.

Statistical analysis, which is defined as the process of making scientific inferences from data that contain variability, has historically played an integral role in advancing nutritional sciences. This tool has gained an increasingly important role in the systems biology era to analyze large, complex data sets generated from genomics, proteomics and metabolomics studies [14–17]. Particularly, analyses of data from the reverse transcriptase-polymerase chain reaction (RT-PCR) as well as microarray, proteomics and other bioinformatics studies requires statistical models to account for various sources of variations [18,19]. Appropriate statistical methods can minimize systemic

☆ This work was supported, in part, by grants from National Institutes of Health (P20RR16481, 2P42 ES007380-09, P20RR020145-01, 1R21 HD049449, and CA57030), King Abdullah University of Science and Technology (KUS-CI-016-04), National Research Initiative Competitive Grants from the Animal Reproduction Program (2008-35203-19120) and Animal Growth & Nutrient Utilization Program (2008-35206-18764) of the USDA National Institute of Food and Agriculture, American Heart Association (#0755024Y), and Texas AgriLife Research (H-8200).
* Corresponding author: Tel.: +1 979 845 1817; fax: +1 979 845 6057.
E-mail address: g-wu@tamu.edu (G. Wu).

0955-2863/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jnutbio.2009.11.007

errors, optimize data analysis, and identify which genes are differentially expressed in the face of substantial biological and technical variations [18–20].

The major objective of this article is to provide guidelines for nutritionists and other biomedical scientists to adequately plan and conduct experiments and to analyze complex data particularly from microarray, RT-PCR, proteomics and other bioinformatics studies.

2. Definitions of general statistical terms

2.1. Hypothesis testing and errors

2.1.1. Null hypothesis and alternative hypothesis

Assume that study subjects in treatment and control groups are drawn randomly from their own populations that follow probability distributions characterized by defined parameters. For example, the parameters may be two normal distributions for body mass index (BMI) with different means and variances, or two Poisson distributions of the number of people having heart attacks. If the study aim is to determine whether dietary supplementation with a nutrient (e.g., L-arginine) will reduce BMI in obese subjects, the hypothesis is to test if the mean BMI of the population receiving the treatment differs from that of the population not receiving the treatment. The null hypothesis (H_0) is that the two populations have equal mean BMI, while the alternative hypothesis (H_a) is that they have unequal mean BMI.

2.1.2. Type I and type II errors

In hypothesis testing, two types of errors may be made when experimental data are analyzed statistically. Type I error (false positive) occurs when a conclusion is drawn in favor of a significant difference while there is no true difference between populations; namely, H_a is claimed to be true while H_0 is true. Type II error (false negative) occurs when the null hypothesis is not rejected while the alternative hypothesis is true; namely, H_0 is claimed to be true while H_a is true. Power of a test is the probability of claiming the alternative hypothesis is true when it is true. Type I error is usually controlled at a very low probability level (e.g., 5% or 1%), which is usually called the significance level of hypothesis testing. The power of a test is the complement probability of type II error and is usually expected to be 80% but may vary from study to study.

2.2. Sample size

A large sample size yields a powerful test to detect a difference between populations. Therefore, sample size calculation is needed to ensure desirable power in hypothesis testing. For this purpose, a difference in the parameters of distributions between study populations needs to be specified, such as a difference of 10 mmHg for mean systolic blood pressure or 5 kg/cm^2 for mean BMI. Based on a desired parameter of biological or clinical significance, a sample size can be calculated on the basis of probability distribution of the measured values with a given significance level (e.g., 5%) and the power of test (e.g., 80%).

2.3. Data collection

Data collection depends on study design. A cross-sectional study may require a survey to collect data with response variables that may reflect the study outcome. A case-control study first identifies subjects (i.e., cases and controls, where controls may or may not match the cases with clinical variables) and then examines the exposure of individual subjects to the risk factors of interest. A randomized controlled trial recruits subjects first and then randomly assigns them to a treatment or control group. A longitudinal study assigns subjects randomly to control and treatment groups, monitors the subjects over time, and collects multiple observations. The data collected from case-control studies may be subjected to large recollection measurement errors and large bias. This is because nutrient intakes of subjects are usually not based on food consumption records but rely on their memory, which can result in large measurement errors and a severe bias toward study aims. Additionally, the recollection of past food consumption by study subjects may be influenced by their knowledge of possible outcomes.

To achieve high accuracy in data collection and ensure high quality of findings, samples may be repeatedly collected from the same subject. These repeated samples may improve the study. However, two kinds of mistakes are often made by investigators. First, repeated samples are treated as independent samples and the correlation between them is ignored in data analysis. This approach may mistakenly yield false positives due to an inflated sample size. Second, repeated measurements from the same samples are averaged out and the averaged values are used for statistical analysis, resulting in false negatives due to the loss of power.

2.4. Statistical modeling and data analysis

Depending on study design and the type of response variables, data will be analyzed with use of different statistical models to reflect the data structure and potential correlation between observations. Categorical response variables are usually analyzed using contingency tables, logistic regressions, or generalized estimation equations (GEE) models. The contingency tables can also be used to test the homogeneity of distributions for categorical response or explanatory variables. In contrast, continuous response variables are analyzed using the *t* test, analysis of variance (ANOVA), correlation, and regression.

Statistical modeling is the data processing step to sort out information from a study. This can be achieved by building a quantitative relationship between the outcome or response variables and the explanatory or independent variables through a mathematical model or equation that characterizes the dependence of the former on the latter. In modeling response variables, their correlations should receive special attention, because the responses (e.g., body weights of the same subject at different time points) are highly correlated and thus the correlation structure should be incorporated into data analysis. Therefore, longitudinal studies should be carefully analyzed for the following reasons: First, subjects are monitored with multiple observations at different time points. Second, the correlation structure between observations affects estimation accuracy and subsequent inference.

Interpretation of statistical analysis results is crucial for making inference and valid conclusion. Particularly, *P* values have played an irreplaceable role in biomedical research. Recent advances in high-throughput technologies have made it possible to simultaneously analyze thousands of genes and identify those that are potentially responsible for the observed differences in the outcome or phenotype of study subjects. This raises a multiple comparison issue in hypothesis testing of multiple genes and thus leads to different criteria for the claim of statistical significance through correction for multiple comparison, such as family-wise error rate (FWER) or false discovery rate (FDR) [21].

2.5. Special concerns over genetic or “omic” data

Genetic or “omic” data are those obtained from genetic, genomic or proteomic studies, such as gene expression data, DNA genetic polymorphism data or protein profiles. Current technologies for conducting genetic and “omic” studies provide measurements of the

intensities of genes or proteins, which represent levels of gene or protein expression. Because of the relatively large noise in microarray data, significant findings usually need to be confirmed experimentally through quantitative RT-PCR (qRT-PCR), which itself also is subjected to variability of other sources. Recent research has led to the study of copy numbers that is more intrinsic to molecular activities and less relies on technologies.

Through high-throughput technologies, a large amount of molecules with known actions (e.g., proteins at the binding site for DNA transcription) can be monitored so that their functions can be studied through measurement of their intensities that may be associated with phenotypes of interest. This approach may provide clues for further study design for causal relationship between risk factors and outcomes. Exceedingly large amounts of data can be obtained from the genetic or "omic" studies, such as thousands or tens of thousands of genes or proteins, as well as millions of single nucleotide polymorphisms (SNP) in genome-wide association studies, which are distinct from the traditional biological or clinical studies. These huge databases provide biologists with opportunities to conduct fine-tune studies with great details of genes or proteins, but also present challenges to quantitative scientists (e.g., statisticians and bioinformaticians) to correctly decipher the data and make meaningful conclusions. In terms of statistical modeling and analysis, the "omic" data are characterized as high dimensional (thousands) or highly correlated (genes cooperate to fulfil biological functions). However, such studies often involve a small sample size, giving rise to the so-called small n — large p problem.

3. Sample size and power calculation

3.1. General considerations

Sample size determination is a major issue in planning quantitative research. Accurate estimation of sample size will not only ensure the planned study to achieve a sufficient power to detect significant differences, but also save time and money by recruiting no more subjects than needed. Many factors affect sample size calculation, including Type I error rate, the power of test and the expected significance of detection. Sample size calculation for studies not involving microarray or other high-throughput technologies can be found in many biostatistics books [22,23]. In this section, we summarize methods for sample size determination for microarray and other studies involving high-throughput technologies.

In microarray studies, experiments involving 20 000 to 30 000 genes or features are conducted at the same time and variability in their expression differs. Thus, a traditional power analysis would result in 20 000 to 30 000 different sample sizes. Genes that have similar, but not the same expression in 2 groups, would require very large sample sizes to detect a minor difference, while genes with dramatic differences can be detected with very small sample sizes. Three per group is usually the minimum sample size for a publication and is reasonable for cell cultures, inbred animals and littermates with small between-subject variability. Human or other studies, where variability is larger, can benefit from at least five to ten subjects per group. Obviously, a larger sample size is always better, but the question is how large is sufficient.

Because the aims of microarray studies include identification of differentially expressed genes between cases and controls, as well as profiling of subjects based on gene expression levels, the main objective of microarray studies is to discriminate cases from controls. Two major classes of statistical models have been studied so far. One class of models focuses on gene expression, including the ANOVA method [24] and the t test-like method, such as significance analysis of microarrays [25]. The other class of models concerns the subject label (case versus control, or receiving study treatment versus standard treatment or control), including logistic regression model, or classification models, such as the Bayesian hierarchical model [26]. Accurate estimation of sample size ensures enough power to: (a) identify differentially expressed genes (DEGs) in the first class of models and (b) allow the selection of genes that will be able to discriminate cases from controls in the second class of models.

Many methods have been proposed to determine sample size when pilot data are available [e.g., 27–35]. In general, two broad approaches (model-based and the direct control of error rate) have been employed. The model-based approach relies rigorously on the models for microarray data analysis, such as the ANOVA model, which may provide accurate estimation of sample size if the proposed model fits the data properly. This approach is similar to the classical approach to sample size determination based on conventional statistical parametric models with specific assumption on distributions of response variables and experimental risk factors, but differs in the special characteristics of the high-throughput data that require error rate adjustment through either FWER or FDR for multiple comparisons. However, it is often difficult to identify a single model that fits experimental data well. In such cases, the direct control of error rate approach is more appropriate and yields accurate estimation. Several authors also provided free software to calculate sample size [32,35]. In addition, free software and public databases of microarray data are also available to support sample size determination for the investigators who have no pilot data for sample size calculation [34]. Such an approach does not require a specific class of model, but rather focuses on the distribution of the P values of single gene expression analysis. This tool makes sample size determination user-friendly and easily accessible.

3.2. The ANOVA model-based approach

This approach rigorously depends on a statistical model for data analysis (i.e., the ANOVA model) where individual gene expression or its transformation (usually a log transformation to ensure the normality of intensity data) is assumed to be normally distributed and analyzed using the ANOVA model. Popular models are one-way or two-way ANOVA, incorporating experimental design factors. See Kerr and Churchill [36] for the global ANOVA model and Wolfinger et al. [37] for a generalization with random effects. Among sample size determination methods, Lee and Whitmore [27] described detailed modeling and calculation based on the classical approach to sample size determination for linear models with adjustments for multiple comparisons through controlling type I error rate, FWER and FDR. They then considered detailed sample size for several standard microarray study designs, including matched-pair data, completely randomized design, and isolated effect design. A sample size table was also provided for each design. This method can be assisted with a software package sizepower in R (see Ref. [35] for details). Similarly, Dobbin and Simon [33] derived sample size calculation based on a model similar to ANOVA taking into consideration more technical details of microarray technology, such as single-label, due-label or dye-swamp microarrays.

Keep in mind that microarray experiments typically have very small sample sizes due to their relatively high costs. Thus, it is essential to keep variability as small as possible, particularly in the one-way or two-way ANOVA analysis. As an example, gene expression is quantified in the liver and heart of mice receiving dietary supplementation with L-arginine or L-alanine (isonitrogenous control). There are two factors (tissue and amino acid), each at two levels. A 2×2 ANOVA with five chips per group would use 20 mice, randomly assigned to each of the four treatment combinations. An alternative design would use 10 mice, randomly assigned to L-arginine or control; RNA from liver and heart of each mouse would be

extracted for mRNA measurements. The second design, in which liver and heart are obtained from the same mouse, will have far less variability than the first design and, therefore, will have much more power to detect true differences in gene expression.

3.3. The direct control of error rate approach

When there is no a priori knowledge of an appropriate model for statistical analysis of experimental data, the required sample size may be calculated based on the direct control of error rate. Muller et al. [28] provided detailed theoretical study on sample size determination of microarray studies based on FDR and false-negative rate, whereas Tsai et al. [29] calculated sample size based on the expected number of false positives using individual comparison-wise error rate. Based on the FDR-control, Jung [31] derived the sample size for a specific number of true rejections while controlling the FDR at a prespecified level and provided an exact formula based on a t-distribution assumption. Alternatively, Pounds and Cheng [32] proposed an anticipated FDR method for sample size determination, controlling the FDR, positive FDR, and conditional FDR. Their method can be easily implemented with R codes available on the web at http://www.stjuderesearch.org/depts/biostats/documents/fdr-library.R. Finally, sample size can be estimated using *t* statistic, FDR, large fold change and other methods [30].

Of note, The PowerAtlas [34] is a power analysis and sample size calculation software package that provides not only sample size calculation, but also needed pilot study data based on publicly available data from previous microarray studies. This sample size calculation method is based on studies of the distribution of *P* values from single gene expression analysis in microarray studies controlling for expected discovery rate. It allows the use of either investigator's own data or publicly available microarray databases already incorporated into the software for sample size calculation. The free software is available on the web site http://www.poweratlas.org/.

Sample size calculation is a critical step in designing microarray studies. Accurate estimation of sample size will not only allow optimal design but also ensure a desirable power to detect significant findings. Because microarray studies present challenges in many different aspects, various methods for sample size calculation make it difficult for investigators to choose an appropriate one. We suggest selection of a method based on study design. If the statistical model for data analysis is known from a pilot study, a more specific method for sample size calculation can be chosen that would best fit data analysis. Otherwise, the FDR-based approach may be used. Additionally, it is possible to take the advantage of the PowerAtlas software to borrow the strength from publicly available microarray study databases.

After microarray data are collected and analyzed, investigators may find it useful to conduct a power analysis if significant findings are not detected for target genes yet there is tendency toward significance. In such a situation, the power of a test can be calculated on the basis of the data and statistical model, as discussed above. The other approach is to determine if enough subjects have been recruited for the study [38]. This approach assumes that the subjects are independently recruited in a serial procedure. The classification model will be updated each time when a subject is recruited and will also be tested on a newly recruited subject. It provides a stopping rule with a pre-specified probability to ensure that, at stopping, the probability of misclassifying the next subject will be less than a pre-determined threshold. A bootstrap approach may be taken for the collected samples so that the needed sample size can be calculated for a stopping time based on the target threshold. Random sampling can also be employed to explore how many more samples will be needed to achieve the pre-determined misclassification level if stopping is not achieved based on experimental data.

4. Statistical analysis of microarray data

4.1. Platform selection

The first decision in a microarray experiment is to pick a platform. The two basic options are one color or two. One color means that only one sample of RNA goes on a chip. With two colors, two RNA samples go on each chip. While more economical, two color chips are often custom-made and require extensive effort in establishing quality control and statistical analysis methods. One color chips are professionally produced and the reliability of chips has already been established. The most common one color (or oligonucleotide) arrays are from Affymetrix or Illumina [39] and will be the focus of our discussion. Most of what is described here applies to two color and oligonucleotide arrays. The flow chart of statistical analysis of microarray data is illustrated in Fig. 1.

4.2. The use of replicates and pooling in microarray analysis

The use of replicates in microarray experiments is under constant debate. Technical replicates are using the same mRNA sample on multiple chips. They are useful for establishing the reliability of the platform, but they cannot be used to increase the sample size for statistical calculations. Biological replicates are where different mRNA samples go on each chip, and thus, they contribute to the overall statistical sample size for the experiment. In general, for professionally produced microarray chips, technical replicates are not useful and the reliability of the platform has already been well established.

A related issue is the use of pooling, which means putting more than one mRNA sample on each microarray chip. This reduces individual variability, and thus increases power, but at the price of not being able to use individual covariates in the statistical model. In the 2×2 ANOVA example discussed previously, it is possible that the weight of the mouse might impact gene expression. If pooling were used, weight could not be used as a covariate in the model. When pooling is adopted, it is essential to extract RNA from every sample and then combine equal amounts of RNA from each sample to go on each chip [40].

4.3. Normalization of gene microarray chips

Once the RNA has been appropriately extracted, hybridized to chips and scanned, it is time to normalize the chips so that data between chips can be "fairly" compared. Although plenty of options are available for chip normalization, the most common are MAS 5 from Affymetrix and gcRMA from Bioconductor (www.bioconductor.org). MAS 5 is the easiest to use and will be the focus of this discussion. gcRMA will typically yield similar results with the exception of the situation where gene expression is very low. In such a case, gcRMA is likely better than MAS 5.

Using MAS 5 results in an output file for each chip that contains the probe set ID, the probe set expression, the presence/absence call, and the presence/absence *P* value. The presence/absence *P* value is used to declare each probe set either "P" for present ($P<.05$), "A" for absent ($P>.065$), or "M" for marginal ($0.05 < P<.065$). The *P* value cutoffs for each label are adjustable. Technically, the assumptions of independence in the statistical test are not satisfied, but the P/A call is still useful as discussed below. MAS 5 can also generate an output called a "change *P*-value" for each probe set on a pair of chips. Change *P* values are statistically wrong and misleading and should never be used in practice.

4.4. Data reduction in microarray analysis

The next step in the statistical analysis is data reduction. If there are probe sets that are not of interest, statistical calculations should not be done on those probe sets. On most, if not all, Affymetrix chips, the first

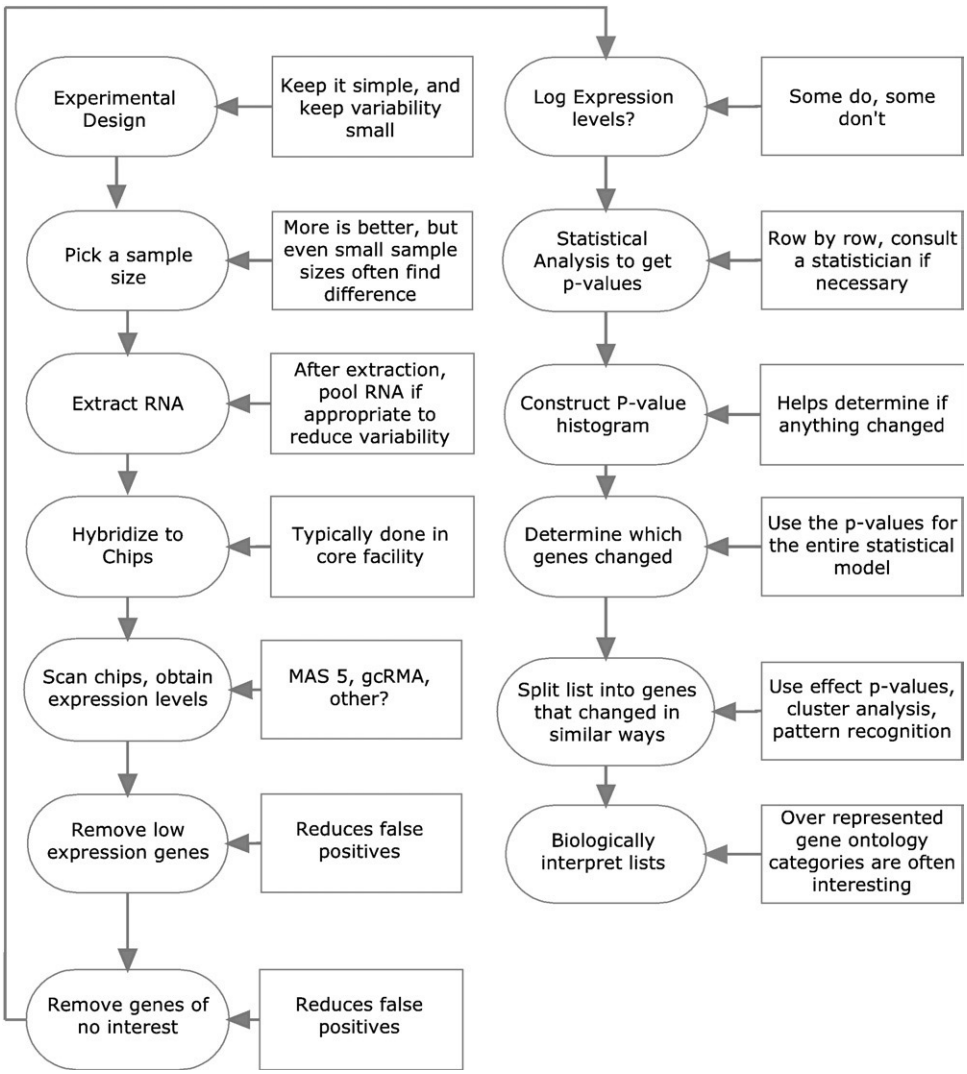

Fig. 1. A flow chart for microarray experiment and data analysis. A microarray experiment involves platform selection, sample size calculation, adequate design, data collection and processing, and normalization of gene chips. Statistical significance in levels of differentially expressed genes among treatment groups is commonly determined by a combination of *P*-value and the false discovery rate. Results of microarray studies are normally verified by quantitative real-time RT-PCR analysis.

approximately fifty probe sets are quality control probe sets used by the MAS 5 software. Typically, there is no need to do statistics on these probe sets. Similarly, many researchers are not interested in Expressed Sequence Tags which are genes that have not been annotated. These should also be removed from the data set if they are not of interest. The final group of probe sets that are typically removed are probe sets that are labeled as absent (A) on all the chips in the experiment. If the P/A call determines that the probe set is not expressed on any chip in the experiment, there is no reason to do statistical analysis on that probe set. If fairly large numbers of chips are involved in the experiment (e.g., a total of 20 or more), the condition of all absent calls on all chips could be relaxed to allow a few marginal or even "present" calls and would still result in being removed prior to statistical analysis. At this stage, researchers should also identify any subsets of the probe sets that are of particular interest (e.g., a particular pathway or annotation feature). These subgroups can be statistically analyzed separately and together with the rest of the probe sets.

4.5. Log transformation of microarray data

The final step prior to statistical analysis of the microarray data is to decide whether or not to take the log transformation of the data. For most microarray data sets, the probe sets with larger expression levels benefit from a log transformation, but the smaller expression levels should not be logged. Most researchers choose to log their data, but many do not. Typically for one color microarrays, there is not much difference in gene lists with or without logging the data.

4.6. Methods of statistical analysis of microarray experiments

The statistical analysis of microarray data is typically done row by row using the analysis appropriate for the experimental design. The most common designs are two sample *t* tests, one- and two-way ANOVA. Almost always, it is assumed that experimental errors are normally distributed. Obviously, this assumption is not true for all probe sets, but the small sample sizes for most microarray data sets make the normality assumption a good choice. The most common error at this point is likely failure to treat dependencies between the chips properly. If RNA is taken from the same subject more than once, a statistical model for repeated measures is needed.

The end result of the statistical analysis will be one or more *P* values for each probe set. The overall *P* value for each row tests whether there are any statistical differences between the rows. A histogram of these *P* values provides useful information. If the

histogram appears to be uniformly distributed (like a rectangle), there may be little if any differences between the treatment groups. On the other hand, a histogram with a large peak for low *P* values indicates that large differences exist between the treatment groups. Histograms with a low or moderate peak for small *P* values indicate that more chips would likely result in smaller *P* values for probe sets that are actually differentially expressed.

The next decision is how to determine the list of probe sets that have changed. Traditionally, a *P* value less than .05 rejects the null hypothesis of no change. In a microarray experiment involving 10,000 tests, using a *P* value cutoff of .05 could mean as many as 0.05×10 000=500 false positives. The FDR of Benjamini and Hochberg [21] chooses the cutoff by a user-specified expected proportion of false positives. Ten or twenty percent are common choices. For experimental conditions which cause differential expression but no large changes, the FDR method may not find any genes that change. As an alternative to using *P*=.05 or the FDR method, many researchers simply use *P*=.01 as the cutoff.

Once the overall *P* values are used to identify DEGs, many researchers attempt to use cluster analysis to determine genes that are responding similarly to the experimental conditions. To avoid excess noise in the gene clusters, be sure to cluster only genes that are determined by statistical methods to be differentially expressed. Many different types of cluster analysis are possible, and they often yield results that are hard to interpret. Statistical pattern matching (e.g., Liu et al. [41]) is an alternative that can be used to divide that list into sublists of genes that change similarly. For example, if two sample t-tests are used to generate the overall *P* values, the list should be sorted into up-regulated and down-regulated genes. The biological interpretation of the resulting list(s) is made by first annotating the gene lists using the manufacturer's Web site. There may be obvious biological conclusions that can be drawn at this point. A more statistical approach is to provide the list of all genes (e.g., the entire chip) to a statistical software package that determines gene ontology categories that are overrepresented on the smaller list compared to the larger list.

5. Statistical analysis of qRT-PCR data

5.1. General considerations

Due to large variability in gene intensity data inherent in the microarray technology, they are subjected to wild noise. Thus, significant findings should be confirmed by a more reliable method. Potential sources of the wild noise in the microarray analysis of gene expression include fluorescent scanning, uneven spray of reagents within arrays, control of environmental factors, and varying experimental conditions for different arrays. These factors lead to large variability and possibly contribute to artifacts. Such problems may not be resolved by the within- and between-array normalization in data preprocessing. The huge number of genes or probes in microarray studies, usually around tens of thousands, may also result in the identification of a large number of false positives. To verify the results of microarray studies, qRT-PCR experiments are often carried out on the DEGs identified by the microarray analysis.

5.2. Threshold cycle in qRT-PCR analysis

qRT-PCR quantifies the amplification of genes and records the real time (a threshold cycle of each gene to achieve a pre-set intensity level). This threshold will be used to calculate the mRNA levels of genes in a biological sample and to compare the values between cases (treatments) and controls in terms of fold or percent change. For this purpose, a reference gene (endogenous) is usually pre-specified to be amplified together with selected genes. To confirm microarray findings, tissues selected from cases and controls will be used for qRT-PCR analysis. Usually, the number of subjects in treatment and control groups is smaller than the corresponding microarray study. Because qRT-PCR experiments generate threshold cycle values for each of the selected genes, data analysis need to be conducted with a proper statistical model. Although mathematical models for PCR experiments have been proposed in the literature, statistical modeling has not received much attention [18].

5.3. Mathematical models for RT-PCR analysis

DNA sequences are amplified in RT-PCR through DNA polymerase. During the exponential amplification phase of RT-PCR, a copy of target gene doubles in one cycle, and then quadruples in the next cycle. Therefore, the amplification is in the power of 2 (exponential amplification), and can be described by an equation $Y_n=Y_0\ 2^n$ with Y_0 being the initial expression level of a target gene, and Y_n being the expression level after *n* cycles. Because the amplification is subjected to variation in experimental conditions and may not be 100% efficient and the amplification process may not end with an exact number of cycles, the above equation can be written generally as $Y_t=Y_0\ (1+e)^t$, where *t* is the duration of the amplification process in continuous number of cycles and *e* is the amplification efficiency, which may depend on many experimental conditions and sequence properties. The amplification efficiency of RT-PCR assays usually ranges between 0 (completely inefficient) and 1 (fully efficient). Thus, it is important to have an endogenous gene to serve as an internal reference to ensure the validity of RT-PCR results. The above equation applies generally to both target genes and the reference gene in cases and controls. Also, the target genes may have different amplification efficiency from the reference gene. Because the samples from both cases and controls are processed simultaneously in one RT-PCR experiment along with the reference gene, the efficiency for samples from cases and controls can be assumed identical [20].

5.4. Statistical models for RT-PCR analysis

So far, only two statistical methods have been proposed: (1) the GEE model — a generalization of the ANOVA model incorporating the correlation between samples within subject [20], and (2) an ANOVA/analysis of covariance (ANCOVA) model that treats within-subject samples using a random effect model [42]. These two models are based on the same mathematical principle as stated above and recognize the within-subject correlation in the modeling. In addition, both papers provide readers with user-friendly SAS program codes [20,42]. However, they employ two different methods to perform data analysis. The GEE model was proposed originally to deal with response variables, either continuous or categorical, to conduct analysis in longitudinal studies with multiple observations from each subject. This model is suitable for qRT-PCR data with repeated samples from each subject and is used by biologists to emphasize the importance of accurate measurements. The ANOVA/ANCOVA model with a random effect is also a good approach to this special type of qRT-PCR data.

However, the SAS program for the ANOVA or ANCOVA model has two major flaws. First, the SAS program does not recognize the correlation between the target gene and the reference gene from the same subject. Because estimating the fold change of each gene between the initial quantity of the cDNA and the final value at the termination of PCR amplification is the goal of the qRT-PCR analysis, the initial quantities of genes from the same sample are highly correlated. Thus, statistical methods that fail to address this correlation will surely lose its power in the detection of significance. Second, the SAS program assumes constant variance for target genes and the reference gene in a biological sample and, therefore, does not

recognize heteroscendasticity, i.e. different variance of expression for different genes. Because expression of genes is known to vary greatly, their variance is not likely to be similar [43]. A statistical model that fails to address this heteroscendasticity issue will surely result in incorrect inference. Therefore, current methods using the ANOVA/ANCOVA model needs to be revised to prevent misleading results. In the GEE model, these two issues can be resolved simply by taking the difference in expression between each target gene and the reference gene [20]. This is the recommended approach in biostatistics to deal with paired observations and resolve the varying variance issue. In addition, the GEE model will yield correlation coefficients between within-subject repeated samples, and the corresponding SAS program is easy to implement [20].

6. Statistical analysis of proteomics data

6.1. General considerations

In many respects, proteomics lags far behind microarray experiments due to several challenges in the collection of mass spectroscopy data. Among these challenges are that (1) identification of peptides is a central feature/difficulty of the analysis (as opposed to microarrays where each gene occupies a "dot" on the array); (2) the quantitation information for peptides requires careful manipulation of possibly overlapping peptide peaks; and (3) missing data is much more of a problem (e.g. in many cases peptides are not identified in an experiment even when they are present) than in microarray experiments. See the recent review by Bantscheff et al. [44] on the experimental side of proteomics studies, including differences in equipment and mechanisms for preparing biological samples.

6.2. Mass spectrometry data

Central to all mass spectroscopy experiments is the preparation of a sample containing a large amount of peptides. This sample is then placed in a tandem mass spectrometry (MS)/MS machine. Ions from the sample then pass through the machine until they reach a detector. At prespecified regular intervals, this first layer of the MS calculates a mass spectrum (measurements of ion abundance on a mass to charge, or m/z, scale) for the ions arriving at that time. This mass spectrum will contain peaks corresponding to the m/z values of abundant peptides in the original sample. After each scan is performed by the first MS layer, the machine selects an m/z region and allows the ions in this region to pass through to the second layer of the tandem MS/MS equipment. These ions are then further broken down and move through the machine to a second detector, which computes a mass spectrum as well. Typically, the region of m/z values chosen for "pass through" is one of the largest peaks in the MS scan, but most equipment will also allow the experimenter to program specific rules for this operation. This selection process is important because always choosing high abundance peaks will fail to identify low abundance, but important, peptides in the original sample [45]. Thus, the main data structure of a tandem MS/MS run is alternating scans, beginning with ion concentrations from the original sample in one scan (MS) and then the ion concentrations from the tandem MS/MS scan. The m/z value is selected to pass ions through to the tandem MS/MS layer. This second layer (the tandem MS/MS scans) is used to identify peptides, while the first layer (the MS scans) is used for quantitation.

6.3. Identification of peptides

The data from the second tandem MS/MS scan is chosen for identification of peptides. Peptide identification is based on the principle that the breakdown mechanism for each possible peptide is known (e.g., trypsin predominantly cleaves peptide chains at the carboxyl side of lysine and arginine, except when either is followed by proline). Therefore, the original mass and the masses of the broken down component result in a "signature" allowing for reconstructing and identifying the original peptide in the first layer (MS) scan. Doing this from scratch without a database of peptide, simply from known masses of amino acids, is called de novo sequencing [46]. Available software for this task includes PEAKS [47], PepNovo [48], AUDENS [49] and NovoHMM [50]. All of these algorithms rely on some heuristic search techniques to search through a set of possible amino acid sequences to reconstruct the original peptide.

More commonly, a peptide is identified through the search of a database of known peptides [51]. Each peptide in the database, combined with the chemical method for protein hydrolysis, results in a "signature" set of peaks in the tandem MS/MS scan. The observed spectrum is compared to each possible peptide in the database and a score is assigned based on the agreement between the observed spectrum and the expected spectrum. If a peptide in the database receives a sufficiently high score, the observed peak is considered to be identified as that peptide from the database. Most software for this purpose reports not only the tentatively identified peptide but a numerical "confidence score" reflecting the uncertainty in that identification. The earliest method, and still a standard, is SEQUEST [52]. SEQUEST works by taking each peptide in the database and determining its expected spectrum (using knowledge of B, A and Y ions). Then, the correlation between the observed spectrum and the expected spectrum is observed. In addition, the correlations between the expected spectrum and shifted versions of the observed spectrum are taken (e.g., all the peaks in the observed spectrum are moved to the left or right in the graph and the correlations recalculated). If the correlation is exceptionally high for the observed spectrum, but moderately low for the shifted spectrum, this indicates that the observed spectrum is aligned well with the expected spectrum because high correlations are expected when the peaks are in the same locations. This calculation is feasible for the number of peptides in a typical database (fortunately processing power has been increasing while the size of the databases has also been increasing). If SEQUEST finds one particular peptide in the database aligns well, it can be concluded that the observed spectrum came from that peptide.

Unfortunately, as databases grow large, it is possible that multiple peptides produce reasonably high correlations. Note that noise in spectra often peaks in the expected spectrum and may not be represented in the observed spectrum, therefore reducing the accuracy of SEQUEST. If so, one may be interested in relative fit, not absolute fit. In other words, a peptide is identified only if the match to the model spectrum is significantly better than others in the database. The program X!Tandem [53] attempts to do this quickly. X! Tandem works by only summing the peaks in the tandem MS/MS spectra which match the model spectra (thus no shifting, which is computationally costly). A scaled version of this quantity is called the Hyperscore. The hyperscores are then calculated for each peptide in the database. Histograms of these hyperscores follow a distribution that would be expected from random matching. This is reasonable considering most peptides in the database should NOT be a match for the observed spectrum. X! Tandem then assigns an "*E*-value" that indicates how good the observed match is compared to what would be expected by chance matching. If the *E*-value is sufficiently low (namely, we would almost never expect this good a match by chance), then the protein is declared to be identified. This calculation is quite similar to a statistical *P*-value. A similar, competing idea to X! Tandem is MASCOT [51]. See Brosch et al. [54] for a comparison of the two. It is possible to combine results from several of these algorithms for even better identification. The MASCOT method has been successfully used in nutritional proteomics research [55,56].

Peptide Prophet provides a method for converting scores (from any identification method) into a probabilistic identification (e.g.,

Peptide Prophet reports a probability that the peptide is correctly identified) [57]. Peptide Prophet begins by producing a discriminant score for each spectrum in the sample. These discriminant scores fortunately fall into two groups — a correctly identified group and an incorrectly identified group. These groups slightly overlap, so Peptide Prophet uses the expectation-maximization (EM) algorithm to fit a mixture distribution and then assigns a probability to each identification [58]. If the discriminant score is firmly within the "identified" group and nonidentified group, the probability values are near 1 and 0, respectively. For discriminant scores in the middle where the overlap occurs, probabilities of identification are anywhere between 0 and 1. Note the user may then choose which peptides to pursue. Only selecting peptides with high identification probabilities results in being surer of the results, but some correctly identified peptides may be missed. Utilizing lower probability scores may be more useful in an exploratory setting to identify peptides of interest. Note that posttranslational modifications of peptides, such as glycosylation or phosphorylation [59], can cause difficulties with these methods. The reason is that these post-translational modifications can alter the chemical structure of the protein and thus change the pattern of its degradation products. This can lead to misidentification of peptides because the peptide is not hydrolyzed according to the "expected signature."

6.4. Isotope-labeled quantification

Quantification experiments in proteomics attempt to determine differences between protein expression across treatment groups (e.g., Alzheimer's versus normal patients). These experiments typically take the form of "shotgun proteomics," where a sample containing a large number of peptides is analyzed with the hope of identifying several differentially expressed proteins for further study. In an isotope labeled experiment, two groups of samples are treated with different isotopes, one heavy and one light. There are a variety of abbreviated heavy methods (ICAT, SILAC, iTRAQ, see Bantscheff et al. [44]). The samples are then mixed together and treated as a single sample for the remainder of the mass spectroscopy run. Statistically, this is important because any variation that can be attributed to handling after the samples are mixed occur equally to all treatment groups. This is in contrast to "label-free" methods, which combine results of multiple mass spectroscopy runs and thus may have differences attributed to experimental variation not present in the labeled experiment.

Quantification is performed through the first set of MS runs. For identified peptides, we have not only an m/z region (the region allowed to pass through for the identification to take place), but also the time when this occurred. The "first MS" scans for those times in a neighborhood around the time of the "identifying" scan will then be collected, as indicated in Fig. 2. The x-axis of the figure contains the scan number, while the y-axis provides the observed intensities for the light isotope (red) and the heavy isotope (blue). The dark red part of the bars indicates overlapping red and blue portions. To acquire these observed intensities, the MS scan where the peptide was identified should be examined. The m/z value that was passed through to the second, tandem MS scan shows us where to look in the MS scans for that peptide. Moreover, we also have to look for m/z values representing the "shift" in mass due to isotope labeling. This is expected to succeed because each isotope should have the same chemical properties and, therefore, the labeling results in different masses between the same peptide across the treatment groups. Furthermore, if the peptide is known, the expected difference in mass between the light and heavy isotope can be determined. Thus, in addition to assessing the expression for the identified peptide, we also look in the mass area where the "companion" peptide from the alternative treatment group is expected to be located. Similarly, for the original peptide, we also acquire a peak and compute the area within that peak to determine the expression of the companion peptide. The ratio of these expressions is then used to quantify the relative expression of the peptide across the treatment groups. Fundamental to this process is the fact that the ion concentrations are measured with noise, and peptides which have relatively low expression (relative to the background variation) are particularly difficult to quantify.

A straightforward but somewhat thorny issue is simply to determine how long before and after the identified peak to add together (Fig. 2). For example, we have to determine how far out to "color" the bars red or blue for inclusion into the calculation. At the center of a good peak, the expression is far above the baseline noise; however, at the extremes the peak begins to slide back into the baseline noise. If a very small window is chosen, fewer values are summed, yielding higher variation in the resulting expression estimates (namely, a smaller sample of points is used). On the other hand, if a very broad window is selected, pure noise is added up into the expression level, resulting in poor estimates of protein expression.

One common method for calculating expression ratios is XPRESS [60], which performs this task using a low pass Butterworth filter and simply taking a sum of the smoothed values. Another is ASAPRatio [61], which combines a Savitzky-Golay filter with a normal (Gaussian) fit of the peak to determine expression. Finally, in RelEx, a Savitzky-Golay filter is applied for 100 scans before and after the identified peak [62]. Then, a linear regression is performed (Fig. 3) on the expressions at each scan. Namely, a data point in the linear regression is the expression for the heavy isotope at that scan plotted against the expression for the light isotope. Ideally (if the expression has similar shapes), these points should form a line, whose slope is the expression ratio. In practice, the points are measured with errors, and thus, the points fall in a linear trend, but not exactly on a line. RelEx is convenient in that the quality of the regression provides a

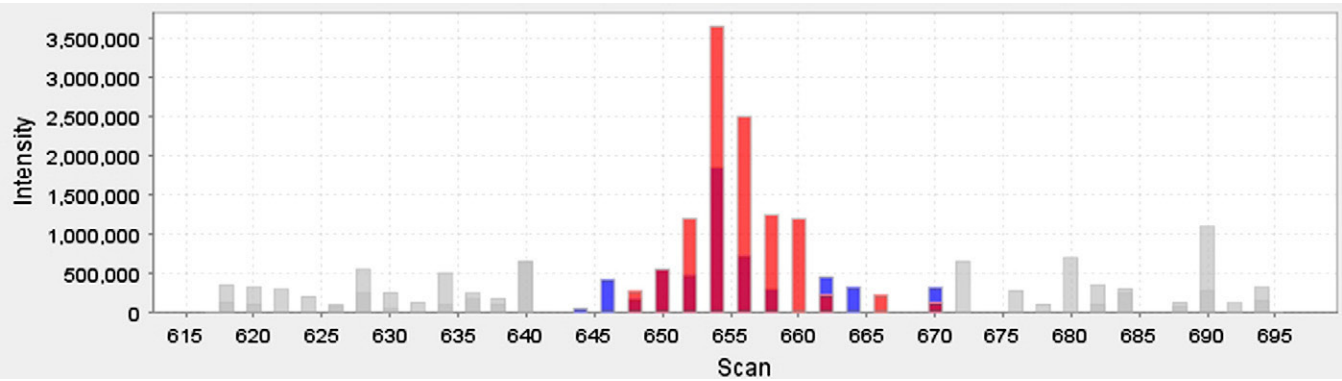

Fig. 2. Isotope-labeled quantification in proteomics analysis. The x-axis contains the scan number while the y-axis contains the observed intensities for the light isotope (red) and the heavy isotope (blue). Dark red areas indicate overlapping bars. The estimated protein expression ratio is simply the sum of the red bars divided by the sum of the blue bars.

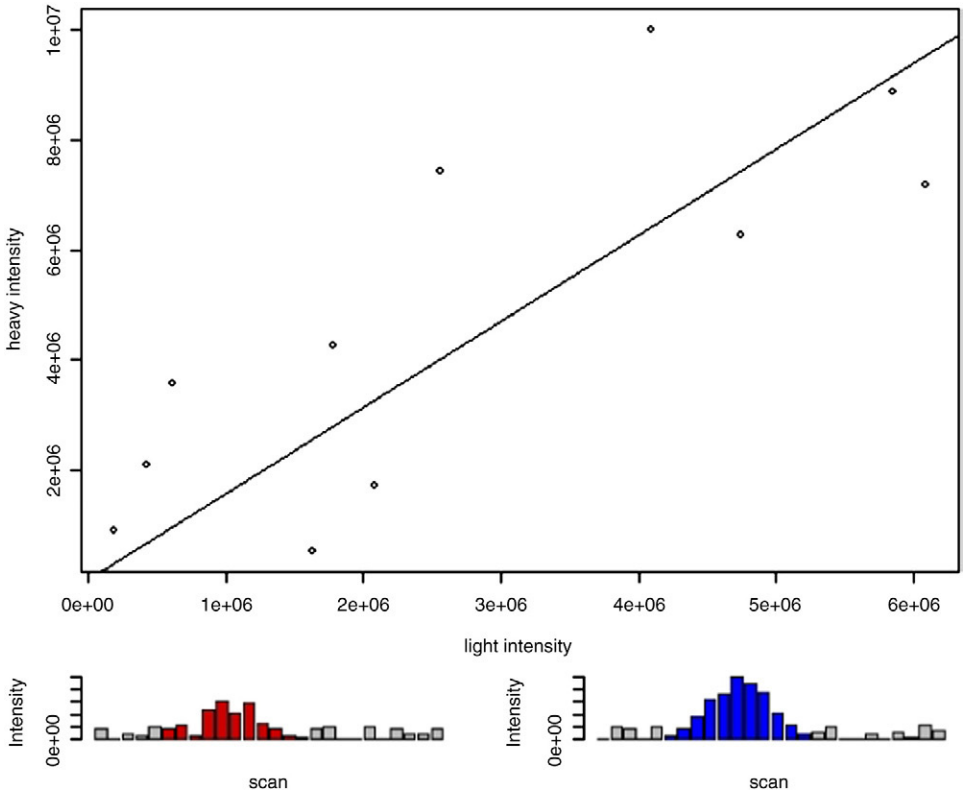

Fig. 3. Idealized version of RelEx. The bottom panes of the figure show the observed light (red) and heavy (blue) ion intensities. The intensities used are colored. The points in the main pane are the light and heavy intensities observed for each scan plotted against each other. The slope of the regression line (determined without an intercept) estimates the expression ratio while the correlation of the points (how well the peaks align) is a measure of confidence in the estimated expression ratio. The full implementation of RelEx adds various smoothers and other enhancements.

quick measure of the confidence in the expression ratio. If the regression diagnostics (e.g., R^2) are not satisfactory this may indicate a poorly identified peptide or other problem.

In these algorithms, expression ratios for peptides, not proteins, are quantified. If a researcher has confidence in a certain list of peptides from a particular protein, protein expression ratios may be obtained by combining the estimated expression ratios for each component peptide. Typically this is done on the log scale as estimates of the log expression ratio are more stable than estimates of the expression ratios themselves, particularly when one of the isotopes is minimally expressed in the denominator of the ratio. Averaging ratios on the log scale is equivalent to taking the geometric mean of the ratios themselves. As averages are normally distributed, a standard error for the protein ratio can be computed. Unfortunately, for various experimental reasons, a researcher also runs into the problem that a peptide is identified for one of the treatment groups but no corresponding peak may be found for the other group. This can occur for reasons other than non-expression in the other group (i.e., it cannot be simply assumed that the peptide was not expressed in the other group). Thus, missing peptides will limit the ability to perform quantification.

6.5. Label-free quantitation

Isotope labeling is a difficult process which is not possible in many areas of clinical investigation, particularly human proteomics. Thus, an active area of research is in “label-free” methods. In a labeled experiment, one mass spectroscopy run is completed with both the heavy and light isotopes together in the input. In a label-free experiment, multiple MS runs are performed. For example, in an experiment attempting to determine protein expression differences between Alzheimer's patients from normal controls, two samples would be prepared; one from the Alzheimer's patients and another from the controls, and each would have a separate MS run. Experimentally, this immediately creates the difficulty that any variation from run to run will now also be included in the estimate of the protein expression ratio, something controlled for in isotope labeled experiments. Among these difficulties is that the peaks are contained in different scans during the experiment. Specifically, the scan numbers containing a particular peptide in one MS run may not be sufficiently close to the scan numbers containing the peptide in another MS run.

Thus, the central statistical difficulty in the analysis of label-free methods is finding corresponding peaks in multiple MS runs. Specifically, one must be able to identify something along the lines of “the peak at such and such m/z value in scan x of the first MS run corresponds to the same peptide as the peak at such and such m/z value in scan Y of the second MS run.” One popular method for handling this problem is PEPPeR (Platform for Experimental Proteomic Pattern Recognition, [45]), a method which heavily utilizes normal mixture models [63] and bootstrapping [64] to identify common peaks which are then used for quantitation. Another possibility is ProtQuant [65], which utilizes the XCorr scores from the SEQUEST identification for quantitation. Label-free quantitation methods is one of the more active research areas, with multiple software packages being developed rapidly

6.6. Software platforms

A wide variety of software is available for the analysis of proteomic data. Some of this software is open source and freely downloadable,

while other software is propriety. We will focus on the open source software. A fundamental problem in proteomics is the reasonable large amount of data generated in a single experiment (many MS and tandem MS scans, each containing ion frequencies at many m/z values, result in large arrays of data). Thus, data storage formats are an initial problem. There are currently several standard formats available [66]. See Droit et al. [66] who discuss variants based on extensible markup language (XML) including mzXML, PepXML, and ProtXML. These data formats are used in a variety of software packages, and thus provide a convenient means for analyzing results with different software packages in addition to a standard format for posting datasets to the internet.

A popular pipeline for the analysis of proteomic database is the Trans-Proteomic Pipeline, freely available from the Seattle Proteome Center (http://tools.proteomecenter.org/software.php). This software includes a large number of tools for data handling (including the conversion of data from a variety of formats), identification of peptides and proteins (PeptideProphet) and quantification tools such as XPRESS and ASAPRatio. A key advantage of any complete open source pipeline is that it may be amended by users, and thus, new analysis methods may be easily compared to older methods [66,67].

One crucial issue where proteomics could be improved is a fuller understanding of the error processes underlying proteomics data. This is a particularly thorny problem due to the detailed interactions present in the chemistry. However, such an understanding is fundamental to proper statistical procedures. Many papers in the literature are light on descriptions of the statistical methods used (in terms of the exact numerical procedure and their error properties) in favor of experimental comparisons. This is compounded by several pieces of software only being available in closed source, propriety forms. Thus, it is essentially impossible to determine exactly how the data is being processed.

A "mundane" statistical procedure like linear regression is not used for analysis of proteomics data because it has performed well only in a few experimental datasets. However, it can be shown that, given certain assumptions (e.g., normally distributed independent errors), linear regression is the optimal way of estimating parameters. These assumptions can then be verified through the use of residual plots.

7. Statistical analysis of other bioinformatics data

7.1. Linkage studies

In addition to microarray and proteomics analysis, bioinformatics research include: (1) linkage studies through analysis of family pedigree, (2) genetic association studies through analysis of SNP genotype, (3) transcription regulatory region studies, (4) copy number variation studies and (5) genome-wide association studies [68]. Linkage studies have been developed extensively for research on genetically inheritable diseases. The requirement of long-term data collection through multiple generations makes it difficult to implement, especially with late onset diseases (e.g., Alzheimer disease). Thus, linkage studies are less attractive and less powerful compared with case-control studies of unrelated individuals. On the other hand, genetic association studies among independent cases and controls offer promise to enhance the detection of the association between specific genes or gene sets and diseases or phenotypes, where the presence of specific allele of genes may alter risks for diseases. Furthermore, human genetic variation studies aid in deciphering the genetics of complex diseases (e.g., hypertension and diabetes) through genome-wide association studies [68] and copy number variation studies [69,70]. Recent developments of copy number estimation methods and SNP genotyping techniques through high density SNP array technologies (e.g., Affymetrix 5.0 and 6.0 SNP arrays [71]) further expedite genome-wide association and copy number variation studies, therefore making it possible to conduct large scale research with millions of SNPs and tens of thousands of subjects [68].

Genetic association studies focus on the identification of: (1) genes or specific alleles of genes; and (2) gene–gene interactions or haplotypes (combination of specific alleles of different genes on the same copy of chromosome) that are highly associated with the disease or phenotype or their interaction with environmental risk factors. The case-control study is the most popular design for genetic association investigations, where a logistic regression model is often employed. While a single gene association model may be straightforward, gene-gene and gene-environment interactions may pose greater challenges and also lead to more significant findings. For example, haplotypes may present special gene-gene interactions but are not observable with double heterozygous SNPs because of an ambiguous phase that needs a special treatment with the statistical EM algorithm for missing data [72–74].

Genetic association studies are based on the genotyping of SNPs, where each SNP generally has two alleles, either the same (homozygous SNP of "AA" or "BB" type) or different (heterozygous SNP of "AB" type) that vary with individuals. Genetic association studies and the most recently developed genome-wide association studies are based on genotype data of a large number of SNPs annotated with high density SNP microarrays, such as Affymetrix 100K SNP arrays, 500K SNP arrays (or 5.0 arrays) and 6.0 SNP arrays. Hence, it is crucial to accurately annotate each SNP from the microarray probe intensity data. So far, a few genotype calling methods have been developed on the basis of Affymetrix high density SNP arrays. In contrast to the single array-based genotype calling methods including the whole genome sample assay and dynamic modeling, the machine learning-based methods (including the RLMM [75], BRLMM [71], CRLMM [76], MAMS [77] and a single array approach GEL [78]) have improved genotyping accuracy. Such techniques usually require multi-array training with the gold-standard annotation of SNP genotype of the HapMap samples. A novel approach based on a robust model of DNA copy numbers requires only one array, which further improves genotyping with high accuracy consistently for different prototypes of arrays and resolves the missing data problem in genotype data [79]. These methods provide a broad spectrum in SNP annotation for genetic association studies, particularly for genome-wide association studies.

7.2. Estimation of DNA copy numbers

DNA copy number variation has been reported to play a critical role in cancer research [80]. Nutritional studies have also been conducted to provide important clues to the development of cancers, including breast cancer, colon cancer, and prostate cancer [81,82]. Several methods have been proposed to estimate DNA copy numbers, including a high-resolution method [83], the CNAG method [84], the CARAT method [85] and the PICR method [79]. Of note, the PICR method provides accurate estimation of copy numbers at each SNP site and thus provides high resolution of copy number detection. This allows for both detection of copy number alteration and SNP genotype calling with high resolution at high accuracy consistently for cross-laboratory studies, even cross-array prototypes.

7.3. Genome-wide association studies

Large scale genome-wide association studies have become increasingly popular in recent years to obtain significant findings for many diseases or phenotypes at an unprecedented speed. Multiple diseases or phenotypes are often studied together. Furthermore, millions of SNPs are annotated to discover major associations

between diseases and genetic risk factors, including gene-gene and gene-environment interactions. The quality of data from such big projects is of great importance because it will lay a solid foundation for powerful detection of associations and accurate assessment of scientific findings. Because a major goal of nutritional studies is to define how the expression of genes is regulated through dietary intervention [1–3], identifying transcription regulator regions (e.g., binding sites) or transcription factors (e.g., promoters and enhancers) plays a major role in advancing the field. Although a variety of methods have been developed in this research area, many of them (e.g., hidden Markov models) have technical limitations and may not be suitable for data from different cell types or animal species. Most recently, a new word-counting method has been shown to be robust for studies of different eukaryotes [86]. Because this approach has few technical limitation, it is applicable to a wide arrange of biological studies [86].

8. Conclusion and perspectives

Statistical analysis is a necessary means to test hypotheses in nutritional and other biomedical research. In the post-genome era, there is increasing interest in quantifying the effects of nutrients on simultaneous expression of thousands of genes and proteins in cells or tissues [87–99]. This has offered new exciting opportunities for nutritionists but also presented technical challenges in experimental designs and valid statistical methods for data analysis. Sample size calculation is a critical step in designing microarray and other high throughput studies. Accurate estimation of sample size will not only allow optimal design and budgeting of the planned research but also ensure the desired power to detect significant findings. It is crucial that complex data obtained from microarray, RT-PCR, proteomics and other bioinformatics studies are subjected to appropriate statistical analysis. In microarray analysis, statistical significance in levels of DEGs among treatment groups is commonly determined by a combination of *P* value and FDR. Likewise, GEE models that properly reflect the structure of data is often employed in statistical models for assessing fold change of gene expression in qRT-PCR experiments. Moreover, a number of software platforms (e.g., MASCOT, Peptide Prophet, Sequence and X!Tandem) are available for identification of proteins in biological samples. Levels of protein expression can be determined using isotope-labeled and isotope-free quantification methods. Finally, bioinformatics tools have been developed for SNP genotyping, genetic linkage or genetic association studies in nutrition research. We anticipate that this article will provide useful guidelines for nutritionists and other biomedical scientists to plan and conduct sound studies at molecular, cellular, tissue and whole-body levels and to employ appropriate statistical methods for analysis of experimental data in the era of systems biology.

Acknowledgments

The authors would like to thank Bill Nelson for help with the Trans-Proteomics Pipeline and Frances Mustcher for office support.

References

[1] Wu G, Bazer FW, Davis TA, Kim SW, Li P, Rhoads JM, et al. Arginine metabolism and nutrition in growth, health and disease. Amino Acids 2009;37:153–68.
[2] Hennig B, Oesterling E, Toborek M. Environmental toxicity, nutrition, and gene interactions in the development of atherosclerosis. Nutr Metab Cardiovas Dis 2007;17:162–9.
[3] Baker DH. Advances in protein-amino acid nutrition of poultry. Amino Acids 2009;37:29–41.
[4] Dekaney CM, Wu G, Yin YL, Jaeger LA. Regulation of ornithine aminotransferase gene expression and activity by all-trans retinoic acid in Caco-2 intestinal epithelial cells. J Nutr Biochem 2008;19:674–81.
[5] Mutch DM, Wahli W, Williamson G. Nutrigenomics and nutrigenetics: the emerging faces of nutrition. FASEB J 2005;19:1602–16.
[6] Wang JJ, Wu G, Zhou HJ, Wang FL. Emerging technologies for amino acid nutrition research in the post-genome era. Amino Acids 2009;37:177–86.
[7] Wu G. Amino acids: metabolism, functions, and nutrition. Amino Acids 2009;37:1–17.
[8] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.
[9] McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, et al. A physical map of the human genome. Nature 2001;409:934–41.
[10] Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 2002;420:563–73.
[11] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–62.
[12] Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004;428:493–521.
[13] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science 1996;274:546–63.
[14] He QH, Kong XF, Wu G, Ren PP, Tang HR, Hao FH, et al. Metabolomic analysis of the response of growing pigs to dietary L-arginine supplementation. Amino Acids 2009;37:199–208.
[15] Desiere F. Towards a systems biology understanding of human health: interplay between genotype, environment and nutrition. Biotechnol Annu Rev 2004;10:51–84.
[16] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrometry Rev 2007;26:51–78.
[17] Karlin S. Statistical signals in bioinformatics. Proc Natl Acad Sci U S A 2005;102:13355–62.
[18] Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA microarray experiments: Biological and technological aspects. Biometrics 2002;58:701–17.
[19] Fu WJ, Haynes TE, Kohli R, Hu J, Shi W, Spencer TE, et al. Dietary L-arginine supplementation reduces fat mass in Zucker diabetic fatty rats. J Nutr 2005;135:714–21.
[20] Fu WJ, Hu J, Spencer T, Carroll R, Wu G. Statistical models in assessing fold changes of gene expression in real-time RT-PCR experiments. Comput Biol Chem 2006;30:21–6.
[21] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Series B 1995;57:289–300.
[22] Rosner B. Fundamentals of Biostatistics. 6th ed. New York (NY): Duxbury Press; 2005.
[23] Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. 3rd ed. New York (NY): Wiley; 2003.
[24] Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol 2000;7:819–37.
[25] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.
[26] Bae K, Mallick BK. Gene selection using a two-level hierarchical Bayesian model. Bioinformatics 2004;20:3423–30.
[27] Lee MLT, Whitmore GA. Power and sample size for DNA microarray studies. Statistics in Medicine 2002;21:3543–70.
[28] Muller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarray. J Am Stat Assoc 2004;99:990–1001.
[29] Tsai CA, Wang SJ, Chen DT, Chen JJ. Sample size for gene expression microarray experiments. Bioinformatics 2005;21:1502–8.
[30] Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005;21:3017–24.
[31] Jung SH. Sample size for FDR-control in microarray data analysis. Bioinformatics 2005;21:3097–104.
[32] Pounds S, Cheng C. Sample size determination for the false discovery rate. Bioinformatics 2005;21:4263–71.
[33] Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005;6:27–38.
[34] Page GP, Edwards JW, Gadbury GL, Ylisette P, Wang J, Trivedi P, et al. The PowerAtlas: a power and sample size atlas for microarray experimental design and research. BMC Bioinformatics 2006;7:84.
[35] Qiu W, Lee MLT, Whitmore GA. Sample size and power calculation in microarray studies using the sizepower package. Technical report, Bioconductor. http://bioconductor.org/packages/2.2/bioc/vignettes/sizepower/inst/doc/sizepower.pdf. 2008.
[36] Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res Camb 2001;77:123–8.
[37] Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8:625–37.
[38] Fu WJ, Dougherty ER, Mallick B, Carroll R. How many samples are needed to train a classifier - a sequential approach. Bioinformatics 2005;21:63–70.
[39] Li H, Wood CL1, Getchell TV, Getchell ML, Stromberg AJ. Analysis of oligonucleotide array experiments with repeated measures using mixed models,. BMC Bioinformatics 2004;5:209.

[40] Peng X, Wood CL, Blalock EM, Chen KC, Landfield PW, Stromberg AJ. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics 2003;4:26.
[41] Liu H, Tarima S, Borders AS, Getchell TV, Getchell ML, Stromberg AJ. Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments. BMC Bioinformatics 2005;6:106.
[42] Yuan JS, Reed A, Chen F, Stewart JR. Statistical analysis of real-time PCR data. BMC Bioinformatics 2006;7:85.
[43] Karlen Y, McNair A, Perseguers S, Mazza C, Mermod N. Statistical significance of quantitative PCR. BMC Bioinformatics 2007;8:131.
[44] Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectroscopy in proteomics: a critical review. Anal Bioanal Chem 2007;389: 1017–31.
[45] Jaffe J, Mani DR, Leptos K, Church G, Gillette M, Carr S. PEPPeR, a Platform for Experimental Proteomic Pattern Recognition. Mol Cell Proteomics 2006;5: 1927–41.
[46] Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X. Performance Evaluation of Existing De Novo Sequencing Algorithms. J Proteome Res 2006;5:3018–28.
[47] Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectroscopy. Rapid Commun Mass Spectrum 2003;17:2337–42.
[48] Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem 2005;77:964–73.
[49] Grossman J, Roos F, Cieliebak M, Liptak Z, Mathis L, Muller M, et al. AUDENS: a tool for automated de novo sequencing. J Proteome Res 2005;4:1768–74.
[50] Fischer B, Roth V, Roos F, Grossman J, Baginsky S, Widmayer P, et al. NovoHMM: a hidden markov model for de novo peptide sequencing. Anal Chem 2005;77: 7265–73.
[51] Perkins D, Pappin D, Creasy D, Cottrell J. Probability based protein identification by searching sequence databases using mass spectroscopy data. Electrophoresis 1999;20:3551–67.
[52] Gentzel M, Köcher T, Ponnusamy S, Wilm M. Preprocessing of tandem mass spectra to support automatic protein identification. Proteomics 2003;3: 1597–610.
[53] Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 2003;75:768–74.
[54] Brosch M, Swamy S, Hubbard T, Choudhary J. Comparison of Mascot and X! Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. Mol Cell Proteomics 2008;7: 962–70.
[55] Wang XQ, Ou DY, Yin JD, Wu G, Wang JJ. Proteomic analysis reveals altered expression of proteins related to glutathione metabolism and apoptosis in the small intestine of zinc oxide-supplemented piglets. Amino Acids 2009;37:209–18.
[56] Wang JJ, Chen LX, Li DF, Yin YL, Wang XQ, Li P, et al. Intrauterine growth restriction affects the proteomes of the small intestine, liver and skeletal muscle in newborn pigs. J Nutr 2008;138:60–6.
[57] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
[58] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Series B 1977;39:1–38.
[59] Walsh C. Posttranslational Modifications of Proteins: Expanding Nature's Inventory. Greenwood Village (Colo): Roberts and Company; 2006.
[60] Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotech 2001;19:946–51.
[61] Li X, Zhang H, Ranish JR, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable isotope dilution and tandem mass spectrometry. Anal Chem 2003;75:6648–57.
[62] MacCoss MJ, Wu CC, Liu H, Sadygov R, Yates III JR. A correlation algorithm for the automated analysis of quantitative "shotgun" proteomics data. Anal Chem 2003;75:6912–21.
[63] McLachlan G, Peel D. Finite mixture models. New York (NY): Wiley; 2000.
[64] Efron B, Tibshirani R. An introduction to the Bootstrap. New York (NY): Chapman and Hall/CRC; 1993.
[65] Bridges S, Magee GB, Wang N, Williams WP, Burgess S, Nanduri B. ProtQuant: a tool for label-free quantitation of MudPIT proteomics data. BMC Bioinformatics 2007;8(Suppl 7):S24.
[66] Droit A, Fillon J, Morissette J, Poirier G. Bioinformatic standards for proteomics-oriented mass spectrometry. Curr Proteomics 2006;3:119–28.
[67] Nelson W, Viele K, Lynn B. An integrated approach for automating validation of extracted ion chromatographic peaks. Bioinformatics 2008;24:2103–4.
[68] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.
[69] Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of DN copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A 2002;99:12963–8.
[70] Rauch A, Ruschendorg F, Juang J, Trautmann U, Becker C, Thiel C, et al. Molecular karyotaping using an SNP array for genomewide genotyping. J Med Genet 2004;41:916–22.
[71] Affymetrix Inc. BRLMM: an improved genotype calling method for the GeneChip human mapping 500K array set. www.affymetrix.com2006.
[72] Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Human Heredity 2003;55:56–65.
[73] Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol 2005;29:108–27.
[74] Cui Y, Fu WJ, Sun K, Romero R, Wu R. Nucleotide mapping of complex binary disease traits with HapMap. Curr Genomics 2007;8:307–22.
[75] Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 2006;22:7–12.
[76] Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 2007;8:485–99.
[77] Xiao Y, Segal MR, Yang YH, Yeh RF. A multi-array multi-SNP genotyping algorithmfor Affymetrix SNP microarrays. Bioinformatics 2007;23:1459–67.
[78] Nicolae DL, Wu X, Miyake K, Cox NJ. GEL: a novel genotype calling algorithm using empirical likelihood. Bioinformatics 2006;22:1942–7.
[79] Wan L, Sun K, Ding Q, Cui Y, Li M, Wen Y, et al. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucl Acids Res 2009, doi:10.1093/nar/gkp559.
[80] Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004;64:3060–71.
[81] Ferguson LR, Philpott M. Nutrition and mutagenesis. Annu Rev Nutr 2008;28: 313–29.
[82] Branda RF, Brooks EM, Chen Z, Naud SJ, Nicklas JA. Dietary modulation of mitochondrial DNA deletions and copy number after chemotherapy in rats. Mutat Res 2002;501:29–36.
[83] Slater HR, Bailey DK, Ren H, Cao M, Bell K, Nasioulas S, et al. High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. Am J Hum Genet 2005;77:709–26.
[84] Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 2005;65:6071–9.
[85] Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, et al. CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 2006;7:83.
[86] Wan L, Li D, Zhang D, Liu X, Fu WJ, Zhu L, et al. Conservation and implications of eukaryote transcriptional regulatory regions across multiple species. BMC Genomics 2008;9:623, doi:10.1186/1471-2164-9-623.
[87] Yin JD, Li XL, Li DF, Yue T, Fang Q, Ni JJ, et al. Dietary supplementation with zinc oxide stimulates ghrelin secretion from the stomach of young pigs. J Nutr Biochem 2009;20:783–90.
[88] Wang JJ, Chen LX, Li P, Li XL, Zhou HJ, Wang FL, et al. Gene expression is altered in piglet small intestine by weaning and dietary glutamine supplementation. J Nutr 2008;138:1025–32.
[89] van Breda SGJ, de Kok MCM, van Delft JHM. Mechanisms of colorectal and lung cancer prevention by vegetables: a genomic approach. J Nutr Biochem 2008;19: 139–57.
[90] Yan GR, He QY. Functional proteomics to identify critical proteins in signal transduction pathways. Amino Acids 2008;35:267–74.
[91] Hu CA, Williams DB, Zhaorigetu S, Khalil S, Wan GH, Valle D. Functional genomics and SNP analysis of human genes encoding proline metabolic enzymes. Amino Acids 2008;35:655–64.
[92] Jobgen W, Fu WJ, Gao H, Li P, Meininger CJ, Smith SB, et al. High fat feeding and dietary L-arginine supplementation differentially regulate gene expression in rat white adipose tissue. Amino Acids 2009;37:187–98.
[93] Liu Y, Zhou DZ, Zhang D, Chen Z, Zhao T, Zhang Z, et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes in the population of mainland China. Diabetologia 2009;52:1315–21.
[94] Zhang J, Zhang F, Didelot X, Bruce KD, Cagampang FR, Vatish M, et al. Maternal high fat diet during pregnancy and lactation alters hepatic expression of insulin-like growth factor-2 and key microRNAs in the adult offspring. BMC Genomics 2009;10:478.
[95] Arzuaga X, Ren N, Stromberg A, Black EP, Arsenescu V, Cassis LA, et al. Induction of gene pattern changes associated with dysfunctional lipid metabolism induced by dietary fat and exposure to a persistent organic pollutant. Toxicol Lett 2009;189: 96–101.
[96] Wang XQ, Wu WZ, Lin D, Li DF, Wu G, Wang JJ, et al. Temporal proteomic analysis reveals continuous impairment of intestinal development in neonatal piglets with intrauterine growth restriction. J Proteome Res 2009, doi:10.1021/pr900747d.
[97] Kim JS, Wilson JM, Lee SR. Dietary implications on mechanisms of sarcopenia: roles of protein, amino acids and antioxidants. J Nutr Biochem 2009;20:917–26.
[98] Bouwman FC, Claessens M, van Baak MA, Noben JP, Wang P, Saris WH, et al. The physiologic effects of caloric restriction are reflected in the in vivo adipocyte-enriched proteome of overweight/obese subjects. J Proteome Res 2009;8: 5532–40.
[99] van Dijk SJ, Feskens EJ, Bos MB, Hoelen DW, Heijligenberg R, Bromhaar MG, et al. A saturated fatty acid-rich diet induces an obesity-linked proinflammatory gene expression profile in adipose tissue of subjects at risk of metabolic syndrome. Am J Clin Nutr 2009;90:1656–64.